ISSN: 2612-226X

Studi AISV 4

IL PARLATO NEL CONTESTO NATURALE

SPEECH IN THE NATURAL CONTEXT

a cura di

Alessandro Vietti, Lorenzo Spreafico, Daniela Mereu, Vincenzo Galatà

AISV
Associazione Italiana Scienze della Voce

IL PARLATO NEL CONTESTO NATURALE

SPEECH IN THE NATURAL CONTEXT

a cura di
Alessandro Vietti, Lorenzo Spreafico, Daniela Mereu,
Vincenzo Galatà

Milano 2018

Studi AISV è una collana di volumi collettanei e monografie dedicati alla dimensione sonora del linguaggio e alle diverse interfacce con le altre componenti della grammatica e col discorso. La collana, programmaticamente interdisciplinare, è aperta a molteplici punti di vista e argomenti sul linguaggio: dall'attenzione per la struttura sonora alla variazione sociofonetica e al mutamento storico, dai disturbi della parola alle basi cognitive e neurobiologiche delle rappresentazioni fonologiche, fino alle applicazioni tecnologiche.

I testi sono sottoposti a processi di revisione anonima fra pari che ne assicurano la conformità ai più alti livelli qualitativi del settore.

I volumi sono pubblicati nel sito dell'Associazione Italiana di Scienze della Voce con accesso libero a tutti gli interessati.

c/o LUSI Lab - Dip. di Scienze Fisiche
Complesso Universitario di Monte S. Angelo
via Cynthia snc
80135 Napoli
email: presidente@aisv.it
sito: www.aisv.it

Edizione realizzata da
Officinaventuno
Via F.lli Bazzaro, 18
20128 Milano - Italy
email: info@officinaventuno.com
sito: www.officinaventuno.com

ISBN edizione digitale: 978-88-97657-28-6
ISSN: 2612-226X

Sommario

ALESSANDRO VIETTI, LORENZO SPREAFICO, DANIELA MEREU, VINCENZO GALATÀ
Prefazione 5

KHALIL ISKAROUS
The encoding of vowel features in Mel-Frequency Cepstral Coefficients 9

CHRISTOPHER CARIGNAN
An examination of oral articulation of vowel nasality in the light of the independent effects of nasalization on vowel quality 19

DALILA DIPINO, CHIARA CELATA
An UTI study of alveolar stops in Italian 41

DANIELA MEREU
Parlato spontaneo e stereotipi locali: il sardo parlato a Cagliari 55

OTTAVIA TORDINI, VINCENZO GALATÀ, CINZIA AVESANI, MARIO VAYRA
Sound maintenance and change: exploring inter-language phonetic influence in first-generation Italo-Australian immigrants 77

CAMILLA BERNARDASCI, STEFANO NEGRINELLI
Analisi fonetiche in due dialetti lombardo-alpini: parlato spontaneo e parlato controllato a confronto 99

GIANCARLO SCHIRRU
Tensione laringea e consonantismo. Il dialetto armeno di Gavar (Nor Bayazet) 117

BARBARA GILI FIVELA, FRANCESCA NICORA
Intonation in Liguria and Tuscany: checking for similarities across a traditional isogloss boundary 131

MARIA CRISTINA PINELLI, CINZIA AVESANI, CECILIA POLETTO
Is it prosody that settles the syntactic issue? An analysis of Italian cleft sentences 157

SIMON WEHRLE, FRANCESCO CANGEMI, MARTINA KRÜGER, MARTINE GRICE
Somewhere over the spectrum: Between singsongy and robotic intonation 179

ANTONIO ORIGLIA, ANTONIO RODÀ, CLAUDIO ZMARICH, PIERO COSI, STEFANIA NIGRIS, BENEDETTA COLAVOLPE, ILARIA BRAI, CRISTIAN LEORIN
Gamified discrimination tests for speech therapy applications 195

CECILIA DI NARDI, ROSANNA TURRISI, ALBERTO INUGGI, NILO RIVA, ILARIA MAURI, LEONARDO BADINO
An automatic speech recognition Android app for ALS patients 217

MARIA DI MARO, SARA FALCONE, FRANCESCO CUTUGNO
Prosodic analysis in human-machine interaction 227

VALENTINA SCHETTINO, ANTONIO ORIGLIA, FRANCESCO CUTUGNO
Dynamic time warping and prosodic prominence 241

ROBERTO GRETTER, MAURIZIO OMOLOGO, LUCA CRISTOFORETTI, PIERGIORGIO SVAIZER
A vocal interface to control a mobile robot 257

DUCCIO PICCARDI, FEDERICO BECATTINI
Voice Onset Time Enhanced User System (VOTEUS): a web graphic interface for the analysis of plosives' release phases 275

PAOLO BRAVI
Prosit: a Praat plug-in for the search and inspection of corpora of annotated audio files 293

Autori 307

Revisori del volume 311

DOI: 10.17469/O2104AISV000000

ALESSANDRO VIETTI, LORENZO SPREAFICO, DANIELA MEREU, VINCENZO GALATÀ

Prefazione

Questo volume prende le mosse dall'attività congressuale del XIV Convegno Nazionale AISV organizzato presso la Libera Università di Bolzano tra il 25 e il 27 gennaio 2018. Le relazioni tenute in quell'occasione vertevano, da diverse prospettive, attorno ai temi della rappresentazione e dello studio del parlato nel contesto naturale.

Per la comunità dei fonetisti, studiare la lingua nel contesto naturale in cui viene usata significa, in primo luogo, allontanarsi dal cosiddetto parlato di laboratorio, ovvero da un dato deliberatamente poco "rumoroso", raccolto da campioni di norma ridotti e, preferibilmente, composti da soggetti socialmente omogenei. Le mutate condizioni tecniche di acquisizione e trattamento di dati linguistici e una nuova consapevolezza teorica e metodologica permettono al fonetista di andare alla ricerca della lingua parlata prodotta in contesti comunicativi reali. Tali contesti sono caratterizzati, per esempio, dalla spontaneità della produzione, dalla natura dialogica dell'interazione, dall'esecuzione in ambienti rumorosi, dalla simultaneità delle fonti di informazione nonché dalla natura dinamica di tali segnali. La "scoperta" dei contesti naturali non si limita evidentemente a questi aspetti, ma comprende anche la possibilità di analizzare in modo strumentale varietà e lingue poco descritte o rappresentate nel panorama della ricerca fonetica e fonologica.

Il volume raccoglie pertanto le diverse declinazioni nelle quali questo ampio tema poteva essere trattato, proponendo una selezione rappresentativa dei contributi presentati durante il convegno. La diversità degli studi e delle riflessioni metodologiche proposte ha reso difficile, o forse superflua, una strutturazione del volume in sezioni tematiche. Si è preferito pertanto ordinare i diversi contributi lungo un *gradatum* all'interno del quale ogni elemento della sequenza condivide localmente alcune proprietà con gli elementi adiacenti. Il volume contiene inoltre anche alcune relazioni a tema libero, ovvero che, secondo lo spirito dei convegni AISV, non trattavano le tematiche suggerite dal convegno.

Apre il volume il contributo di Khalil Iskarous che ha proposto, a partire dalla sua relazione su invito[1], una riflessione sul tema "The encoding of vowel features in Mel-Frequency Cepstral Coefficients". L'articolo di Iskarous affronta il tema del convegno da una prospettiva metodologica. All'interno della ricerca fonetica la parametrizzazione delle caratteristiche acustiche del parlato è avvenuta principalmente attraverso la selezione di indici all'interno del dominio della frequenza. Tuttavia,

[1] La seconda relazione su invito dal titolo "Biosignal-based Spoken Communication", presentata da Tanja Schulz, non è contenuta in questo volume.

questi indici, pur essendo molto informativi, sono allo stesso tempo estremamente sensibili alle caratteristiche fisiche del parlante e alle perturbazioni indotte dall'ambiente nel quale il segnale si propaga. Al contrario, l'analisi del segnale vocale adottata nell'ambito delle tecnologie del parlato si basa da molti anni su LPC (*Linear Predictive Coding*) e MFCC (*Mel-Frequency Cepstral Coefficients*) che si sono mostrate fonti di informazione più robuste e meno condizionate dalle caratteristiche individuali dei parlanti. L'obiettivo dell'articolo di Iskarous è quello di esplorare il significato linguistico degli MFCC attraverso uno studio di caso sulle vocali, ovvero osservare quali relazioni sussistano tra i coefficienti e la geometria del tratto vocale.

La raccolta di contributi prosegue con due studi che mirano a valutare l'apporto differenziale degli articolatori nella produzione di parlato: da un lato, la batteria di esperimenti condotti da Carignan mostra quali siano gli effetti indipendenti della nasalizzazione sulla struttura spettrale delle vocali; dall'altro, l'indagine presentata da Dipino e Celata fornisce una prima descrizione del ruolo della configurazione linguale nella realizzazione di contrasti di durata e sonorità nelle occlusive dell'italiano toscano.

A questi primi due studi seguono tre contributi distinti per obiettivi di ricerca, ma senz'altro accomunabili per il generale interesse per la variazione e il cambiamento linguistico, lo studio di varietà e dialetti romanzi e la prospettiva metodologica orientata all'analisi di parlato spontaneo. Mereu, nel suo studio sulla palatalizzazione di /k, g/ nella varietà di sardo di Cagliari, sottolinea l'importanza di osservare dati ecologicamente validi per poter elicitare variabili sociofonetiche marcate diastraticamente. Tordini, Galatà, Avesani e Vayra mettono in evidenza il ruolo del contatto linguistico nella produzione di fricative coronali nella varietà di veneto bellunese parlata da emigrati veneti in Australia. Infine, Bernardasci e Negrinelli utilizzano la distinzione tra parlato spontaneo e controllato come variabile esplicativa in due processi di cambiamento in atto che colpiscono, da un lato, il contrasto tra affricata alveo-palatale e occlusiva palatale e, dall'altro, la distribuzione delle vocali medie anteriori in due dialetti lombardo-alpini.

Il contributo di Schirru presenta invece un'indagine strumentale del consonantismo nel dialetto armeno di Gavar. Lo studio mostra, attraverso un'analisi acustica, la reale natura fonologica dei contrasti all'interno del sistema delle occlusive. L'analisi rivela infatti che l'opposizione tra la serie delle sorde aspirate e quella delle sonore deve essere piuttosto reinterpretata come una distinzione di tensione laringea.

I tre studi che seguono sono contraddistinti dallo studio dell'intonazione come dominio di indagine. In particolare, la ricerca condotta da Gili Fivela e Nicora ha tra i propri obiettivi quello di determinare la natura dei *pattern* intonativi in zone di transizione tra aree dialettali nelle quali sistemi intonativi distinti entrano in contatto e, non di rado, in conflitto. Il secondo contributo, a opera di Pinelli, Avesani e Poletto, si concentra sull'interfaccia tra sintassi e prosodia nelle frasi scisse dell'italiano. Le autrici esaminano la presenza e natura di indizi prosodici in grado di testare l'ipotesi che tali costruzioni condividano in maggior grado i tratti strutturali

delle frasi mono-clausola. Il terzo elemento in questa sotto-sequenza prosodica è lo studio di Wehrle, Cangemi, Krüger e Grice che propone un metodo per misurare la nozione, altrimenti impressionistica, di intonazione "monotonica" (*robotic*) e "canterina" (*singsongy*). Il metodo illustrato si basa sul cambiamento dinamico di F0 nel tempo e arricchisce in questo modo il repertorio dei metodi di misurazione della distribuzione di F0 su domini molto estesi, ovvero eccedenti i costituenti prosodici più ampi.

Il contributo di Origlia, Rodà, Zmarich, Cosi, Nigris, Colavolpe, Brai e Leorin discute dello sviluppo di un sistema ludicizzato per la somministrazione di un test di discriminazione fonologica ad apprendenti di italiano L1. Il sistema è capace di tenere traccia delle prestazioni dei parlanti e, quindi, di regolarsi dinamicamente generando nuove serie di stimoli alla luce sia delle risposte sino a quel punto ricevute, sia dell'utilità e del valore informativo che gli stimoli possono avere per valutare le abilità di discriminazione fonologica del bambino.

Il lavoro di Di Nardi, Turrisi, Inuggi, Riva, Mauri e Badino è invece dedicato alla presentazione di un sistema di riconoscimento automatico dell'italiano parlato da soggetti affetti da sclerosi laterale amiotrofica. Il contributo mostra chiaramente come il tipo di architettura e il paradigma di addestramento della rete neurale possano incidere sulla precisione del sistema di riconoscimento, in particolare come ciò interessi più il parlato atipico (disartrico) che non quello tipico.

La ricerca di Di Maro, Falcone e Cutugno riguarda l'analisi delle richieste di utenti umani che interagiscono con agenti virtuali. Gli autori confrontano alcune caratteristiche fonetiche di narrazioni semi-spontanee con quelle rintracciabili nelle interazioni rivolte dagli stessi parlanti a due distinti agenti conversazionali. I risultati mostrano che mentre alcuni parametri fonetici sono modificati in egual misura e indipendentemente dall'agente con cui si interagisca (per esempio la velocità di eloquio), altri (per esempio la riconfigurazione dello spazio vocalico) subiscono l'influenza della familiarità dell'utente con il tipo di agente virtuale e sono quindi variabilmente modificati.

L'articolo di Schettino, Origlia e Cutugno tratta di prominenza prosodica e suggerisce un miglioramento della scala di valutazione nota come metodo PromDrum. In particolare, gli autori mostrano come ricorrendo all'algoritmo Dynamic Time Warping sia possibile non solo ricomprendere nelle analisi sulla prominenza file (solitamente scartati, seppur altamente informativi) in cui il numero di battiti annotati non equivalga alle sillabe attese, ma anche recuperare le informazioni sulle metriche di distanze tra battiti e sillabe per valutare la qualità del lavoro del valutatore nonché delle singole valutazioni di un valutatore.

Il contributo di Gretter, Omologo, Cristoforetti e Svaizer descrive lo sviluppo di una piattaforma robotica a basso costo con cui gli utenti possono interagire a distanza. Il sistema, sempre in ascolto, funziona in tempo reale e senza necessità di connessione a servizi remoti. La gestione dell'interazione dialogica è basata su una tecnologia di riconoscimento vocale che sfrutta un modello di Markov nascosto (HMM). La piattaforma è dotata di una serie di microfoni per la localizzazione del

parlante e una migliore acquisizione del segnale parlato anche in ambiente rumoroso, tant'è che le verifiche sperimentali dimostrano che le prestazioni del sistema sono soddisfacenti sia in termini di velocità di riconoscimento che di comprensione, anche quando l'utente si trovi a diversi metri di distanza dal robot.

Gli ultimi due contributi del volume presentano soluzioni che semplificano il processo di annotazione e analisi di dati fonetici, rendendoli accessibili anche a utenti non esperti. In particolare, Piccardi e Becattini propongono VOTEUS, un'interfaccia grafica concepita per assistere il ricercatore durante il processo di allineamento, annotazione ed estrazione delle durate del tempo di attacco della sonorità. Bravi invece presenta Prosit, un plug-in per Praat progettato per permettere ai ricercatori di gestire raccolte di file audio contenenti annotazioni TextGrid, effettuarvi delle ricerche e, infine, ascoltare, visualizzare e analizzare i segmenti restituiti dalla ricerca.

Questa breve introduzione mostra la varietà e la ricchezza di temi, approcci e finalità degli autori che hanno deciso di trattare del parlato in contesto naturale come da invito degli organizzatori del convegno e curatori del volume. Questi ultimi, da parte loro, non avrebbero mai potuto svolgere il lavoro di supervisione tanto del convegno quanto del volume che ne è scaturito se non fossero stati supportati nelle varie fasi del processo dal fattivo contributo dei membri del comitato scientifico, ovvero Cinzia Avesani, Leonardo Badino, Chiara Bertini, Maria Grazia Busà, Silvia Calamai, Francesco Cangemi, Chiara Celata, Piero Cosi, Conceição Cunha, Francesco Cutugno, Maria Paola D'Imperio, Silvia Dal Negro, Anna De Meo, Mauro Falcone, Barbara Gili Fivela, Mirko Grimaldi, Phil Hoole, Khalil Iskarous, Constantijn Kaland, Paolo Mairano, Pietro Maturi, Daniela Müller, Maurizio Omologo, Antonio Origlia, Elisa Pellegrino, Marianne Pouplier, Irene Ricci, Antonio Romano, Luciano Romito, Stephan Schmid, Jim Scobbie, Antonio Stella, Jane Stuart-Smith, Mario Vayra, Daniela Veronesi, Claudio Zmarich, Enrico Zovato. A ciascuno di loro va il più sentito ringraziamento dei curatori.

DOI: 10.17469/O2104AISV000001

KHALIL ISKAROUS

The encoding of vowel features in Mel-Frequency Cepstral Coefficients

Most work on acoustic phonetics uses formant frequencies as the parameterization of the phonetic signal for understanding the acoustic difference between the sounds of the world's languages. Work in speech technology, however, has relied for several decades on Linear Prediction Coefficients (LPC) and Mel-Frequency Cepstral Coefficients (MFCC's), due to their greater invariance to physical differences between speakers. This paper explores the phonetics of the MFCC's, asking whether these coefficients can be used by phoneticians to develop a greater understanding of the phonetic nature of speech segments. This is done through an analysis of the ability of individual coefficients to distinguish between American English vowels in the Hillenbrand database.

Keywords: Mel-Frequency Cepstral Coefficients, vowel features.

1. *Introduction*

For the last seventy years, the most popular parametrization of the speech signal in the field of Linguistic Phonetics has been the frequency spectrum (linear, bark, or mel-tranformed), whether parameterized in terms of its formants or moments (e.g., Potter, Kopp & Green, 1947; Joos, 1948; Odden, 1991; Boersma, Escudero & Hayes, 2003; Forrest, Weismer, Milenkovic & Dougall, 1988; Labov, 1994). The spectrum has been used to characterize vowel and consonant inventories across languages of the world, sociophonetic differences, and the influence of prosody on segmental production. In contrast, the field of speech technology started to move away from the spectrum and its formant peaks about 50 years ago for the main reason that the shape of the spectrum varies enormously across speakers, especially when children are included. This is for the simple reason that the spectrum, even when bark or mel-transformed, is highly sensitive to vocal tract length which can vary from 5-10 cm in children to 15-18 cm in adults. Early work in speech coding established Linear Prediction Coefficients (LPC) and Reflection Coefficients (RC), methods based on multiple and partial regression analysis (Wakita, 1973), as parameterizations of the speech signal that are robust to speaker variability. Later work identified the closely related Mel-Frequency Cepstral Coefficients (MFCCs) (Mermelstein, 1976) to be especially robust to large variation in speakers, and became the most popular parametrization for speech recognition, since the mid 1980's. Despite a few exceptions (e.g. discrete cosine coefficients, Harrington & Cassidy, 1999), the fields of Linguistic Phonetics and Speech Technology have proceeded largely in parallel over the last several decades, each using its separate characterization of speech acoustics. This situation is perplexing, since speaker-independence should not be a concern only for the speech tech-

nologist, but also for linguists interested in the acoustic distinctions across dialects and across prosodic domains within a dialect, *regardless of the specific physical characteristics of the speaker*. A possible reason for this parallel procession of highly inter-related fields is that since Joos (1948), we have known, at least for vowels, how to phonetically characterize the spectrum: the Front/Back contrast is characterized largely by F2, the High/Low Contrast is characterized largely by F1, and the Round/Unround contrast is characterized by the lowering of both F1 and F2. In contrast, little, if anything, is known about how vowel features are encoded in the MFCC coefficients. The goal of this paper is to initiate an understanding of the phonetics of MFCC's, and specifically how the basic vowel features of Front/Back, High/Low, and Round/Unround are encoded in these coefficients.

One hypothesis is that the vowel features are encoded in a highly distributed way across the MFCC's, so that no single coefficient codes for a single vowel feature. The competing hypothesis is that the vowel features are encoded in specific coefficients. Which of these possibilities holds makes the project of understanding the phonetics of MFCC's of possible interest to speech technologists, not just to linguists. The reason is that a chief motivation in the field of speech technology is to express the information in the speech signal that is speaker-independent with as few bits as possible. So, if the linguists have been correct in the hypothesis that the differences between vowel segments are specifiable with very few pieces of information, the feature settings (e.g. Front/Back, High/Low, and Round/Unround), and if it is the case that individual MFCC's code these vowel contrasts, then vowels could be specified with a smaller number of coefficients, those that specify the featural contrasts. That is, we would establish a hierarchy of importance amongst the MFCC's. Therefore, investigating the phonetics of MFCCs is possibly of interest to linguists and technologists. After all, these two groups are basically interested in the same thing, efficient encoding of speech, whether it's for the purpose of efficiently describing linguistic systems or for enabling efficient technologies. The overall aim of this program, therefore, is to bridge a gap between two fields that have a common interest in speech invariants and speech variation, which started five decades ago. The paper will concentrate on the Front/Back, High/Low, and Round/Unround features only, and will leave features such as Tense/Lax, Nasal/Oral, and the consonant features for future work. Also left for future research are time-varying effects, such as coarticulation.

2. *MFCCs*

The aim of this section is to highlight the meaning of MFCC's. One aspect is the mel-frequency transformation, emphasizing low frequencies, and averaging across higher frequencies, and is already quite familiar to linguists as it is a common transformation of the spectrum (Ladefoged, 1996). Since formant peaks for vowels are quite narrow in bandwidth, emphasizing the low frequencies, allows for higher resolution necessary for vowel identification. High Frequency burst peaks for stops and sibilant noise have higher bandwidth, therefore averaging across large spans of higher frequency still allows for consonant identification. The Cepstral aspect of MFCC's part is less familiar, except in the particular application to F0 extraction, which will not be dis-

cussed in this paper. Cepstral Coefficients have a long history in the signal processing and geophysics literatures (Robinson, 1954; Bogert, Healy & Turkey, 1963), which was reviewed recently in Oppenheim, Schafer (2004).

Before presenting the formula for how to compute MFCC's, it is useful to understand *why* they were invented. The title of the first paper to present Cepstral coefficients in detail, Bogert et al. (1963), is quite telling "The quefrency analysis of time-series for echoes". The purpose of the cepstrum was to find echos in time series. In geophysics, where they were developed, the interest is in seismogram time series, where signal reflections from the earth are processed to find echos from significant structures like oil or earthquakes (Silvia, Robinson, 1978). The idea of the cepstrum is to reveal important information about *where* the crucial echoing structures are in the medium from which the signal emerges, by processing the signal in a particular way (to be discussed momentarily). If we regard the vocal tract as an echoing chamber where glottal and supra-laryngeal sound signals are reflected in constrictions and the glottal and lip ends of the vocal tract (Wakita, 1972), with the speech signal emerging at the lips after all these echoing events, we could see how the same cepstral technique could be useful for revealing useful information about the constrictions in the vocal tract, potentially revealing information about Vowel Constriction Location and Degree, which are basically equivalent to vowel features like Front/Back, High/Low, and Round/Unround. The basic finding of Bogert et al. (1963) is that the crucial information about the echo generating structures can be found from taking the Fourier Transform of the (Log of) the Fourier Transform (i.e., the spectrum that has been at the center of acoustic phonetics). The reason why this works is based on the idea of homomorphic signal processing, and was soon recognized by Schafer and Oppenheim (Schafer, 1968; Oppenheim, 1964). The basic intuition can be seen in two steps: 1) the source-filter theory explains the spectrum of vowels as the multiplication of the transfer function of the vocal tract filter and the spectrum of the glottal source in the frequency domain, since the independent variables of the functions are frequency; 2) taking the Fourier Transform of the Log of the Fourier Transform yields a replacement of the spectrum by an *addition*, instead of a multiplication, of the cepstrum of the vocal tract filter and the cepstrum of the glottal wave. Since the spectrum of the glottal wave is rich in high frequencies (harmonics) it yields a spike in higher (quefrency) of the cepstrum, and can be *liftered* out, leaving the low quefrency vocal tract contribution. Since the technique is closely related to the traditional spectrum, with its amplitude and phase, Bogert et al. (1963) used a word game to develop names for the cepstral quantities: the *spectrum* became the *cepstrum*, the independent variable of *frequency* became *quefrency*, the *magnitude* of the spectrum became the *gamnitude,* the phase became *saphe,* and filtering became *liftering*. The algorithm for measuring the MFCC's is quite simple:

a. Obtain the Fourier spectrum of a portion of the speech signal (e.g. 25 ms).
b. Mel-transform the frequency scale, emphasizing low frequencies.
c. Obtain the logarithm of the mel-transformed Fourier spectrum.
d. Obtain the Fourier spectrum of the previous result.
e. Lifter out the effect of the glottal wave.

Due to the ubiquity of MFCC's in speech technologies, many computer languages have libraries for computing them. In this paper, the Python python_speech_features library was used.

3. *Methods*

An ideal data set for this investigation is the Hillenbrand vowel database (Hillenbrand, Getty, Clark & Wheeler, 1995). This data set contains utterances by 45 men, 48 women, 27 boys, and 21 girls producing hVd for all the American English vowels. The large number of speakers, the inclusion of different ages, and the near staticness of the vowel in the h_d context, allows us to understand the phonetics of MFCC's for static vowels in data with high speaker variability, but little temporal variability. The vowels were classified by the author, uncontroversially, for their settings of their basic vowel features (Table 1). As mentioned earlier the features Tense/Lax which could distinguish [i] and [ɪ] for instance, are not being investigated in this paper.

Table 1 - *Vowels of American English and their feature specifications*

	Front/back	*High/Low*	*Round/Unround*
i	Front	High	Unround
ɪ	Front	High	Unround
e	Front	NA	Unround
ɛ	Front	Low	Unround
æ	Front	Low	Unround
ə	NA	NA	NA
u	Back	High	Round
ʊ	Back	High	Round
o	Back	NA	Round
ɔ	Back	Low	Round
ɑ	Back	Low	Unround
ɚ	NA	NA	NA

The database contains the beginning, steady state, and end of each vowel. 25 ms were extracted from middle of each steady state, a hamming window was applied, and the resulting speech waveform was preemphasized. The python_speech_features *mfcc* function was applied, yielding 13 MFCC's. The results for two sampling rates will be presented:16,000 Hz and 8,000 Hz. The results to be presented are sensitive to these specific choices of sampling rates, a problem that will be discussed later in Section 5. There were a total of 1668 measurements (540 men vowels, 576 women vowels, 324 boy vowels, 228 girl vowels).

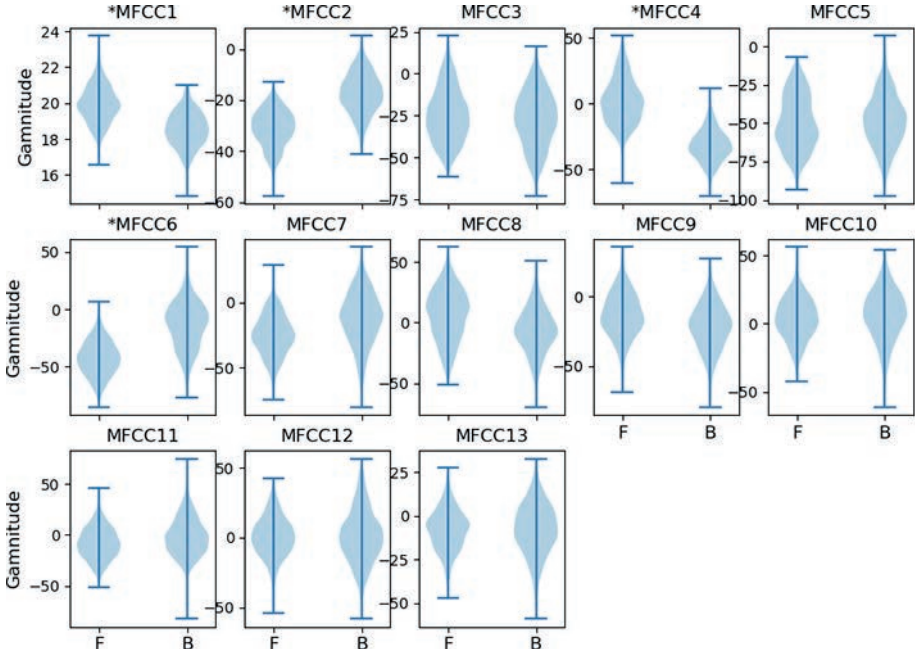
4. *Results*

The goal of this work is to understand how each of the basic vowel features is encoded in the MFCCs. To achieve this goal, for each feature, we compared the vowels specified oppositely for that feature on each of the MFCC's. The results for 16,000 Hz will be discussed before the results for 8,000 Hz. Figure 1 shows the results for the Front/Back feature. Each panel shows how Front and Back vowels compare on each of the MFCC's by showing the distribution of that coefficient's data for all the speakers. To evaluate whether each of the coefficients captures the Front/Back distinction, an ANOVA was run for each of the coefficients, and the effect was deemed significant, if $p < .001$, a low significance level, due to Bonferroni correction for 39 (3 Features x 13 coefficients) tests. In addition, to directly test the distance between the distributions for each coefficient, the Cohen's d effect size measure was used for measuring the distance between distributions A and B:

$$\text{Cohen's d (A, B)} = \frac{\text{Mean(A)} - \text{Mean(B)}}{\sqrt{\frac{\text{sd(A)}^2 + \text{sd(B)}^2}{2}}}$$

Cohen's d estimates the distance between means of the distributions in units of standard deviations. In Figure 1, significance of a distributional contrast is indicated by an asterisk before the word MFCC at the top of each panel. Significance was determined through both $p < .001$ for the ANOVA test and Cohen's d being larger than 1 standard deviation in magnitude.

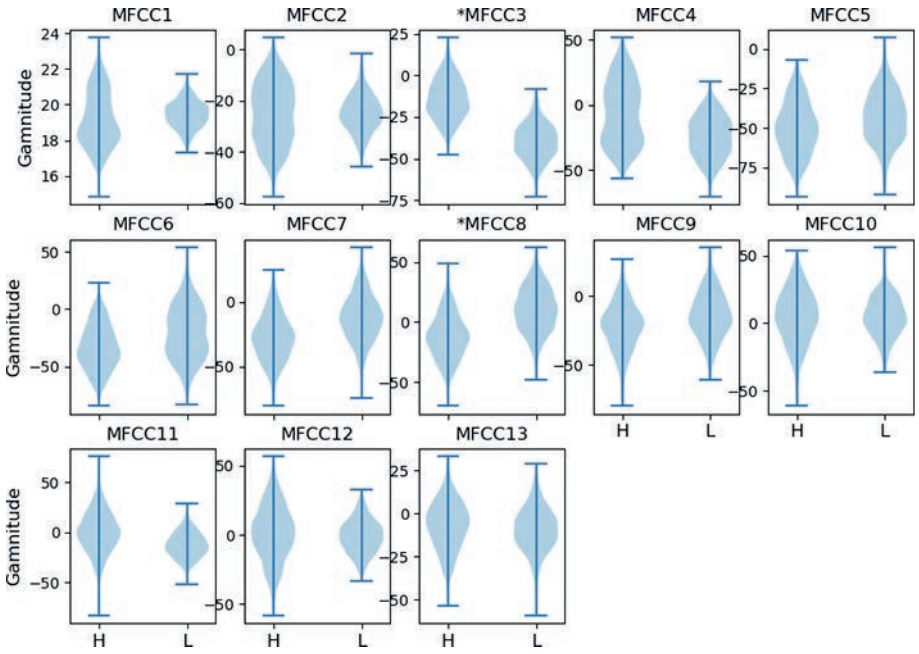
Figure 1 - *Front/Back encoding by MFCC's*



As can be seen from Figure 1, MFCC's 1, 2, 4, and 6 *directly* and significantly distinguish between Front and Back vowels. This does not mean that the other MFCC's do not code for vowel frontness, since they could do so in complex combinations with each other. But it does mean that at least one of the MFCC's, indeed 4, directly encode one of the most important features used by linguists to describe vowel systems.

Figure 2 shows the analogous situation for vowel height. As can be seen, MFCC 3 and 8 *directly* code for the height contrast. Two important things emerge from this result. One is that the MFCC's for distinguishing Front/Back (1, 2, 4, 6) do not overlap with those that distinguish High from Low (3, 8). This conforms to the idea from linguistic phonetics that the features describe different aspects of a segment (Jakobson, Fant & Halle, 1952).

Figure 2 - *High/Low encoding by MFCC's*

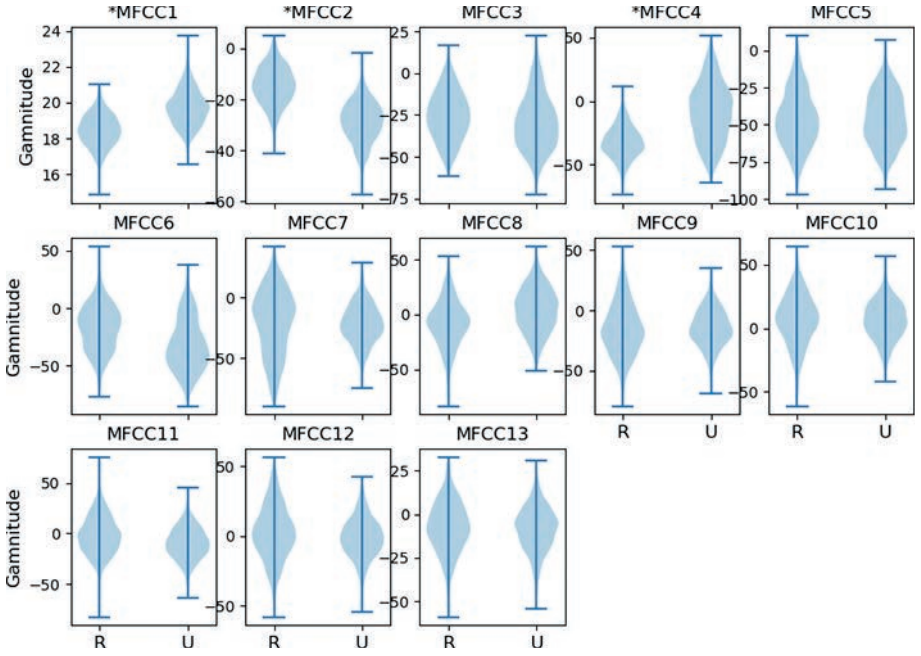


The other thing that emerges is that the number of MFCC's that directly encode Front/Back is significantly larger than the number of MFCC's encoding High/Low significantly. There are several possible reasons for this. One is an inheritance from the spectrum, in which F1, the spectral indicator of Height, ranges over a smaller frequency range (approximately 200-1000 Hz) than F2 (approximately 800-2800 Hz), the spectral indicator of Frontness. Since the cepstrum is based on the spectrum, one would expect major aspects of spectral structure to affect cepstral structure. Another possibility is that if we relax the significance criteria, more MFCC's would show significant results, therefore

the difference would not be indicative of anything substantial about the phonetic encoding of MFCC's.

Figure 3 presents the results for the Round/Unround feature.

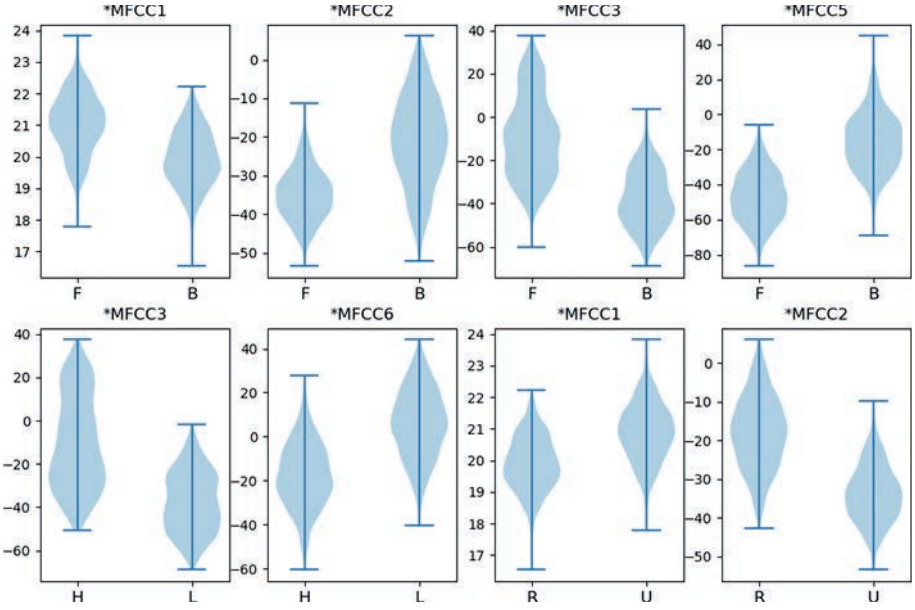
Figure 3 - *Round/Unround encoding by MFCC's*



The major thing that emerges is that the direct indicators of Round/Unround are a subset of those for Front/Back, something that phoneticians have usually refer to for English as "backness predicts rounding". Note also that the direction of the effect supports the linguistic fact that Front vowels and Unrounded vowels are similar, whereas Back vowels and Rounded vowels are similar. To determine if MFCC's can directly encode rounding in a non-overlapped manner with frontness, one must use datasets for languages that have front rounded vowels (like French) or back unrounded vowels (like Japanese), which is left for future research.

Figure 4 presents the significant coefficients for the same dataset and features, when the sampling frequency was halved at 8,000 Hz. Front/Back is encoded via MFCC's 1, 2, 3, 5. High/Low is encoded via MFCC's 3 and 6. Round/Unround is encoded via MFCC's 1 and 2. Therefore at this sampling rate, there is indeed overlap between the encoding of Front/Back and High/Low in MFCC 3. However, MFCC 6 encodes High/Low, and not Front/Back. Further investigation will need to take place to reveal why the same MFCC significantly encodes two linguistic distinctions. The other result emerging from Figure 4 is that Rounding is redundant, as was seen with the 16,000 Hz data.

Figure 4 - *Significant encodings of vowel features for 8,000 Hz sampling rate*



5. *Discussion and conclusion*

In this paper, it's been shown that there are *individual* MFCC's that code the linguistically motivated vowel features used for many decades. The main drawback of this work is that we get different results for different sampling rates, which is something unfamiliar to linguists investigating speech acoustics, where statements about F1 and F2 of the spectrum are made regardless of the sampling rate used. What is even more problematic is that for the higher sampling rate, 16,000 Hz, we had no overlap in the MFCC's encoding Front/Back and High/Low, whereas at the lower sampling rate, 8,000 Hz, we get some partial overlap.

The first issue, having to qualify statements about acoustic phonetics with information about the sampling rate used, is not as problematic as it may seem, however. The reason that there is sensitivity to sampling rate in MFCC, as well as other parameterizations such as Linear Prediction Coefficients and Reflection Coefficients, is that these coefficients encode information about the reflectivity of acoustic waves in the vocal tract, and sampling rate is directly related to the assumed number of tubes we assume the vocal tract to consist of (Wakita, 1972), and inversely related to the assumed average length of the vocal tract through the formula: $SamplingRate = Mc/2l$, where M is the number of tubes, c is the speed of sound in air, and l is the assumed average length of the vocal tract. The higher the sampling rate, the higher the assumed number of tubes. Therefore, it should not be surprising that the information in these coefficients is sensitive to sampling rate.

The question then arises as to which sampling rate to use. Familiarity with the spectrum for many years has taught us that if we care only about vowels, a sampling rate of about 8,000 Hz is sufficient, since the first three formants are below 4,000 Hz. However, if we care about consonants as well, and there is no general reason to not care about them, then 16,000 to 20,000 Hz is sufficient, since the distinctions between sibilants for instance can be detected between 5,000 and 10,000 Hz. If we sample at higher rates, e.g. the current default of 44,100 Hz, it is quite easy to downsample to these rates. Therefore, it is possible to pick an optimal sampling rate like 16,000 or 20,000 Hz, and make acoustic phonetic statements, assuming whichever of them the field settles on. The use of a higher sampling rate like 16,000 Hz also seems to solve the other problem we saw, which is overlap between the phonetic encodings of Front/Back and High/Low, a problem which needs further investigation since 8,000 Hz is sufficient for the investigation of vowels, which is what was being done in this paper.

Another issue that arises is redundancy. All the features seem to be encoded by more than one MFCC. This is somewhat surprising, since MFCC's are usually assumed to be highly uncorrelated, but it is not really a problem, since one can use multiple indicators of the same feature, leading to more robust classifications. Further investigation of how Tense/Lax is encoded, and whether the encoding of this feature overlaps with the other features, should reveal how the MFCC's separately and jointly represent linguistic distinctions.

Bibliography

Boersma, P., Escudero, P. & Hayes, R. (2003). Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories. *Proceedings of the 15th International Congress of Phonetic Sciences*, 1013-1016.

Bogert, B.P., Healy, M.J.R. & Turkey, J.W. (1963). The quefrency analysis of time series for echoes. In Rosenblatt, M. (Ed.), *Proc. Symp. Time Series Analysis*. New York: Wiley, 209-243.

Forrest, K., Weismer, G., Milenkovic, P. & Dougall, R. (1988). Statistical analysis of word initial voiceless obstruents: Preliminary data. In *The Journal of the Acoustical Society of America*, 84, 115-123.

Harrington, J., Cassidy, S. (1999). *Techniques in Speech Acoustics*. Dordrecht: Kluwer Academic.

Hillenbrand, J., Getty, L.A., Clark, M.J. & Wheeler, K. (1995). Acoustic characteristics of American English vowels. In *The Journal of the Acoustical Society of America*, 97, 3099-3111.

Jakobson, R., Fant, G. & Halle, M. (1952). *Preliminaries to speech analysis: The distinctive features and their correlates*. Technical Report 13. Massachusetts: Acoustics Laboratory, MIT.

Joos, M. (1948). *Acoustic Phonetics*. Issue 23 of Language Monographs, Linguistic Society of America Language Supplement.

Labov, W. (1994). *Principles of Linguistic Change, Volume 1: Internal Factors*. Wiley-Blackwell.

Ladefoged, P. (1996). *Elements of Acoustic Phonetics.* The University of Chicago Press.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Proceedings of the Joint Workshop on Pattern Recognition and Artificial Intelligence*, Held at Hyannis, Massachusetts, June 1-3, 1976.

Odden, D. (1991). Vowel geometry. In *Phonology,* 8, 261-289.

Oppenheim, A. (1964). *Superposition in a class of nonlinear systems*. Ph.D. Dissertation, MIT.

Oppenheim, A.K., Schafer, R.W. (2004). From frequency to quefrency: a history of the cepstrum. *IEEE Signal Processing Magazine*, 21, 95-106.

Potter, R., Kopp, G. & Green, H. (1947). *Visible Speech.* New York: Van Nostrand.

Robinson, E.A., (1954). *Predictive Decomposition of Time Series with Applications to Seismic Exploration*. Ph.D. Dissertation, MIT.

Schafer, R. (1968). *Echo removal by discrete generalized linear filtering*. Ph.D. dissertation, MIT.

Silvia, M., Robinson, E. (1978). Use of the kepstrum in signal analysis. In *Geoexploration,* 16, 55-78.

Ulrych, T.J. (1971). Application of homomorphic deconvolution to seismology. In *Geophysics*, 36(4), 650-660.

Wakita, H. (1973). Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. In *IEEE Transactions on Audio and Electroacoustics*, 21, 417-427.

DOI: 10.17469/O2104AISV000002

CHRISTOPHER CARIGNAN

An examination of oral articulation of vowel nasality in the light of the independent effects of nasalization on vowel quality

In this paper, a summary is given of an experimental technique to address a known issue in research on the independent effects of nasalization on vowel acoustics: given that the separate transfer functions associated with the oral and nasal cavities are merged in the acoustic signal, the task of teasing apart the respective effects of the two cavities seems to be an intractable problem. The results obtained from the method reveal that the independent effects of nasalization on the acoustic vowel space are: F1-raising for high vowels, F1-lowering for non-high vowels, and F2-lowering for non-front vowels. The results from previous articulatory research performed by the author on the production of vowel nasality in French, Hindi, and English are discussed in the light of these independent effects of nasalization on vowel quality.

Keywords: vowel nasality, vowel quality, articulation, acoustics, sound change.

1. *Introduction*

A traditional characterization of vowel nasality adopts a seemingly binary classification of vowel sounds based on the relative height of the velum: nasal vowels are produced with a low velum position (and, thus, air radiation from both the oral and nasal cavities), whereas oral vowels are produced with a high velum position (and, thus, air radiation from the oral cavity alone). While it is unquestionably true that nasal vowels are produced with a lowered velum, this traditional characterization carries an implicit assumption about the state of the oral cavity for the production of a nasal vowel, i.e., that the nasal vowel maintains the same articulatory characteristics as its non-nasal counterpart in all aspects except for the height of the velum. This binary view is arguably strengthened by the use of a diacritic marker to denote nasality as a secondary feature in International Phonetic Alphabet transcriptions, i.e., [˜] or /˜/. For example, the transcription [ɛ̃] implies that the vowel quality is the same as for [ɛ], and that the only articulatory difference between the two sounds is the relative height of the velum. This articulatory assumption necessarily carries a corresponding acoustic assumption: if the oral articulation of a nasal vowel is the same as its oral vowel counterpart, then any acoustic differences observed between the two vowels would be assumed to be due to nasalization itself.

These assumptions are problematic for phonetic and phonological research of vowel systems. A common practice in this research is to separate and characterize vowel categories based on acoustically measurable features, most notably the frequencies of the first two spectral formants, F1 and F2. This practice is relatively straightforward for oral vow-

els, since the formant frequencies are considered to arise from resonances associated with the singular oro-pharyngeal cavity beginning at the glottis and ending at the lips. Changes to the shape and/or length of this cavity result in relatively predictable formant frequency modulations, and the mapping from changes in articulation to changes in formant frequencies is mostly well understood, e.g., F1 is generally negatively correlated with tongue/jaw height, F2 is generally positively correlated with tongue anteriority, and lip rounding is positively correlated with lowering of all formants (Stevens, 2000; Johnson, 2003). However, when the velum is lowered, the coupling of the nasal cavity to the oro-pharyngeal cavity introduces additional resonances and anti-resonances to the acoustic spectrum, which are predicted to affect the spectral properties of nasalized vowels (including formant frequencies). Since the separate acoustic transfer functions associated with the two cavities are merged in the acoustic signal, changes to formant frequencies that arise independently from the two cavities cannot be teased apart when using acoustic measurements alone. Thus, given some observed difference in formant frequencies between an oral and a nasal vowel in a given language, how can one ascertain whether the difference arises from velum lowering or from a change in oral vowel quality?

In this paper, I summarize a novel methodological approach to determining the independent effects of vowel nasalization on F1 and F2 frequencies, as well as the resulting observation of the modification to the vowel space created by velum lowering. Subsequently, I present an overview of recent articulatory experiments that I have carried out on the production of vowel nasality in French, Hindi, and English, as well as how the language-specific results from these studies can be interpreted in the light of the independent effects of nasalization on the acoustic vowel space.

2. *Determining the independent effects of nasalization on the acoustic vowel space*

2.1 Data collection

The method – described in full in Carignan (2018) but summarized here – uses ultrasound and nasalance technologies to predict the effect of lingual configuration on formant frequencies of nasalized vowels, account for acoustic variation due to changing lingual posture, and exclude its contribution to the acoustic signal. Data collection took place at two sites in Sydney, Australia: the MARCS Institute for Brain, Behaviour and Development (Western Sydney University) and Macquarie University. Native speakers of six different languages/dialects participated in the study (American English, Australian English, Mandarin, Cantonese, French, and Hungarian): four males and two females, with a mean age of 31.3 (SD 7.5). All speakers were either graduate students or professional academics in phonetics and/or phonology.

The speakers were instructed to produce 20 sustained repetitions of each of the 11 vowels /i ɪ e ɛ æ a ɑ ɔ o ʊ u/. For each repetition, the speaker was instructed to sustain phonation of an oral quality of the vowel, then subsequently lower the velum during the sustained phonation while attempting to maintain tongue posture[1]. During the sus-

[1] The method does not necessarily require constant tongue posture, since the resulting metric accounts

tained vowel productions, nasalance data and ultrasound data related to tongue shape in the midsagittal plane were co-collected. Nasalance data were captured using a Glottal Enterprises H-SEP-MU nasalance plate, consisting of two microphones separated by a baffle that surrounds the speaker's upper lip. Ultrasound images were generated using a GE Logiq e laptop, and a GE 8C-RS transducer was positioned between the speaker's mandible and larynx and held in place with an elastic headset (Derrick, Best & Fiasson, 2015). An example of the experimental setup is shown in Figure 1. Ultrasound video was captured in real time from the GE Logiq e VGA video output using an Epiphan VGA2USB Pro video grabber. The nasalance audio and ultrasound video data were co-registered on a dedicated computer, using FFmpeg software to record a continuous .AVI file at 30 fps with embedded audio sampled at a rate of 44.1 kHz.

Figure 1 - *An example of the experimental setup used to determine the independent effects of nasalization on the vowel space, including a hand-held Glottal Enterprises nasalance device and an ultrasound probe holder headset (Derrick et al., 2015). Reproduced from Carignan (2018)*

2.2 Data analysis

Analysis of the synchronized nasalance data was carried out using Praat (Boersma, Weenink, 2015). The sustained vowel productions were segmented manually according to the broadband spectrogram and corresponding waveform. The average duration for the segmented vowels was 1.77 s (SD 0.57 s). Separate amplitude tracks for the oral

for formant variation that is due to tongue shape. However, maintaining tongue posture helps to ensure that the ultrasound image variance used to predict formant values falls within the range of image variance used to map the articulation to the acoustics.

and nasal signals were created, and nasalance was derived by calculating the proportional nasal amplitude, i.e., nasal amplitude over total amplitude. The time points associated with the minimum and maximum nasalance in each token were located automatically; these time points correspond to the most oral and most nasal parts of the token and will be referred to as the "oral point" and "nasal point", respectively. Formant estimation was performed on the combined audio from the stereo nasalance channels. Two-formant estimation at the oral and nasal time points was carried out using the Burg LPC method, with optimized parameters for each speaker and vowel, derived from a semi-automated procedure similar to Escudero, Boersma, Rauber & Bion (2009). The suitability of these optimized parameters was verified manually for each speaker and vowel via inspection of the formant tracks against a broadband spectrogram.

The indices of all of the ultrasound frames located between the oral and nasal points of the vowel tokens in each recording were logged[2], and these ultrasound images were subsequently filtered and processed separately for each speaker in MATLAB (The Mathworks Inc., 2015) using Temporally Resolved Articulatory Configuration Tracking of Ultrasound software (TRACTUS; Carignan, 2014b): images were downsized via bi-cubic interpolation to 20% of the original resolution (in order to reduce dimensionality), a region of interest (RoI) around the bounds of the movement of the tongue surface was selected, and the down-sampled pixels in the RoI were used as dimensions in principal components analysis (PCA) modeling. PCs which independently explained at least 1% of the image variance were retained, yielding between 12 and 15 total PCs for each speaker.

Due to the orthogonal nature of the components, the PC scores are able to be used as independent variables in regression models. Thus, two separate regression models (for F1 and F2) were created for each speaker. Each model included formant values as the dependent variable and the ultrasound PC scores related to the oral time point of each token as independent variables, effectively mapping the lingual articulation to the formant structure when the velum is closed. These linear models were subsequently used to predict formant values for the corresponding nasal point of each token, using the ultrasound PC scores from the nasal acoustic time points as predictor variables. The result represents formant values (in Hz) that are predicted by tongue posture alone, without any acoustic influence of nasalization. Thus, differences between predicted and measured formant values at the nasal time point can be assumed to be independent effects of nasalization. However, it is nevertheless possible that a portion of the difference between predicted and measured formant values at the nasal time point might be due to model error or to formant frequency modifications that arise from non-lingual oral articulation (e.g., labial configuration). To control for these possible sources of error, formant predictions and measurements were also made at the oral time point of each token, in order to obtain baselines of error for the oral models. The oral model errors were averaged for each vowel category, and these vowel-specific error baselines were subtracted from the measured formant values at the nasal point of the corresponding vowel tokens. In this way, each data point was corrected for vowel-specific error in the linear mapping, yielding a more conservative estimate of differences

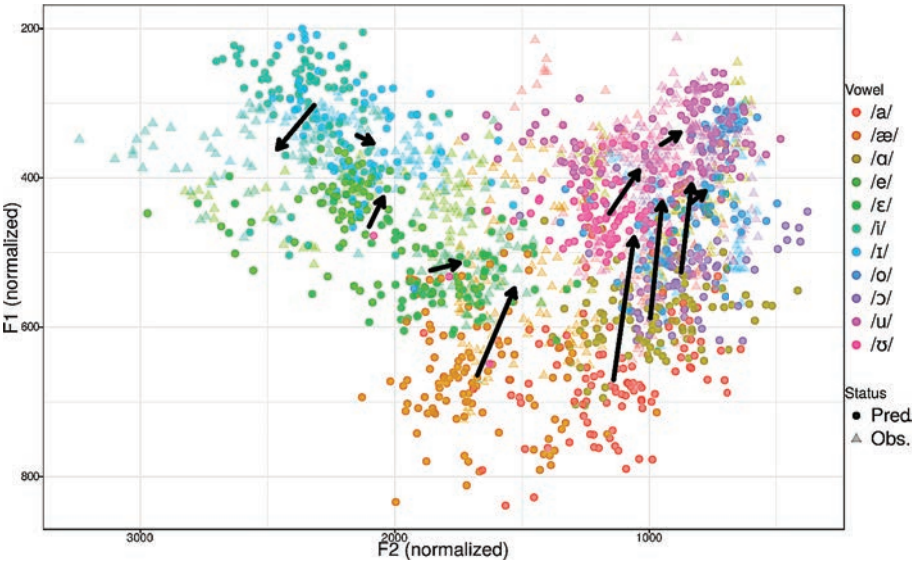
[2] The average number of total frames for each speaker was 4791 (SD 1372).

between predicted and measured formant values observed for the nasal time point. Finally, in order to combine the vowel spaces of the different speakers, formant values were normalized for each speaker via Lobanov transformation before translation back to Hz using the grand mean and average SD, in order to preserve the within-speaker normalized structure while retaining interpretability of the results.

2.3 Results

Figure 2 displays the vowel-corrected formant values measured at the nasal time points of the tokens, obtained using the method described above. The opaque colored dots are formant values predicted by tongue shape, while the transparent colored triangles are formant values that were actually observed. The arrows connect the means of the predicted values (start of the arrow) and observed values (end of the arrow). The overall pattern suggests that the independent effects of nasalization on the acoustic vowel quadrilateral are F1-raising of high vowels (vowels with F1 ≤ 350 Hz), F1-lowering of non-high vowels (vowels with F1 ≥ 350 Hz), and F2-lowering of non-front vowels (vowels with F2 ≤ 2000 Hz). The cumulative effect of these formant frequency modifications resembles a counter-clockwise chain shift: low vowels raise and retract in the vowel space, encroaching on the acoustic space of the mid-back vowels, which also raise and retract, encroaching on the acoustic space of the high-back vowels, which also raise and retract.

Figure 2 - *Acoustic vowel space of speaker-normalized and vowel-corrected formant values for nasalized productions. Opaque colored dots represent formant values predicted by lingual ultrasound images; transparent triangles represent actual measured values. Arrows connect the means of the predicted and measured categories. Reproduced from Carignan (2018)*



It is reasonable to question whether these systematic modifications to the formant structure of nasalized vowels might be perceived by listeners as changes in vowel

quality. Indeed, perceptual evidence suggests that this may be the case. With regard to F1, Beddor, Krakow & Goldstein (1986) and Krakow, Beddor & Goldstein (1988) observed that listeners can attribute an increase in F1 for nasalized high vowels to either a lower tongue position or an increase in degree of nasalization, and they can attribute a decrease in F1 for nasalized low vowels to either a higher tongue position or an increase in degree of nasalization. Similarly, Wright (1975, 1986) found that listeners perceived nasalized [ĩ] as lower and more retracted than oral [i] and nasalized [ã] as higher than oral [a]. With regard to F2, Delvaux (2009) has shown that F2-lowering alone is sufficient to trigger the percept of nasality on synthesized vowels in French, and Beddor (1993) suggests that the increased F1-F2 proximity of non-front nasal vowels observed for Hindi, Turkish, Igbo, and English (Beddor, 1982) should result in perceptual retraction compared to their oral counterparts – she notes, however, that this retraction is not necessarily well supported in the perceptual vowels spaces of Wright (1986).

By and large, these perceptual effects of nasalization mirror the independent acoustic effects of nasalization on the vowel space observed above: F1-raising of high vowels, F1-lowering of low vowels, and (arguably) F2-lowering of non-front vowels. This suggests that velum lowering creates an acoustic pressure on the vowel space, and that this pressure can be perceived by listeners as systematic shifts in vowel quality. It is, thus, of great interest to determine whether this acoustic-perceptual pressure can lead to subsequent changes in the production of nasal vowel quality by speakers, i.e., whether speakers make modifications to the *oral* articulation of nasal(ized) vowels in response to the acoustic-perceptual modifications that arise from velum lowering – either enhancing the acoustic-perceptual shifts or compensating for them. In the following sections, the results from studies on the oral articulation of phonological and phonetic vowel nasality in French, Hindi, and English will be summarized and discussed in the light of these independent effects of nasalization on the F1/F2 vowel space.

3. *Oral articulation of phonological vowel nasality*

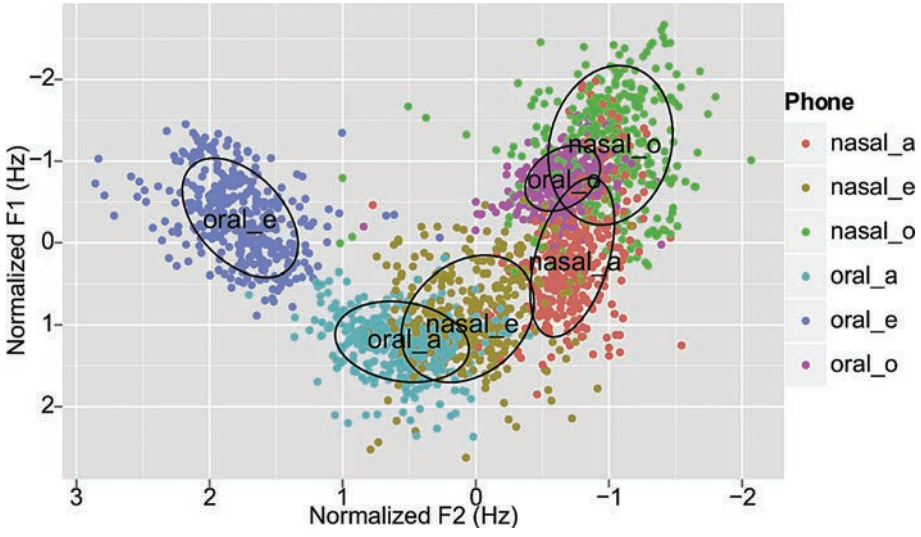
3.1 French

Northern Metropolitan French (NMF) – defined as the variety of French spoken in urban areas of France, north of the Midi–Provence southern line – is a compelling language variety for the study of vowel nasality for a variety of both historical and synchronic reasons. In particular, the phonological three nasal vowel system /ɛ̃ ɑ̃ ɔ̃/ of modern NMF is said to be undergoing a "push chain shift" (Fónagy, 1989; Maddieson, 1984; *inter alia*). Specifically, impressionistic reports claim that /ɛ̃/ is lowered and retracted, nearing the space of [ɑ̃]; that /ɑ̃/, in turn, is "pushed", retracting and raising near the space of [ɔ̃]; and that /ɔ̃/, in turn, is raised, becoming more [õ]-like. To a large extent, this counter-clockwise chain shift resembles the vowel quality modifications due to nasalization outlined in Section 2.3, with the exception of the lowering of /ɛ̃/, which is expected to undergo slight F1-lowering

due to nasalization (see Figure 2). Understanding the specific oral articulatory configurations involved in the manifestation of this chain shift would help determine if the impression of a chain shift in the NMF nasal vowel system arises primarily from the acoustic pressure on the vowel space created by velum lowering (i.e., whether the reported chain shift is merely acoustic-perceptual), or whether NMF speakers actually modify their oral articulations of the nasal vowels in a way that mirrors the acoustic modulations due to velum lowering.

In Carignan (2014a), articulatory and acoustic data relating to the productions of three oral–nasal vowel pairs /a/-/ɑ̃/, /ɛ/-/ɛ̃/, and /o/-/ɔ̃/ were recorded from 12 female NMF speakers. Lobanov normalized formant measurements for the speakers' productions of 10 repetitions of French CV lexical items containing these six vowels are shown in Figure 3. The advanced nature of the reported nasal vowel chain shift is evidenced clearly: the realization of /ɛ̃/ is lowered and retracted to such a degree that it occupies an acoustic space that is posterior to the realization of /a/, and the realization of /ɔ̃/ is raised to such a degree that it occupies an acoustic space even higher than the realization of /o/. The results suggest that the (acoustic) realization of the NMF nasal vowel system is actually closer to [ɐ̃ ɔ̟̃ õ̟] than implied by the IPA transcriptions [ɛ̃ ɑ̃ ɔ̃], transcriptions that are traditionally used to describe these vowels in NMF.

Figure 3 - *Speaker-normalized formant values for the vowels /ɑ/ ('oral_a'), /ɑ̃/ ('nasal_a'), /ɛ/ ('oral_e'), /ɛ̃/ ('nasal_e'), /o/ ('oral_o'), and /ɔ̃/ ('nasal_o'). Reproduced from Carignan (2014a)*



Articulatory data related to tongue and lip posture were recorded using electromagnetic articulography systems made by Carstens Medizinelektronik GmbH: the AG500 Electromagnetic Articulograph, located in the Speech Dynamics Laboratory in the Beckman Institute at the University of Illinois at Urbana-Champaign, Illinois, USA, and the AG200 Electromagnetic Midsagittal Articulograph, located at

Grenoble Images Parole Signal Automatique (GIPSA-lab) at l'Université Stendhal, Grenoble, France. Two speakers were recorded using the AG500 in Illinois, and 10 speakers were recorded using the AG200 in Grenoble. For each speaker (on both systems), three sensors were adhered along the midsagittal line of the tongue at even intervals, beginning ≈ 1 cm behind the tip of the tongue, ending as far back along the tongue as could comfortably be reached, with a sensor at the midpoint between these two. The vertical and horizontal positions of these sensors were used to measure tongue height and anteriority, respectively. Additionally, two sensors were placed on the lips: one on the vermilion border of the upper lip, and the other one on the vermilion border of the lower lip, in order to measure the degree of labial aperture and lip protrusion.

With regard to the vowel pair /ɛ/-/ɛ̃/, the independent acoustic effect of velum lowering observed in Section 2.3 suggests that F1 and F2 should both lower slightly for the nasalization of [ɛ]. The acoustic results from Carignan (2014a) showed that /ɛ̃/ was realized with a higher F1 and lower F2 compared to /ɛ/ for all 12 speakers. With regard to oral articulation, /ɛ̃/ was produced with a lower and more retracted tongue position than /ɛ/ for all 12 speakers, with no consistent overall labial articulatory differences for /ɛ̃/ compared to /ɛ/. In light of the independent acoustic effects of nasalization for [ɛ], these results suggest that in NMF: (1) the acoustic realization of F1 for /ɛ̃/ is not, in fact, due to nasalization, but to tongue height; and (2) the acoustic realization of F2 for /ɛ̃/ is due to a combination of nasalization and tongue retraction.

With regard to the vowel pair /a/-/ã/, the independent acoustic effect of velum lowering suggests that F1 should lower to a large degree, and that F2 should lower slightly, for the nasalization of [a]. The acoustic results from Carignan (2014a) showed that /ã/ was indeed realized with a lower F1 and F2 compared to /a/ for all 12 speakers. With regard to oral articulation, very few speakers produced /ã/ with a higher tongue position compared to /a/; in fact, most speakers produced /ã/ with a *lower* tongue position. However, 11/12 speakers produced /ã/ with a more retracted tongue position compared to /a/, and all 12 speakers produced /ã/ with greater lip rounding (via lip protrusion and/or smaller lip aperture) than /a/. In light of the independent acoustic effects of nasalization for [a], these results suggest that in NMF: (1) the acoustic realization of F1 for /ã/ is due to a combination of nasalization and lip rounding (but not tongue height); and (2) the acoustic realization of F2 for /ã/ is due to a combination of nasalization, tongue retraction, and lip rounding.

With regard to the vowel pair /o/-/ɔ̃/, the independent acoustic effect of velum lowering suggests that both F1 and F2 should lower slightly for the nasalization of [o]. The acoustic results from Carignan (2014a) showed that /ɔ̃/ was indeed realized with a lower F1 for just over half of the speakers (8/12) and a lower F2 for the majority of speakers (10/12), in comparison to /o/. However, only two speakers produced /ɔ̃/ with a higher tongue position than /o/; on the contrary, the majority of speakers produced /ɔ̃/ with a lower tongue position compared to /o/. Moreover,

only six of the 10 speakers who realized /ɔ̃/ with a lower F2 than /o/ manifested any evidence of lingual retraction for /ɔ̃/ compared to /o/. With regard to labial articulation, just over half of the speakers (8/12) produced /ɔ̃/ with a smaller labial aperture than /o/. However, five speakers produced /o/ with greater lip protrusion than /ɔ̃/. These results suggest that both /o/ and /ɔ̃/ are characterized by some degree of lip rounding, but that the articulatory strategies used to produce this rounding are different for the two vowels (i.e., greater labial protrusion for /o/ vs. smaller labial aperture for /ɔ̃/), and that these strategies vary across speakers. In light of the independent acoustic effects of nasalization for [o], these results suggest that in NMF: (1) the acoustic realization of F1 for /ɔ̃/ is due to a combination of nasalization and lip rounding (speaker dependent); and (2) the acoustic realization of F2 for /ɔ̃/ is due to a combination of nasalization, tongue retraction (speaker dependent), and lip rounding (speaker dependent).

3.1.1 Summary of articulatory results in the light of the acoustic effects of nasalization

The results from Carignan (2014a) reveal that the independent acoustic effects of nasalization are observed in the NMF nasal vowel system in the majority of cases: the only exception is a higher F1 for /ɛ̃/ compared to /ɛ/, which is posited to be due to a lower tongue position for /ɛ̃/. In some cases, adjustments in oral articulation did not yield the predicted corresponding acoustic adjustments, e.g., a lower tongue position for /ɑ̃ ɔ̃/ yet lower measured F1, which is posited to be due (at least partially) to nasalization. In many other cases, oral articulatory adjustments were observed that are predicted to yield acoustic adjustments which mirror (and perhaps enhance) the acoustic effects of nasalization, e.g., lip rounding for /ɑ̃ ɔ̃/ and more retracted tongue position for /ɛ̃ ɑ̃ ɔ̃/. Taken together, these results suggest that the acoustic chain shift in the NMF nasal vowel system is due in part to the independent acoustic effects of nasalization on the vowel space, and that in some cases oral articulatory configurations are involved in ways that enhance these acoustic modulations.

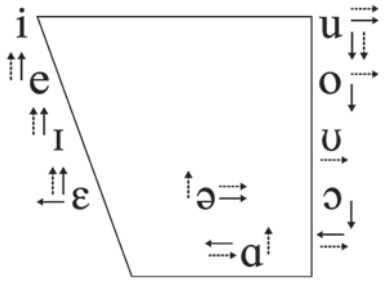
3.2 Hindi

In comparison with the three nasal vowel system of NMF, vowel nasality in Hindi involves a richer system that occupies a much larger area of the vowel space. In fact, the 10 phonemic oral vowels /i e ɪ ɛ ə ɑ ɔ ʊ o u/ each have corresponding phonemic nasal counterparts: /ĩ ẽ ɪ̃ ɛ̃ ə̃ ɑ̃ ɔ̃ ʊ̃ õ ũ/ (Ohala, 1999; Sharma, 1958). This allows for a more complete comparison between the acoustic and oral articulatory modifications in Hindi and the independent acoustic effects of nasalization on the larger vowel space. In Shosted, Carignan & Rong (2012), articulatory and acoustic data relating to the productions of nonce words created in Devanagari script were collected from four bilingual Hindi-English speakers (three females). The target items included all 20 phonemic oral and nasal vowels. Articulatory data were collected using the Carstens AG500 in Illinois, with the same sensor placement as described for Carignan (2014a). In addition to the two sensors placed on the upper and lower

lips, two additional sensors were placed at the corners of the mouth in order to obtain more complete information about labial aperture.

Unlike for NMF, where vowel nasality was often observed to involve increased lip rounding, no differences in labial configuration were observed for any of the 10 Hindi nasal vowels in comparison to their oral counterparts. However, both acoustic and lingual articulatory differences were observed; a summary of these differences is shown in Figure 4, with dashed arrows representing shifts in F1/F2 and solid arrows representing shifts in vertical/horizontal tongue position of the nasal vowels in comparison with their oral counterparts. Overall, the tongue was generally observed to be lower for back nasal vowels, more anterior for low nasal vowels, and higher for front nasal vowels, in comparison with their oral counterparts. These movements were generally supported by corresponding changes in formant frequencies that are predicted by these specific lingual shifts, e.g., a higher tongue position for the front nasal vowels is accompanied by concomitant lower F1 values.

Figure 4 - *Summary of acoustic shifts (dashed arrows) and lingual articulatory shifts (solid arrows) of Hindi nasal vowels in comparison to their oral vowel counterparts. Reproduced from Shosted et al. (2012)*



However, there are a number of cases in which acoustic differences were observed without any differences in tongue posture, as well as some cases in which lingual articulatory differences were observed without any acoustic changes. With regard to F1, /ə ɑ̃/ were realized with lower F1 values than /ə ɑ/ but no differences in tongue height, and /ɔ̃ õ/ were produced with a lower tongue position than /ɔ̃ õ/ but no differences in F1. With regard to F2, /ɑ̃ ɔ̃ ʊ̃ õ/ were realized with lower F2 values than /ɑ ɔ ʊ o/, respectively, but either no differences in horizontal tongue position (for /ʊ̃ õ/) or a more anterior tongue position (for /ɑ̃ ɔ̃/). Morever, /ɛ̃/ was produced with a more anterior tongue position than /ɛ/ but no difference in F2.

3.2.1 Summary of articulatory results in the light of the acoustic effects of nasalization

Overall, the results from Shosted et al. (2012) suggest that a clockwise lingual articulatory shift may be in progress for the nasal vowel system of Hindi: back vowels are lowered, low vowels are fronted, and front vowels are raised. This pattern is directly contrary to the pattern observed in NMF, wherein the nasal vowel system is undergoing a counter-clockwise chain shift. Moreover, the pattern of these articula-

tory modifications in Hindi is generally opposed to the pattern of the independent effect of nasalization on the acoustic vowel space observed in Section 2.3. For the most part, the lingual articulatory shifts observed in Hindi were accompanied by concomitant shifts in acoustic vowel quality, although there are a number of cases in which discrepancies between articulation and acoustics were observed. Strikingly, in each of these cases, the pattern of the discrepancy matches the pattern of the independent acoustic effect of nasalization on the vowel space: F1 was lower than predicted by tongue height for /ə̃ ɑ̃ ɔ̃ o̰/ – either a lower F1 was observed without any change in tongue height or a lower tongue position was observed without an accompanied rise in F1 – and F2 was lower than predicted by tongue anteriority for /ɛ̃ ɑ̃ ɔ̃ o̰ õ/ – a lower F2 was observed without any change in horizontal tongue position, a more fronted tongue position was observed without an accompanied rise in F2, or a more fronted tongue position was observed along with a lower F2. Finally, lower F2 values were observed for all non-front vowels compared to their oral vowel counterparts, which is the precise pattern that was observed in Section 2.3 for the independent effect of nasalization on F2 frequency. Taken together, these results for Hindi suggest that, although lingual articulatory shifts were observed that arguably oppose the effect of nasalization on formant frequencies (unlike the pattern observed in NMF), whenever a discrepancy was observed between these articulatory shifts and the measured acoustics, the discrepancy matches the pattern of the independent acoustic effect of nasalization on the vowel space in every case.

4. *Oral articulation of phonetic vowel nasality*

The studies presented in the previous section provide evidence of oral articulatory modifications for the production of vowel nasality in two typologically unrelated languages, French and Hindi. Since vowel nasality is phonologically contrastive in both of these languages, it might be argued that these articulatory modifications arose diachronically through a process of co-phonologization with velum lowering. In other words, as nasality became part of the phonological representation of the vowels in these languages, the oral articulatory modifications became part of the phonological representation as well, in ways that either mirror the natural effects of nasalization on vowel quality (e.g., French) or oppose the natural effects of nasalization on vowel quality (e.g., Hindi). This suggests that these oral articulatory modifications may have existed at some stage before nasality was phonologized, when vowel nasality in these languages was merely contextual. However, any possible oral articulations that may have, at one time, been phonetic responses to contextual nasalization have since been phonologized (in the sense of Hyman, 2008). Thus, it is important to explore the possibility of oral articulatory shifts not only in phonemic vowel nasality, but in phonetic vowel nasality as well. In this section, two studies on the oral articulation of phonetic vowel nasality in North American English (NAE) are described. The results from these studies suggest that the two somewhat opposing patterns observed

previously for French and Hindi both have synchronic analogues in contextual vowel nasality of NAE.

4.1 Articulatory compensation for vowel nasality in North American English

In Carignan, Shosted, Shih & Rong (2011), articulatory and acoustic data were used to observe whether American English (AE) /i/ and /ɑ/ manifest different degrees of tongue height when they are nasalized, i.e., when they are followed by tautosyllabic nasal consonants. These two vowels were chosen specifically to test the possibility of articulatory modifications in response to the effect of nasalization on F1 at the two extreme ends of the vowel height dimension (i.e., F1-raising for high vowels and F1-lowering for non-high vowels). Accordingly, if AE speakers adjust tongue height in order to enhance the effect of nasalization on F1 frequency, a lower tongue position is expected for nasalized vs. oral /i/, and a higher tongue position is expected for nasalized vs. oral /ɑ/. However, if AE speakers adjust tongue height in order to compensate for the effect of nasalization on F1 frequency, a higher tongue position is expected for nasalized vs. oral /i/, and a lower tongue position is expected for nasalized vs. oral /ɑ/. Finally, if AE speakers do not adjust tongue height in response to the effect of nasalization on F1 frequency, then the predicted acoustic effects are expected to be observed for the contextually nasalized variants of both vowels, i.e., higher F1 for nasalized vs. oral /i/, and lower F1 for nasalized vs. oral /ɑ/.

4.1.1 Data collection and analysis

In order to test these hypotheses, data related to tongue position, nasal airflow, and acoustics were collected from four male native AE speakers' productions of nonce words containing either /i/ or /ɑ/. 108 CVC nonce words were used as stimuli, with three randomized blocks in the experiment. The tokens had two types of nuclei (/i ɑ/, represented orthographically in the stimuli by 'ee' and 'ah', respectively), six types of onset consonant (/p b t d k g/), and nine types of coda consonant (/p b m t d n k g ŋ/). Tongue position data were obtained using the Carstens AG500 in Illinois, with three sensors adhered to the tongue mid-line in the same way as previously described for French and Hindi. In order to measure nasal flow, participants wore a vented Scicon NM-2 nasal mask (Scicon R&D, Inc., Beverly Hills, CA), connected to a Biopac TSD160A pressure transducer (Biopac Systems, Inc., Goleta, CA). The signal was digitized at 1 kHz and recorded using custom-written scripts (Sprouse, 2006) running in MATLAB. The EMA and aerodynamic data were synchronized using the pulse signal generated by the Sybox-Opto4 unit from the AG500 system. Audio data were captured using a Countryman Isomax E6 directional microphone (Countryman Associates, Inc., Menlow Park, CA) positioned 4-5 cm from the corner of the mouth.

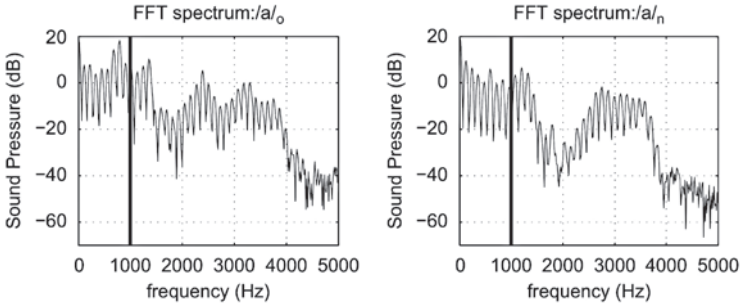
The nasal flow signal was filtered using a 75 Hz 5th order low-pass Butterworth filter, and it was used to segment the nasalized portion of the contextually nasalized vowels. For each nasalized token, the beginning of the segment was defined by

the onset of anticipatory nasalization in the filtered nasal flow signal, and the end of the segment was defined as the end of voicing in the acoustic signal. The three repetitions of each speaker's nasalized token (e.g. Speaker 1's /kim/) were used to calculate an average proportion of the vowel that was nasalized. This average proportion was then applied to the vowels of the matching oral tokens (e.g. Speaker 1's three repetitions of /kib/ and /kip/), and tongue height measurements were made at the temporal midpoint of this portion of both the nasalized and oral vowels. These time points were also used for taking acoustic measurements of F1 center of gravity (COG), which was calculated in the band 0-1000 Hz for /ɑ/ and 0-500 Hz for /i/.

4.1.2 Results

With regard to /ɑ/, no difference in tongue height was observed between the nasalized and the oral contexts. However, F1 COG was found to be significantly lower for nasalized /ɑ/ compared to oral /ɑ/. An example of the spectral shift for nasalized /ɑ/ can be seen in Figure 5. These results suggest that, since no lingual adjustments were made in the nasalized context, the F1-lowering effect of nasalization on the realization of /ɑ/ was observed. With regard to /i/, a somewhat opposing pattern was found: although no difference in F1 COG was observed between the oral and nasalized contexts, /i/ was produced with a higher tongue position in the contextually nasalized context compared to the oral context. These results suggest that, precisely *because* a lingual adjustment was made in the nasalized context – specifically, a raising of the tongue body, which is predicted to result in a lowered F1 frequency – the F1-raising effect of nasalization on the realization of /i/ was offset by the lingual adjustment. In other words, no net acoustic change was observed because the separate acoustic effects of the two articulations (i.e., F1-raising due to velum lowering and F1-lowering due to tongue raising) effectively counteracted one another.

Figure 5 - *Acoustic spectra of oral /ɑ/ (left) and nasalized /ɑ/ (right) from a speaker of American English. The vertical line represents the frequency cutoff for calculation of center of gravity measurements. Reproduced from Carignan et al. (2011)*



4.1.3 Summary of articulatory results in the light of the acoustic effects of nasalization

The results from Carignan et al. (2011) suggest that speakers of AE employed compensatory adjustments of tongue height during the production of contextual vowel nasality, but only in restricted contexts. Previous studies suggest that there is as

much as twice the variation in F1 for AE /a/~/ɑ/ vs. /i/ (Hillenbrand, Getty, Clark & Wheeler, 1995; Perkell, Nelson, 1985). The reduced variation for /i/ may be due to the proximity of its acoustic neighbor /ɪ/: increased variation in F1 for either of these two categories could result in acoustic overlap and, subsequently, a possible merger. However, /ɑ/ has no near acoustic neighbor of this type in AE; therefore, increased variation in F1 for /ɑ/ may be acceptable without any phonological consequence. Taking this into account, it is possible that the AE speakers produced /i/ with a slightly higher tongue body in the context of anticipatory nasalization as a way of compensating for F1-raising due to velum lowering, thus helping to prevent an oral-nasal phonemic split and/or prevent an acoustic merger with /ɪ/. However, the same speakers did not employ such compensatory articulation for nasalized /ɑ/, since there is no immediate phonological consequence for the resulting decrease in F1 frequency that arises from nasalization.

The fact that the AE speakers employed a compensatory articulatory adjustment in nonce words suggests that this adjustment is a phonetic and "purely synchronic" action (i.e., an online cognitive response), rather than a lexicalized articulatory modification. As a point of comparison to the observations from Carignan et al. (2011), the following section investigates a phonological process of North American English that, itself, involves lingual adjustments, in order to investigate whether the acoustic effects of contextual vowel nasalization might have any impact on the manner in which these lingual adjustments have become lexicalized in nasal vs. oral environments.

4.2 Articulatory enhancement of vowel nasality in North American English

Most dialects of North American English exhibit acoustic raising/tensing of the low vowel /æ/ in at least some phonological contexts, including a raising-falling trajectory before nasals (e.g., [bɐən] *ban*) over much of North America, and a less widespread raising pattern with a rising trajectory before /g/ (e.g., [bɐjg] *bag*). Previous studies have argued that the acoustic manifestation of pre-nasal raising, specifically, could be due to acoustic consequences of nasalization in some speakers, rather than to lingual dynamics (De Decker, Nycz, 2012; Baker, Mielke & Archangeli, 2008). In order to explore the lingual articulatory basis of /æ/-raising across North American English dialects, Mielke, Carignan & Thomas (2017) collected acoustic and ultrasound data from a regionally diverse group of 22 native English speakers (14 males, age range 20-72) from geographic regions of the United States and Canada known to exhibit distinct patterns of /æ/-raising – as well as a male speaker from Newfoundland and a female speaker from the United Kingdom, where /æ/-tensing is not expected to occur, as a basis for comparison.

4.2.1 Data collection and pre-processing

The stimuli consisted of 170 English words and English-like nonwords, each of which was presented three times in the experiment. These included 41 stimuli with /æ/ followed by a range of consonants, and in most cases preceded by a labi-

al consonant or no consonant. Additional stimuli were included as distractors for the purpose of this study; these stimuli included /æ/ as well as the other vowels along the front diagonal of the vowel space /ɑ ɛ ej ɪ i/. Data collection occurred at two sites: 20 people participated at the Phonology Laboratory at North Carolina State University in Raleigh, North Carolina, USA, and four speakers participated at the Sound Patterns Laboratory at the University of Ottawa in Ottawa, Ontario, Canada. At both labs, data collection occurred inside a sound-attenuated booth, with ultrasound image acquisition at 60 fps occurring on a Terason t3000 ultrasound machine, running Ultraspeech 1.2 (Hueber, Chollet, Denby, & Stone, 2008), and using a microconvex array transducer (8MC3 in Raleigh and 8MC4 in Ottawa). Articulate Instruments headsets were used for probe stabilization (Scobbie, Wrench, & van der Linden, 2008), and audio was recorded in Audacity using a head-mounted omnidirectional microphone and SoundDevices USBPre2 preamplifier.

Phone-level segmentations of the audio recordings were made using the Penn Phonetics Lab Forced Aligner (P2FA; Yuan, Liberman, 2008); closure intervals of stops were hand-corrected as necessary. After segmention, the frequencies of the first three formants were measured at 5 ms intervals during all vowel intervals using a Praat script that automatically selected the best measurement parameters for each vowel token based on the similarity of the measured formant frequencies and bandwidths to a set of previous measurements from the Raleigh Corpus of interviews (Dodsworth, Kohn, 2012). Formant frequencies were normalized by speaker using Lobanov transformation.

4.2.2 Articulatory signal generation and analysis

Both the ultrasound images and the formant measurements were used to create time-varying lingual articulatory signals for each speaker. These signals were based on the front diagonal of the acoustic vowel space (normalized F2 - normalized F1, or Z2-Z1, where "Z" refers to *z*-scores), since much of the acoustic variation between raised and un-raised /æ/ falls along this axis (Labov, Rosenfelder & Fruehwald, 2013). The articulatory signal related to acoustic Z2-Z1 will be referred to as "lingual Z2-Z1" because it represents the lingual component of movement along the front diagonal of the vowel space.

In the same manner as outlined in Section 2.3, the ultrasound images were processed separately for each speaker using TRACTUS (Carignan, 2014b), yielding PC scores representing independent axes of variation within each speaker's ultrasound image set. 20 PCs were retained for each speaker, which explained a total of 66%-80% (mean: 73.95%) of the variance in each speaker's image set. For each speaker's data set, a linear regression was performed with dependent variable Z2-Z1 and independent variables PCs 1-20. The data included every frame in the interval of the vowels [ɑ æ ɛ ej ɪ i]. The coefficients from the linear regression model were used to transform the articulatory PC score matrix to match the articulatory diagonal, resulting in a lingual posture signal composed of a single score for each ultrasound frame. For any given ultrasound frame, the higher the score is the more raised

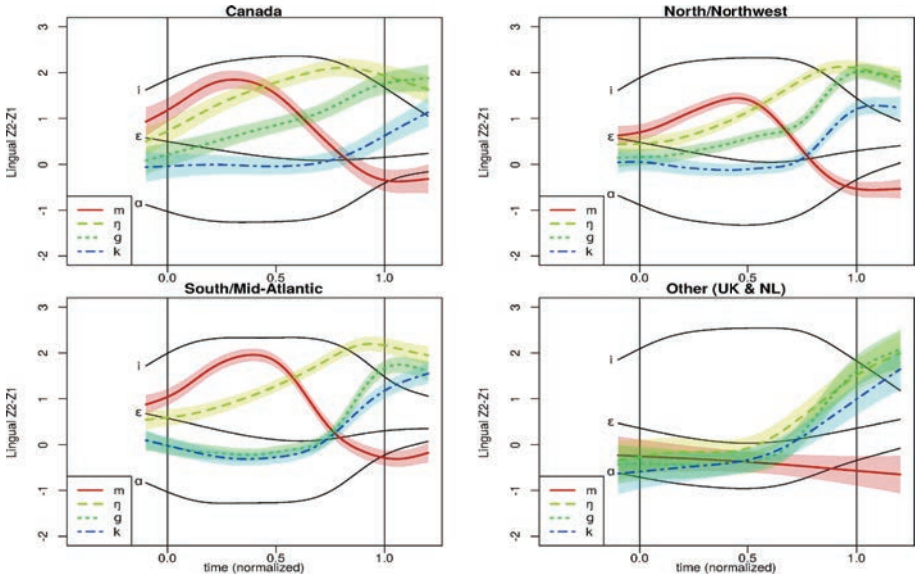
and fronted the tongue body is. Since it is derived from ultrasound images instead of acoustic data, the articulatory signal is continuous throughout the recording, even during consonant and silence intervals.

To compare lingual Z2-Z1 trajectories for groups of speakers in various segmental contexts, a generalized additive mixed model (GAMM) was created using the *bam* function from the R package MGCV (Wood, 2017). Time was normalized with the start and end of the vowel interval as (0,1), and time points within the interval (–0.1,1.2) were included in the model in order to incorporate a portion of the preceding and following consonants. The GAMM included an interaction variable (Region × Context). The dependent variable lingual Z2-Z1 was modeled with an intercept for region/context, a smooth for normalized time by region/context, and a random smooth for subject. To provide comparisons for the /æ/ lingual Z2-Z1 trajectories, another GAMM was created with /ɑ ɛ i/ data instead of /æ/ and a smooth for region/vowel instead of region/context. The mean trajectories for these three vowel models are provided as reference for contextualizing the degree of tensing in the /æ/ trajectories.

4.2.3 Results

Figure 6 shows lingual Z2-Z1 trajectories (GAMM predictions) for /æ/ before /m ŋ g k/. For the Other speakers (for whom /æ/-tensing was not predicted) we can see that tongue raising only occurs before the velar consonants, and only towards the end of the vowel interval, reaching its peak within the consonant closure. This raising pattern is, thus, indicative of anticipatory co-articulation with the following velar closure. However, for the speakers in the other three regions, tongue raising is also evidenced to various degrees in other contexts. For all three regions, a raising-falling pattern is evidenced for /æ/ before /m/; the peak of this gesture occurs at or just prior to the temporal midpoint of the vowel. The same pattern was also observed before /n/ (not shown in this figure), except that the tongue remains slightly raised at the end of the vowel for the alveolar closure. Tensing before /g/ is not evidenced for the South/Mid-Atlantic region, however it is for both of the northern regions (i.e., Canada and North/Northwest): tongue raising begins early in the vowel interval and increases in a linear fashion until the end of the vowel, where the gesture reaches its peak for the velar consonant closure. The same pattern can be observed for /æ/ before /ŋ/, except that the magnitude of tongue raising is higher throughout the entire vowel interval compared to /æ/ before /g/ for all three tensing regions, even those which display pre-/g/ tensing.

Figure 6 - *Comparisons of lingual Z2-Z1 trajectories created from GAMM predictions for /æ/ before /m/ and velars /k g ŋ/. Reproduced from Mielke et al. (2017)*



4.2.4 Summary of articulatory results in the light of the acoustic effects of nasalization

The results from Mielke et al. (2017) suggest that, not only does /æ/-tensing vary across dialectal regions of North American English, but it also varies across phonological environments with regard to the temporal characteristics and magnitude of the tensing gesture. Firstly, tensing before the anterior nasals /m n/ manifests as a rising-falling lingual gesture that is particular to these contexts and not observed for other tensing environments. Since the distinctive gesture is observed for both the coronal and labial contexts, this suggests that the gesture is not due to anticipatory co-articulation (i.e., there is no lingual setting required for the following /m/, yet the gesture is still observed in the vowel). Secondly, tensing in pre-velar contexts reveals a distinction between the nasalized /æ/ (i.e., before /ŋ/) and the oral /æ/ (i.e., before /g/, but not before /k/). This suggests as well that the gesture is not due to anticipatory velar co-articulation – the gesture is not evidenced before /k/ – and that the magnitude of tensing has been been lexicalized to different degrees for the pre-oral and pre-nasal contexts – a larger magnitude has been lexicalized for /æ/ before /ŋ/.

The independent acoustic effects of velum lowering observed in Section 2.3 suggests that F1 should lower to a large degree, and that F2 should lower slightly, for the nasalization of [æ]. In stressed context, the nearest NAE vowel categories in the direction of this acoustic shift lie along the front diagonal of the vowel space (i.e., in the direction of acoustic tensing; see Figure 2). Thus, it is plausible that, at some point in the evolution of NAE: (1) the acoustic consequence of velum lowering on the quality of /æ/ was mis-perceived by listeners as acoustic tensing and, subsequently, produced with a higher tongue position; and/or (2) speakers began producing

/æ/ before nasal consonants with a higher tongue position as way of enhancing the natural acoustic effect of nasalization on the vowel. In any case, the articulatory observations from Mielke et al. (2017) suggest that tongue raising/fronting of /æ/ has been lexicalized for varieties of NAE in certain contexts (i.e., it is no longer due to anticipatory co-articulation), and that the magnitude of the articulatory gesture involved in this lexicalization is greater for contextually nasalized contexts.

5. *Discussion*

The traditional characterization of vowel nasality – as well as the use of a diacritic marker to designate nasality in phonetic and phonological transcription – carries the implicit assumption that the only articulatory distinction between oral and nasal(ized) vowels is the relative height of the velum. On the contrary, the results from the studies surveyed in this manuscript reveal that both contrastive and contextual vowel nasality can be realized with modifications to the shape and length of the oral cavity in addition to velum lowering, even in typologically unrelated languages.

In some cases, these oral articulatory modifications are predicted to result in formant frequency modulations that enhance the independent effects of nasalization on the vowel space, e.g., the counter-clockwise chain shift of the nasal vowel system of Northern Metropolitan French and /æ/-tensing in North American English. It is possible that these articulatory modifications arose diachronically due to listeners (mis-)attributing the acoustic-perceptual vowel quality change arising from velum lowering to changes in oral articulation, subsequently producing these vowels with the corresponding changes in oral articulation. Alternatively, it is possible that speakers began producing the contextually nasalized vowels with changes in oral articulation whose acoustic effects mimic the acoustic effects of velum lowering on vowel quality, as a way of enhancing the natural acoustic consequence of nasalization.

In other cases, these oral articulatory modifications are predicted to result in formant frequency modulations that oppose the independent effects of nasalization on the vowel space, e.g., the clockwise chain shift of the nasal vowel system of Hindi and compensatory tongue body raising of nasalized /i/ in American English. It is possible that these articulatory modifications act as a way of counteracting and compensating for the acoustic-perceptual vowel quality change arising from velum lowering, in order to maintain vowel categories and prevent phonemic splits and mergers. Such a response is arguably more likely for Hindi than French, due to the larger and more acoustically crowded nasal vowel system (10 vowels in Hindi vs. 3 vowels in French); likewise, it is arguably more likely for American English /i/ than /ɑ/, due to the relative proximity of the acoustic neighbor /ɪ/ in the vowel space.

6. *Conclusion*

The results from this manuscript suggest that an accurate description of the production of vowel nasality in a given language (be it phonetic or phonological vowel nasality) cannot be ascertained without knowledge of the configuration of the entire vocal tract, and not simply the articulatory state of the velum. Moreover, this knowledge cannot be properly assessed using the acoustic signal alone, since velum lowering has been shown here to result in independent modifications to acoustic vowel quality throughout the entirety of the vowel space. Thus, the articulatory cause of changes to acoustic vowel quality cannot be determined from the acoustic signal alone. The results from the experimental method summarized here (and presented in detail in Carignan, 2018), as well as the overview of oral articulatory research on vowel nasality provided in this manuscript, therefore serve as both a caution and a challenge to linguists: perhaps it is time to question the validity of the use of acoustic measurements for researching vowel nasality, and perhaps it is time to question the traditional practice of characterizing vowel nasality in a binary fashion.

Acknowledgements

This research was funded by DFG grant HA 3512/15-1 "Nasal coarticulation and sound change: a real-time MRI study" (J. Harrington & J. Frahm).

Bibliography

Baker, A., Mielke, J. & Archangeli, D. (2008). More velar than /g/: Consonant coarticulation as a cause of diphthongization. In *Cascadilla Proceedings Project, Somerville, MA*, http://www.lingref.com/cpp/wccfl/26/paper1656.pdf. Accessed 01.07.17.

Beddor, P.S. (1982). Phonological and phonetic effects of nasalization on vowel height. Ph.D. thesis, University of Minnesota. Distributed by the Indiana University Linguistics Club.

Beddor, P.S. (1993). The perception of nasal vowels. In Huffman, M.K., Krakow, R.A. (Eds.), *Nasals, Nasalization, and the Velum* (Phonetics and Phonology 5). Academic Press: San Diego, 171-196.

Beddor, P.S., Krakow, R.A. & Goldstein, L.M. (1986). Perceptual constraints and phonological change: A study of nasal vowel height. In *Phonology Yearbook*, 3, 197-217.

Boersma, P., Weenink, D. (2015). Praat: doing phonetics by computer. *[Computer software program] available from http://www.praat.org/*.

Carignan, C. (2014a). An acoustic and articulatory examination of the 'oral' in 'nasal': The oral articulations of French nasal vowels are not arbitrary. In *Journal of Phonetics*, 46, 23-33.

Carignan, C. (2014b). TRACTUS (Temporally Resolved Articulatory Configuration Tracking of UltraSound) software suite. *[Computer software program] available from http://christophercarignan.github.io/TRACTUS*.

Carignan, C. (2018). Using ultrasound and nasalance to separate oral and nasal contributions to formant frequencies of nasalized vowels. In *Journal of the Acoustical Society of America*, 143(5), 2588-2601.

Carignan, C., Shosted, R., Shih, C. & Rong, P. (2011). Compensatory articulation in American English nasalized vowels. In *Journal of Phonetics*, 39, 668-682.

De Decker, P.M., Nycz, J.R. (2012). Are tense [æ]s really tense? The mapping between articulation and acoustics. In *Lingua*, 122, 810-821.

Derrick, D., Best, C.T. & Fiasson, R. (2015). Non-metallic ultrasound probe holder for co-collection and co-registration with EMA. In *Proceedings of 18th International Congress of Phonetic Sciences (ICPhS)*, 1-5.

Dodsworth, R., Kohn, M. (2012). Urban rejection of the vernacular: The SVS undone. In *Language Variation and Change*, 24, 221-245.

Escudero, P., Boersma, P., Rauber, A.S. & Bion, R.A.H. (2009). A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. In *Journal of the Acoustical Society of America*, 126(3), 1379-1393.

Fónagy, I. (1989). Le français change de visage. In *Revue Romane*, 24(2), 225-254.

Hillenbrand, J., Getty, L.A., Clark, M.J. & Wheeler, K. (1995). Acoustic characteristics of American English vowels. In *Journal of the Acoustical Society of America*, 97(5), 3099-3111.

Hueber, T., Chollet, G., Denby, B. & Stone, M. (2008). Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. In *Proceedings of the 8th International Seminar on Speech Production*, 365-369.

Hyman, L.M. (2008). Enlarging the scope of phonologization. In *UC Berkeley phonology lab annual report*, 382-408.

Johnson, K. (2003). *Acoustic and Auditory Phonetics*. Blackwell: Oxford.

Krakow, R.A., Beddor, P.S. & Goldstein, L.M. (1988). Coarticulatory influences on the perceived height of nasal vowels. In *Journal of the Acoustical Society of America*, 83(3), 1146-1158.

Labov, W., Rosenfelder, I. & Fruehwald, J. (2013). One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. In *Language*, 89, 30-65.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.

Mielke, J., Carignan, C. & Thomas, E.R. (2017). The articulatory dynamics of pre-velar and pre-nasal /æ/-raising in English: an ultrasound study. In *Journal of the Acoustical Society of America*, 142(1), 332-349.

Ohala, M. (1999). Hindi. In *Handbook of the International Phonetic Association*. Cambridge University Press: Cambridge, 100-103.

Perkell, J.S., Nelson, W.L. (1985). Variability in production of the vowels /i/ and /a/. In *Journal of the Acoustical Society of America*, 77, 1889-1895.

Scobbie, J.M., Wrench, A.A. & van der Linden, M. (2008). Head-probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement. In *Proceedings of the 8th International Seminar on Speech Production*, 373-376.

Sharma, A. (1958). *A Basic Grammar of Modern Hindi*. Central Hindi Directorate, Ministry of Education and Social Welfare: New Delhi.

Shosted, R., Carignan, C. & Rong, P. (2012). Managing the distinctiveness of phonemic nasal vowels: Articulatory evidence from Hindi. In *Journal of the Acoustical Society of America*, 131(1), 455-465.

Sprouse, R. (2006). DAQSession (Version 0.3). *[Computer software program]*. University of California, Berkeley: Berkeley, CA.

Stevens, K.N. (2000). *Acoustic phonetics*. MIT Press, Cambridge, MA.

The Mathworks Inc. (2015). *Matlab 2015a*. Natick, Massachusetts, United States.

Wright, J.T. (1975). Effects of vowel nasalization on the perception of vowel height. In Ferguson, C.A., Hyman, L.M. & Ohala, J.J. (Eds.), *Nasálfest: Papers from a symposium on nasals and nasalization*. Stanford University Language Universals Project: Palo Alto, CA, 373-388.

Wright, J.T. (1986). The behavior of nasalized vowels in perceptual vowel space. In Ohala, J.J., Jaeger, J.J. (Eds.), *Experimental Phonology*. Academic Press: New York, 45-67.

Yuan, J., Liberman, M. (2008). Speaker identification on the SCOTUS corpus. In *Proceedings of Acoustics '08*, 5687-5690.

Wood, S.N. (2017). mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. [Computer software program] available from https://cran.r-project.org/package=mgcv.

DOI: 10.17469/O2104AISV000003

DALILA DIPINO, CHIARA CELATA

An UTI study of alveolar stops in Italian

We investigate lingual articulation in alveolar stops produced by 5 native Tuscan Italian speakers and varying for voicing and phonological length. Both constriction location and overall tongue configuration are evaluated. The results suggest uniformity in constriction location of singleton and geminate stops as well as in voiced and voiceless stops. On the contrary, overall tongue configuration shows different patterns for length and voicing. Moreover, the subjects show individual preferences as far as constriction location is concerned. The findings are discussed with reference to cross-linguistic patterns of articulatory variation as a function of changes in duration and glottal activity associated to the production of alveolar stops.

Keywords: alveolar stops, voicing, gemination, Italian, Ultrasound Tongue Imaging.

1. *Introduction*

Voicing distinctions in obstruents are mostly realized through different laryngeal configurations during the closure phase (as in so-called 'true voicing' languages) or different timing patterns between closure at the oral level and vibratory activity at the glottal level (as in languages where the voice onset time is the major correlate of stop distinctions); languages differ for the relative weight they assign to either laryngeal strategy. Recent studies suggest that differences at the glottal level for obstruent voicing distinction may also have an impact at the level of oral articulators. On a different domain, length distinctions in obstruents (as well as in sonorants and vowels) are mostly realized at the durational level, but may additionally involve duration-related changes at the supraglottal levels. This paper is the first investigation of the lingual characteristics of Italian alveolar stops varying in voicing and length. Its aims are those of documenting patterns of potential variation in target articulations due to either or both phonological distinctions and to relate them to findings available for other languages.

Supraglottal articulatory correlates of the voicing contrast in obstruents have long since been observed in various languages. For instance, /d/ is realized with more extended central contact than /t/ in English (e.g. Dagenais, Lorendo & McCutcheon, 1994) and Japanese (Matsumura, Kimura, Toshino, Tachimura & Wada, 1994); it is produced with lower and more retracted tongue tip, less extensive contact, occasionally incomplete closure and a lower target position for the jaw in German (Fuchs, Perrier, 2003), British English and Norwegian (Moen, Simonsen, 1997), Czech (Skarnitzl, 2013), Japanese (Kochetov, 2014; Kochetov, Kang, 2017) and Moroccan Arabic (Zeroual, Esling & Crever-Buchman, 2008). The tongue body starts lower

and ends lower in /g/compared to /k/ in English and Swedish, and this gesture has greater amplitude and higher velocity (Löfqvist, Gracco, 1994). Active oral cavity enlargement has also been detected (Westbury, 1983; Kent, Moll, 1969). These articulatory characteristics are associated to shorter duration of the voiced compared to the voiceless consonant in all the investigated languages, consistently with the observation that stop voicing and stop closure have conflicting aerodynamic requirements (Ohala, 1983; 2011). The articulatory maneuvers mentioned above would then prevent quickly rising intraoral pressure from challenging the maintenance of vocal fold vibration during the closure period.

Articulatory changes in the geminates compared to corresponding singletons have also been investigated. For Italian, average electropalatographic data in Payne (2006: 90) show that seal contact, when present, is limited to the first front row in /t/ and /d/, whereas it is always present and may extend to the second front row in their geminate counterparts. The data also suggest that /t/ and /t:/ have more extended front contact than /d/ and /d:/, but since the focus of the paper is gemination, potential voicing distinctions are not thoroughly discussed. Zmarich, Gili Fivela, Perrier, Savariaux & Tisato (2006; 2009) and Gili Fivela, Zmarich, Perrier, Savariaux & Tisato (2007) additionally show that geminate stops are not only realized with longer consonantal gestures than singletons, but also differ for being articulated with longer and wider constriction and release gestures; additionally, release gestures are faster, thus suggesting that the kinematic differences between singletons and geminates are stronger at consonantal offset than at consonantal onset. Similar findings are discussed by Fujimoto, Funatsu & Hoole (2015) with respect to Japanese, where /t/ is shown to reach its kinematic peak at about half of the closure whereas for /t:/ the peak is reached far later and closer to the consonantal offset than to the consonantal onset. Longer and more extended contact in word- and utterance-initial geminate coronal stops has also been reported for Swiss German (Kraehenmann, Lahiri, 2008), Tashlhyit Berber (Ridouane, 2007; Ridouane, Hallé, 2017, which also includes utterance-medial word-initial contexts) and Cypriot Greek (Armosti, 2009). Intervocalic alveolar and velar Japanese geminates are also articulated with longer and more extended linguo-palatal contact (Kochetov, 2012; Kochetov, Kang, 2017), coupled with slower tongue movements (Löfqvist, 2007). On the other hand, longer tongue tip contact with no differences in tongue tip target position is reported for Moroccan Arabic alveolar geminates (Zeroual et al., 2008). Bilabial stops have been investigated to uncover potential differences in the coordination between the lip gesture for the consonant and the tongue gesture for the flanking vowels. These studies suggest differences in the lip closing gesture more consistently than in the opening gesture. For instance, Šimko, O'Dell & Vainio (2014) show that the lip closing gesture starts earlier with respect to the lingual movement in geminates than in singletons in Finnish; Türk, Lippus & Šimko (2017) show that the lip closure gesture in Estonian geminates is longer and larger, while maintaining the same average velocity of singletons.

In sum, both voicing and length significantly impact the way in which oral articulators move to realize the consonantal gesture and coordinate with other gestures. The

available evidence suggests that some changes are robust cross-linguistically while other are more likely to be differently implemented across languages and phonetic contexts. Voicing and length may also interact, as shown by e.g. Tashlhiyt Berber /t/-/tː/ and /d/-/dː/ contrasts, where the increased linguo-palatal contact of geminates compared to singletons is more evident in the voiceless than in the voiced pair (Ridouane, Hallé, 2017). Most of the reviewed studies are based on tongue or lip movement tracking through electromagnetic articulography (EMA) or measurements of linguo-palatal contact through electropalatography (EPG).

In this paper we investigate the lingual correlates of voicing and length in alveolar stops in Italian by tracking the midsagittal tongue contour through ultrasound tongue imaging (UTI) (Stone, 2005). The study aims to uncover potential differences in tongue tip gesture as well as tongue body configuration during the realization of /t tː d dː/. Based on the findings on other languages reviewed above, we predict that voicing entails a lower and/or more retracted tongue tip if voicing affects the target constriction location, a lower tongue body if voicing affects the overall tongue configuration. On the other hand, length is expected to entail higher tongue tip and/or higher tongue body depending on whether increased closure duration influences more the constriction location or the overall tongue configuration. Our analysis will be limited to average lingual configuration; we reserve the investigation of the dynamic properties of the consonantal gesture (e.g., closing and opening gestures) to a future study.

2. *Methodology*

2.1 Stimuli, participants and procedure

The stimuli analysed here are a subset of a longer list that was recorded in the context of a research project on the production of various consonantal contrasts of Italian and Austrian German (Celata, Meluzzi, Moosmüller, Hobel & Bertini, 2017).

The stimuli were 12 paroxytone dysillabic real words. The target consonants were /t d tː dː/ preceded by stressed /a/ or /ɔ/ and followed by unstressed /a/ or /o/. Word-initial consonants, when present, were bilabials in order to avoid lingual coarticulation effects. The list of stimuli is given in Table 1. The contexts with preceding /a/ and those with preceding /ɔ/ were distinguished in the analysis.

Table 1 - *Experimental stimuli*

C	V1	Singleton	Geminate
Voiceless	/a/	Bata /ˈbata/	batta /ˈbatːa/
	/ɔ/	mota /ˈmɔta/	motto /ˈmɔtːo/
Voiced	/a/	Ada /ˈada/	Adda /ˈadːa/
	/ɔ/		bodda /ˈbɔdːa/

The stimuli were elicited according to a multi-repetition reading task. Three male and two female native Tuscan Italian speakers aged between 19 and 32 years were recorded in the anechoic chamber of the linguistics laboratory of the Scuola Normale Superiore di Pisa. None of them reported current or past speech or hearing disorder. The participants were asked to read aloud each word upon appearance on the screen of a computer, after a hardware pulse, while keeping their speech rate as uniform as possible. Each participant performed the reading task alone on a dedicated session.

A microconvex ultrasound transducer (Mindray 65EC10EA 6.5 MHz) was placed under the chin of the participants and blocked by a stabilizing headset (Scobbie, Wrench & Van der Linden, 2008). The ultrasound signal was collected at 30 Hz (corresponding to 60 Hz after de-interlacing) by the Standard Mindray DP6600 system. The acoustic signal was captured through a unidirectional dynamic Shure microphone. The ecographic signal and the acoustic signal were synchronously acquired by the Articulate Assistant Advanced (AAA) software, version 2.16.16 (Articulate Instruments Ltd).

Each word was repeated three times by four speakers and four times by a fifth speaker. The experimental corpus thus includes a total of (7 words × 4 speakers × 3 repetitions) + (7 words × 1 speaker × 4 repetitions) = 112 tokens.

2.2 UTI data preprocessing

Once completed, the recordings were exported from AAA in *.wav format and imported into Praat (Boersma, Weenink, 2016) for the manual segmentation and annotation process. For each stimulus, the target consonant and the preceding vowel were annotated. Phoneme boundaries were established on the basis of the oscillogram and the broadband spectrogram of the acoustic signal. The VOT was included in the consonantal interval.

The annotated acoustic signal was then reimported into AAA for the semi-automatic tracing of the mid-sagittal tongue profiles. A gross recognition of the brightest point of the ultrasound image potentially corresponding to the tongue profile was automatically run in the AAA environment. Then, careful manual correction was carried out frame by frame. The fan set up (i.e., the search area within which the software operates the first gross tongue profile tracking) was customized for each speaker.

The hard palate of each speaker was also tracked according to the same semi-automatic procedure and then superimposed to the lingual profiles for reference. The palate images were obtained from the ultrasound frames relative to the moment of swallowing some water. For each speaker, the palate was traced from its most visible image chosen from different swallowing moments. For Speaker 5, palate location and tracing was problematic, especially in its most anterior region; therefore Speaker 5's productions were evaluated with reference to the rear of the palate only (see below, §3).

The analysis of tongue configuration was done based on average tongue profiles and the area of standard deviation for each relevant context. For instance,

the average tongue profile of /t/ as produced by a given speaker was calculated by averaging the position of all tongue profiles included in each annotated interval of each relevant /t/ produced by that speaker.

2.3 Analysis

The acoustic duration of segments was analysed to verify if the target consonant and the preceding vowel varied in duration as a function of stop's voicing and phonological length (two-sample t-tests for the comparison of the means; SPSS 22.0.0).

The differences in tongue configuration across groups of items were evaluated via inspection of tongue mean and standard deviation profiles for each speaker separately. The analysis of profiles was done in the AAA environment.

3. *Results*

3.1 Acoustic duration

Table 2 reports mean and standard deviation values for the duration of each consonant and the preceding vowel. The effect of voicing was found to be significant for the target consonant, voiceless stops being significantly longer than voiced stops ($t = 2.129$, $p < .05$) consistently with what is generally reported in the literature (see above, §1). The effect of voicing was significant also for the preceding vowel, with longer vowels before voiced stops and shorter vowels before voiceless stops ($t = -2.307$, $p < .05$), consistently with the lengthening-before-voicing effect reported for different languages including Italian (Celata, Calamai, 2011).

Phonological length had the expected effect on the duration of the target consonants ($t = -16.871$, $p < .001$) as well as of the preceding vowel ($t = 5.930$, $p < .001$), vowels before geminates being significantly shorter than vowels before singletons (Bertinetto, 1981).

Table 2 - *Means and standard deviations of the duration of consonants and preceding vowels (in ms) averaged across speakers*

Target Consonant	V		C	
	Mean	St. Dev.	Mean	St. Dev.
/t/	165,9	20,9	110,2	27,0
/d/	186,0	18,8	84,9	13,0
/t:/	133,1	31,6	224,0	28,5
/d:/	151,2	25,6	177,1	19,0

3.2. Tongue profiles

Figure 1 shows the mean tongue profiles and the standard deviation of the tongue for the voiceless-voiced comparison, separately for each speaker and for the two phonetic contexts (after /a/ and /ɔ/). The blue lines represent the tongue profile in voiceless consonants, whereas the red lines represent the voiced consonants. Whenever the two profiles are distant enough to show no overlap of the area defined by upper and lower limits of the standard deviation, the tongue configuration in that region can be considered to be significantly different.

Figure 1 shows that there are significant differences in both phonetic contexts for speakers 1, 2 and 5, whereas the differences are significant in the /a/ context only for speaker 3; speaker 4 does not show any significant difference between voiceless and voiced stops in any of the two contexts. In all of the cases, the tongue profile for producing the voiced consonant is lower than for producing the voiceless consonant; this difference is mostly visible in the regions of tongue dorsum and post-dorsum, and in the predorsum in one case only (speaker 1, /ɔ/ context).

There are no significant differences between voiced and voiceless stops as far as the position of the tongue tip is concerned.

Figure 1 also shows that the constriction location can be different across speakers. Speakers 1 and 2 make a constriction with the the tongue tip approaching the upper part of the alveolar ridge, whereas speakers 3 and 4 show a comparatively more fronted constriction location, with the tongue tip approaching the lowest part of the alveolar ridge, closer to the teeth. The constriction location of Speaker 5 is difficult to ascertain because of the reported problems in tracing the speaker's palate.

Figure 2 shows the tongue mean profiles and the standard deviation of the tongue for the singleton-geminate comparison, again separately for each speaker and for the two phonetic contexts (after /a/ and /ɔ/).

In this graph, the blue lines represent the tongue profile in singleton consonants, whereas the red lines represent the geminate consonants.

According to Figure 2, geminate consonants tend to show a higher overall tongue position compared to singletons, although the difference is significant only in the /ɔ/ context for speakers 1 (tongue dorsum), 2 (dorsum-predorsum) and 3 (dorsum) and in the /a/ context for speaker 4 (mostly in the dorsum and predorsum) and possibly of speaker 5 (the tongue root is more posterior in the geminate than in the singleton).

The position of the tongue tip is unaffected by the phonological length of the consonant, with the potential exception of speaker 2 in the /ɔ/ context, where the geminate appears to have a higher tongue tip compared to the singleton.

The cross-subject differences in constriction location observed in Figure 1 for the voiceless-voiced contrast were consistently observed also in Figure 2 for the singleton-geminate contrast, with Speakers 1 and 2 showing a more retracted constriction location than Speakers 3 and 4.

Figure 1 - *Mean (thick line) and standard deviation (thin line) of voiceless (blue) and voiced (red) target consonants as a function of preceding vowel (columns) and speaker (rows). Black upper lines represent the speaker's palate*

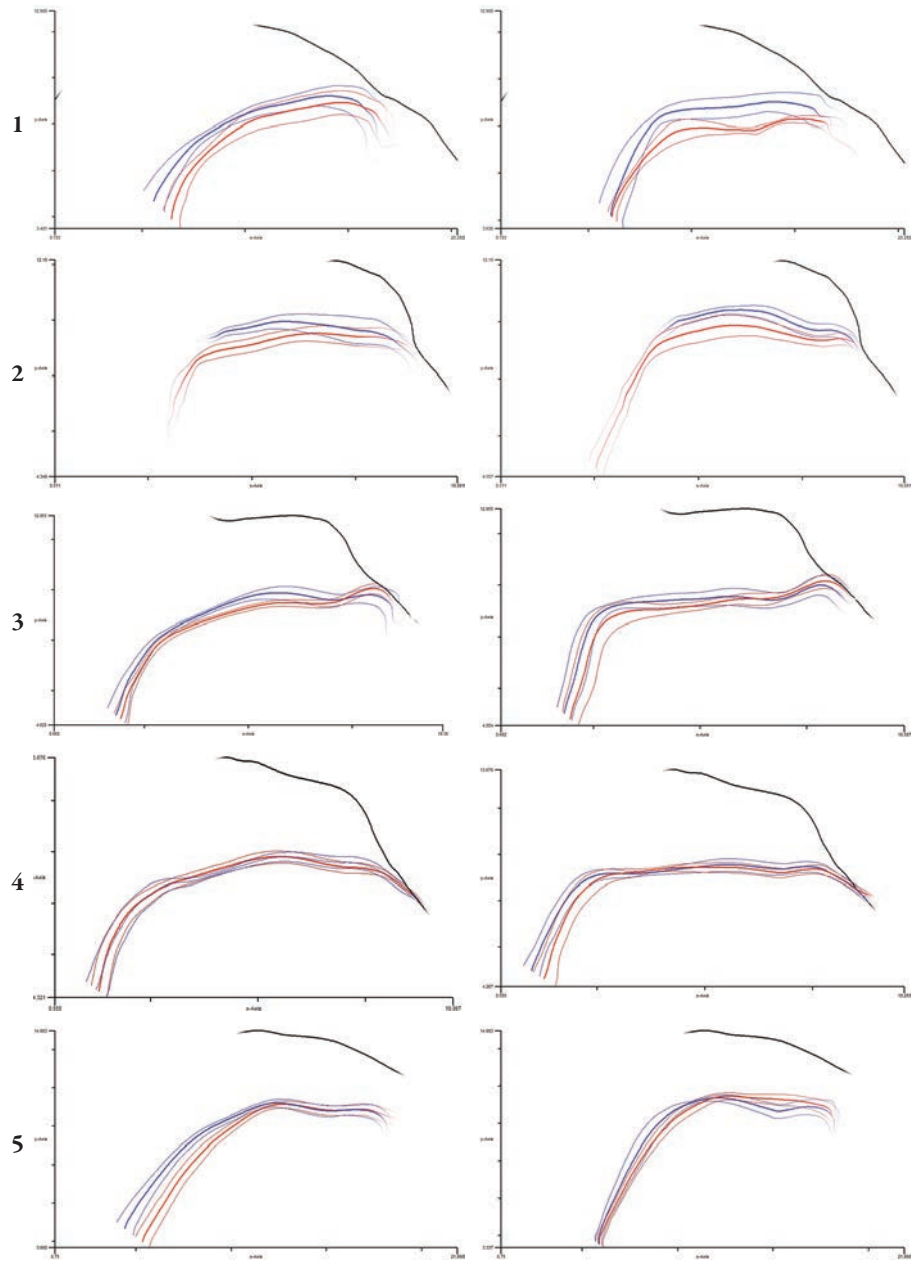
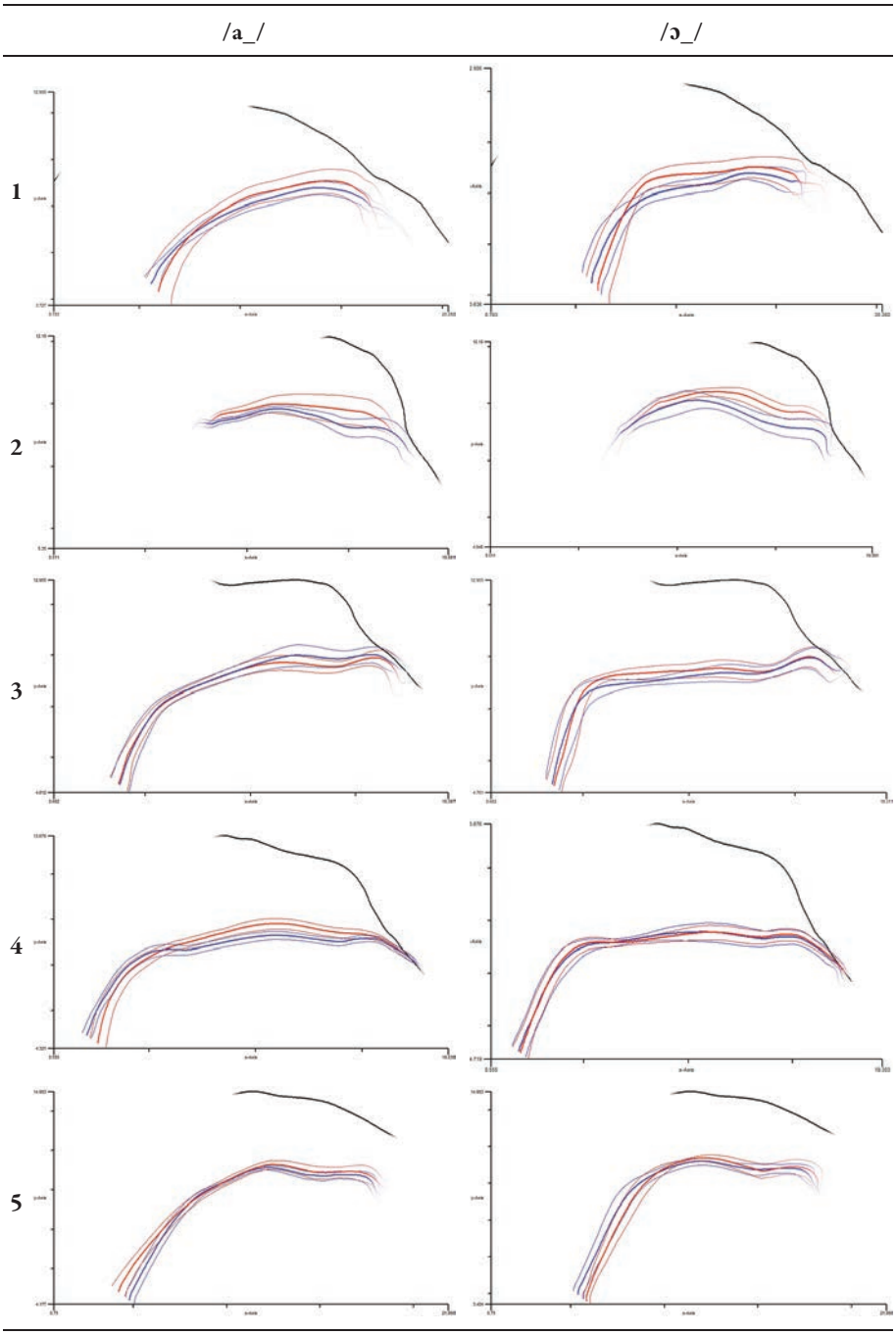


Figure 2 - *Mean (thick line) and standard deviation (thin line) of singleton (blue) and geminate (red) target consonants as a function of preceding vowel (columns) and speaker (rows). Black upper lines represent the speaker's palate*



4. *Discussion*

The findings of the present experiment show that both stop voicing and phonological length may have an impact on the lingual configuration that is used to produce alveolar stops in /a/ and /ɔ/ contexts by Tuscan Italian speakers. These effects are subtle and do not concern the entire profile of the tongue, but only selected portions of it. Nonetheless, they are consistently present in many of the speakers and in both phonetic contexts, thus suggesting that both voicing and length have an impact on tongue configuration in this language. Therefore, this study provide additional evidence in support of the existence of supraglottal consequences of stop voicing as well as of modifications in oral configuration as a function of temporal variations. In the relevant literature, such effects are mostly reported for linguo-palatal contact or articulators' movement (see the relevant literature in §1); by contrast, the present study shows that these effects are also visible through inspection of midsagittal lingual profiles.

With respect to the voicing distinction, the fact that the tongue was found to be lower in voiced stops compared to voiceless stops, particularly as far as tongue dorsum and post-dorsum are concerned, might be consistent with the observation that the jaw is lower and the linguo-palatal contact is less complete in voiced stops compared to voiceless in languages such as English (e.g. Dagenais et al., 1994), German (Fuchs, Perrier, 2003), Swedish (Moen, Simonsen, 1997) or Arabic (Zeroual et al., 2008). In those studies, a lower jaw and tongue tip configuration is interpreted as a strategy to enlarge the oral cavity during the production of a voiced stop, thus preventing intra-oral pressure to rise and vocal fold vibration to extinguish during stop closure. It might be hypothesized that among the consequences of such articulatory manoeuvres to keep voicing during closure is a lowering of the back of the tongue. This lowering might either be passive, i.e. a direct consequence of jaw lowering, or active, i.e. to provide more room for the airflow in the oral cavity and consequently reduce intraoral pressure. However, the current study does not allow answering such a question.

The absence of significant changes in the position of the tongue tip might be interpreted as evidence of the fact that the constriction location of alveolar stops does not change as a function of stop voicing. Under this hypothesis, the Italian data presented here would therefore differ from what is reported for other languages, where voiced and voiceless stops differ in constriction extension and location (e.g. Fuchs, Perrier, 2003 for German; Dagenais et al., 1994 for English). However, it must be recalled that the UTI technique only provides evidence on the midsagittal contour of the tongue and no direct information can be deduced about the contact with the palate or the position of the lateral regions of the tongue. Moreover, the tracing of the tongue tip from ultrasound images is subject to empirical errors because of the potential shadowing of the tongue exerted by the mandibular bone in the most advanced region of the oral cavity (Stone, 2005; Davidson, 2012). Therefore, no conclusive data about the constriction location of alveolar stops can be drawn from UTI analysis. The current data indicate that no visible differences in the tongue

tip raising gesture are consistently related to voicing distinctions. More research, possibly including linguo-palatal contact data, is needed in order to definitively rule out the hypothesis that voicing differences induce a variation in the constriction location of alveolar stops.

The results of the singleton vs geminate analysis further suggested that the constriction location of singletons and geminates is not significantly different, but still with the caveats expressed above about the potential incompleteness of UTI data. However, geminates were found to be variably articulated with a higher tongue dorsum than singletons, especially in the context of a preceding /ɔ/; in some cases, increased tongue height for geminates also extends to predorsum. Taken together, these data only indirectly support the view that geminates are produced with more extended constriction than singletons, as proposed in the context of EPG and EMA studies on different languages (e.g. Kraehenmann, Lahiri, 2008 for Swiss German; Ridouane, 2007 for Tashlhyit Berber; Kochetov, 2012 for Japanese) as well as on Italian (Payne, 2006). However, the findings of the current study are consistent with the view that geminates are articulated with a wider tongue gesture, aiming at an overall higher lingual target (e.g. Gili Fivela et al., 2007).

Finally it is worth noticing that, as usual in articulatory studies, the lingual strategies used to achieve a given phonological target may change across individuals. The speakers of the present sample varied in the exact location (alveodental or alveolar) where the tongue approaches the palate surface, although being internally consistent in that choice across stimuli and phonetic contexts. The speakers also differed in the way they realized the phonological contrasts under investigation, with some of them enhancing both contrasts by means of secondary tongue configuration differences, some others showing enhancement in only one of the two contrasts and/or in a subset of the phonetic contexts, and finally others showing no secondary difference in any of the phonological contrasts.

In conclusion, this study has shown that, in the Tuscan Italian variety investigated here, the voicing distinction in alveolar stops is often associated to changes in lingual configuration involving a lower tongue body in voiced as opposed to voiceless stop. Similarly, the distinction between singletons and geminates can be associated to non-durational variations in tongue configuration involving a higher tongue dorsum in geminates as opposed to singletons. Both results are interpretable as articulatory strategies used to facilitate voicing during closure (in the first case) or as articulatory correlates of increased temporal extension (in the second case). Further analysis will have to clarify if voiceless and geminate stops are produced with a different, possibly more extended, linguo-palatal contact than voiced and singleton stops, respectively, and if changes in articulatory configurations also imply changes in the dynamic properties of lingual gestures.

Acknowledgements

This study has been realized with the financial support of Scuola Normale Superiore di Pisa, grant SNS15C_A to the second author. The study presented here has been jointly developed by the two authors; however, for academic purposes, DD takes responsibility for the segmentation, annotation, pre-processing and analysis of the dataset and for drafting §2 and §3; CC takes responsibility for drafting §1; §4 was jointly written. The contribution of Irene Ricci and Chiara Meluzzi in the recording sessions is also acknowledged.

Bibliography

Armosti, S. (2009). The phonetics of plosive and affricate gemination in Cypriot Greek. PhD dissertation, University of Cambridge.

Articulate Instruments Ltd (2016). *Articulate Assistant Advanced User Guide: Version 2.16.16.* Edinburgh, UK: Articulate Instruments Ltd.

Bertinetto, P.M. (1981). *Strutture prosodiche dell'italiano: accento, quantità, sillaba, giuntura, fondamenti metrici.* Firenze: Accademia della Crusca.

Boersma, P., Weenink, D. (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0. 15 (2016). http://www.praat.org.

Celata, C., Calamai, S. (2011). Timing in Italian VNC sequences at different speech rates. In *Proceedings of 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, 28-31 August 2011.

Celata, C., Meluzzi, C., Moosmueller, S., Hobel, B. & Bertini, C. (2017). The acoustic and articulatory bases of speech timing: a cross-linguistic study. Paper presented at the *XIII Convegno Nazionale AISV*, SNS, Pisa, Italy, 25-27 January 2017.

Dagenais, P.A., Lorendo, L.C. & McCutcheon, M.J. (1994). A study of voicing and context effects upon consonant linguopalatal contact patterns. In *Journal of Phonetics*, 22(3), 225-238.

Davidson, L. (2012). Ultrasound as a tool for speech research. In Cohn, A.C., Fougeron, C. & Huffman, M.K. (Eds.), *The Oxford Handbook of Laboratory Phonology,* Oxford: Oxford University Press, 484-495.

Fuchs, S., Perrier, P. (2003). An EMMA/EPG study of voicing contrast correlates in German. In *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, 3-9 August 2003, 1057-1060.

Fujimoto, M., Funatsu, S. & Hoole, P. (2015). Articulation of single and geminate consonants and its relation to the duration of the preceding vowel in Japanese. In *18th International Congress of Phonetic Sciences,* Glasgow, UK, 10-14 August 2015.

Gili Fivela, B., Zmarich, C., Perrier, P., Savariaux, C. & Tisato, G. (2007). Acoustic and kinematic correlates of phonological length contrast in Italian consonants. In *16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, 6-10 August 2007, 469-472.

Kent, R.D., Moll, K.L. (1969). Vocal Tract Characteristics of the Stop Cognates. In *The Journal of the Acoustical Society of America*, 46(6B), 1549-1555.

Kochetov, A. (2012). Linguopalatal contact differences between Japanese geminate and singleton stops. In *Canadian Acoustics*, 40(3), 28-29.

Kochetov, A. (2014). Voicing and Tongue-Palate Contact Differences in Japanese Obstruents. In *Journal of the Phonetic Society of Japan*, 18(2), 63-76.

Kochetov, A., Kang, Y. (2017). Supralaryngeal implementation of length and laryngeal contrasts in Japanese and Korean. In *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 62(1), 18-55.

Kraehenmann, A., Lahiri, A. (2008). Duration differences in the articulation and acoustics of Swiss German word-initial geminate and singleton stops. In *The Journal of the Acoustical Society of America*, 123(6), 4446-4455.

Löfqvist, A. (2007). Tongue movement kinematics in long and short Japanese consonants. In *The Journal of the Acoustical Society of America*, 122(1), 512-518.

Löfqvist, A., Gracco, V.L. (1994). Tongue body kinematics in velar stop production: Influences of consonant voicing and vowel context. In *Phonetica*, 51(1-3), 52-67.

Matsumura, M., Kimura, K., Yoshino, K., Tachimura, T. & Wada, T. (1994). Measument of palatolingual contact pressure during consonant productions using strain gauge transducer mounted palatal plate. In *Third International Conference on Spoken Language Processing*.

Moen, I., Simonsen, H.G. (1997). Effects of voicing on /t, d/ tongue/palate contact in English and Norwegian. In *5th European Conference on Speech Communication and Technology*.

Ohala, J.J. (1983). The origin of sound patterns in vocal tract constraints. In MacNeilage, P. F. (Ed.), *The Production of Speech*. Heidelberg and Berlin: Springer Verlag, 189-216.

Ohala, J.J. (2011). Accommodation to the aerodynamic voicing constraint and its phonological relevance. In *Proceedings of the 17th International Congress of Phonetic Sciences*, 64-67.

Payne, E.M. (2006). Non-durational indices in Italian geminate consonants. In *Journal of the International Phonetic Association*, 36(1), 83-95.

Ridouane, R. (2007). Gemination in Tashlhiyt Berber: an acoustic and articulatory study. In *Journal of the International Phonetic Association*, 37(2), 119-142.

Ridouane, R., Hallé, P. (2017). Word-initial geminates: from production to perception. In Kubozono, H. (Ed.), *The Phonetics and Phonology of Geminate Consonants* (Vol. 2). Oxford: Oxfrod University Press.

Scobbie, J.M., Wrench, A.A. & van der Linden, M. (2008). Head-Probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement. In *Proceedings of the 8th International Seminar on Speech Production,* 373-376.

Šimko, J., O'Dell, M. & Vainio, M. (2014). Emergent consonantal quantity contrast and context-dependence of gestural phasing. In *Journal of Phonetics*, 44, 130-151.

Skarnitzl, R. (2013). Asymmetry in the Czech alveolar stops: An EPG study. In *AUC Philologica* 1/2014, *Phonetica Pragensia*, 101-112.

Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. In *Clinical linguistics & phonetics,* 19(6-7), 455-501.

Türk, H., Lippus, P. & Šimko, J. (2017) Context-dependent articulation of consonant gemination in Estonian. In *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 8(1): 26, 1-26.

Westbury, J.R. (1983). Enlargement of the supraglottal cavity and its relation to stop consonant voicing. In *The Journal of the Acoustical Society of America*, 73(4), 1322-1336.

Zeroual, C., Esling, J.H. & Crevier-Buchman, L. (2008). The contribution of supraglottic laryngeal adjustments to voice: phonetic evidence from Arabic. In *Logopedics Phoniatrics Vocology*, 33(1), 3-11.

Zmarich, C., Gili Fivela, B., Perrier, P., Savariaux, C. & Tisato, G. (2006). Consonanti scempie e geminate in Italiano: studio acustico e cinematico dell'articolazione linguale e bilabiale. In Giordani, V., Bruseghini, V. & Cosi, P. (Eds.), *Scienze vocali e del linguaggio. Metodologie di valutazione e risorse linguistiche. Atti del III Convegno Nazionale AISV*, Trento, Italy, 29-30 November/11 December 2006, Torriana (RN): EDK, 151-163.

Zmarich, C., Gili Fivela, B., Perrier, P., Savariaux, C. & Tisato, G. (2009). L'organizzazione temporale dei gesti vocalici e consonantici nelle consonanti scempie e geminate dell'italiano. In Romito, L., Galatà, V. & Lio, R. (Eds.), *La Fonetica Sperimentale: Metodo e Applicazioni. Atti del IV Convegno Nazionale AISV,* Arcavacata di Rende (CS), Italy, 3-5 December 2007. Torriana (RN): EDK, 89-104.

DOI: 10.17469/O2104AISV000004

DANIELA MEREU

Parlato spontaneo e stereotipi locali: il sardo parlato a Cagliari

The aim of this paper is to present a sociophonetic research based on spontaneous speech of Cagliari Sardinian. In particular, thanks to this kind of data elicited by means of semi-structured ethnographic interviews, it has been possible to study a local stereotype, the palatalization of /k, g/ in front of /a/, e.g. *cani* ['kʲani] 'dog', *gatu* ['gʲattu] 'cat'. The stylistic analysis allowed to explore the social meanings of the different variants of the variable under investigation, identified during the acoustic analysis.

Keywords: Cagliari Sardinian, spontaneous speech, local stereotypes, sociophonetics, palatalization.

1. *Introduzione*

L'obiettivo di questo lavoro è quello di mostrare una delle possibili modalità di elicitazione e di analisi di variabili sociofonetiche che rappresentano degli stereotipi linguistici in senso laboviano. A tal fine, si presenterà lo studio di una variabile sociofonetica del sardo cagliaritano, la palatalizzazione delle occlusive velari sorda e sonora /k, g/ di fronte a vocale centrale aperta /a/, fenomeno che rappresenta uno stereotipo locale della varietà di sardo parlata a Cagliari.

Per quanto riguarda le metodologie di escussione dei dati, si darà conto di come l'intervista etnografica possa essere considerata un valido strumento per la registrazione di tratti sociofonetici stigmatizzati, mentre sul versante strettamente analitico, considerata la natura marginale degli stereotipi, e quindi anche la loro bassa frequenza d'uso, verrà messo in luce come l'analisi stilistica consenta di approfondire lo studio dei significati sociali di questo tipo di variabili.

Per definizione, uno stereotipo sociolinguistico è un tratto altamente stigmatizzato e soggetto ad aperti commenti da parte dei parlanti nativi. «Stereotypes are referred to and talked about by members of the speech community; they may have a general label, and a characteristic phrase which serves equally well to identify them» (Labov, 1972: 314). Data la loro stigmatizzazione, che ne rende difficile l'elicitazione in sede di raccolta dati, durante la pianificazione della ricerca occorre valutare quali compiti escussivi possano rivelarsi i più adatti per la registrazione del loro uso nel parlato degli intervistati. L'obiettivo del lavoro sul campo in questo tipo di ricerca diventa dunque quello di creare una situazione comunicativa che induca i soggetti intervistati a produrre un parlato dialogico di tipo naturale, ovvero semi-spontaneo.

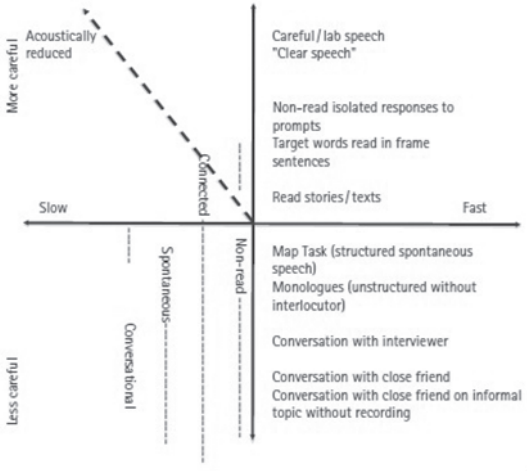
In sociofonetica gli studi sul parlato spontaneo sono di gran lunga inferiori rispetto a quelli incentrati sul parlato di laboratorio, a causa dei numerosi fattori che interferiscono sulla registrazione:

> materials gathered in the field [...] can be difficult because of impaired technical quality, the unpredictable returns of spontaneous data (overlapping speech, the lack of sufficient tokens of the features of interest), and because analysis needs to cater for many potential factors which may influence phonetic forms (Foulkes, Scobbie & Watt, 2010: 728).

Tuttavia, al fine di condurre un'analisi che voglia essere anche sociolinguisticamente fondata è preferibile che una parte della base dei dati sia costituita da interazioni più vicine al parlato spontaneo, rispetto a quelle eseguite in laboratorio. Il parlato spontaneo elicitato rappresenta infatti una finestra dalla quale analizzare i correlati linguistici del parlato di tipo colloquiale[1], sia da un punto di vista fonetico-interazionale, sia da un punto di vista sociolinguistico: da una parte permette di registrare tutti i fenomeni di coarticolazione tipici del parlato di tipo connesso e dall'altra rappresenta quella dimensione in cui possono essere individuati i tratti linguistici connessi ai registri più informali. In altri termini, l'importanza dell'analisi del parlato spontaneo è da ricondurre sia al piano fonetico, e quindi al livello linguistico di analisi di riferimento, sia all'ambito sociolinguistico, ovvero, più propriamente, al punto di vista teorico di riferimento. Nello specifico, rispetto al primo punto, esso dà conto dei fenomeni di riduzione e più in generale di ipoarticolazione tipici del parlato connesso[2], mentre dal punto di vista sociolinguistico, a livello stilistico, si caratterizza per l'informalità e la trascuratezza.

Un'articolata discussione sul parlato spontaneo e su come questo si posizioni rispetto al grado di attenzione al parlato, alla velocità di elocuzione e alla riduzione fonetica è presente in Warner (2012: 624) ed è sintetizzata in uno schema che rende conto di come queste tre dimensioni separate si intersechino nel parlato (Fig. 1).

Figura 1 - *Rappresentazione schematica delle tre dimensioni (attenzione al parlato, velocità di eloquio e riduzione acustica) entro le quali il parlato può ricadere (Warner, 2012: 624)*



[1] Per una distinzione tra correlati linguistici funzionali e correlati sociolinguistici del parlato si rimanda a Voghera (2017: 35).

[2] Per il parlato ipoarticolato si veda Savy, Cutugno (1997).

> Spontaneous speech often includes sequences with such strong reduction phenomena that one could never have predicted them and is rather surprised to see them when one examines the spectrogram [...] One could establish a continuum of carefulness or naturalness. On one end might be vowels or nonsense monosyllables read in isolation [...]. At the other end might be informal conversation among family or friends, perhaps at home with no microphone present (Warner, 2012: 622).

Inoltre, l'autrice sottolinea che le etichette di parlato veloce e ridotto non si posizionano sullo stesso *continuum* dell'attenzione al parlato: il parlato informale, letto o sorvegliato possono essere veloci o lenti e la velocità di elocuzione può essere misurata acusticamente, diversamente dalla spontaneità (Warner, 2012: 623).

In questa sede il parlato spontaneo verrà sfruttato per le sue potenzialità sociolinguistiche più che per quelle fonetiche, nel senso che l'obiettivo che si intende perseguire non è quello di dar conto dei fenomeni di ipoarticolazione presenti nei dati raccolti, sebbene talvolta si farà loro riferimento, ma di studiare la variabilità stilistica del fenomeno oggetto di interesse. In particolare, si mostrerà come il ricorso al parlato spontaneo permetta sia di registrare i tratti sociofonetici stigmatizzati sia di fornire una base di dati adatta per lo studio della variabilità intra-parlante.

Al fine di condurre lo studio dei *pattern* di variazione *intra-speaker* sono stati presi in considerazione due livelli di analisi differenti: un livello basato sul *topic* di discussione e l'altro su parametri di tipo conversazionale. Nel primo caso si è fatto riferimento al criterio della suddivisione delle interviste per argomento e micro-genere salienti per ogni specifico gruppo di parlanti, identificati grazie all'analisi etnografica, che ha rivolto particolare attenzione a una confraternita religiosa cagliaritana, l'*Arciconfraternita della Solitudine,* associazione che si occupa principalmente dell'organizzazione dei riti della Settimana Santa. Tra i *topic*/micro-generi individuati nelle interviste degli informanti appartenenti a questo sodalizio citiamo per esempio: "storia della confraternita", "descrizione dei riti religiosi" e "rivalità tra confraternite"; mentre per il gruppo dei soggetti non appartenenti a essa alcune delle categorie identificate sono: "problemi sociali del quartiere", "descrizione delle feste in città" e "lamentela".

I lavori pionieristici sul modo in cui il cambio dell'argomento di conversazione possa influenzare la variazione del dettaglio fonetico (*fine-grained phonetic topic shifts*) hanno riguardato inizialmente l'uso delle diverse varianti in relazione alla scelta di codice tra diverse varietà linguistiche. Rientrano in questo settore gli studi di Ervin-Tripp (1964) e Blom & Gumperz (1968), ma anche i più recenti Becker (2009) e Mendoza-Denton, Hay & Jannedy (2003). Meno frequenti sono invece le ricerche che si focalizzano sulla variazione stilistica in funzione del *topic* nell'ambito di un'unica varietà linguistica. Alcuni esempi in questa direzione sono gli studi di Lawson (2009), Love, Walker (2012) e Hay, Foulkes (2016), che hanno evidenziato l'attivazione di una determinata variante sociofonetica in specifici argomenti di discussione.

Per quanto riguarda il secondo livello di analisi, si è ricorso invece all'individuazione di categorie stilistiche basate su indizi di tipo conversazionale. In questo modo, l'approccio di tipo etnografico è stato integrato con un altro di carattere conversazionale più generale. Nello specifico, nel parlato dei soggetti intervistati sono state identificate due diverse categorie sulla base di elementi di tipo conversazionale-interazionale, così come suggerito da Milroy (1980), la quale distingue nel parlato delle sue interviste tra *Interview Style* e *Spontaneous Style*. Il primo stile si definisce: a) per una chiara struttura a due parti (intervistatore che pone la domanda e intervistato che risponde); b) per il fatto che uno dei due partecipanti, ovvero l'intervistatore, detiene il controllo dell'interazione, nel senso che sceglie in che modo elicitare il parlato e seleziona gli argomenti di discorso. Viceversa, lo stile spontaneo si contraddistingue per la mancanza di una chiara struttura del discorso a due parti. In questo caso, i segmenti di parlato spontaneo non possono facilmente essere analizzati come risposte date all'intervistatore sugli argomenti introdotti. Sulla base di questi criteri, nel nostro *corpus* quei segmenti che costituivano delle chiare risposte alle domande poste dalla ricercatrice e nelle quali veniva mantenuto costante l'argomento proposto sono stati fatti rientrare nell'*Interview Style*, mentre le digressioni, le battute di spirito, i dialoghi con gli altri partecipanti all'intervista, gli aneddoti, ovvero tutte quelle parti in cui la struttura canonica dell'intervista è venuta a mancare sono stati considerati come *Spontaneous Style*[3]. Un'analisi di questo tipo risponde alla necessità di dare conto anche degli aspetti legati alla struttura del discorso, in modo da verificare il peso dei diversi *topic* di riferimento in relazione anche alle diverse parti dell'intervista, intesa come un'interazione tra due partecipanti. In particolare, l'obiettivo che quest'analisi stilistica si propone di perseguire consiste nel valutare se la variante socialmente marcata verso il basso ricorra maggiormente in particolari *topic* o se la sua distribuzione sia regolata (o almeno influenzata) dalla struttura del discorso, nel senso che, a prescindere dall'argomento di conversazione, essa sia comunque più frequente nel parlato più spontaneo (*Spontaneous Style*) rispetto al parlato più sorvegliato (*Interview Style*).

La varietà linguistica di riferimento per questo lavoro è il sardo cagliaritano, dialetto che negli studi di linguistica sarda non ha ricevuto grande attenzione, in quanto varietà urbana (vd. Mereu, 2017) e ritenuta pertanto, secondo i paradigmi della dialettologia tradizionale, di minore interesse rispetto ai dialetti rurali perché più soggetta agli influssi linguistici esterni. Il dialetto cagliaritano rappresenta al momento attuale una varietà a forte rischio di estinzione (Loporcaro, Putzu, 2013: 205), visto che i suoi parlanti sono molto pochi non solo tra le fasce più giovani, ma anche tra quelle più anziane della popolazione[4].

[3] Queste due categorie corrispondono a ciò che Labov (1966) ha identificato come *careful speech* e *casual speech*.

[4] Per una descrizione della situazione sociolinguistica a Cagliari si rimanda a Mereu (2018).

2. *La palatalizzazione di* /k, g/ *di fronte ad* /a/

Come anticipato, la variabile sociofonetica oggetto di analisi è la palatalizzazione di /k, g/ di fronte ad /a/, es. *cani* ['kʲani] 'cane', *gatu* ['gʲattu] 'gatto'.

Il termine 'palatalizzazione', in ambito fonetico, è usato per indicare due fenomeni distinti (Spinu, 2007: 303):

1. l'effetto delle vocali anteriori o dell'approssimante [j] sulla consonante precedente, che determina il cambiamento del luogo di articolazione;
2. la presenza di un'articolazione secondaria caratterizzata dall'innalzamento del corpo della lingua verso la parte anteriore del palato (Ladefoged, Maddieson, 1996: 355).

Il tipo di palatalizzazione esaminato in questa sede può essere definita un'articolazione secondaria e, nello specifico, un tipo di palatalizzazione 'di transizione' (*transitional palatalization*), in quanto la costrizione dell'articolazione basica è rilasciata mediante un'approssimazione palatale della punta e della lamina della lingua, come parte della transizione al segmento seguente. La palatalizzazione di transizione si differenzia da quella 'simultanea' (*simultaneous palatalization*) per il fatto che in quest'ultima la modificazione della posizione della lingua avviene contemporaneamente agli altri gesti articolatori del segmento.

Dato che questo processo articolatorio coinvolge sia il segmento consonantico sia quello vocalico, dal punto di vista acustico la palatalizzazione può essere studiata prendendo in considerazione parametri consonantici e vocalici. Ní Chiosáin, Padgett (2012), nella loro ricerca sulla palatalizzazione nella varietà di irlandese del Connemara, individuano alcune fondamentali proprietà acustiche che riflettono le condizioni articolatorie del fenomeno. I principali indizi acustici sono rappresentati da un valore più alto della frequenza della transizione formantica consonante-vocale, una durata più lunga, un'intensità maggiore e un centro di gravità (CoG) del rilascio consonantico più alto.

Per quanto riguarda lo studio di questa variabile nel sardo cagliaritano, le poche informazioni disponibili riguardano il versante dialettologico. Ugo Pellis registrò questo fenomeno nelle sue indagini condotte in Sardegna per l'*Atlante Linguistico Italiano* (ALI). Come tutte le inchieste dialettologiche del periodo, anche questa si basava su pochissimi informatori (due). Pellis (1934: 60) scrive:

> Ho notato pure che non di rado, in ispecie davanti ad *a*, il suono gutturale (*k*, *g*) tende verso il palatale (quasi *k'*, *g'*): *baska* (caldo, afa) varia con *bask*^{a}, *bask*^{e}, *bask*^{ɨ}. Accanto a *bukka* (bocca) si ha *bukk*^{a}, *buk'k*^{a}; acc. a *grò̀ga* (gialla) si ha *grò̀g*$^{'a}$; accanto a *tokkà̀ŭ* (toccato) c'è *tok'kà̀ŭ*; *čirkanta* (cercano): *čirk'ant*^{a}; *mù̀sk*^{a} (mosca): *mù̀sk*$^{'a}$; *su-*^{g}*à̀ne* (il cane): *su-gà̀ne*; ecc. ecc.

Virdis (1978, 2013) attribuisce in modo più preciso la palatalizzazione delle velari di fronte ad /a/ alla parlata popolare di Cagliari. Quanto alle indagini di carattere sociolinguistico disponibili, questo fenomeno è stato preso in esame anche nel recentissimo studio sul repertorio cagliaritano condotto da Rattu (2017), che si è

basato su un *corpus* raccolto prevalentemente mediante lo strumento del questionario[5]. Nelle risposte ai questionari da parte degli informanti questo tratto non è stato rilevato (se non in un unico caso e in seconda risposta a una domanda posta dal raccoglitore). Tuttavia, risulta molto interessante notare come invece esso sia stato colto durante le conversazioni libere. Come sottolinea l'autore, la quasi totale assenza di questo tratto trova ragione nel tipo di inchiesta svolta e, in modo particolare, nell'argomento sul quale il questionario era incentrato, ovvero la lingua sarda, che ha determinato un maggiore grado di attenzione e controllo da parte dell'informatore nei confronti della lingua (Rattu, 2017: 162-163). Tali risultati confermano come lo studio di un tratto dotato dello statuto di stereotipo necessiti del ricorso a uno strumento di escussione diverso da quello del questionario sociolinguistico che si dimostri capace di generare un tipo di parlato colloquiale.

3. *Metodologia della raccolta dati*

Considerato che l'obiettivo della raccolta dati era quello di riuscire a elicitare un tipo di parlato quanto più vicino a quello spontaneo, la tecnica di elicitazione ritenuta più adatta è stata quella dell'intervista etnografica semi-strutturata (Vietti, 2003; Abete, 2012). Si tratta di una forma di intervista flessibile che lascia spazio agli intervistati di parlare liberamente, ma che prevede che sia l'intervistatore a indurre l'intervistato ad affrontare una serie di argomenti di carattere etnografico. Più precisamente, al fine di ottenere un parlato semi-spontaneo sono stati impiegati alcuni espedienti, tra cui citiamo: un certo grado di insincerità da parte della ricercatrice (Vietti, 2003), che ha rivelato il fine linguistico dell'intervista solamente alla fine della registrazione, in modo che il soggetto intervistato si concentrasse sull'argomento di discussione e non sul proprio modo di parlare; domande incentrate su argomenti di interesse per gli informanti, che hanno permesso di registrare interviste molto lunghe (che durano anche due ore e mezza, nel caso delle interviste di gruppo); conversazioni di gruppo, che si sono rivelate i contesti ideali per la documentazione di dati molto ricchi dal punto di vista della variazione sociofonetica; la posizione dell'intervistatrice come apprendente (Labov, 1984).

Gli argomenti scelti per questa ricerca sono stati distribuiti in due diversi modelli di intervista: il primo, somministrato al gruppo di parlanti appartenenti alla confraternita religiosa, comprendeva prevalentemente domande relative al sodalizio confraternale a cui appartenevano, mentre il secondo, riservato a un gruppo di parlanti non riconducibili a questa confraternita, è stato incentrato su argomenti riguardanti la vita nel quartiere di residenza dei parlanti, i problemi sociali della

[5] In particolare, il questionario citato si componeva di due parti. La prima sezione, di tipo macro-sociolinguistico, era volta alla raccolta dei dati socio-demografici sui parlanti e di informazioni relative ad autovalutazioni riguardanti le competenze e l'uso del sardo nei diversi contesti, mentre la seconda parte, di carattere micro-sociolinguistico, era finalizzata all'elicitazione di produzioni linguistiche prevalentemente mediante l'uso di traduzioni.

città e così via[6]. I dati sono stati registrati con un registratore Zoom H5, con una campionatura a 44.100 Hz e la digitalizzazione a 16-bit. Il *corpus*, raccolto tra il 2015 e il 2016 a Cagliari e costituito da circa 10 ore di parlato, è composto da interviste singole e di gruppo. I materiali raccolti sono stati annotati mediante il *software* Elan con una trascrizione di tipo ortografico. Inoltre, al fine di svolgere l'analisi stilistica, i dati sono stati etichettati anche in funzione dell'argomento trattato e della tipologia stilistica. Questo lavoro di annotazione ha permesso in fase di analisi di osservare la distribuzione delle varianti in base alle categorie stilistiche individuate. Nel complesso, i parlanti analizzati sono 13 (9 uomini e 4 donne); tuttavia, al fine di interpretare al meglio i risultati derivanti dall'analisi stilistica, come è stato già anticipato, i soggetti intervistati sono stati suddivisi in due gruppi: un gruppo appartenente alla confraternita e l'altro esterno a essa. Il numero ridotto dei locutori riflette le difficoltà incontrate nel reperimento dei parlanti durante la raccolta dati, difficoltà che a loro volta possono essere considerate diagnostiche della poca vitalità di cui oggi gode il sardo a Cagliari.

Considerato che l'obiettivo della ricerca era anche quello di documentare la varietà sarda cagliaritana e constatato che senza un'indicazione specifica gli informanti tendevano a utilizzare l'italiano come lingua per l'interazione, è stato esplicitamente chiesto ai soggetti intervistati di parlare in sardo. Il codice prevalentemente usato nelle conversazioni registrate è quindi il sardo, con presenza di alternanze di codice da sardo a italiano[7].

4. *Analisi dei dati*

4.1 Analisi acustica

Di seguito si riportano due esempi tratti dal *corpus* di riferimento. La Figura 2 rappresenta lo spettrogramma della realizzazione di *s'America* [s aˈmɛrikʲa] 'l'America', prodotta dal parlante LaVM57[8].

È possibile osservare sin da subito che la seconda formante della vocale [a] risulta molto alta (circa 2300 Hz), indice della presenza del fenomeno di palatalizzazione. Analoga considerazione può essere fatta per il secondo esempio spettrografico (Fig. 3), prodotto dalla parlante VF49, *gana* [ˈgʲana] 'voglia', in cui l'*onset* della seconda formante di [a] si posiziona a circa 2500 Hz.

[6] Una descrizione dettagliata delle metodologie impiegate per la raccolta dati è presente in Mereu (2018).

[7] I fenomeni di alternanza di codice non sono stati presi in considerazione per questo lavoro, ma saranno oggetto di studio nel prossimo futuro.

[8] Le etichette associate ai soggetti sono composte da una sigla relativa al quartiere di appartenenza (C: Castello; V: Villanova; M: Marina; LaV: LaVega; SA: Sant'Avendrace; IsM: Is Mirrionis; SE: Sant'Elia; SB: San Benedetto; S: Stampace), seguita dal sesso e dall'età. Per motivi di economia di spazio e considerato il fine specifico del presente studio, non si ritiene necessario qui dare conto di ulteriori informazioni relative all'occupazione lavorativa e al grado di istruzione dei parlanti. Tuttavia, per un quadro macrosociolinguistico completo sul campione dei soggetti intervistati si rimanda a Mereu (2018).

Figura 2 - *Spettrogramma di* s'America *[s aˈmɛɾikja] 'l'America' prodotto dal parlante LaVM57*

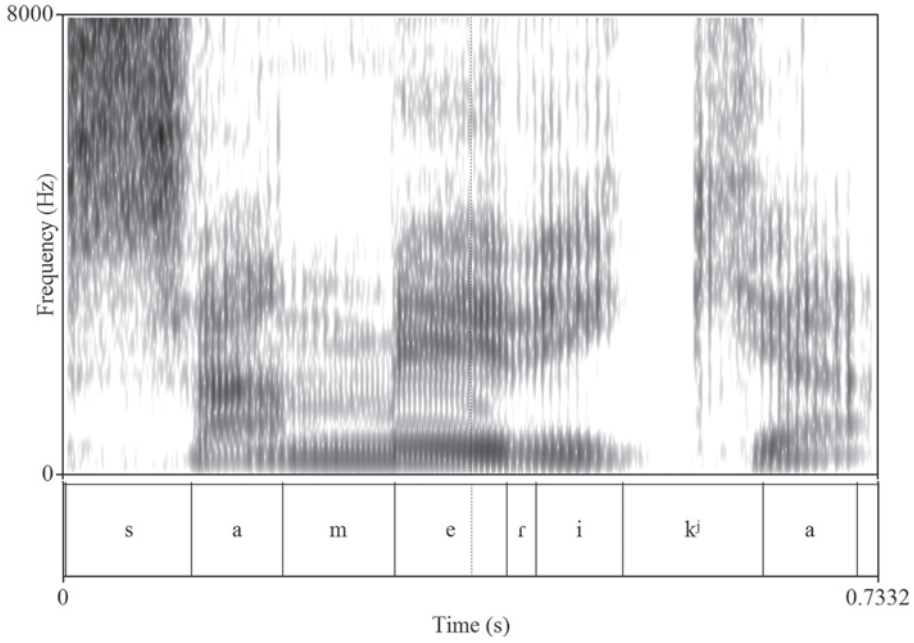
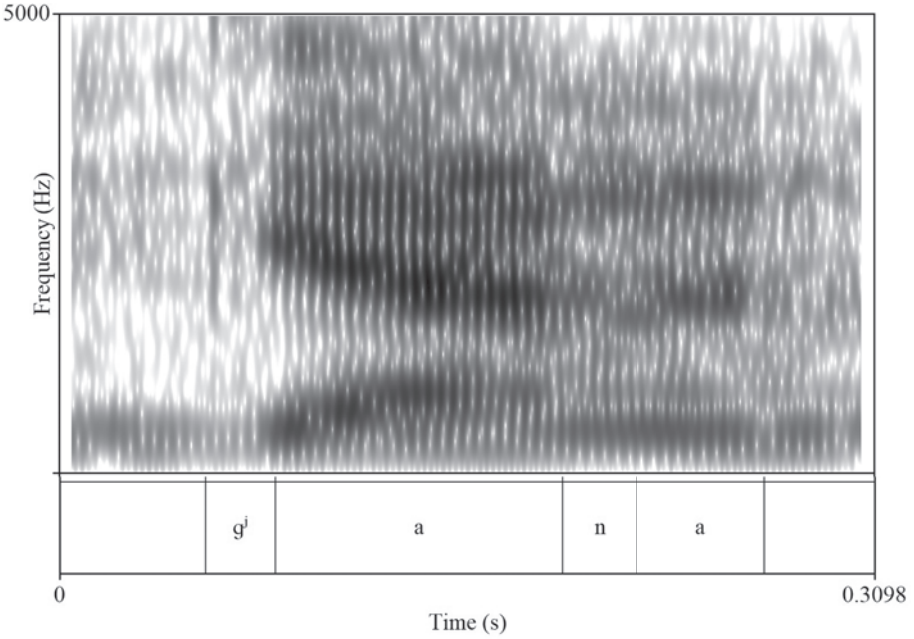


Figura 3 - *Spettrogramma della realizzazione di* gana *[ˈgʲana] 'voglia' prodotta dalla parlante VF49*



Tutte le occorrenze (1704) sono state segmentate manualmente con Praat (Boersma, Weenink, 2017) su base spettroacustica. Per poter valutare con precisione e in modo completo il fenomeno fonetico, sono state prese in esame le occorrenze contenenti i suoni /k, g/ seguiti non solo da /a/, ma anche da tutte le altre vocali, /e, i, o, u/. In questo modo si è potuto verificare l'eventuale coinvolgimento di altre vocali nel processo fonetico. Per quanto riguarda la segmentazione, per ogni file sonoro, sulla *TextGrid* di Praat sono stati creati due diversi segmenti, uno per la consonante e uno per la vocale, visto che le informazioni acustiche dovevano essere estratte sia dal segmento vocalico sia da quello consonantico. Tutti i *tokens* sono stati etichettati su base sia percettivo-uditiva sia spettrografica. Occorre tuttavia ricordare che, in questa varietà di sardo, /k, g/ sono soggette a processi di lenizione: in particolare, /k/ al confine di parola e in contesto intervocalico subisce lenizione e viene quindi realizzata come una fricativa velare sonora [ɣ], mentre /g/ subisce lo stesso processo quando si trova in posizione interna di parola, sempre in contesto intervocalico. Pertanto, in questi particolari contesti, al posto di un'occlusiva viene realizzata una fricativa [ɣ]. Tuttavia, l'etichettatura, oltre alle varianti attese, ha fatto emergere numerosi altri allofoni. Per questo motivo, in una prima fase si è resa necessaria un'etichettatura molto dettagliata, al fine di esplorare altre eventuali varianti sociofonetiche non previste. Con questo procedimento attento al minimo dettaglio fonetico sono state individuate 11 diverse varianti di /k, g/: [k], occlusiva velare sorda; [g], occlusiva velare sonora; [ɣ], fricativa velare sonora; [g^{j}], occlusiva velare sonora palatalizzata; [k^{h}], occlusiva sorda aspirata; [ɣ^{j}], fricativa velare sonora palatalizzata; [k^{j}], occlusiva velare sorda palatalizzata; [k^{hj}], occlusiva velare sorda aspirata palatalizzata; [kx], affricata velare sorda; [kx^{j}], affricata velare sorda palatalizzata; [x], fricativa velare sorda. La ricchezza allofonica registrata è da attribuire alla natura del parlato registrato, dialogico e semi-spontaneo, caratterizzato in alcune sue parti da alta velocità di eloquio e da una considerevole presenza di fenomeni di riduzione fonetica e assimilazioni. Per semplificare l'analisi, considerato che non sono emerse evidenze che facevano ipotizzare che la maggiore aspirazione o affricazione potesse rappresentare un indice sociofonetico, si è deciso di raggruppare alcuni degli allofoni in classi di suoni, per cui [k], [k^{h}], [kx] sono state fatte rientrare nella categoria [k], mentre [k^{j}], [k^{hj}], [kx^{j}] sono state considerate come [k^{j}].

Prima di focalizzare l'attenzione sull'analisi acustica, è bene sottolineare che mentre la segmentazione e l'etichettatura sono state condotte sulle occorrenze contenenti le velari di fronte a tutte le vocali, l'analisi acustica e quella sociolinguistica, invece, hanno coinvolto solamente le produzioni del fenomeno nei contesti di fronte ad /a/. Le fasi propriamente analitiche hanno riguardato quindi un numero minore di occorrenze, 544, di cui solamente 30 rappresentano le varianti palatalizzate, mentre le varianti non palatalizzate sono 514.

Seguendo il metodo di analisi di Ní Chiosáin, Padgett (2012), dal segmento consonantico sono stati estratti i valori del CoG e della durata del rilascio conso-

nantico, mentre per la vocale i valori dell'*onset* di F2[9]. In particolare, le varianti palatalizzate dovrebbero essere caratterizzate da un CoG del rilascio consonantico più alto, una durata maggiore del rilascio consonantico e un valore più alto dell'*onset* di F2 della vocale successiva. Nel caso delle realizzazioni fricative, data la loro natura, sono stati estratti solo i valori vocalici. Analogamente, anche nel caso di [g] non è stato possibile analizzare i parametri consonantici perché per le occlusive sonore il concetto di rilascio consonantico è controverso (cfr. Nakai, Scobbie, 2016; Stuart-Smith, Sonderegger, Rathcke & Macdonald, 2015).

Partendo dai fondamentali studi sull'argomento (Lisker, Abramson, 1964; Cho, Ladefoged, 1999), in questa sede si è deciso di considerare il *Voice Onset Time* (VOT) come l'intervallo di tempo che intercorre tra il primo chiaro segnale di rilascio consonantico e il primo segno di periodicità, sulla base della forma d'onda congiuntamente all'informazione spettrografica (cfr. Nakai, Scobbie, 2016). Visto che non sono stati presi in considerazione i VOT negativi, l'analisi ha compreso solamente le occlusive sorde. Le misurazioni, quando possibile, sono state prese facendo riferimento all'oscillogramma e nei punti di ampiezza con valore zero. Quanto alla parte vocalica, l'*offset* della vocale è stato fatto coincidere con la fine del segnale di periodicità, ovvero con l'ultimo ciclo periodico.

Prima dell'estrazione dei valori acustici, al fine di rimuovere gli effetti di sonorità, è stato applicato un filtro passa alto (*high pass filter*). Poiché il *corpus* comprendeva suoni sia occlusivi sia fricativi, e visto che per le fricative il filtro generalmente applicato è di 500 Hz (Munson, 2001; Stuart-Smith, 2007), mentre per le occlusive si fa generalmente ricorso a un filtro di 200 Hz (vd. Sundara, 2005; Chodroff, Wilson, 2014), si è optato per un filtro di 300 Hz.

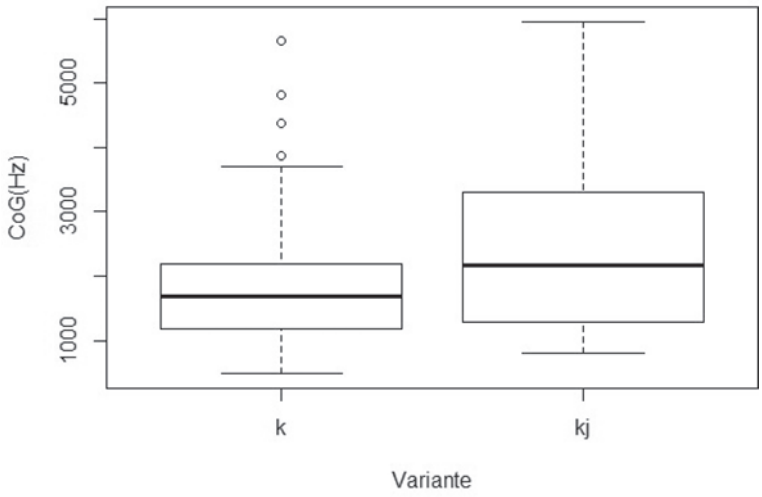
Per quanto riguarda la parte consonantica, come anticipato, i valori estratti sono stati il CoG e la durata del rilascio consonantico. Per l'estrazione dei valori del CoG è stato utilizzato lo *script* di Christian Di Canio (2013)[10]. Tuttavia, in ragione della durata molto variabile dei segmenti, che comprende intervalli molto brevi per le occlusive e segmenti più lunghi per le realizzazioni fricative, lo *script* è stato modificato in modo da poter essere utilizzato per il nostro *corpus* di dati. In particolare, è stata ridotta la larghezza della finestra d'analisi (10 ms) e il numero delle finestre prese in esame (3).

Per esigenze di semplificazione, riportiamo di seguito solamente i risultati dell'analisi del CoG (Fig. 4) delle varianti già accorpate in categorie. Dato che si tratta dei valori estratti dal rilascio consonantico, le varianti esaminate sono solamente le occlusive. Inoltre, considerato che le varianti palatalizzate sono state registrate quasi esclusivamente tra gli informanti uomini (solo un'occorrenza è stata attestata nel gruppo delle donne), i risultati delle analisi sono circoscritti solamente a questo gruppo.

[9] Lo studio citato ha compreso anche l'intensità del rilascio consonantico. Tuttavia, considerato il tipo di parlato esaminato in questa sede, si è preferito non includere questo parametro.

[10] http://www.acsu.buffalo.edu/~cdicanio/scripts/Time_averaging_for_fricatives_2.0.praat.

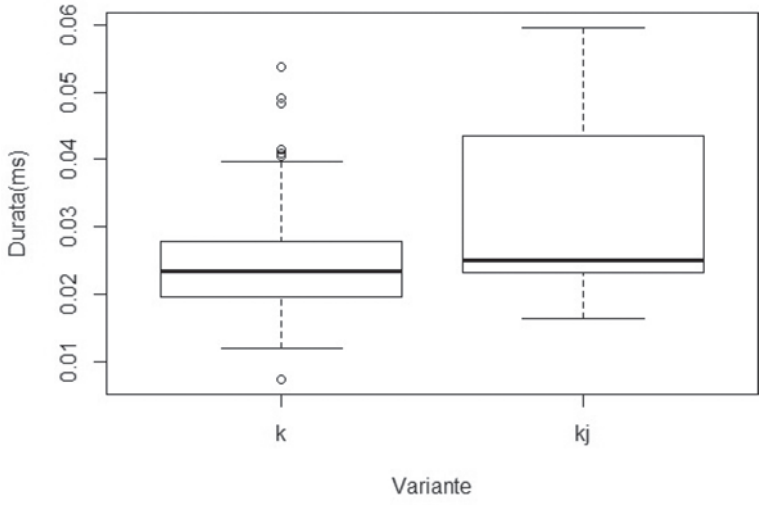
Figura 4 - *Boxplot raffigurante i risultati dei valori del CoG per le varianti semplificate*



La differenza dei valori del CoG tra [k] e [kʲ] è risultata statisticamente significativa al t-test (t = 2.0353, df = 15.748, p-value < 0.05).

Dal *boxplot* riportato di seguito (Fig. 5), che mostra i risultati ottenuti dall'analisi della durata del rilascio consonantico, possiamo osservare come anche i valori di questo secondo parametro analizzato rispecchino le attese.

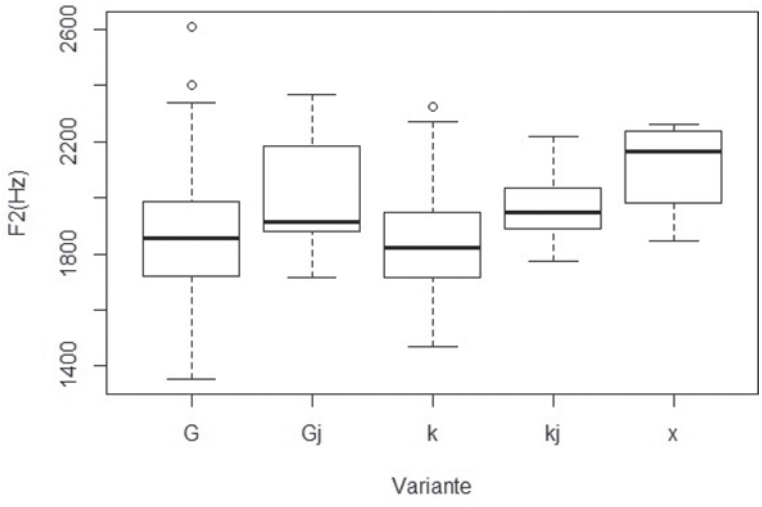
Figura 5 - *Boxplot raffigurante i risultati dei valori della durata del rilascio consonantico per le varianti semplificate*



Anche in questo caso, un t-test ha dimostrato che la differenza tra la durata del rilascio consonantico di [k] e quella di [kʲ] è statisticamente significativa (t = 2.3268, df = 15.688, p-value < 0.05).

L'ultimo parametro esaminato è l'*onset* di F2, ovvero il segmento iniziale della seconda formante della vocale successiva all'occlusiva (o fricativa, nel caso in cui la produzione sia il risultato della lenizione dell'occlusiva). I valori della seconda formante, estratti utilizzando un altro *script* di Di Canio (2012)[11], sono stati misurati nel punto del segmento che costituisce il 20% dell'intero intervallo. Dal grafico (Fig. 6) è possibile constatare che i valori estratti dall'*onset* di F2 confermano l'individuazione delle varianti palatalizzate come tali. In particolare, è possibile notare come il gruppo *G* (che corrisponde a [ɣ]) e il gruppo *k* (che comprende [k, kʰ, kx]) mostrino valori più bassi dell'*onset* di F2 rispetto alle loro controparti palatalizzate, *Gj* ([ɣʲ]) e *kj*, [kʲ, kʰʲ, kxʲ] rispettivamente.

Figura 6 - *Boxplot raffigurante i risultati dei valori dell'onset di F2 per le varianti semplificate*

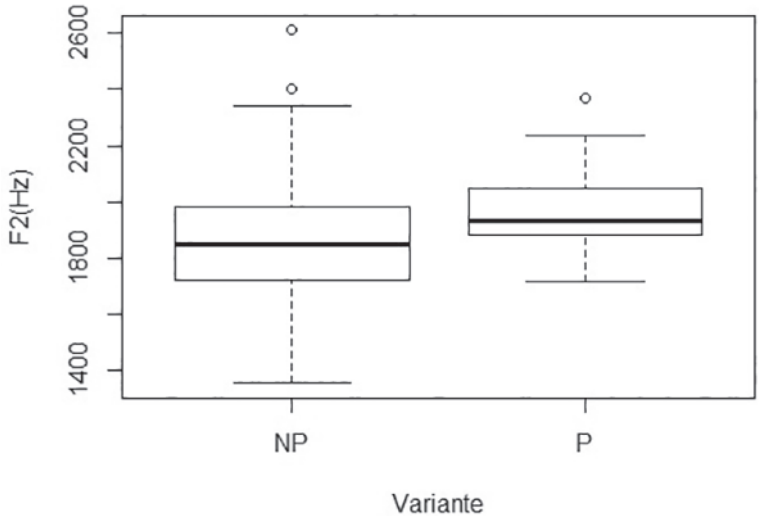


In questo caso, visto che questo parametro si riferisce al segmento vocalico, le varianti possono essere ulteriormente accorpate in due gruppi: realizzazioni palatalizzate e non palatalizzate. Il risultato dell'accorpamento è dato nel *boxplot* in figura 7.

Un t-test ha confermato che la differenza tra i valori dell'*onset* di F2 delle varianti palatalizzate e quelli delle varianti non palatalizzate è estremamente significativa (t = 3.644, df = 35.514, p-value < 0.001). Alla fine dell'analisi acustica siamo pertanto in grado di affermare che tutti e tre i correlati acustici della palatalizzazione hanno risposto alle attese anche da un punto di vista statistico.

[11] http://www.acsu.buffalo.edu/~cdicanio/scripts/Vowel_Acoustics.praat.

Figura 7 - *Boxplot raffigurante i risultati dei valori dell'onset di F2 per le varianti palatalizzate (P) e non palatalizzate (NP)*



4.2 Analisi dei vincoli linguistici

Per quanto riguarda l'analisi dei contesti linguistici, è possibile fare solo delle considerazioni provvisorie a causa del numero di occorrenze molto basso. Innanzitutto, occorre fare presente che, grazie all'esame dei contesti relativi anche alle altre vocali, /e, i, o, u/, è stato possibile registrare alcuni casi di palatalizzazione anche di fronte alla vocale /e/, per es. *piciocheddu* [pitʃɔ'kʲɛɖɖu] 'ragazzino' e un solo caso di fronte a /u/: *cumenti* [kʲu'mɛnti] 'come'.

Focalizzandoci ora sulla palatalizzazione di /k, g/ di fronte ad /a/, dalle analisi svolte possiamo mettere in luce qualche tendenza, seppur blanda. In primo luogo, il fenomeno è favorito dalla posizione interna di parola rispetto a quella esterna e più specificatamente dal contesto fonologico C.CV, es. *conca* 'testa', seguito poi da V.CV, es. *pigamus* 'prendevamo'[12].

Per quanto riguarda la struttura e la posizione della sillaba in relazione all'accento lessicale, la palatalizzazione è leggermente favorita dalla sillaba aperta e, in misura ancora più significativa, dalla sillaba tonica. I test esatti di Fisher mettono in risalto valori non significativi per la prima variabile (struttura sillabica) e valori significativi (p-value < 0.05) per l'accento lessicale. Come è stato sottolineato, per il momento si tratta di tendenze che, seppure interessanti da segnalare, non possono certamente costituire delle generalizzazioni sui vincoli linguistici che regolano la realizzazione di questa variabile.

[12] Per ragioni di semplificazione dei risultati da questo momento in poi, per consentire una chiara rappresentazione grafica dei *pattern* della variabile e dei suoi vincoli, tutte le varianti palatalizzate saranno raggruppate in un'unica categoria etichettata con P, mentre le varianti non palatalizzate saranno indicate con NP.

4.3 Analisi sociolinguistica

Il numero molto ridotto di occorrenze delle varianti marcate testimonia la marginalità di questo fenomeno e conferma la natura di stereotipo locale della palatalizzazione. Come già anticipato, si tratta infatti di una forma altamente stigmatizzata, collocata verso il polo basso sia nella dimensione diastratica sia in quella diafasica. Questo fenomeno inoltre è avvertito dai parlanti come caratteristico della parlata del capoluogo sardo, se non addirittura esclusivo. Tuttavia, in merito alla sua diffusione, considerato che non esistono indagini che documentino tale esclusività, sembra opportuno non avanzare alcuna ipotesi in questo senso. La realizzazione palatalizzata è oggetto di discussione da parte dei parlanti, anche nelle discussioni sui *social network* ed è associata alla parlata cagliaritana al punto tale che il nome del gruppo che su *Facebook* raccoglie gli utenti originari di Cagliari si basa proprio su questa peculiarità fonetica della parlata locale, "Sei di Cagliari...se dici Chiagliari"[13] (Fig. 8). "Chiagliari", tra l'altro, rappresenta anche la parola target generalmente usata quando si intende fare riferimento a questo particolare tratto linguistico.

Figura 8 - *Immagine del nome del gruppo Facebook 'Sei di Cagliari se... dici Chiagliari'*

Prima di passare all'analisi stilistica, nonostante il numero dei soggetti intervistati sia esiguo, può essere utile mostrare come le varianti si distribuiscono tra i diversi parlanti (Figg. 9 e 10).

[13] Il nome del gruppo è in italiano, così come la maggior parte delle occorrenze che si rintracciano su *Facebook,* visto che questo fenomeno è passato alla varietà di italiano regionale locale.

Figura 9 - *Grafico a mosaico raffigurante la distribuzione delle varianti palatalizzate (P) e non palatalizzate (NP) per parlante nel gruppo della confraternita.*

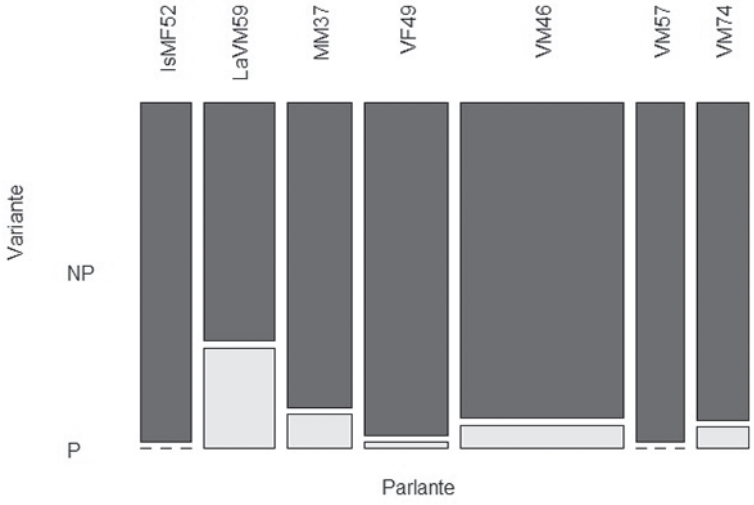
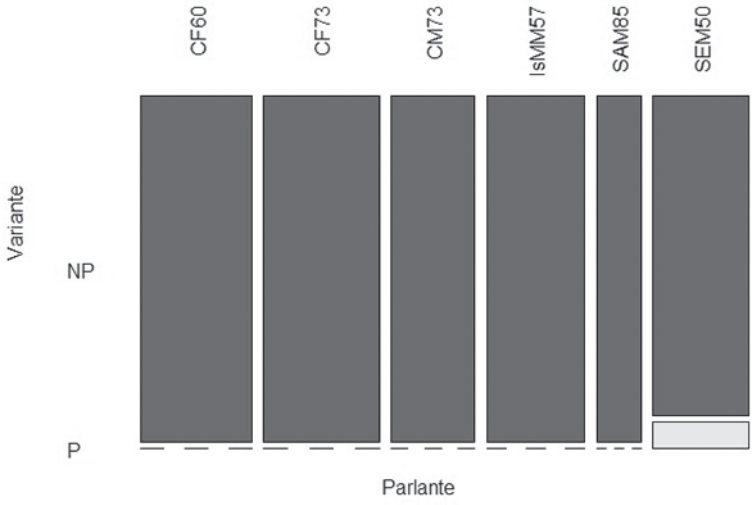


Figura 10 - *Grafico a mosaico raffigurante la distribuzione delle varianti palatalizzate (P) e non palatalizzate (NP) per parlante nel gruppo dei parlanti esterni alla confraternita.*



Se consideriamo i due gruppi nel complesso, notiamo che le quattro donne presenti nel campione (IsMF52, VF49, CF60 e CF73) mostrano una maggiore tendenza verso la norma linguistica e quindi verso l'uso delle varianti non marcate, con l'eccezione di un'unica realizzazione prodotta dalla parlante VF49. Inoltre, vale la pena sottolineare che quest'unica occorrenza attribuibile a una locutrice – che non è stata considerata ai fini acustici, perché il confronto tra i parametri acustici di tutte le varianti non marcate con una sola variante palatalizzata non sarebbe stato né metodologicamente corretto né analiticamente coerente – è stata prodotta in un punto ben

preciso dell'intervista. In particolare, le due condizioni che sembrano essere state determinanti per la realizzazione di questa occorrenza sono state: 1) il momento temporale in cui essa è stata prodotta, ovvero alla fine di un'intervista durata 40 minuti circa, e 2) il suo inserimento nel racconto di un aneddoto molto divertente e particolarmente imbarazzante per l'intervistata. Il fatto che sia stata prodotta un'unica occorrenza contenente la variante palatalizzata da parte delle donne può essere considerato un indizio del comportamento linguistico più normativo delle donne, che tendono a utilizzare le varianti più prestigiose e a evitare le deviazioni dalla norma che sono apertamente censurate. Analogamente a quanto registrato in Mereu (2017) per l'arretramento di /s/ in posizione preconsonantica, anche in questo caso il *pattern* di variazione di genere riflette la tendenza riassunta dal principio laboviano secondo cui, nel caso di variabili sociolinguistiche stabili, gli uomini tendono a usare più delle donne le varianti socialmente marcate verso il basso (Labov, 1990). Più precisamente, i risultati riscontrati – indicativi solo di una tendenza, considerato il numero esiguo di parlanti – sembrano sposare il corollario del primo principio secondo il quale nei cambiamenti dall'alto, che talvolta mostrano proprio stereotipi come stabili variabili sociolinguistiche, le donne prediligono rispetto agli uomini l'uso delle forme di maggiore prestigio.

4.3.1 Analisi stilistica

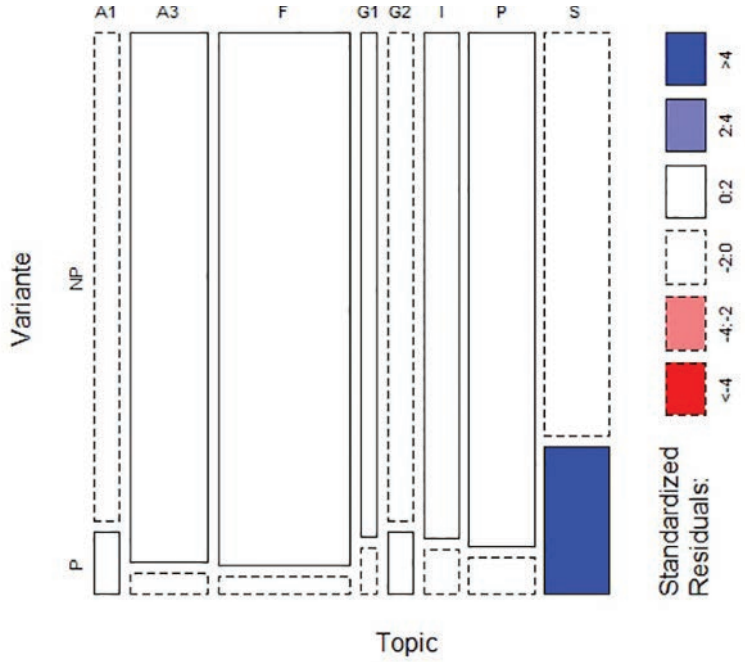
Come anticipato, l'analisi stilistica è stata affrontata separatamente per il gruppo dei parlanti appartenenti alla confraternita e per quello esterno a essa, in modo da dare rilievo ai significati sociali locali della variabile studiata. Per i parlanti del gruppo della confraternita il *topic* esercita un'influenza significativa sulla distribuzione delle varianti, come ha dimostrato l'applicazione del test esatto di Fisher (p-value < 0.05). I risultati dell'analisi possono essere rappresentati nel grafico a mosaico con l'indicazione dei residui standardizzati.

La categoria che si è mostrata più rilevante è l'argomento 'Rivalità tra confraternite' (S). Vale la pena riportare che risultati analoghi a questi sono stati registrati per questo stesso gruppo di parlanti nell'analisi di un'altra variabile, l'arretramento di /s/ in posizione preconsonantica, che costituisce anch'essa uno stereotipo locale della parlata cagliaritana (cfr. Mereu, 2017).

Per capire la correlazione esistente tra questo particolare *topic* e l'uso della forma marcata è necessario fare riferimento al lavoro etnografico svolto sul campo, grazie al quale è stato possibile comprendere come questo particolare argomento abbia una salienza particolare per i membri del sodalizio religioso. Dalle interviste effettuate e dalla partecipazione agli eventi della confraternita è emersa l'esistenza di un'altra confraternita che assolve gli stessi compiti dell'Arciconfraternita della Solitudine, ovvero l'accompagnamento in cattedrale nello stesso giorno (il Venerdì Santo) dei simulacri religiosi. La rivalità tra le due associazioni è dovuta al medesimo ruolo svolto in relazione al compimento dei riti religiosi e all'origine comune delle due Masse (cori di cantori)[14].

[14] L'antropologa Chiara Solinas riassume in modo molto efficace questa realtà con le seguenti parole:

Figura 11 - *Grafico a mosaico raffigurante la distribuzione delle varianti palatalizzate (P) e non palatalizzate (NP) in funzione del topic e del micro-genere per il gruppo dei parlanti appartenenti alla confraternita. Legenda: A1: aneddoto divertente; A3: aneddoto personale; F: descrizione; G1: dialogo con l'intervistatore; G2: dialogo tra informanti; I: lamentela; P: problemi del quartiere; S: rivalità tra confraternite*



Questa rivalità, che genera numerose discussioni su quale delle due sia la confraternita più autentica, è incrementata dal fatto che i fedeli cagliaritani durante la processione del Venerdì Santo si distribuiscono tra le due confraternite, decidendo così quale dei due gruppi seguire. È facile capire quindi il motivo per cui tale argomento suscita nei confratelli intervistati un grande coinvolgimento emotivo.

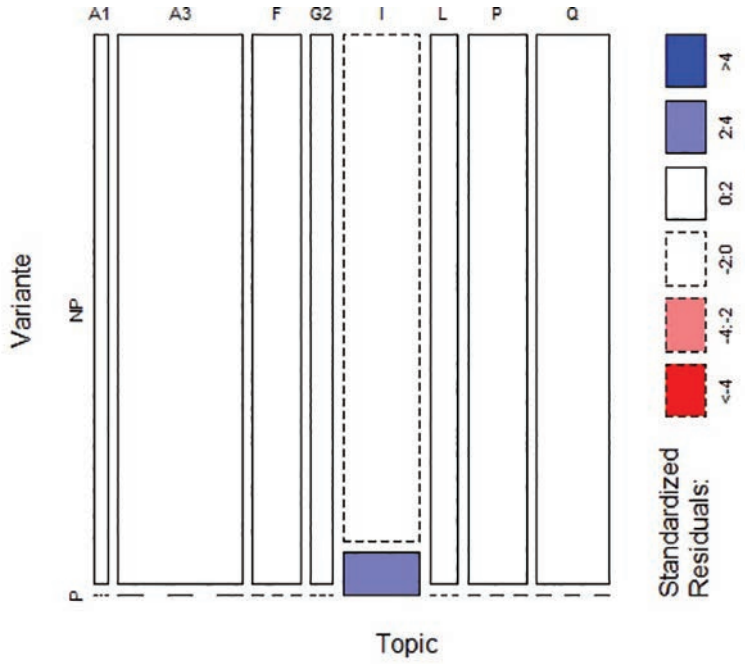
I segmenti di parlato in cui non sono mai state prodotte varianti palatalizzate sono gli argomenti 'Lingua sarda' (L) e 'Quartiere' (Q), che non compaiono nel grafico per ragioni legate alle analisi di tipo statistico[15]. La totale mancanza di forme marcate in queste due sezioni dell'intervista risulta molto significativa e, al contempo, anche facilmente interpretabile: i parlanti usano in modo categorico le varianti non marcate in queste due sezioni in quanto sia l'argomento di tipo metalinguistico (relativo alla lingua sarda) sia il genere descrittivo (descrizione del quartiere) sono associati in genere a un registro sorvegliato.

«gli elementi più caratteristici della Settimana Santa cagliaritana da prendere in considerazione sono essenzialmente: un quartiere, due chiese, due confraternite, due gruppi di cantori, due rituali analoghi, un solo repertorio» (Solinas, 2007: 138).

[15] Per le analisi statistiche la matrice dei dati non deve presentare delle celle vuote.

Nel gruppo dei parlanti non appartenenti alla confraternita (Fig. 12) il segmento in cui le palatalizzazioni trovano la loro manifestazione più piena – e anche l'unico – è la lamentela (I).

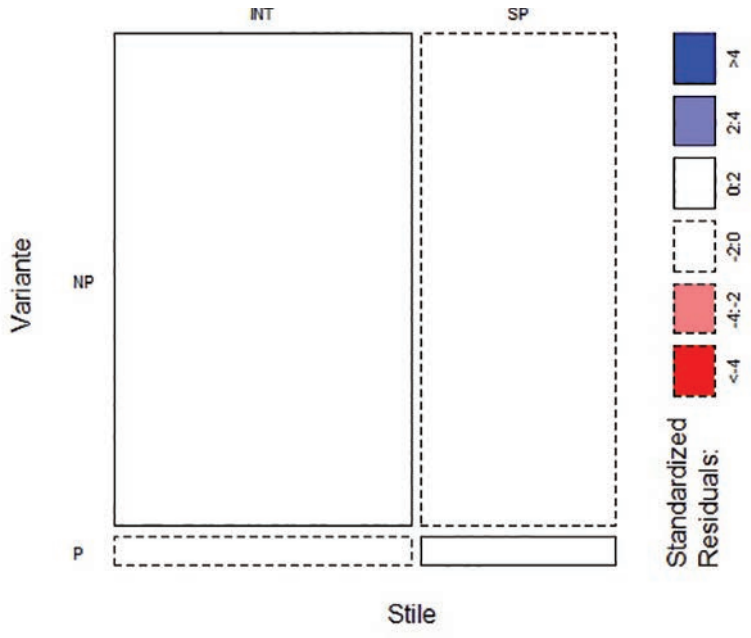
Figura 12 - *Grafico a mosaico raffigurante la distribuzione delle varianti palatalizzate (P) e non palatalizzate (NP) in funzione del topic e del micro-genere per il gruppo dei parlanti non appartenenti alla confraternita. Legenda: A1: aneddoto divertente; A3: aneddoto personale; F: descrizione; G2: dialogo tra informanti; I: lamentela; L: lingua sarda; P: problemi del quartiere; Q: quartiere*



Dall'analisi dei due gruppi di soggetti intervistati è evidente come l'attivazione della variante marcata sia connessa a quei segmenti di parlato in cui vengono affrontati argomenti particolarmente coinvolgenti per i locutori. Tale coinvolgimento determina infatti un'attenzione maggiore nei confronti del contenuto a scapito del parlato stesso (cfr. Mereu, 2017).

Per quanto riguarda il secondo livello di analisi, basato su criteri di tipo conversazionale, i due gruppi di parlanti sono stati studiati nel loro complesso, in quanto i parametri considerati sono legati alla struttura dell'intervista e non al contenuto trattato. L'individuazione e l'attribuzione dei due stili di intervista 'Interview Style' e 'Spontaneous Style' ai diversi segmenti di parlato ha permesso di mettere in luce i seguenti risultati (Fig. 13).

Figura 13 - *Grafico a mosaico raffigurante la distribuzione delle varianti palatalizzate (P) e non palatalizzate (NP) in funzione dello stile conversazionale (INT: Interview Style; SP: Spontaneous Style)*



Dal test esatto di Fisher la variabile 'Stile' risulta non significativa, così come si può evincere anche dal grafico a mosaico con i residui standardizzati. In questo caso il peso delle due categorie stilistiche è pressoché identico per quanto riguarda la distribuzione delle varianti palatalizzate, quindi sembra che la posizione all'interno dell'evento dell'intervista in cui le occorrenze vengono realizzate sia indifferente. Considerato quindi che il *topic* / micro-genere ha invece avuto un ruolo significativo nell'uso delle diverse varianti, sembrerebbe che a guidare la scelta della produzione di una variante invece di un'altra sia proprio il contenuto affrontato da chi parla e non il tipo di interazione in atto con l'intervistatrice.

5. *Conclusioni*

Dai risultati emersi dall'analisi possono essere tratte alcune riflessioni generali, di carattere metodologico e analitico, sul parlato spontaneo come base di dati su cui operare. Il mezzo di elicitazione adottato durante la raccolta dati, l'intervista etnografica semi-strutturata, ha permesso di registrare un numero sufficiente di occorrenze contenenti la variabile oggetto di interesse, nonostante questa costituisca uno stereotipo locale e quindi un fenomeno molto marginale e difficilmente elicitabile, se non in contesti comunicativi altamente informali. Un ulteriore vantaggio di questa tecnica di escussione dei dati consiste nella sua efficacia nel registrare una grande quantità di parlato che può essere sfruttata ai fini di un'analisi stilistica. Nel caso

specifico, lo studio dei *pattern* di variazione della variabile per *topic* ha permesso di evidenziare l'attivazione della variante palatalizzata in determinati argomenti di conversazione. Le forme linguistiche marcate sono emerse laddove l'argomento di discussione affrontato era di particolare salienza per i parlanti, come nel caso del *topic* "Rivalità tra confraternite" o del micro-genere "Lamentela".

Il fatto che questo tratto rappresenti uno stereotipo molto diffuso nei *social network* e passato anche all'italiano regionale lo rende un tratto bandiera caratterizzato da prestigio coperto. Il valore di prestigio coperto, spesso associato alle forme stigmatizzate (cfr. Trudgill, 1978; Labov, Cohen, Robins & Lewis, 1968), è sintomatico della percezione che i parlanti hanno di queste forme, avvertite come simbolo di lealtà nei confronti delle norme e dei valori locali o come simbolo di mascolinità (Labov, 2001). In questo caso, la forma palatalizzata è passata all'italiano regionale per segnalare l'appartenenza alla comunità cagliaritana, ma viene evitata in sardo nel parlato sorvegliato. Tuttavia la mancanza di studi sociolinguistici sull'italiano regionale finalizzati all'indagine delle modalità d'impiego di questo fenomeno non ci consente di andare oltre l'ipotesi generale di prestigio coperto.

Riferimenti bibliografici

Abete, G. (2012). Aspetti metodologici per lo studio della variazione fonetica nel parlato dialettale. In Bianchi, P., De Blasi, N., De Caprio, C. & Montuori, F. (Eds.), *La variazione nell'italiano e nella sua storia. Varietà e varianti linguistiche testuali. Atti dell'XI Congresso SILFI Società Internazionale di Linguistica e Filologia Italiana (Napoli, 5-7 ottobre 2010)*. Firenze: Franco Cesati editore, 827-836.

Becker, K. (2009). /r/ and the construction of place identity on New York City's Lower East Side. In *Journal of Sociolinguistics,* 13(5), 634-658.

Blom, J.P., Gumperz, J.J. (1968). Social meaning in linguistic structures: code switching in Norway. In Gumperz, J.J., Hymes, D.H. (1972) (Eds.), *Directions in Sociolinguistics*. New York: Holt, Rinehart and Winston, 407-434.

Boersma, P., Weenink, D. (2017). *Praat: Doing Phonetics by Computer*. http://www.praat.org/.

Cho, T., Ladefoged, P. (1999). Variation and universals in VOT: Evidence from 18 languages. In *Journal of Phonetics* 27(2), 207-229.

Chodroff, E., Wilson, C. (2014). Burst spectrum as a cue for the stop voicing contrast in American English. In *The Journal of the Acoustical Society of America,* 136(5), 2762-2772.

Eckert, P. (2000). *Linguistic Variation as Social Practice. The Linguistic Construction of Identity in Belten High*. Malden / Oxford: Blackwell.

Ervin-Tripp, S. (1964). An analysis of the interaction of language, topic and listener. In *American Anthropologist*, 66, 86-102.

Foulkes, P., Scobbie, J. & Watt, D. (2010). Sociophonetics. In Hardcastle, W.J., Laver, J. & Gibbon, F.E. (Eds.), *Handbook of Phonetic Sciences*. Oxford: Blackwell, 703-754.

Hay, J., Foulkes, P. (2016). The evolution of medial (-t-) in real and remembered time. In *Language*, 92(2), 298-330.

Labov, W. (1966). *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1984). Field Methods of the Project on Linguistic Change and Variation. In Baugh, J. & Sherzer, J. (Eds.), *Language in Use*. Englewood Cliffs, NJ: Prentice-Hall, 28-53.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. In *Language Variation and Change*, 2(1990), 205-254.

Labov, W. (1994). *Principles of Linguistic Change. Vol. I: Internal Factors*. Oxford: Blackwell.

Labov, W. (2001). *Principles of Linguistic Change. Vol. II: Social Factors*. Oxford: Blackwell.

Labov, W., Cohen, P., Robins, C. & Lewis, J. (1968). A study of the non-standard English of Negro and Puerto Rican Speakers in New York City. *Cooperative Research Report 3288*, vols I-II. Philadelphia: U.S. Regional Survey.

Ladefoged, P., Maddieson, I. (1996). *The Sounds of the World's Languages*. Malden: Blackwell.

Lawson, R. (2009). Sociolinguistic constructions of identity among adolescent males in Glasgow. PhD dissertation, University of Glasgow.

Lisker, L., Abramson, A.S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. In *Word*, 20(3), 384-422.

Loporcaro, M., Putzu, I.E. (2013). Variation in auxiliary selection, syntactic change, and the internal classification of Campidanese Sardinian. In Paulis, G., Pinto, I. & Putzu, I.E. (Eds.), *Repertorio plurilingue e variazione linguistica a Cagliari.* Milano: Franco Angeli, 200-244.

Love, J., Walker, A. (2012). Football versus football: Effect of topic on /r/ realization in American and English sports fans. In *Language and Speech*, 56(4), 443-460.

Mendoza-Denton, N., Hay, J. & Jannedy, S. (2003). Probabilistic sociolinguistics: Beyond variable rules. In Bod, R., Hay, J. & Jannedy, S. (Eds.), *Probabilistic Linguistics.* Cambridge: MIT Press, 97-139.

Mereu, D. (2017). Arretramento di /s/ nel sardo cagliaritano: uno studio sociofonetico. In Bertini, C., Celata, C., Lenoci, G., Meluzzi, C. & Ricci, I. (Eds.), *Fattori sociali e biologici nella variazione fonetica. Social and Biological Factors in Speech Variation*. Collana Studi AISV, Milano: Officinaventuno, 45-65.

Mereu, D. (2018). Il sardo parlato a Cagliari: uno studio sociofonetico. Tesi di Dottorato, Università degli Studi di Bergamo.

Milroy, L. (1980). *Language and Social Networks.* Oxford: Blackwell.

Munson, B. (2001). A method for studying variability in fricatives using dynamic measures of spectral mean. In *Journal of the Acoustical Society of America*, 110(2), 1203-1206.

Nakai, S., Scobbie, J.M. (2016). The VOT category boundary in word-initial stops: Counter-evidence against rate normalization in English spontaneous speech. In *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 7(1):13, 1-31.

Ní Chiosáin, M., Padgett, J. (2012). An acoustic and perceptual study of Connemara Irish palatalization. In *Journal of the International Phonetic Association*, 42(2), 171-191.

Pellis, U. (1934). *Cinquanta inchieste linguistiche in Sardegna*. Udine: Società Filologica Friulana «G.I. Ascoli».

Rattu, R. (2017). Repertorio plurilingue e variazione sociolinguistica a Cagliari: i quartieri di Castello, Marina, Villanova, Stampace, Bonaria e Monte Urpinu. Tesi di Dottorato, Università di Cagliari.

Savy, R., Cutugno, F. (1997). Ipoarticolazione, riduzione vocalica, centralizzazione: come interagiscono nella variazione diafasica?. In Cutugno, F. (Ed.), *Fonetica e fonologia degli stili dell'italiano parlato. Atti delle VII Giornate di Studio del Gruppo di Fonetica Sperimentale*. Roma: Esagrafica, 177-194.

Solinas, C. (2007). Una ricerca antropologico musicale in ambito urbano. I canti e i riti della Settimana Santa a Cagliari. In *Portales,* 9, 135-141.

Spinu, L. (2007). Perceptual properties of palatalization in Romanian. In Camacho, J., Flores-Ferran, N., Sanchez, L., Deprez, V. & Cabrera, M.J. (Eds.), *Romance Linguistics 2006: Selected Papers of the 36th Linguistic Symposium on Romance Languages (LSRL)*. Amsterdam: John Benjamins, 303-317.

Stuart-Smith, J. (2007). Empirical evidence for gendered speech production: /s/ in Glaswegian. In Cole, J. & Hualde, J.I. (Eds.), *Laboratory Phonology 9*. Berlin: Mouton de Gruyter, 65-86.

Stuart-Smith, J., Sonderegger, M., Rathcke, T. & Macdonald, R. (2015). The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. In *Laboratory Phonology*, 6(3-4), 505-549.

Sundara, M. (2005). Acoustic-phonetics of coronal stops: A cross-language study of Canadian English and Canadian French. In *Journal of the Acoustical Society of America*, 118(2), 1026-1037.

Trudgill, P. (1978). Sex, covert prestige, and linguistic change in the urban British English of Norwich. In *Language in Society*, 1, 179-96.

Vietti, A. (2003). Come costruire una intervista 'ecologica': per una interpretazione contestualizzata dei dati. In Valentini, A., Molinelli, P., Cuzzolin, P. & Bernini, G. (Eds.), *Ecologia linguistica. Atti della Società di Linguistica Italiana.* Roma: Bulzoni, 161-184.

Virdis, M. (1978). *Fonetica del dialetto sardo campidanese.* Cagliari: Edizioni della Torre.

Virdis, M. (2013). La varietà di Cagliari e le varietà meridionali del Sardo. In Paulis, G., Pinto, I. & Putzu, I. (Eds.), *Repertorio plurilingue e variazione linguistica a Cagliari.* Milano: Franco Angeli, 165-180.

Voghera, M. (2017). *Dal parlato alla grammatica. Costruzione e forma dei testi spontanei*. Roma: Carocci.

Warner, N. (2012). Methods for studying spontaneous speech. In Cohn, A., Fougeron, C. & Huffman, M. (Eds.), *Handbook of Laboratory Phonology*. Oxford: Oxford University Press, 612-633.

DOI: 10.17469/O2104AISV000005

OTTAVIA TORDINI, VINCENZO GALATÀ, CINZIA AVESANI, MARIO VAYRA

Sound maintenance and change: Exploring inter-language phonetic influence in first-generation italo-australian immigrants[1]

The present study explores the phonetic influence exerted by late-acquired L3 English on the native dialect of four first-generation Italo-Australian speakers from Belluno, Northern Veneto, who moved to Sydney, Australia in the mid-late 1950s. We map phenomena of attrition, maintenance and/or loss in their spoken L1 (Veneto dialect) by investigating the fine phonetic details of selected voiceless coronal obstruents: the interdental [θ], shared with English L3 but absent in the phonological inventory of their L2 (Standard Italian, SI); [s] and [ʧ], present in all the three repertoires. Their dialectal productions are compared to those of four control speakers who were born and live in the same area of origin of the four first-generation Italo-Australian speakers.

Keywords: sociophonetics, dialect, language variation and change, heritage languages, Italo-Australian immigrants, fricatives, spectral moments.

1. *Introduction*

This contribution builds upon previous work carried out within the IRIAS project, which has already been presented and thoroughly discussed in Avesani, Galatà, Vayra, Best, Di Biase, Tordini & Tisato (2015) and Avesani, Galatà, Best, Di Biase, Vayra & Ardolino (2017). In this study, we acoustically analyse dialectal L1 productions of four first-generation Italo-Australian speakers included in the IRIAS Corpus who originate from Northern Veneto (specifically from the areas of Feltre and Cadore, Belluno province), and of four control-group Italian speakers, who were born and currently live in the very same places of origin of Italo-Australian immigrants. Our purpose is to test whether Italo-Australian speakers from Feltre and Cadore have maintained the fine-grained phonetic features of their L1 voiceless coronal obstruents /θ, s, ʧ/ in a long-standing contact with similar – but not phonetically identical – /θ, s/ of Australian English (AusEng); and to examine whether the fricative consonants and the fricative portion of the post-alveolar affricate have undergone any phonetic change with respect to same phones currently spoken by the control group of Italian speakers in Veneto.

[1] Authorship note: while the paper is the result of a joint collaboration and discussion between the four authors, main responsibility for this paper is divided as follows: § 1, 3, 5.1, 5.3, 6: Tordini, Avesani, Galatà; § 2: Tordini; § 4: Tordini, Avesani; § 5.2: Tordini, Galatà; § 5.4: Galatà; § 7: Avesani, Tordini.

Both the Italo-Australians and the control speakers have the local dialect as L1 and learned the Veneto regional variety of Italian when they entered the elementary schools at 6 years of age. They can be considered as early sequential bilinguals, but it should be reminded that the local dialect and regional Italian are two varieties that stand in a diglossic relationship within the speakers' linguistic repertoire: the local dialect is the language of everyday communication, within and outside the family, while regional Italian is used in more formal exchanges.

The Italo-Australian speakers were exposed to English upon arrival in the new country and learned it as their third language spontaneously, by immersion in the new society. While the local dialect kept being used within the family and the heritage community (due to a chain type of immigration), Italian widened its use to become the vehicular language of the larger community of Italian immigrants, converging towards shared forms and being permeated by English lexical items imported within an Italian morphological frame (Avesani et al., 2017). Given the length of residency of the Italo-Australian speakers in Australia exceeded five decades at the time of interview, they represent an interesting group, albeit small, to study processes of linguistic contact, cross-linguistic influence (CLI) and attrition.

The paper proceeds as follows: in § 2, we contextualize this study within the scope of primary language attrition, focusing on its manifestations in migration settings. In § 3, we offer a sociolinguistic overview on Italian immigrant communities in Australia and a brief description of the speakers' native Veneto dialect. In § 4, we present an influential theoretical framework relevant to the interpretation of our results, the Speech Learning Model and the predictions we can make based on that model. In § 5, the data collection procedure and the methodology employed for the acoustic analysis are described and in § 6 we will provide the outcome of the spectral analysis and the statistical results. Discussion and conclusions are presented in § 7.

2. *Primary language attrition: an insight into migration settings*

Many studies report evidence of primary language (L1) attrition[2] caused by a persistent contact with an L2 in a specific communication setting (Nagy, 2015). Nevertheless, it has been demonstrated that enduring contact and use of L2 are not the only factors triggering an impoverishment in the knowledge of a given linguistic code (see, for example: Andersen, 1982; Weltens, de Bot & van Els, 1986; Seliger, Vago, 1991; de Bot, Clyne, 1994; Schmid, 2011; Schmid, Köpke, 2013). Linguistic

[2] The notion of "attrition" is frequently included within the concept of "language loss", which is used as its cover term (Opitz, 2011: 10). While language loss is generally employed to indicate the phenomenon of change or reduction of linguistic skills, attrition specifically indicates «the loss of a language by a healthy individual (that is, loss which is not caused by brain injury or some pathological condition, such as aphasia or dementia)». Following the classification suggested in Weltens, de Bot & van Els (1986), this work will solely address phenomena of primary language loss in a second language (L2) environment, e.g. loss of specific native language features experienced by migrants.

and sociolinguistic attrition both within individuals and within language communities, in fact, occur frequently as a result of language shift, that is, when speakers consistently reduce (or abandon) the use of their L1 (Schmid, 2011). In general, the decline in competence caused by an asymmetric interaction between the linguistic systems normally can derive either from a functional reduction ("shifting") or a structural reduction ("attrition") in L1 (Schmid, Köpke, Keijzer & Weilemar, 2004; Celata, Cancila, 2010).

There is also evidence that migration settings offer a unique perspective to observe primary language attrition phenomena. The new social and linguistic environment is characterised by an extensive use of L2, which is permanently active and gradually becomes the medium of privileged communication in daily life (Kroll, Bobb & Wodniecka, 2006; Köpke, Schmid, Keijzer & Dostert, 2007: 3), by a decrease in native language use in everyday exchanges, and by a dramatic reduction of constant L1 input (Olshtain, 1989: 151). We can observe from a significant amount of studies that being immersed in an L2 environment has profound effects on adult migrants (Seliger, Vago 1991; Yağmur, 1997; Schmid et al. 2004; Köpke et al., 2007; De Leeuw, 2008; De Leeuw, Schmid & Mennen, 2010, among others). Namely, adult migrants who have a full native language proficiency and have learned an L2 after puberty (namely, late bilinguals) can reach a high proficiency also in the L2 and often reverse their language dominance in favour of the L2 (Mägiste, 1979; Opitz, 2011): these individuals are defined as "late L1 attriters" and are opposed to "child attriters", whose process of L1 acquisition is arrested or reversed (Opitz, 2011: 13).

Phenomena of attrition in communities of immigrants are generally related to several extralinguistic factors: the age of arrival in the host country (AoA) or the length of residence (LOR); the amount, frequency and context of input and exposure to the foreign language(s) (Schmid, Köpke, 2013). It is undebatable that the process of L2 acquisition in adults is also led by other factors, such as: motivation to integrate in the host country's social and professional community, aptitude, time and effort that the speaker employs in the language learning process, cultural identity. However, age of acquisition (AOA) of a second language has been acknowledged as a main factor in assessing bilinguals' competence and has therefore received considerable attention in this area of research in recent years (see e.g. studies by DeKeyser, Larsen-Hall, 2005; Hyltenstam, Abrahamsson; 2003; Kroll, de Groot; 2005; Ahn, Chang, DeKeyser & Lee-Ellis 2017, a.o., on age-related maturational constraints). However, it is still unclear whether the age at which a decreased L1-L2 contact begins can predict the extent of L1 attrition (Ahn et al., 2017). Moreover, the extent of attrition as a consequence of the predominant use of L2 does not seem to progress linearly over longer time-periods (e.g. de Bot, Clyne, 1994), thus making it difficult to describe its effects in a straightforward manner.

3. *The community of Veneto immigrants in Australia and their dialects*

As can be seen from a large number of studies on Italian immigrants in Australia (e.g. Bettoni, 1981; Tosi, 1991; Bettoni, 2000; Bettoni, Rubino, 1996; Rubino, 2006; Campolo, 2009; Caruso, 2010; Gallina, 2011; Rubino, 2014), there is general consensus about the social and linguistic composition of Italian communities. In the post-WWII years, a considerable amount of Italians left from regions in the south (Sicily, Calabria, Campania, Abruzzo) and from few regions in the north, notably from Veneto. While South America was the most common destination of Veneto immigrants, Australia attracted many of them due to a bilateral agreement for assisted immigration signed in 1951 between Australia and Italy (for further details, see Campolo, 2009; see also Avesani et al., 2015; Avesani et al., 2017).

Statistical surveys conducted after WWII in Italy revealed that the population's linguistic behaviour was strongly oriented towards dialectophony: only 35% of Italian citizens declared a regular and daily use of the national language, while the remaining 65% was divided between those who employed Italian only occasionally and those who made an exclusive use of the local dialect (De Mauro, 1972: ISTAT census of 1951).

Italians who left Veneto in the mid 40s-early 60s moving to Australia were mainly workers or farmers who mostly spoke their local dialect as L1 in everyday communication, both within the family and within the larger local community. They had learnt (regional) Italian at school, as an L2, and used it only in formal circumstances. Subsequently, they learnt AusEng as L3 only upon arrival in the country, almost exclusively by immersion (Tosi, 1991; Bettoni, Rubino, 1996) while maintaining the local dialect to communicate with the family and the network of other immigrants coming from the same areas of Veneto[3]. Once in Australia, the input from the same regional variety of Italian dramatically diminished, while at the same time the speakers were exposed to different varieties of regional Italian as spoken by the variegated Italian community. Italian gradually lost its status of "high" language used mainly for writing and spoken only in formal situations, and became the spoken vehicular language of the great Italian community of immigrants, acquiring in the process new traits that were absent in the original varieties spoken in Italy (for the development of the community language known as *Italo-Australiano* see for example Bettoni, Rubino, 1996; Tosi, 1991; Campolo, 2009). Progressively, English-L3 became the dominant language, but the native local dialect continued to be used in the family and with friends as long as they were available.

The issues of maintenance or change of the heritage languages, i.e. local dialect and regional Italian, by Italian immigrants in Australia have been addressed in the past by many studies, which were mainly focused on the effects of English-L3 on the native repertoire at a morphosyntactic and at a pragmatic level (e.g. Bettoni,

[3] Italian immigration to Australia, especially to rural areas, has followed the pattern of a "chain migration", which creates "districts in which emigrants are bound together by shared kinship ties based on a specific village" (Tosi, 1991: 337; Corazza, Grigoletti & Pellegrini, 2012).

1981; Bettoni, 2000; Bettoni, Rubino, 1996; Rubino, 2006; Caruso, 2010; Rubino, 2014)[4]. However, the regionally-differentiated speech characteristics of Italo-Australian speakers have been less often considered, although they are strong markers of their linguistic identity and their "foreign accent" (Avesani et al., 2015; Avesani et al., 2017).

In this paper, we explore whether a set of coronal obstruents of the Bellunese dialect have been maintained by emigrants who left their villages in Cadore and in the area of Feltre in the late fifties-early sixties. The specific aim of this study is to verify if and to which extent the fine phonetic properties of the native language can resist attrition after more than fifty years of contact with an L3 learned in adulthood, when the native and the late-learned languages share the same set of consonants but differ for their phonetic content.

The dialect spoken in the province of Belluno belongs to Northern Veneto, one of the five linguistic sub-system in which the Veneto dialects are traditionally divided (e.g., Zamboni, 1974; 1988). The set of coronal obstruents of this system includes the unvoiced obstruents /θ, s, ʧ/[5]. Examples are: [θimiˈtɛːro] ("cemetery", Standard Italian [ʧimiˈtɛːro]); [ˈsapa] ("hoe", Standard Italian [ˈtsapːa]), [ˈʧeːza] ("church", Standard Italian [ˈkjeːza]). Note that none of the Veneto dialects presents the post-alveolar fricative /ʃ/. /s/ and /ʧ/ are shared by all languages, while /θ/ is shared only by English and Belluno dialect[6] and is missing in Standard Italian. Phonetically, [s] is an apico-alveolar fricative in Belluno dialect while is a lamino-denti-alveolar in Standard Italian. In a previous work (Avesani et al., 2015) we analysed the dialect of two speakers from Belluno and we found empirical evidence that singleton [s] in intervocalic position has a retracted place of articulation: it sounds more like [ʃ] and the acoustic properties of its frication noise (Center of Gravity) do not differ from those of the fricative release of the postalveolar affricate [ʧ]. Such a retraction is attested also in the regional Italian spoken in those areas, as indicated by the detailed auditory analysis of Canepari (1984). The phonetic details of Bellunese /s/ also differ from those of Australian English /s/. In fact, /s/-retraction is a gradient sound change that has taken place in several varieties (i.e. "dialects") of English, but not in Received Pronunciation nor Australian English (Baker, Arcangeli & Mielke, 2011). It is also worth mentioning recent studies by Stevens, Harrington (2016) and Stuart-Smith, Sonderegger, McAuliffe, Mcdonald, Mielke, Thomas & Dodsworth (2018), which have demonstrated that the precursors of such retraction lie in the lower frequency spectral energy for /s/ in /str/ than

[4] With the notable exception of Horvath (1985), who conducted a sociophonetic analysis on the Sydney speech community including groups of Italian immigrants. However, the origin of the Italian speakers is not specified, thus hindering a proper comprehension of their linguistic and sociolinguistic features and the impact of English on their native variety.

[5] In the present study, as in the previous ones, we do not analyze allophones of /ʃ/ but the fricative release [ʃ] of the postalveolar affricate /ʧ/, and we will refer to it as [(t)ʃ]. Such a choice has been induced by the scarcity of occurrences of /ʃ/ in the corpus.

[6] Note that among the five Veneto subsystems [θ] occurs only in the Northern one.

in singleton /s/ and that the phonetic bases of /s/-retraction are subject to "dialectal" and social factors[7].

The afore-mentioned interlinguistic phonetic differences will enable us to make predictions about maintenance or attrition of L1 coronal sounds in contact with similar phonetic categories in L3.

4. *Cross-language phonetic interference: a theoretical framework*

Within the broader phenomena of attrition, L1 change due to extensive use of L2 is traditionally identified through headings such as "cross-language interaction/influence" (CLI), "reverse interference", "convergence" and is reported at all linguistic levels (Pavlenko, 2004; Schmid, Köpke, 2013). With reference to the sound structure, it has been clearly demonstrated that L2 experience can exert a considerable influence on L1 oral productions to the point of triggering a phonological restructuring of its elements (e.g., Chang, 2012; Flege, 1987, 1995, 2007; Flege, McKay & Meador, 1999; Flege, Schirru & McKay, 2003; Major, 1992, 2010). In the last few decades, issues related to cross-language phonetic interaction have been addressed by a significant amount of studies. Recognition of L2 influence on L1 speech has been extensively discussed in the works of Flege (1987, 1995), within the theoretical framework of the Speech Learning Model (SLM). Other influential models have addressed CLI, such as the Perceptual Assimilation Model (PAM) and its extension PAM-L2 (Best, 1995; Best, Tyler, 2007) and the Second Language Linguistic Perception model (L2LP: Escudero, 2005). However, PAM-L2 did not specifically address the effects of L2 on L1 and Escudero's model focuses on vowel learning.

The starting point of SLM is that even proficient late bilinguals are likely to experience restructuring in the L1 as a consequence of L2 experience at a phonetic level (Schmid, 2011)[8]. Contrary to the hypothesis of a critical period, the SLM assumes that the capacity for speech learning remains intact across the life span (Flege et al., 2003) such that a specific L1 phonetic category «[...] continues to develop in adulthood under the influence of all sounds identified with that category» (Chang, 2012: 250). According to Flege's model, elements making up the L1 and L2 phonetic subsystems of a bilingual exist in a "common phonological space", and so will necessarily influence one another (see also Bergmann, Nota, Sprenger & Schmid, 2016). Namely, L1 and L2 sounds are posited to exist in a shared system in which the sounds interact through two distinct mechanisms. The first mechanism, "category assimilation", is thought to operate when a new L2 category fails to be established despite audible differences between it and the closest L1 speech sound. The formation of a new phonetic category in L2 will be blocked if instances of an L2 speech category continue to

[7] The phenomenon of /s/-retraction has also been investigated by Mereu (2017) in the Sardinian variety spoken in Cagliari. The author demonstrates that the realization of /s/ as the local stereotype [ʃ], i.e. the substandard variant, is correlated to stylistic variation.

[8] A similar assimilatory process towards L2 phonetic settings has recently been found even in the L1 of beginner learners (Chang, 2012).

be identified as instances of an L1 category, either because it is perceived as the same sound or as similar to an existing sound in the native system (i.e. more or less deviant exemplar of a L1 phone) via a mechanism of "equivalence classification". In both cases «a single phonetic category will be used to process perceptually linked L1 and L2 sounds» (Flege, 1995: 239; Flege et al., 2003; de Leeuw, 2008; Bergmann *et al*., 2016). SLM predicts that a "merged" category will develop over time, and that it will subsume the phonetic properties of the perceptually linked L1 and L2 speech sounds. On the production side, given that a single, merged L1–L2 category is used to produce corresponding speech sounds in the L1 and L2, the SLM predicts that «[...] the more a bilingual approximates the phonetic norm for an L2 speech sound, the more [his/] her production of the corresponding L1 speech sound will tend to diverge from L1 phonetic norms» (Flege et al., 2003: 469-470).

The second mechanism through which L1 and L2 phonetic segments interact is called "phonetic category dissimilation". It operates when a new category has been established for an L2 speech sound, that is when a newly encountered L2 sound is perceived to be sufficiently dissimilar from the nearest L1 sound ("dissimilar sounds" were originally named as "new" in Flege, 1987). The newly established L2 phonetic category will shift away from the closest L1 sound by a mechanism of category dissimilation, because «bilinguals strive to maintain phonetic contrast between all of the elements in their combined L1-L2 phonetic space in the same way that monolinguals strive to maintain phonetic contrast among the elements making up their (L1-only) phonetic space» (Flege et al., 2003: 470). The predictions on the production side are that the phonetic properties of a new L2 category and the closest L1 category will diverge from one another and that the productions of bilinguals will display values that are more extreme than in monolinguals. Summarizing, L1 and L2 sounds are posited to exist in a shared system, where there is a general pressure to keep them distinct, and to be related to each other on an allophonic, rather than a phonemic, basis.

In the present study, we will explore if regressive CLI occurs from L3-English on L1-Belluno dialect. Phonetic/phonological influence exerted by a third language on the native system has so far received insufficient attention, with respect to the larger amount of studies concerning patterns of L2 interaction with L1 in late bilinguals (Cabrelli Amaro, 2012: 32; see also De Angelis, 2007; Hammarberg, 2001; Rothman, Cabrelli Amaro & de Bot, 2013, for a review). Moreover, the distinction between L3/*Ln* and L2 acquirers has been often neglected, although the former display a larger phonological awareness, as well as wider linguistic repertoire (Gut, 2010). Consequently, their multilingual competence makes it more likely for CLI to occur. As Cabrelli Amaro (2017) further highlights, most studies on CLI have addressed progressive transfer from the L1 and/or L2 to the L3, to the detriment of L3 regressive transfer (i.e. when L3 affects the L2 and/or L1).

Extending the predictions of the SLM to involve also L3, we could posit that the phonetic elements of all languages of our multilingual speakers, L1 dialect, L2 Veneto Italian and L3 AusEng exist in a common phonological space and are related to one another on an allophonic basis. In our specific case, the phonological subset

of coronal obstruents we investigate /θ, s, ʧ/ is shared by the three languages, but we know that at least the phonetic properties of /s/ differ in L1-dialect and L3-English. /s/ in Belluno dialect (as in other dialectal varieties of the Veneto region) has a retracted place of articulation such that it sounds similar to the postalveolar /ʃ/.

Based on the assumptions and empirical results of the SLM, we can predict that if the Veneto immigrants have failed to create a new L3 phonetic category for English /s/, either because they perceived it as the same sound or because they perceived it as deviant with respect to the Belluno /s/ but still classified it as equivalent to the/s/ of their L1, the acoustic properties of the frication noise of /s/ in the Belluno dialect should be intermediate between those of the native category and those of English /s/, as the immigrant speakers merged the L1 and L3 category. However, as we can assume that after 50 years of residency in the English-speaking country the immigrant speakers have approximated the L3 phonetic norms, we expect that in time their productions diverged from the L1 norms; that is, we expect that the /s/ of the multilingual speakers will be *less* retracted than in the speech of monolingual Veneto speakers living in Veneto. Conversely, if they have formed a new L3 category, we expect that due to the mechanism of phonetic category dissimilation the properties of the frication noise of /s/ as spoken in Belluno dialect will shift away from the properties of English /s/, and therefore that /s/ produced by the Belluno immigrants in their L1 dialect will be *more* retracted than in Italian monolingual speakers.

5. *Materials and methods*

5.1 Origins of the informants

The data used in the present study stem from the *Italian Roots in Australian Soil* corpus (IRIAS[9]; Avesani et al., 2015; Avesani et al., 2017; Galatà et al., in prep.) containing elicited speech samples of dialect L1, Italian L2 and Australian English L3 and providing a significant contribution to the linguistic situation of an Italian immigrant community in an English-speaking country.

Our specific aim here is to carry out a sociophonetic investigation on Italian immigrants from the areas of Cadore and Feltre (in the province of Belluno, Veneto), with the purpose to detect the degree of attrition exerted by the L3 on their native dialect, by considering the variability of selected voiceless coronal fricatives (Avesani et al., 2015). To allow an accurate contrastive analysis, coronal obstruents produced by immigrants are compared to the same target consonants produced by a control group of Italians born and living in their very same villages of origin. Data for the control group have been collected along the same lines and with the same procedure described in Avesani et al. (2015) and Galatà et al. (in prep.).

Sociolinguistic information for both groups of speakers can be found in Table 1.

[9] See the project's webpage at http://irias.filefaustralia.org/.

For the Italo-Australian group (Ita-Au), 4 first-generation Italo-Australian speakers from the area of Belluno have been selected from the IRIAS speech corpus: two male speakers (GPZ and MZN) and two female speakers (CZM and ACS)[10]. From Table 1, we see that they are balanced with respect to: age (from 72 to 82 years), local dialect as L1 (Cadorino and Feltrino varieties), number of years of experience of English as expressed by length of residence (LoR) in Australia (range in years = 51-57). Moreover, the age they started to acquire regional Italian as L2 corresponds for all of them with the beginning of primary school in Italy at age 6. The age of arrival (AoA) in Australia also represents the age they began to learn English as L3 (AoA Eng) by spontaneous immersion. Similarly, the Italian control group (Ita) is composed of two males (ALM and SPR) and two females (BCL and RDP). Their age ranges from 59 to 75, and they match the Italo-Australians for L1-dialect (Cadorino and Feltrino varieties) and competence of Italian as L2 (learned in primary school since age 6). As for Age, while the Italian control speakers from Feltre are comparable to the immigrants from the same area (75 and 74 years old, respectively) the 2 control speakers from Cadore (BCL and ALM) are younger than their Italo-Australian counterparts (on average 60.5 and 78 years old, respectively) and report to have a higher level of education (secondary and middle schools vs primary and middle schools).

Table 1 - *Italo-Australian speakers' sociolinguistic information: Age = age (in years) at time of recording; LoR = Length of Residence in Australia; AoA Eng = Age of Arrival and onset of Acquisition of English; NA = not available*

Group	*ID*	*Sex*	*Age*	*AoA Eng*	*LoR*	*Dialect L1*	*L2*	*L3*	*Level of education*	*Profession*
Ita-Au	CZM	F	74	17	57	Cadorino	Ita	Eng	primary school	housewife
	GPZ	M	82	29	53	Cadorino	Ita	Eng	primary school	craftsman
	ACS	F	78	23	55	Feltrino	Ita	Eng	NA	NA
	MZN	M	72	21	51	Feltrino	Ita	Eng	NA	NA
Ita	BCL	F	62	NA	NA	Cadorino	Ita	NA	secondary school	employee
	ALM	M	59	NA	NA	Cadorino	Ita	NA	middle school	farmer
	RDP	F	73	NA	NA	Feltrino	Ita	NA	primary school	nanny
	SPR	M	75	NA	NA	Feltrino	Ita	NA	primary school	craftsman

5.2 Data selection and preparation

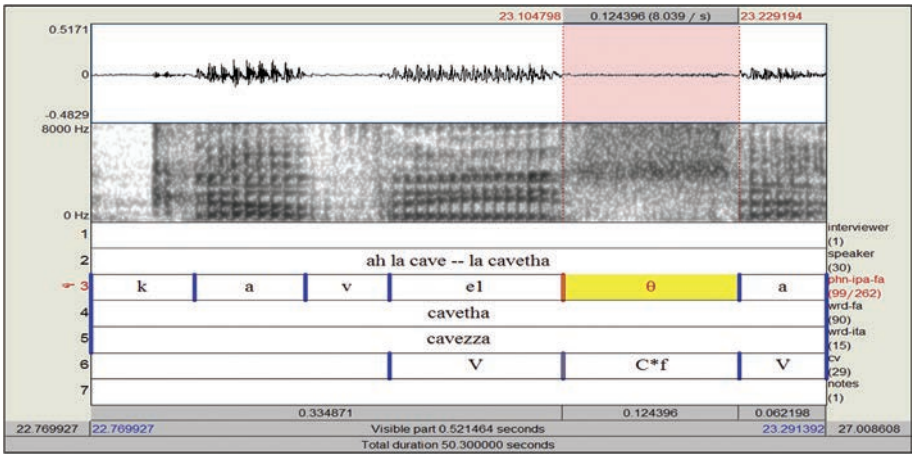
In the present study, data preparation was achieved using a different strategy with respect to the previous works (cfr. Avesani et al., 2015 and Avesani et al., 2017).

First, an orthographic transcription of the audio files (about 2 hours per speaker) was performed by one of the authors by means of ELAN (version 4.9.4). The resulting transcribed *.eaf files were first processed through the Chunk Preparation tool (Reichel,

[10] Dialectal productions of 2 out of 4 participants (e.g. CZM and GPZ) have been already analyzed in Avesani et al. (2015).

Kisler, 2014) to generate derived tiers as input for the Forced Alignment (FA) procedure[11]. From the FA, we obtained two tiers, respectively containing an orthographic word level segmentation (*wrd-fa*) and an IPA phone level segmentation (*phn-ipa-fa*). Then, we added two more tiers: the Italian translation of the target dialectal word (*wrd-ita*) and the target consonant's manner of articulation (*cv*) as well as its preceding or following phonetic context. An example of the resulting *.TextGrid structure is shown in Figure 1 (for more in-depth details, see Galatà et al., in prep).

Figure 1 - *A screenshot of the resulting data organization on the different tiers*



5.3 Acoustic analysis

Consistently with previous work (Avesani et al., 2015; Avesani et al., 2017) and to create a homogeneous set of data, in this study we performed an acoustic analysis only on the set of dialectal voiceless coronal fricatives [θ], [s] and on the fricative release of [ʧ] (from now on [(t)ʃ]) by using an adapted version of a Praat script by Di Canio (2013)[12].

The spectral features of a fricative sound are given by the shape and the size of the oral cavity in front of the constriction (see e.g., Shadle, 1985). Voiceless fricatives /s/ and /ʃ/ are produced through a constriction in the upper anterior portion of the oral cavity between the tongue-tip or tongue-blade (Jones, McDougall, 2009: 280),

[11] The Chunk Preparation and the WebMAUs (Munich AUtomatic Segmentation system) tool employed for the FA procedure are both available from: https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface.

[12] The original script by Di Canio (retrieved from: http://www.acsu.buffalo.edu/~cdicanio/scripts.html) extracts the first four spectral moments (center of gravity, standard deviation, skewness, and kurtosis), global intensity and duration for each fricative. Discrete Fourier Transformations (DFTs) are averaged for each token using time-averaging (based on Shadle, 2012). Prior to the analysis, a 300 Hz low pass cut-off filter was applied to all the recordings to remove any F0-related influence. Then CoG, SDev, Skew and Kurt were computed over the central 80% of the fricative segment's duration using 5 DFTs with an analysis window set to 10ms.

whereas the post-alveolar is produced with a more posterior place of articulation. Therefore, realizations of /ʃ/ are supposed to involve lower frequency concentrations of energy with respect to /s/ (respectively, with a spectral peak at around 2.5-3 kHz and around 4 to 6 kHz, according to Jongman, Wayland & Wong, 2000). On the other hand, the interdental /θ/ generally displays a broad range of peaks above 5 kHz, which is attributed to the relatively short front cavity (Fuchs, Toda & Żygis, 2010; Narayanan, Alwan, 2000).

In this work, we performed a spectral moment analysis[13], in which the power spectrum is treated as a probability distribution (Li, Edwards & Beckman, 2009), to classify the dialectal productions of target coronal fricatives. As previously illustrated in Avesani et al. (2015), we measured: duration, spectral moments (M1=Center of Gravity, M2=Standard Deviation, M3=Skewness, M4=Kurtosis). The Center of Gravity (CoG)[14] provides information about where the energy is concentrated: a higher CoG mirrors a more advanced place of articulation. The Standard Deviation (SDev) measures the variance in the energy distribution and indirectly indicates the degree of laminality: the higher the SDev, the higher the laminality of a given fricative. Skewness indicates the (a)symmetry of the distribution of energy around the average, and is related to the CoG, since the tilt in spectrum correlates with the location of the constriction. Kurtosis indicates the peakedness/flatness of the spectrum and correlates with the degree of laminality of the fricative.

5.4 Data cleaning, data exploration and statistical analysis

Data cleaning, data exploration and statistical analysis were carried out in R (R Core Team, 2018). Complying with the literature (e.g. Fant, 1960; Shadle, Mair, 1996, among others) that reports how lip-rounding affects fricative spectra, the vocalic context that would mostly affect the fricative spectra, i.e. /u/, was excluded from the current analysis and we selected only those fricatives that occurred before /a/, /e, ɛ/ and /o, ɔ/, with the purpose of balancing the anticipatory coarticulatory influence of the following vowel (see also Avesani et al., 2015). From a total of 1518 observations in the selected contexts, we retained 1443 observations after removing a few evident outliers (due mainly to overlapping noise in the recording), all the word final fricatives and all those fricatives with a duration shorter than 37 ms. The final dataset is summarized in Table 2 and Table 3.

For the statistical analysis, linear mixed-effects models (LMMs) were fitted using the *lmer* function of the *lme4* package and the *lmerTest* package in R. We built

[13] Spectral moment analyses are commonly found in literature for the identification and description of stable acoustic cues of fricative noises (e.g., Hughes, Halle, 1956; Stevens, 1960; Shadle, 1985, 2012; Jongman et al., 2000; Harrington, 2010). Yet, Spinu, Lilley (2016) show that cepstral coefficients allow to classify the fricatives for place of articulation with a 10% more accuracy than spectral moments (95%). However, the analysis based on spectral moments in the latest studies provide an accuracy rate in the classification of fricatives that can be as high as 85%.

[14] For an accurate description of spectral moments, see e.g. Forrest, Weismer, Milenkovic & Dougall (1988); Jongman et al. (2000); Jones, McDougall (2009); Li et al. (2009).

up the full model by adding one predictor at a time from a baseline model (*null. model*) including only the intercept as predictor. The baseline model was fitted for each of the dependent variables (e.g. CoG, SDev, Kurt and Skew) by entering the factor *speaker* as random effect with *phonelabel* ([θ] vs [s] vs [(t)ʃ]) nested within *speaker* (to account for the repeated measures design). The additive models were fitted one by one using R's *update()* function by adding potential predictors as fixed effects and their interactions. Models were compared with the *anova()* function from the package *stats4* in R and goodness of fit of each model was assessed by means of Akaike's Information Criterion (AIC). *P*-values of overall effects were determined using Likelihood Ratio Tests (L.Ratio), as implemented in the *anova()* function. Both baseline and additive models were fitted and compared using maximum likelihood (ML) method. Deviations from homoscedasticity or normality have been checked by visual inspection of residual plots. For comparability purposes, the same structure of the model fitted for the dependent variable CoG was used for the other dependent variables SDev, Kurt and Skew. The chosen models were re-fitted to the data using residual maximum likelihood (REML) estimation to obtain unbiased estimates of the covariance parameters (West, Welch & Galecki, 2014: 334). Further inspection followed with pairwise post-hoc analysis (Tukey adjusted) using the *emmeans* package with a 95% confidence level and Kenward-Roger correction of degrees-of-freedom. For the post-hoc analysis we report on the outcomes by providing the estimated marginal mean and the associated standard error (±SE).

Table 2 - *Number of observations grouped by phone and speaker*

	speaker							
	ACS	CZM	BCL	RDP	MZN	GPZ	ALM	SPR
θ	86	32	57	33	48	35	43	38
s	173	23	79	18	114	16	70	112
ʃ	82	37	83	12	93	44	55	60

Table 3 - *Number of observations grouped by phone, speakers group and dialect*

		dialect	
	group	Cadorino	Feltrino
θ	Italians (Ita)	100	71
	Italo-Australians (Ita-Au)	67	134
s	Italians (Ita)	149	130
	Italo-Australians (Ita-Au)	39	287
ʃ	Italians (Ita)	138	72
	Italo-Australians (Ita-Au)	81	175

6. *Results*

In Table 4, we report the results of each model fitted for the four dependent variables CoG, SDev, Kurt and Skew. The predictors entered in each model are the following: *gender* (female vs. male); *phonelabel* ([θ] vs. [s] vs. [(t)ʃ]); *group* (Italians (Ita) vs. Italo-Australians (Ita-Au)); *dialect* (Cadorino vs. Feltrino); three two-way interaction terms *phonelabel*group, phonelabel*dialect* and *group*dialect*.

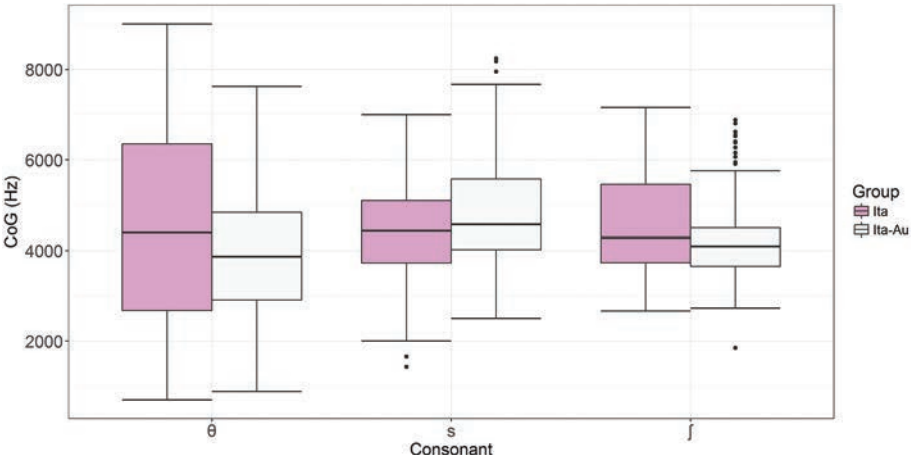
Table 4 - *Results of the four LMMs fitted for the dependent variables CoG, SDev, Kurt and Skew with b estimates and standard errors in parentheses and significance level p for significant predictors in the analysis*

	CoG	SDev	Kurt	Skew
Intercept	5821.199***	3103.086***	-1.615	-0.039
	(309.289)	(197.388)	(5.086)	(0.532)
gender male	-1215.106***	-303.291*	6.511	1.094**
	(188.265)	(119.684)	(4.179)	(0.417)
phonelabel [s]	-378.155	-1080.293***	2.944	0.415
	(394.642)	(252.176)	(3.501)	(0.446)
phonelabel [(t)ʃ]	-374.246	-1026.882***	2.742	0.331
	(394.946)	(252.285)	(3.504)	(0.446)
group Ita-Au	-1155.245**	-23.765	7.302	0.986
	(376.861)	(239.378)	(6.364)	(0.661)
dialect Feltrino	-1854.285***	-461.418	6.496	1.261
	(376.499)	(239.253)	(6.362)	(0.661)
phonelabel [s] * *group* Ita-Au	191.858	-409.641	6.807	0.554
	(461.576)	(293.234)	(4.091)	(0.520)
phonelabel [(t)ʃ] * *group* ItAu	-31.288	-442.589	8.701*	0.815
	(460.749)	(292.975)	(4.085)	(0.519)
phonelabel [s] * *dialect* Feltrino	1291.722**	479.538	-5.033	-0.841
	(461.674)	(293.251)	(4.091)	(0.520)
phonelabel [(t)ʃ] * *dialect* Feltrino	1212.977**	487.923	-6.555	-0.949
	(460.735)	(292.971)	(4.084)	(0.519)
group Ita-Au * *dialect* Feltrino	1535.140***	562.708*	-13.250	-1.434
	(376.875)	(239.433)	(8.358)	(0.834)
Observations	1443	1443	1443	1443
Log Likelihood	-11987.310	-10899.310	-5152.900	-2051.968
Akaike Inf. Crit.	24002.620	21826.630	10333.800	4131.936
Bayesian Inf. Crit.	24076.460	21900.470	10407.640	4205.779
Note:	*p <0.05; **p <0.01; ***p <0.001			

Examining the results in Table 4, and starting with Center-of-Gravity (CoG), we find a significant effect of *gender* ($F(1, 12.76) = 41.657$, $p < .0001$). This is in line with what reported in the literature (e.g., Jongman et al., 2000), and the results of

the fitted model show that male speakers have overall lower CoG values as compared to female speakers (male = 3678.5 ± 132.3 SE; female = 4893.6 ± 134.5 SE). Concerning the interaction *phonelabel*group*, single pairwise comparisons between the Italians and the Italo-Australians reveal to be non-significant. At a first glance, these results might suggest that the place of articulation of each target consonant does not present relevant differences across the two groups (Figure 2).

Figure 2 - *Boxplots for CoG values (in Hz) per group and type of consonant*



Nonetheless, a significant difference emerges for the interaction *group*dialect* (F(1, 12.85) = 16.592, p = .0013): a post-hoc analysis reveals that there is a marginal difference within the speakers originating from Cadore, such that the group of Italian speakers has CoG values globally higher than the Italo-Australian ones ($Ita_{Cadorino}$ = 4962.8 ± 184.6; $Ita\text{-}Au_{Cadorino}$ = 3861.1 ± 193.4; p = .0735). Within the group of Italians, there is also a marginal difference between the Cadorino and Feltrino speakers (p = .0908), with CoG values globally higher in the fricatives of the speakers from Cadore ($Ita_{Cadorino}$ = 4962.8 ± 184.6; $Ita_{Feltrino}$ = 3943.5 ± 193.8).

The *phonelabel*dialect* interaction represented in Figure 3 is significant ($F(1, 12.82) = 4.934$, $p = .0257$) and a post-hoc analysis shows that: the CoG for [θ] in the Cadorino speakers is significantly higher compared to the Feltrino speakers ($\theta_{Cadorino}$ = 4636.0 ± 231.1; $\theta_{Feltrino}$ = 3549.3 ± 229.4; $p = .0509$); for Feltrino speakers, [θ] is only marginally different from [s] in that the first one is lower than the second one ($\theta_{Feltrino}$ = 3549.3 ± 229.4; $s_{Feltrino}$ = 4558.8 ± 227.7; $p = .0878$); for the Cadorino speakers, no significant difference is detected among the three fricatives.

As for Standard Deviation, we found a significant effect of *gender* ($F(1, 12.70) = 6.422$; $p = .0253$), *phonelabel* ($F(2, 12.68) = 32.671$; $p < .0001$) and a significant interaction *group*dialect* ($F(1, 12.73) = 5.523$; $p = .0358$). The other predictors and interactions such as *phonelabel*group* and *phonelabel*dialect* are non-significant.

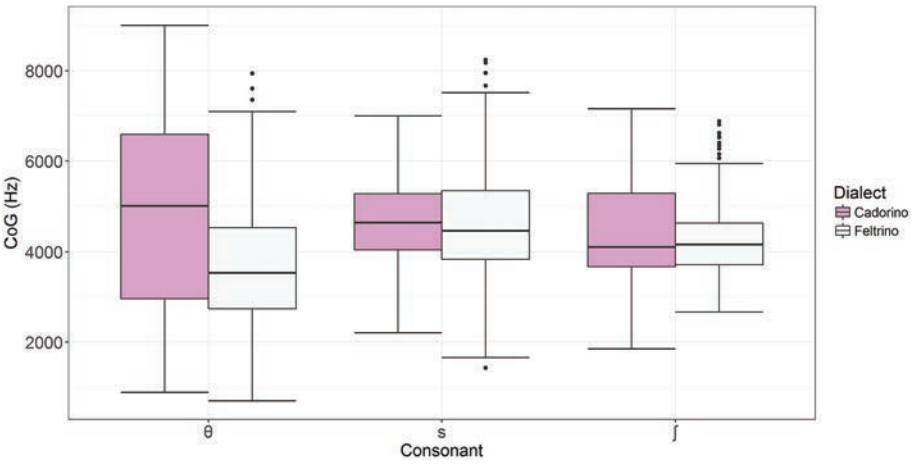
The results of a post-hoc analysis show that: the difference in SDev for *gender* is due to females having higher SDev values compared to males (female = 2317.9 ± 85.1;

male = 2014.7 ± 84.3; p = .0854); for *phonelabel,* [θ] differs from [s] (p < .0001) and from [(t)ʃ] (p < .0001) with higher SDev values for [θ] (2849.5 ± 103.5) compared respectively to [s] (1804.2 ± 103.8) and [(t)ʃ] (1845.3±103.7). SDev for [s] is not significantly different from [(t)ʃ] (p = .9578); despite the significant *group*dialect* interaction (F(1, 12.73) = 5.523; p = .0356), none of the pairwise comparisons results significant.

For Skewness, a significant effect of *gender* was found (F(1, 2.95) = 6.8925; p = .0802) with female speakers having lower Skew values (0.9 ± 0.3) compared to males (1.9 ± 0.3) while all other main effects and interactions are non-significant.

As for Kurtosis, no significant main effects are found.

Figure 3 - *Boxplots for CoG values (in Hz) per type of consonant and dialect (Cadorino vs Feltrino)*



7. *Discussion and conclusions*

In this paper, we acoustically explored spoken dialectal productions of Italo-Australian immigrants from the areas of Cadore and Feltre, Veneto, compared to those of Veneto speakers born and living in their same villages of origin. Specifically, we performed analyses on the four spectral moments (CoG, SDev, Skew, Kurt) of the frication noise of selected voiceless coronal obstruents: the interdental [θ], occurring both in their native dialectal variety (Cadorino and Feltrino, respectively), and AusEng L3 but absent in the phonological inventory of their L2 (the regional variety of Standard Italian); [s] and the fricative portion of ʧ, present in all the repertoires. Our purpose was to identify possible phenomena of attrition, maintenance and/or loss in their fine-grained speech features, after decades of persistent contact with English.

Statistical analyses on acoustic results revealed that the factor *gender* is highly significant for three out of four spectral moments: fricatives as spoken by females

have higher values of CoG and lower values of Skew, both indexing a smaller size of their vocal tract, and higher values of SDev.

As for the fricatives' identity (factor *phonelabel*), our data show that the sibilant fricative [s] and the fricative release of [(t)ʃ] in the dialect of either group of speakers do not differ for any spectral moment, with mean CoG values for [s] that approximate those for [(t)ʃ], indicating a clear retraction of [s] (the estimated mean CoG for [s] and [ʃ] is respectively 4554 and 4518 Hz for Italians, and 4358 and 4099 Hz for Italo-Australians). On the contrary, Tabain (2001) and Jones, McDougall (2009) report that in Australian English [s] has significantly higher CoG values than [(t)ʃ] (such values are similar in other varieties of English: see Harrington, 2010; Shadle, 2012).

Moreover, a significant SDev and a significant post-hoc analysis indicate that [θ] differs from [s] ($p < .0001$) and from [(t)ʃ] ($p < .0001$) for the variance in the energy distribution. That is, [θ] has a low intensity and a spread spectrum, as well as a higher SDev with respect to [s] and [(t)ʃ] (similarly to what reported for AusEng by Tabain, 2001; Jones, McDougall, 2009, and for other varieties of English by Jongman et al., 2000). Coherently with EPG data obtained by Tabain (2001) for Australian English, [θ] reveals a greater acoustic instability and a greater articulatory variability than the sibilant fricatives: in fact, we observe a greater dispersion for [θ] as compared to [s] and [(t)ʃ], both across groups (Figure 2) and across dialects (Figure 3).

A post-hoc test on the significant interaction *phonelable*dialect* has revealed a difference induced by the local variety of dialect on the spectral properties of [θ], such that the dental fricative of speakers originating from Cadore has higher values of CoG than speakers originating from Feltre. One possible source of such acoustic difference could be related to the lower age of the Italian speakers from Cadore (averagely, 60.5 y.o.) with respect to the Italo-Australian speakers from the same area (averagely, 78 y.o.). Italian and Italo-Australian speakers from Feltre are balanced in age (respectively, 75 y.o. and 74 y.o. on average) and do not show such acoustic difference. It could be hypothesized that either the local dialect of the Italians in Cadore has undergone a change after the Italo-Australians left the region or, more likely, that we are facing individual differences. However, we are conscious that the limited number of subjects here analyzed does not allow to assess whether these results are representative of a more general trend, or whether they are due to an idiosyncratic linguistic behaviour of the two younger Veneto speakers from Cadore. The analysis of more speakers will help in solving the question.

Two are the conclusions that can be drawn from the present study. First, on the methodological side, the spectral moment analysis – despite the limitations recently shown by Spinu, Lilley (2016) – remains a valid tool to study the spectral properties of fricative sounds, as in our data SDev successfully separates the sibilant from the non-silbilant fricatives. Second, on the theoretical side, the results on our limited set of speakers do not confirm the predictions based on the SLM about L3 influence on native L1 dialect. The Italo-Australian speakers have not formed

a new category for Australian English [s], as the CoG values do not indicate that dissimilation has taken place between the native [s] and the Australian English [s]. Had it been the case, we would expect that the dialect [s] would be pronounced by the Italo-Australian speakers even more backward in the vocal tract, showing lower CoG values, than by Italian speakers. Second, Italo-Australian speakers have not assimilated their native [s] to the Australian English [s] either, as on the one hand CoG values of Italo-Australian [s] are not intermediate between the native and the corresponding Australian English [s]; nor, on the other hand, are they closer to the Australian English values as we could have expected if they had approximated the L3 norm after so many years of contact with Australian English.

What we can argue is that the phonetic properties of the native [s] have been maintained as such by the Italo-Australian speakers and that there is no evidence from production that the L1 and L3 consonants are represented in a shared phonetic space in the mind of these speakers.

As a final note, we believe that the homogeneity encountered in the sociolinguistic features of Italo-Australian speakers might play a relevant role in explaining their linguistic behavior. In fact, these subjects report that their social networks are generally circumscribed to their families and other members of the immigrant community, with limited external interactions. Moreover, both males and females reveal in their interviews that they feel more comfortable in employing their native dialect, rather than English, in everyday communication. Ultimately, it is worth reminding that none of them has received a formal education in Australia. Arguably, this has implied a substantial limitation in the amount of L3 to which they have been exposed through the years.

Yet, these analyses should be extended to a wider number of subjects, in order to verify whether the results obtained so far could be fully reliable and representative of the overall linguistic situation of the Bellunese community in Australia.

Bibliography

Ahn, S., Chang., C., DeKeyser, R. & Lee-Ellis, S. (2017). Age effects in first language attrition: Speech perception by Korean-English bilinguals. In *Language Learning*, 67(3), 694-733.

Andersen, R.W. (1982). Determining the Linguistic Attributes of Language Attrition. In Lambert, R.D., Freed, B.F. (Eds.), *The Loss of Language Skills*. Rowley: Newbury House, 83-117.

Avesani, C., Galatà, V., Vayra, M., Best, C., Di Biase, B. & Ardolino, F. (2017). Phonetic details of coronal consonants in the Italian spoken by Italian-Australians from two areas of Veneto. In Bertini C., Celata C., Lenoci, G., Meluzzi, C. & Ricci, I.(Eds.), *Sources and Functions of Speech Variation. Disentangling the Role of Biological and Social Factors*. Collana Studi AISV, Milano: Officinaventuno, 281-306.

Avesani, C., Galatà, V., Vayra, M., Best, C., Di Biase, B., Tordini, O. & Tisato, G. (2015). Italian roots in Australian soil: coronal obstruents in native dialect speech of Italian-

Australians from two areas of Veneto. In Vayra, M., Avesani, C. & Tamburini, F. (Eds.), *Il farsi e disfarsi del linguaggio. Acquisizione, mutamento e destrutturazione della struttura sonora del linguaggio*. Milano: Studi AISV, 61-86.

Baker, A., Arcangeli, D. & Mielke, J. (2011). Variability in American English s-retraction suggests a solution to the actuation problem. In *Language Variation and Change*, 23, 347-374.

Bergmann, C., Nota, A., Sprenger, S.A. & Schmid, M.S. (2016). L2 immersion causes non-native-like L1 pronunciation in German attriters. In *Journal of Phonetics*, 58, 71-86.

Best, C. (1995). A direct realistic view of Cross-Language Speech Perception. In Strange, W. (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. Baltimore: York Press, 171-204.

Best, C., Tyler, M. (2007). Nonnative and second-language speech perception. In Bohn, O.S., Munro. M.J. (Eds), *Nonnative and Second-language Speech Perception: Commonalities and Complementarities, Language Experience in Second Language Speech Learning. In Honor of James Emil Flege*. Amsterdam: John Benjamins, 13-45.

Bettoni, C. (1981). *Italian in North Queensland. Changes in the speech of first and second generation bilinguals*. Townsville: Capricornia.

Bettoni, C. (2000). La terza generazione italiana all'estero. In *Italiano e oltre*, 15(1), 50-54.

Bettoni, C., Rubino, A. (1996). *Emigrazione e comportamento linguistico. Un'indagine sul trilinguismo dei siciliani e dei veneti in Australia.* Galatina: Congedo editore.

Cabrelli Amaro, J. (2017). Testing the Phonological Permeability Hypothesis: L3 phonological effects on L1 versus L2 systems. In *International Journal of Bilingualism*, 21(6), 698-717.

Canepari, L. (1984). *Lingua italiana nel Veneto*. Padova: CLESP.

Celata, C., Cancila, J. (2010). Phonological attrition and the perception of geminate consonants in the Lucchese community of San Francisco (CA). In *International Journal of Bilingualism*, 14(2), 185-209.

Campolo, C. (2009). L'italiano in Australia. In *Italiano LinguaDue*, 1, 128-141.

Caruso, M. (2010). *Italian Language Attrition in Australia. The Verb System*. Milano: Franco Angeli.

Chang, B.C. (2012). Rapid and multifaceted effects of second-language learning on first-language speech production. In *Journal of Phonetics*, 40, 249-268.

Corazza, A., Grigoletti, M. & Pellegrini, E. (2012). *Australia solo andata. Un secolo di emigrazione veronese nella terra dei sogni*. Verona: Cierre Edizioni.

De Angelis, G. (2007). *Third or Additional Language Acquisition*. Clevedon: Multilingual Matters.

De Bot, K., Clyne, M. (1994). A 16-year longitudinal study of language attrition in Dutch immigrants in Australia. In *Journal of Multilingual and Multicultural Development*, 15(1), 17-28.

DeKeyser, R.M., Larson-Hall, J. (2005). What does the critical period really mean?. In Kroll J.F., de Groot, A.M.B. (Eds.), *Handbook of Bilingualism: Psycholinguistic Approaches.* New York: Oxford University Press, 88-108.

DE LEEUW, E. (2008). When your native language sounds foreign: a phonetic investigation into first language attrition. PhD Dissertation, Queen Margaret University Edinburgh.

DE LEEUW, E., SCHMID, M.S. & MENNEN, I. (2010). The effects of contact on native language pronunciation in an L2 migrant setting. In *Bilingualism: Language and Cognition*, 13(1), 33-40.

DE MAURO, T. (1972). *Storia Linguistica dell'Italia Unita*. Bari: Laterza.

DI CANIO, C. (2013). *Spectral moments of fricative spectra script in Praat*. http://www.acsu.buffalo.edu/~cdicanio/scripts.html.

FANT, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

FLEGE, J.E. (1987). The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. In *Journal of Phonetics,* 15, 47-65.

FLEGE, J.E. (1995). Second language speech learning: Theory, findings, and problems. In STRANGE, W. (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. Baltimore: York Press, 233-277.

FLEGE, J.E., MACKAY, I.R.A. & MEADOR, D. (1999). Native Italian speakers' perception and production of English vowels. In *Journal of Acoustical Society of America*, 106(5), 2973-2987.

FLEGE, J.E., SCHIRRU, C. & MACKAY, I.R.A. (2003). Interaction between the native and second language phonetic subsystems. In *Speech Communication*, 40, 467-491.

FLEGE, J.E. (2007). Language contact in bilingualism: Phonetic system interactions. In COLE, J. HUALDE, J. (Eds.), *Laboratory Phonology 9*. Berlin: Mouton de Gruyter, 353-380.

FORREST, K., WEISMER, G., MILENKOVIC, P. & DOUGALL, R.N. (1988), Statistical analysis of word-initial voiceless obstruents: Preliminary data. In *Journal of the Acoustical Society of America*, 84, 115-123.

FUCHS, S, TODA, M. & ŻYGIS, M. (Eds.) (2010). *Turbulent Sounds: An Interdisciplinary guide*. Berlin: Mouton De Gruyter.

GALATÀ, V., AVESANI, C., BEST, C., DI BIASE, B. & VAYRA, M. (in preparation). *The Italian Roots In Australian Soil multilingual speech corpus*.

GALLINA, F. (2011). *Australia e Nuova Zelanda*. In VEDOVELLI, M. (Ed.), *Storia linguistica dell'emigrazione italiana nel mondo*. Roma: Carocci, 429-475.

GUT, U. (2010). Cross-linguistic influence in L3 phonological acquisition. In *International Journal of Multilingualism, 7(*1), 19–38.

HAMMARBERG, B. (2001). Roles of LI and L2 in L3 production and acquisition. In CENOZ, J., HUFEISEN, B. & JESSNER, U. (Eds.), *Cross-Linguistic Influence in L3 Acquisition: Psycholinguistic Perspectives*. Clevedon-Philadelphia: Multilingual Matters, 21-41.

HARRINGTON, J. (2010) Acoustic Phonetics. In HARDCASLE, W., LAVER, J. & GIBBON, F. (Eds.), *The Handbook of Phonetic Sciences*. Chichester: Wiley - Blackwell, 81-129.

HORVATH, B. (1985). *Variation in Australian English: The Sociolects of Sydney*. Cambridge: Cambridge University Press.

HUGHES, G.W., HALLE, M. (1956). Spectral Properties of Fricative Consonants. In *Journal of the Acoustical Society of America*, 28(2), 303-310.

Hyltenstam, K., Abrahamsson, N. (2003). Maturational constraints in SLA. In Doughty, C.J, Long, M.H. (Eds.), *Handbook of second language acquisition*. Oxford: Blackwell, 539-588.

Jones, M., McDougall, K. (2009). Acoustic character of fricated /t/ in Australian English: a comparison with /s/ and /ʃ/. In *Journal of the International Phonetic Association*, 29(3), 265-289.

Jongman, A., Wayland, R. & Wong, S. (2000). Acoustic characteristics of English fricatives. In *Journal of the Acoustical Society of America,* 108, 1252-1263.

Köpke, B., Schmid, M.S., Keijzer, M. & Dostert, S. (2007). *Language Attrition. Theoretical perspectives*. Amsterdam-Philadelphia: John Benjamins Publishing Company.

Kroll, J.F., Bobb, S.C. & Wodniecka, Z. (2006). Language selectivity is the exception, not the rule: arguments against a fixed locus of language selection in bilingual speech. In *Bilingualism: Language and Cognition*, 9(2), 119-135.

Kroll, J.F., de Groot, A.M.B. (Eds.) (2005). *Handbook of Bilingualism: Psycholinguistic Approaches*. New York: Oxford University Press.

Li F., Edwards, J. & Beckman, M. (2009). Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers. In *Journal of Phonetics*, 37, 111-124.

Mägiste, E. (1979). The competing language systems of the multilingual: a developmental study of decoding and encoding processes. In *Journal of verbal learning and verbal behaviour*, 18(1), 79-89.

Major, R.C. (1992). Losing English as a first language. In *Modern Language Journal*, 76(2), 190-208.

Major, R.C. (2010). First language attrition in foreign accent perception. In *International Journal of Bilingualism*, 14(2), 163-183.

Mennen, I, Scobbie J.M., de Leeuw, E., Schaeffler, S. & Schaeffler, F. (2010). Measuring language-specific phonetic settings. In *Second Language Research,* 26(1), 13-41.

Mereu, D. (2017). Arretramento di /s/ nel sardo cagliaritano: uno studio sociofonetico. In Bertini C., Celata C., Lenoci, G., Meluzzi, C. & Ricci, I. (Eds.), *Sources and Functions of Speech Variation. Disentangling the Role of Biological and Social Factors*. Collana Studi AISV, Milano: Officinaventuno, 45-65.

Nagy, N. (2015). A sociolinguistic view of null subjects and VOT in Toronto heritage languages. In *Lingua*, 164(2), 309-327.

Narayanan, S., Alwan, A. (2000). Noise source models for fricative consonants. In *IEEE transactions on speech and audio processing*, 8(2), 328-344.

Olshtain, E. (1989*).* Is second language attrition the reversal of second language acquisition? In *Studies in Second Language Acquisition*, 11, 151-165.

Opitz, C. (2011). First language attrition and second language acquisition in a second language environment. PhD Dissertation, Trinity College Dublin.

Pavlenko, A. (2004). L2 influence and L1 attrition in adult bilingualism. In Schmid, M.S., Köpke, B., Keijzer, M. & Weilemar, L., *First Language Attrition: Interdisciplinary Perspectives on Methodological Issues*. Amsterdam-Philadelphia: John Benjamins.

R Core Team (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Reichel, U.D., Kisler, T. (2014). Language-independent grapheme-phoneme conversion and word stress assignment as a web service. In Hoffmann, R. (Ed.) *Elektronische Sprachverarbeitung. Studientexte zur Sprachkommunikation,* vol. 71, Dresden: TUDpress, 42-49.

Rothman, J., Cabrelli Amaro, J. & De Bot, K. (2013). Third language acquisition. In Herschensohn, J., Young-Scholten, M. (Eds.), *The Cambridge Handbook of Second Language Acquisition*. Cambridge: Cambridge University Press, 372-393.

Rubino, A. (2006). Linguistic practices and language attitudes of second-generation Italo-Australians. In *International Journal of the Sociology of Language*, 180, 71-88.

Rubino, A. (2014). L'italiano in Australia tra lingua immigrata e lingua seconda. In De Meo, S., D'Agostino, M., Iannaccaro, G. & Spreafico, L. (Eds.), *Varietà di contesti di apprendimento linguistico*, Studi AItLA 1, Milano: AItLA, 241-261.

Schmid, M.S. (2011). *Language attrition*. Cambridge: Cambridge University Press.

Schmid, M.S., Köpke, B. (2013). *First Language Attrition*. Amsterdam-Philadelphia: John Benjamins.

Schmid, M.S., Köpke, B., Keijzer, M. & Weilemar, L. (Eds.) (2004). *First Language Attrition: Interdisciplinary Perspectives on Methodological Issues*. Amsterdam/Philadelphia: John Benjamins.

Seliger, H.W., Vago, R.M. (1991). *First Language Attrition*. Cambridge: Cambridge University Press.

Shadle, C.H. (1985). The acoustics of fricative consonants. PhD Dissertation, MIT.

Shadle, C.H. (2012). On the acoustics and aerodynamics of fricatives. In. Cohn A. C, Fougeron, C. & Huffman, M. (Eds.), *The Oxford Handbook of Laboratory Phonology*. Oxford: Oxford University Press, 511-526.

Shadle, C.H., Mair, S.J. (1996). Quantifying spectral characteristics of fricatives. *Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia, 3-6 October 1996, 1521-1524.

Spinu, L., Lilley, J. (2016). A comparison of cepstral coefficients and spectral moments in the classification of Romanian fricatives. In *Journal of Phonetics,* 57, 40-58.

Stevens, P. (1960). Spectra of fricative noise in human speech. In *Language and Speech*, 3(1), 32-49.

Stevens, M., Harrington, J. (2016). The phonetic origins of /s/-retraction: Acoustic and perceptual evidence from Australian English. In *Journal of Phonetics*, 58, 118-134.

Stuart-Smith, J., Sonderegger, M., McAuliffe, M., Mcdonald, R., Mielke, J., Thomas, E. & Dodsworth, R. (2018). Dialectal and social factors affect the phonetic bases of English /s/-retraction. Poster presented at the *16th Conference on Laboratory Phonology – LabPhon16*, Lisbon, 19-22 June 2018.

Tabain, M. (2001). Variability in fricative production and spectra: Implications for the hyper – and hypo – and quantal theories of speech production. In *Language and Speech*, 44(1), 57-94.

Tosi, A. (1991). *L'italiano d'oltremare. La lingua delle comunità italiane nei Paesi anglofoni*. Firenze: Giunti.

Weltens, B., De Bot, K. & Van Els, T. (Eds.) (1986). *Language Attrition in Progress.* Dordrecht: Foris.

West, B., Welch, K. & Galecki, A. (2014). Linear Mixed Models. New York: Chapman and Hall/CRC.

Yağmur, K. (1997). *First language Attrition Among Turkish Speakers in Sydney*. Tilburg: Tilburg University Press.

Zamboni, A. (1974). *Veneto*. Pisa: Pacini Editore.

Zamboni, A. (1988). Italienisch: Areallinguistik IV a) Venezien. In Holtus, G., Metzeltin, M. & Schmitt, C. (Eds.), *Lexicon der Romanistischen Linguistik*. Tubingen: Niemeyer Verlag, vol. IV, 517-538.

DOI: 10.17469/O2104AISV000006

CAMILLA BERNARDASCI, STEFANO NEGRINELLI

Analisi fonetiche in due dialetti lombardo-alpini: parlato spontaneo e parlato controllato a confronto

This study compares results obtained from analyses of spontaneous and controlled speech in the dialects of Cavergno and Olivone, two small villages situated in the Italian-speaking part of Switzerland. In Cavergno, the contrast between the voiceless palatal plosive [c] and the voiceless palato-alveolar affricate [ʧ] was studied using the 'center of gravity' (CoG) as an acoustic correlation of the place of articulation. In Olivone, we focused on the contextual distribution of the mid front vowels ([e] in paroxytone words, [ɛ] in oxytone words). Thanks to the collaboration of six informants (three in each village), data from a sample of spontaneous speech and from a questionnaire of controlled speech were collected for both locations to verify if spontaneous speech, unlike controlled speech, could detect an incipient language change.

Keywords: Lombard dialects, spontaneous speech, controlled speech, palatal obstruents, stressed mid front vowels.

1. *Introduzione*

1.1 Scopo dello studio

Lo scopo dello studio è quello di confrontare risultati ottenuti da analisi condotte su parlato controllato prodotto in un ambiente con condizioni paragonabili a quelle di un laboratorio e su parlato spontaneo registrato in un contesto comunicativo reale[1]. In particolare si cercherà di capire se il parlato spontaneo, a differenza del parlato controllato, possa essere rilevante per individuare e studiare un mutamento linguistico incipiente.

Come noto, in un contesto comunicativo naturale la qualità della trasmissione del segnale e l'interazione tra le istanze comunicative divergono sensibilmente rispetto alle condizioni che si trovano in un contesto controllato: nel primo caso, infatti, si ha un'interazione dialogica tra più parlanti in un ambiente rumoroso, mentre nel secondo contesto ha luogo un'interazione singola in un ambiente silenzioso tra informatore

[1] Il presente lavoro è stato concepito e redatto congiuntamente (l'ordine degli autori è stato selezionato esclusivamente su base alfabetica). Tuttavia, a fini accademici, i §§ 1.3, 2 e 3.2 si devono a CB, i §§ 1.1, 1.2, 3.1 e 4 vanno attribuiti a SN. Gli autori ringraziano i sei informatori (tre di Cavergno e tre di Olivone) che hanno gentilmente accettato di partecipare alle inchieste svolte tra il novembre e il dicembre del 2017, così come Pietro Martini, che ha fatto da intermediario e ci ha guidato per le vie di Cavergno. Desideriamo inoltre ringraziare Stephan Schmid e i due revisori anonimi per i loro consigli e suggerimenti. Siamo infine grati a Dieter Studer-Joho e a Volker Dellwo per l'aiuto nell'elaborazione degli *script* in *Praat* e a Chiara Zanini e Marie-Anne Morand per il supporto nel corso delle analisi statistiche.

e intervistatore. Inoltre, il contesto controllato permette l'utilizzo di un questionario il quale, in fase di allestimento, può tenere conto e bilanciare i contesti in cui il fenomeno studiato ricorre, mentre nel parlato spontaneo un controllo dei diversi contesti risulta essere molto più difficoltoso. Questo studio ha quindi lo scopo di valutare l'incidenza dell'ambiente in cui si svolgono le registrazioni sui risultati di un'analisi fonetica.

Come si espliciterà più nel dettaglio nel § 2, i dati del *corpus* su cui si basano le nostre analisi provengono da due dialetti della Svizzera italiana, quello di Cavergno (459 m/sm) e quello di Olivone (902 m/sm), due località situate in fondo a due valli del Sopraceneri (la valle Maggia e la valle di Blenio) in cui si parlano dialetti lombardo-alpini (cfr. Salvioni, 1907: 156 [724]).

Immagine 1 - *Le due località d'inchiesta: Cavergno e Olivone*

1.2 Le occlusive palatali a Cavergno

In un suo studio su *La risoluzione palatina di* k *e* ĝ *nelle Alpi Lombarde* Carlo Salvioni (1898: 93-94 [1-2]) rimarcava, come già l'Ascoli (1873: 249-316) prima di lui in un capitolo dei *Saggi ladini*, che gli esiti [c] e [ɟ] derivanti dalla palatalizzazione delle occlusive velari davanti ad A non fossero un tratto esclusivo, sul versante alpino, delle sole varietà ladine[2]: anche un nutrito numero di parlate italo-romanze infatti presentava,

[2] Si ricordi, infatti, che questo tratto è citato proprio dall'Ascoli (1873: 337) come primo elemento a supporto della tesi sull'unità ladina.

sebbene con una distribuzione differente, questo tratto così caratteristico. Citando numerosi esempi raccolti di prima mano, Salvioni presenta una panoramica che da ovest a est annovera accanto a varietà delle valli dei bacini della Toce e dell'Adda anche alcune parlate ticinesi[3], tra cui quella di Cavergno che è scelta in rappresentanza delle parlate della valle Maggia (Salvioni, 1898: 103 [11]).

Successivamente lo stesso Salvioni (1935; 1936; 1937) offrirà un commento linguistico più dettagliato allestito sulla base di svariati testi (in prevalenza poesie, con l'aggiunta della *Parabola del figliol prodigo* e di una novella del Boccaccio) raccolti e trascritti di prima mano anni prima dal dialettologo bellinzonese (Salvioni, 1905).

In posizione iniziale il dialetto di Cavergno distingue il contesto atono (cfr. (1)), non palatalizzante (> [k]), da quello tonico (cfr. (2)), palatalizzante (> [c])[4]:

(1) [kaˈval] 'cavallo', [kaɰˈrin] 'capretto' (Salvioni, 1935: 440 [24])[5], [kaˈvi] 'capelli', [kɐlˈtsɛt] 'calzino', [koˈlɔstro] 'colostro';

(2) [ca] 'casa', [cɐɲ] 'cane', [ˈcɐɰra] 'capra', [cɐmp, cimp] 'campo, campi', [cøl] 'collo', [ˈcynta] 'racconta'.

Tuttavia, come ricorda il Salvioni (1935: 439-440 [23-24]), la regola è tutt'altro che sistematica: accanto agli esempi citati troviamo infatti casi come [ˈkara] 'cara', [kar] 'carro' o [kas] 'caso'.

Sfuggono a tale alternanza, e in questo caso in modo più regolare, i proparossitoni, che mantengono un esito velare (cfr. (3)):

(3) [ˈkodiga] 'cotenna', [ˈkalkoɫ] 'calcolo'.

Per i contesti all'interno o in fine di parola, troviamo un'occlusiva palatale sorda in alcuni nessi consonantici (cfr. (4)), come esito di una geminata (cfr. (5)) e in posizione finale (cfr. (6)):

(4) [c] /s, ʀ _ V: [fraˈʃca] 'frasca', [ʃcyr] 'scuro', [ˈmuʃca] 'mosca', [ˌpɛʃcaˈdua] 'pescatore' (Salvioni, 1935: 440 [24]), [ˈforcɐ] 'forca', [la biˈforca] 'il ramo forcuto';

(5) -cc- > [c]: [ˈsɛca] 'secca', [ˈvaca] 'vacca';

(6) [c] /_#: [lɐːrc] 'largo', [byːʃc] 'boschi'.

Di origine diacronica differente sono i numerosi casi di affricata sorda [ʧ]. In posizione iniziale e interna si ha (-)c+ᴇ, ɪ > [ʧ] (cfr. (7)). Tale esito lo si dà anche come risultato dei nessi latini -ct- (cfr. (8)), (-)cʟ- (cfr. (9)) e -tj- (cfr. (10)), in linea dunque con gran parte delle varietà lombarde:

(7) (-)c+ᴇ, ɪ > [ʧ]: [ʧeːnt] 'cento', [diˈʧembre] 'dicembre', [ˈʧiʎo] 'ciglio';

[3] Per una più ampia panoramica sull'attuale distribuzione delle occlusive palatali nell'arco alpino si confrontino gli studi di Molino, Romano (2004), Romano, Molino & Rivoira (2005) e da ultimo Negrinelli (2018), dove si dànno anche alcune brevi indicazioni sulla situazione in antico.

[4] Se non indicato diversamente, gli esempi sono tratti dal *corpus* raccolto per questo studio.

[5] Quando necessario le trascrizioni fonetiche degli esempi tratti dalla bibliografia sono state adattate all'alfabeto IPA sulla base della tabella di corrispondenza presentata da Barbato (2008: 139-141).

(8) -CT- > [ʧ]: [leʧ] ‘latte’, [feʧ] ‘fatto’, [leʧ] ‘letto’, [ˈpeʧan] ‘pettine’;

(9) (-)CL- > [ʧ]: [ʧef] ‘chiave’, [ʃpeʧ] ‘specchio’, [øʧ] ‘occhio’, [ʧaˈma] ‘chiamare’;

(10) -TJ- > [ʧ]: [ˈbɛʃʧa] ‘molto’.

L’interesse per il fenomeno di palatalizzazione descritto prende spunto da alcuni contributi che negli ultimi anni hanno indagato la realtà acustica di questi suoni e in particolare si sono concentrati sull’effettivo contrasto tra occlusive palatali e affricate palato-alveolari. Queste ricerche hanno infatti dimostrato che tale distinzione, ancora salda in gran parte dell’arco alpino nella prima metà del secolo scorso, oggigiorno non è più riscontrabile a livello acustico per alcune varietà. È il caso ad esempio del vallader (Schmid, 2010) e dello jauer (Schmid, Negrinelli, 2015), dove i due foni sono stati neutralizzati in favore dell’affricata palato-alveolare, con la distinzione mantenuta unicamente dalla diversa resa grafica. La stessa evoluzione è avvenuta oramai da tempo, come noto, anche nel friulano, dove «in varietà di tipo innovativo [...] si hanno /ʧ, ʤ/ da /c, ɟ/ (/ˈʧan, ˈʤat/ ‘cane, gatto’ < /ˈcan, ˈɟat/)» (Miotti, 2015: 382), mentre per il ladino dolomitico Kramer (1981²: 108 e 109) registrava una variazione di tipo allofonico per le varietà fodom, gardenese e fassano: «die phonetische Realisierung des Phonems [č] erfolgt in b. g. f. [fodom, gardenese e fassano, SN], wo es [...] kein Phonem [’ć] mehr gibt, fakultativ als [č] oder [’ć]»; «Heute ist /’ć/ nur noch im Gadertal [badiotto, SN] ein eindeutig von /č/ zu scheidenes Phonem». Sempre per l’area dolomitica nel già citato studio di Schmid, Negrinelli (2015) si è dimostrato come nel marebbano, la varietà dolomitica parlata in Val Marebbe, la distinzione sia mantenuta solo dai parlanti più anziani (oltre i 70 anni), mentre è stata persa dai più giovani, in modo analogo dunque ai casi engadinesi e friulani sopra presentati. Lo stesso non sembra però valere per le Alpi occidentali: gli studi di Molino et al. (2004), Romano et al. (2005) e Romano (2007) hanno dimostrato su base acustica che le varietà dell’Ossola e della Val Sesia mantengono ancora molto bene la distinzione tra le due realizzazioni. Per le varietà ticinesi l’indagine è ancora alle porte: in un recente studio (Negrinelli, 2018) si sono presentate delle prime analisi su tre località lombardo-alpine (tra le quali figura anche Cavergno) situate in valli distinte, dove sembra ancora mantenersi una distinzione tra le diverse realizzazioni, anche se non sempre col medesimo grado.

In virtù dunque di tali risultati è sembrato opportuno analizzare, accanto a dati raccolti in situazione di parlato controllato, anche del materiale registrato in contesto spontaneo, così da poter ottenere indicazioni su un eventuale mutamento incipiente e sulla sua direzione e cogliere possibili fenomeni di accomodamento.

1.3 Le vocali medie anteriori a Olivone

Nel dialetto di Olivone le vocali toniche latine Ĭ, Ē e Ĕ in sillaba originariamente chiusa hanno esiti diversi a seconda della struttura sillabica (cfr. Sganzini, 1928: 157-59; Vicari, 1992: 38). Si vedano gli esempi seguenti:

(11) > [e] in parole parossitone: [ˈbeːlɐ] ‘bella’;

(12) > [ɛ] in parole ossitone: [vɛʧ] 'vecchio'.

Nelle parole parossitone, in cui la vocale tonica si trova in posizione interna romanza, l'esito è quindi medio-alto (cfr. (11)), mentre nelle parole ossitone, in cui la vocale romanza si trova in posizione finale (non assoluta), l'esito è medio-basso (cfr. (12)).

Vi sono inoltre dei contesti in cui, indipendentemente dalla struttura sillabica romanza, si hanno determinati esiti vocalici:

(13) > [ɛ] /_ʀ + C: [fɛrm] 'fermo', [ˈtɛːrɐ] 'terra';

(14) > [e] /_ɴ, ᴍ + C_{occl}: [ˈʃeːndrɐ] 'cenere', [ʧeːnt] 'cento'.

Se la vocale tonica è seguita da un nesso composto da /r/ + C l'esito della vocale sarà sempre medio-basso (cfr. (13)), mentre il contesto di nasale + C_{occl} provoca l'innalzamento al timbro medio-alto (cfr. (14)).

Un terzo esito si trova nei continuatori dei suffissi latini -ĔLLU, -ĔLLI, i quali si svilluppano in una vocale anteriore alta:

(15) -ĔLLU, -ĔLLI > [il]: [kaˈvil] 'capello'.

La distribuzione timbrica basata sulla struttura sillabica illustrata in (11) e (12), che rappresenta la caratteristica più peculiare del sistema vocalico di Olivone, ricorre anche in altre parti del sistema in questione (cfr. Vicari, 1992: 36-40)[6]: osservando gli esiti della vocale bassa, infatti, si trova l'esito palatalizzato [ɛ] solo in posizione interna romanza (cfr. (16a), sillaba tonica non finale) e finale assoluta di parola (cfr. (16b))[7], mentre si ha l'esito non palatalizzato (> [a]) in posizione finale romanza (cfr. (17), sillaba tonica finale)[8].

(16a) [ˈʃkɛːtrɐ] 'scatola', [ˈlɛːna] 'lana', [mʊnˈtɛːɲɐ] 'montagna', [ˈvɛːka] 'vacca';

(16b) [ʒɐˈrɛː] 'gelare', [sɪˈgɛː] 'falciare', [ʃtɛː] 'stare';

(17) [vak] 'vacche', [sak] 'sacco', [laʃ] 'laccio'.

A Olivone «si trovano tracce» di tale distribuzione anche negli gli esiti delle vocali alte /i/ e /y/. In posizione finale romanza, infatti, si ha una pronuncia aperta delle vocali (cfr. (18)) che si oppone alla loro realizzazione chiusa in posizione interna romanza (cfr. (19)), «purché la cons.[onante, CB] successiva (caduta o tuttora presente) non abbia provocato l'allungamento della tonica» (Vicari, 1992: 40):

(18) [gɐˈlɪŋ] 'galline', [vʏŋ] 'uno';

[6] Secondo Vicari (1992: 37), l'evoluzione della distribuzione degli esiti della vocale bassa «va fatta risalire alla differenza di struttura fra voci con ᴀ in posizione finale e rispettivamente interna romanza», quindi allo stesso criterio che regola la distribuzione delle vocali medie anteriori.

[7] A Olivone si ha palatalizzazione sia da sillaba chiusa, sia da sillaba aperta latina.

[8] Ad essere correlata alla struttura sillabica vi è naturalmente anche la durata della vocale tonica: come afferma Sganzini (1928: 152-153) «[l]e prime [le voci ossitone, CB] hanno la tonica breve, le seconde [le voci parossitone, CB] lunga; in questa differenza di quantità sta il motivo del diverso esito della tonica: l'[aː] degli esiti parossitoni ha avuto la possibilità di svolgersi fino a [ɛ]».

(19) [kyˈziːna] ‘cugina’, [mɐˈdyːrɐ] ‘matura’.

A differenza delle vocali basse e di quelle medie anteriori, però, nel caso delle vocali alte questo fenomeno si manifesta con minor sistematicità, come dimostrano, ad esempio, gli esiti seguenti:

(20) [ʃkuˈʂɪːnɐ] ‘ricotta’, [ˈʏltim] ‘ultimo’;

(21) [prim] ‘primo’, [dʒyŋ] ‘giugno’.

È possibile ipotizzare che la distribuzione complementare dei timbri osservabile ancora oggi per le vocali medie e basse fosse in origine più sistematica anche per le vocali alte (come suggerisce Vicari, 1992: 40). Nell’analisi fonetica qui condotta si vuole verificare, per le vocali medie, se ci sono differenze nella distribuzione timbrica (medio-alta o medio-bassa) e nella realizzazione (aree di esistenza) legate al contesto di produzione (parlato controllato e parlato spontaneo). Nel caso vi fossero delle differenze, ci si aspetterebbe che in un contesto comunicativo reale la confusione timbrica sarebbe più probabile che non nel parlato spontaneo di laboratorio. L’analisi contrastiva dei dati raccolti nei diversi contesti potrebbe quindi essere rilevante per cogliere lo stato di un mutamento fonetico in atto, che nel caso del vocalismo tonico del dialetto di Olivone potrebbe consistere nella distribuzione non (più) sistematica di [e] e [ɛ] sulla base della struttura sillabica, come si è visto negli esempi (18-21) per le vocali alte anteriori.

2. *Dati e metodo*

2.1 Dati

I dati su cui si basa l’analisi di questo contributo sono stati raccolti a Cavergno e Olivone tra il novembre e il dicembre del 2017. Nella Tabella 1 sono riportati i dati personali dei parlanti registrati e il numero di occorrenze per ogni informatore.

Tabella 1 - *Parlanti analizzati e numero di occorrenze*

Località	*Sigla*	*Sesso*	*Anno di nascita*	*Professione*	*Nr. di occorrenze*
Cavergno	IvD	m	*1941	contadino	200 (88/112[9])
	RoD	m	*1953	insegnante	151 (50/101)
	TiD	m	*1961	muratore	149 (52/97)
Olivone	GiC	m	*1947	ex segretario comunale	124 (62/62)
	GMB	m	*1945	contadino	102 (51/51)
	MaT	m	*1948	contadino	76 (38/38)

[9] A sinistra si trova l’indicazione per le occorrenze nel palato spontaneo, a destra quella per il parlato controllato.

Il *corpus* di Cavergno è composto da 72 tipi lessicali; in totale sono state analizzate 500 occorrenze, 190 per il parlato spontaneo e 310 per quello controllato. In contesto di parlato controllato il numero di lemmi resta costante (72), mentre il più alto numero di occorrenze è dovuto all'eventuale produzione di forme flesse. In contesto di parlato spontaneo invece il numero di tipi lessicali è risultato essere leggermente minore (60), mentre a crescere sono state le ricorrenze di alcuni lemmi ad alta frequenza (come ad esempio 'casa', 'anche', 'campo', 'cane', 'vacca', i participi passati di 'fare, dire, dare', ecc.). Di questi 60 tipi lessicali 16 sono rappresentati da almeno un'occorrenza anche in parlato controllato.

Il *corpus* di Olivone è composto da 151 parole di parlato spontaneo e da 151 parole di parlato controllato, per un totale di 302 occorrenze. Le parole sono state inserite nel *corpus* a coppie, così da avere, per ogni contesto rappresentato nel parlato spontaneo[10], un'occorrenza corrispondente nel parlato controllato. Per ogni parlante sono state inoltre inserite nel *corpus* a titolo di controllo 12 parole (6 in parlato spontaneo e 6 in parlato controllato) con A tonica in contesto non palatalizzante.

2.2 Metodo: raccolta ed elaborazione dei dati acustici

In entrambe le località d'inchiesta si è adottata la stessa procedura per la raccolta dei dati.

2.2.1 Parlato controllato

Durante l'inchiesta ci si è dapprima concentrati sulla raccolta del parlato controllato: a ogni parlante è stato sottoposto un questionario che indagasse il fenomeno oggetto di studio. Per Cavergno è stato elaborato un questionario di 72 entrate nel quale le tre consonanti /ʧ c k/ si trovano prevalentemente in posizione iniziale, seguite da vocali il più possibile diverse al fine di neutralizzare eventuali effetti di coarticolazione. Per Olivone invece è stato allestito un questionario di 136 entrate composto da una lista di parole con vocale tonica media o bassa in diversi contesti. Agli informatori è stato chiesto di tradurre dall'italiano al dialetto le parole che comparivano sullo schermo di un computer portatile attraverso il programma *SpeechRecorder* (versione 2.10.16). Le voci sono state elicitate in posizione isolata, senza l'ausilio di frasi cornice. Per le registrazioni (in formato .wav) ci si è avvalsi di un'interfaccia audio (USBPre 2) alla quale si è collegato un microfono a cravatta Sennheiser MKE 2 (direttività onnidirezionale, gamma di frequenza di 20-20.000 Hz ± 23dB e coefficiente di trasmissione a vuoto di 10 mV/Pa ±2.5 dB). Le inchieste sono state svolte in un locale tranquillo, anche se non insonorizzato (dove si è comunque tentato di minimizzare i rumori esterni in modo da non compromettere la qualità delle registrazioni) alla presenza dei due intervistatori.

2.2.2 Parlato spontaneo

In un secondo momento si sono raccolti i dati di parlato spontaneo in contesto comunicativo reale: i tre parlanti precedentemente intervistati singolarmente sono stati invitati ad accomodarsi attorno a un tavolo e a parlare liberamente in dialetto

[10] Si è tenuto conto in particolare del modo e del luogo di articolazione della consonante postonica.

di temi a loro scelta[11]. In questo modo, per ognuno dei sei parlanti si è registrato sia del parlato controllato, sia del parlato spontaneo.

Le porzioni di parlato spontaneo sono state registrate tramite un registratore Zoom H2n con quattro microfoni integrati (registrazione stereo, formato .wav). Dopo aver salvato i documenti sonori su computer, tramite il programma *Praat* (versione 6.0.35) sono stati separati i due canali: sulla base della posizione del registratore e della disposizione delle persone attorno al tavolo, la prima traccia è stata utilizzata per segmentare, annotare e analizzare il parlato spontaneo di un parlante, mentre la seconda è stata selezionata per compiere le stesse analisi per gli altri due informatori. Per Olivone sono stati esclusi dall'analisi i passaggi di parlato spontaneo in cui vi erano forti rumori di sottofondo o sovrapposizioni importanti tra gli interlocutori. Una volta portata a termine questa operazione di selezione, si sono avuti a disposizione 1376 secondi di registrazione totali (771 s per GiC, 200 s per MaT, 405 s per GMB). Per i materiali di Cavergno invece si è deciso per una soluzione in parte diversa: la segmentazione dei *file* audio (in tutto 2081 secondi) è stata fatta selezionando per l'analisi tutte le occorrenze pertinenti incontrate durante l'ascolto, avendo cura di tralasciare i casi registrati in regime di sovrapposizione di più voci.

Le consonanti occlusive, affricate e velari del dialetto di Cavergno e le vocali toniche e le consonanti postoniche del dialetto di Olivone sono state segmentate manualmente con l'ausilio di *Praat* mantenendo i parametri di *default* e tenendo conto in particolare degli oscillogrammi, degli spettrogrammi e dell'andamento delle formanti[12]. Per l'analisi dei dati di Olivone è stato creato per ogni registrazione un *TextGrid* organizzato su tre livelli contenenti la trascrizione fonetica dei singoli foni, l'etichettatura della vocale tonica (v) e della consonante postonica (c) e la parola nella sua interezza. La misurazione e l'esportazione dei valori formantici è stata eseguita tramite uno *script*. Per i dati di Cavergno invece il *TextGrid* creato prevedeva solamente due livelli, uno con la trascrizione fonetica dei singoli foni interessati ([ʧ], [c] o [k]) e uno con l'etichettatura della parola. Per analizzare la natura acustica delle diverse consonanti sono stati estratti in *Praat* tramite *script* i valori dei diversi momenti spettrali, il primo dei quali è stato utilizzato come correlato acustico del luogo di articolazione ('media spettrale' o 'Centro di gravità' (CdG)), parametro ampiamente adottato negli studi di fonetica contrastiva sulle ostruenti (cfr. ad es. Forrest, Weismer, Milenkovic & Dougall, 1988; Gordon, Barthmaier & Sands, 2002)[13]: un valore più alto del CdG risulta da una maggiore intensità dell'energia

[11] Si tenga presente che i tre informatori selezionati per entrambe le località si conoscono bene e sono legati da un sentimento di amicizia, motivo per cui non è stato necessario l'intervento dell'intervistatore per guidare la conversazione che, grazie a queste condizioni ottimali, è risultata essere davvero spontanea.

[12] Nei casi di segmentazione dubbia, ad esempio quando dopo una vocale tonica ricorre una consonante liquida o nasale (ma non solo), ci si è basati su una verifica uditiva.

[13] Per completezza si possono richiamare anche altre metodologie per l'analisi dei correlati acustici dei luoghi di articolazione. In prima istanza va citata la classica teoria dei *loci* elaborata da Delattre, Liberman & Cooper (1965) e ulteriormente sviluppata da Sussman, Hoemeke & Farhan (1993) attraverso l'applicazione di rette di regressione, al fine di studiare le transizioni formantiche; lo stesso

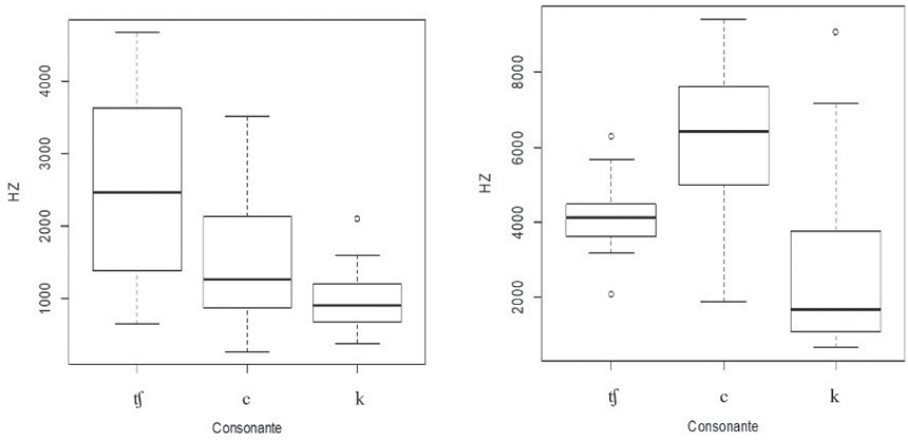
spettrale nelle bande di frequenza superiori, tipica di un rumore consonantico 'acuto' prodotto nella zona anteriore del tratto vocale.

3. *Risultati*

3.1 Le occlusive palatali a Cavergno

Prima di passare al commento dei dati e presentare i risultati dell'analisi è utile rappresentare graficamente per mezzo di diagrammi a scatola (Box-plot 1-6) tutti i valori di CdG misurati per ogni parlante nei due contesti[14].

Box-plot 1-2 - *Valori del Centro di Gravità (in Hz) per* [ʧ, c, k] *in parlato spontaneo (sinistra) e in parlato controllato (destra), parlante IvD*[15]

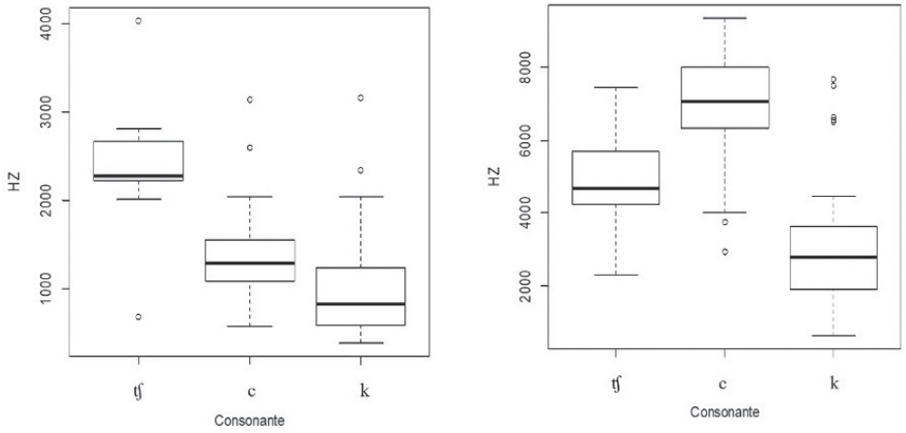


metodo è alla base ad esempio delle analisi svolte nel già citato studio di Romano et al. (2005). Un accenno infine alla *Discrete Cosine Transformation* (DCT; cfr. Watson, Harrington, 1999) utilizzata di recente da Jannedy, Weirich & Helmeke (2015) nella discriminazione del luogo di articolazione di consonanti fricative.

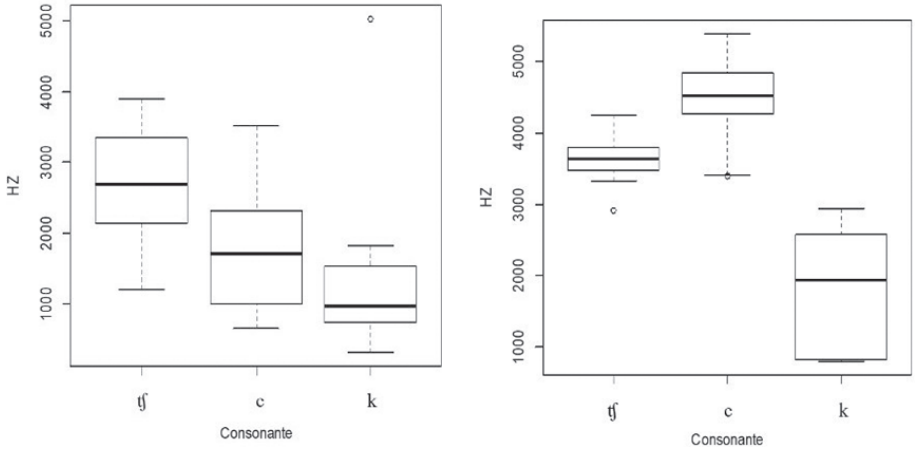
[14] Per l'analisi statistica si è considerato il rapporto tra il valore di CdG di [c] e [ʧ] quale variabile dipendente; le variabili indipendenti sono costituite invece del tipo di consonante ('fono') e dalla modalità di eloquio ('parlato spontaneo' e 'parlato controllato'). A fini descrittivi si riportano le seguenti medie e le deviazioni standard: [c] (media = 4200.61, dev. st. = 2527.54), [ʧ] (media = 3582.54, dev. st. = 1184.56), parlato spontaneo (media = 2031.475, dev. st. = 1043.93), parlato controllato (media = 4969.50, dev. st. = 1631.16). L'analisi della varianza (ANOVA) per i tre parlanti con due variabili indipendenti mostra un effetto non significativo per 'fono' ($F(1, 2) = 3.74, p > 0.05$), ma significativo per 'modalità di eloquio' ($F(1, 2) = 22.29, p < 0.05$)). Si rivela però un'interazione significativa tra i due ($F(1, 2) = 33.89, p < 0.05$).

[15] Un t-test non accoppiato ha rilevato una differenza significativa tra i valori di [c] e [ʧ] unicamente in contesto controllato $t(51.25) = 6.4, p < 0.05$: in parlato spontaneo infatti si ottengono i valori $t(67.6) = -4.1, p > 0.05$, che indicano la non significatività statistica della differenza.

Box-plot 3-4 - *Valori del Centro di Gravità (in Hz) per* [ʧ, c, k] *in parlato spontaneo (sinistra) e in parlato controllato (destra), parlante TiD*[16]



Box-plot 5-6 - *Valori del Centro di Gravità (in Hz) per* [ʧ, c, k] *in parlato spontaneo (sinistra) e in parlato controllato (destra), parlante RoD*[17]



Il confronto tra le misurazioni del CdG in dati raccolti in contesto di parlato spontaneo e in dati raccolti in contesto di parlato controllato può fornirci diverse informazioni. In particolare, concentrandoci solo sulle analisi di parlato controllato (diagrammi a destra), possiamo osservare come tutti i parlanti mantengano una distinzione tra le tre consonanti: come si nota le sovrapposizioni sono molto limitate (in particolare per RoD e TiD,

[16] Un t-test non accoppiato ha rilevato una differenza significativa tra i valori di [c] e [ʧ] in entrambi i contesti: in contesto controllato abbiamo $t(63.65) = 6.4$, $p < 0.05$: in parlato spontaneo invece si ottengono i valori $t(19.63) = -3.84$, $p < 0.05$.

[17] Un t-test non accoppiato ha rilevato una differenza significativa tra i valori di [c] e [ʧ] in entrambi i contesti: in contesto controllato abbiamo $t(62.45) = 11.1$, $p < 0.05$: in parlato spontaneo invece si ottengono i valori $t(19.44) = -2.85$, $p < 0.05$.

meno per IvD). La situazione appare invece meno nitida per i diagrammi dei dati di parlato spontaneo (a sinistra)[18]: in questi casi infatti si nota una sovrapposizione maggiore tra le scatole per tutti e tre i parlanti, benché sembra salva la distinzione tra le tre consonanti (cosa che conferma peraltro la percezione uditiva avuta durante la raccolta dati). Inoltre dal confronto tra le due situazioni balza all'occhio come vi sia uno spostamento notevole nel luogo di articolazione dell'occlusiva [c]: difatti la realizzazione in parlato spontaneo è molto più arretrata rispetto a quella in parlato controllato, e va quasi a confondersi con quella velare (si vedano in particolare TiD e RoD) o con l'affricata (IvD), risultato coerente peraltro con il fenomeno d'accomodamento che era dato attendersi nell'analisi di materiale raccolto in parlato spontaneo.

Sul piano generale la differenza di realizzazione dell'occlusiva [c] messa in luce appunto dal confronto tra una situazione di parlato spontaneo e una di parlato controllato potrebbe essere rilevante, come ipotizzato *supra*, per identificare un eventuale mutamento linguistico: la più ampia sovrapposizione dei valori di CdG è infatti indice di una maggior vicinanza nei luoghi di articolazione delle consonanti, fatto che potrebbe spingere in direzione di una neutralizzazione.

A differenza però di quanto avvenuto in altre varietà (cfr. *supra* § 1.2), dove l'occlusiva [c] ha finito per neutralizzarsi con l'affricata [ʧ], nel nostro caso il mutamento sembra procedere in una direzione differente, vale a dire verso l'occlusiva velare [k][19]: sembra, perché a tale proposito sarebbe utile ampliare la quantità di dati a disposizione e forse approfondire la distinzione tra i diversi contesti. Pare infatti che il mutamento che porta all'eventuale perdita delle occlusive palatali possa essere sensibile al contesto: in posizione iniziale di parola o in posizione interna intervocalica si potrà andare in direzione di una restaurazione della velare (sotto pressione della koinè, che presenta nei medesimi tipi lessicali una velare[20], o del sistema stesso, che alterna forme atone con la velare a forme toniche con l'occlusiva palatale e affianca a quest'ultime casi di velare in posizione tonica, esito inoltre dei proparossitoni e dei prestiti recenti dall'italiano); in posizione interna dopo consonante o in posizione finale invece sembra diminuire l'effetto della koinè e mantenersi più saldamente la distinzione con i casi del tipo [feʧ] 'fatto' (cfr. *supra* (8)).

3.2 Le vocali medie a Olivone

In generale, a Olivone il confronto delle aree di esistenza delle vocali realizzate in un contesto comunicativo reale con quelle delle stesse prodotte, invece, in forma di parlato controllato permette di osservare, come c'era da attendersi, una maggiore estensione delle aree nel primo che non nel secondo contesto. Inoltre, come emerge dai grafici (1-6) riportati sotto, le aree di esistenza delle vocali medio-alte sono distinte sia nel parlato

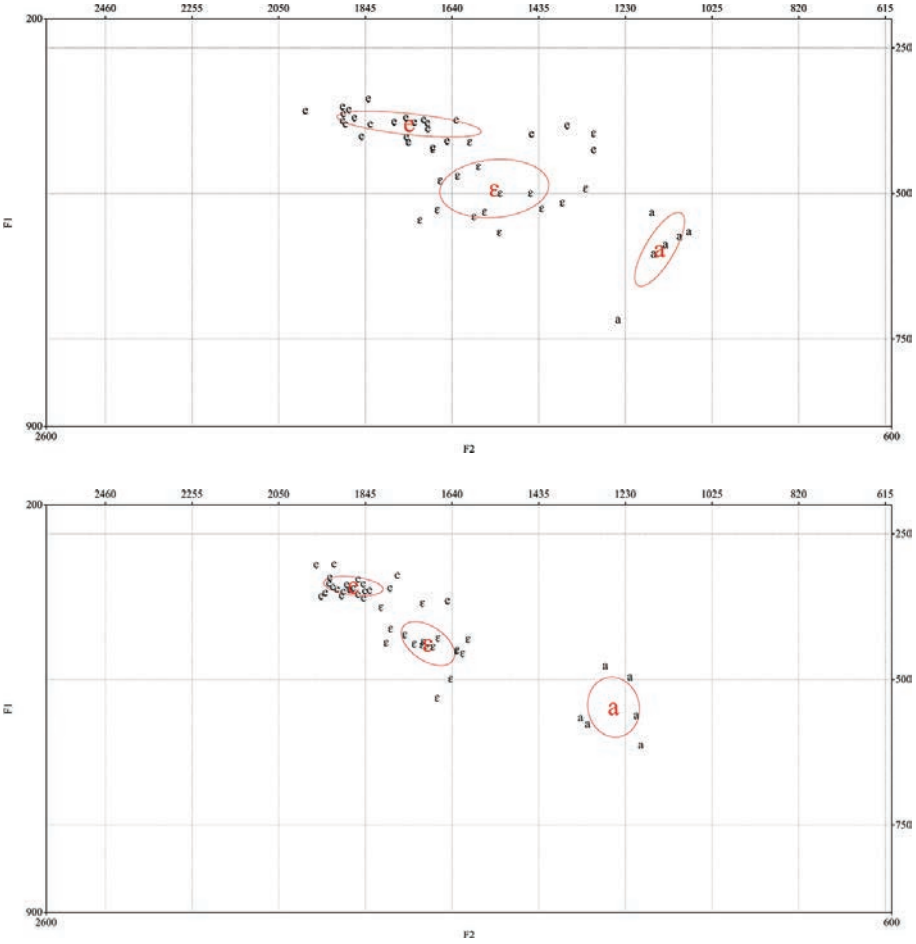
[18] Come sembra indicare il test statistico per IvD, cfr. n. 15.

[19] Tale tendenza sarebbe in linea con quanto ipotizzato da Petrini (1988: 131) in relazione all'influsso della koinè ticinese: «Siamo portati a concludere che la modificazione delle velari in senso palatale tenda a sparire in quanto assente nella lingua "tetto" e dai dialetti degli agglomerati che ne sono da tempo "coperti"».

[20] Cfr. ad esempio, per il corrispettivo nella koinè ticinese di alcuni dei casi sopra citati, il *RID* s.v. 'campo' (1, 235), 'cane' (1, 238) e 'vacca' (2, 719) dove si riportano rispettivamente *can*, *camp* e *vaca* (il grafema <c> equivale a un'occlusiva velare sorda).

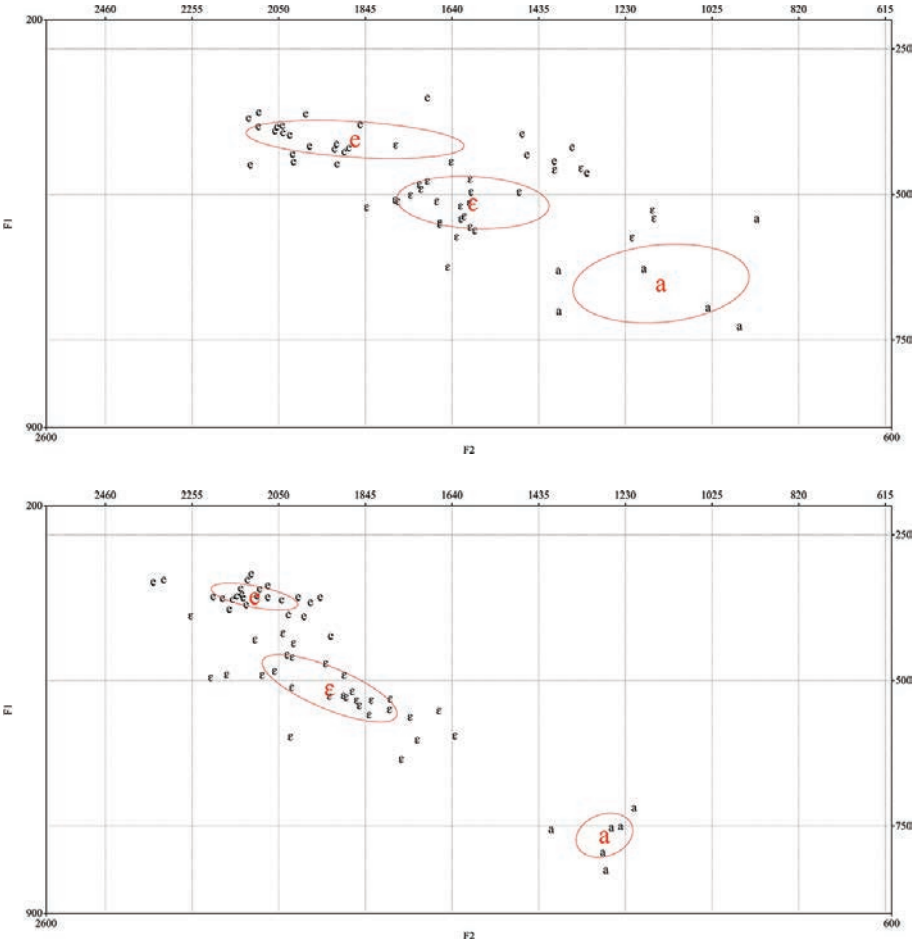
spontaneo sia in quello controllato da quelle delle medio-basse. Nel parlato controllato la distinzione tra i timbri è inequivocabile, mentre nel parlato spontaneo si può osservare, per alcune occorrenze, una tendenza alla centralizzazione. Tuttavia, non si registra un'interazione significativa tra la modalità di eloquio e la realizzazione della vocale[21].

Grafici 1-2 - *Aree di esistenza delle vocali toniche in parlato spontaneo (in alto) e parlato controllato (in basso), parlante GMB*



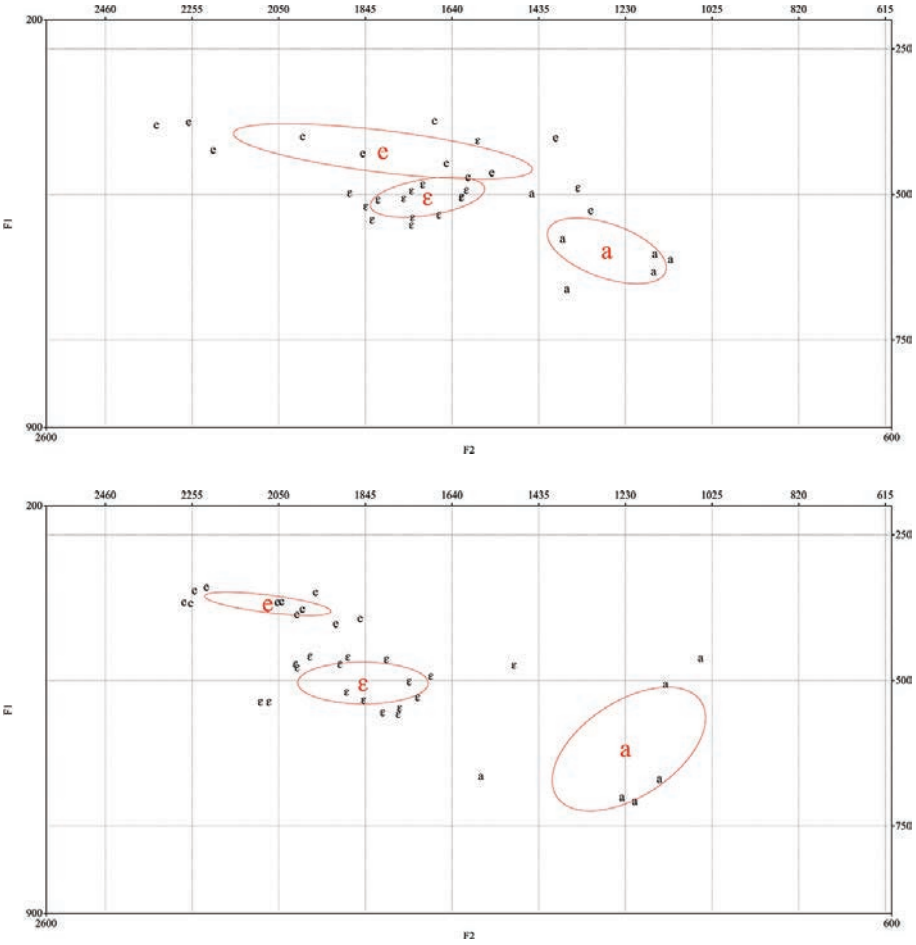
[21] Per l'analisi statistica si è considerato il rapporto tra F1 e F2 delle due vocali medie ([e] e [ɛ]) quale variabile dipendente; le variabili indipendenti sono costituite invece dal tipo di vocale ('fono') e dalla modalità di eloquio ('parlato spontaneo' e 'parlato controllato'). A fini descrittivi si riportano le seguenti medie e le deviazioni standard: [e] (media = 0.202, dev. st. = 0.045), [ɛ] (media = 0.293, dev. st. = 0.050), parlato spontaneo (media = 0.272, dev. st. = 0.066), parlato controllato (media = 0.224, dev. st. = 0.056). L'analisi della varianza (ANOVA) per i tre parlanti con due variabili indipendenti mostra un effetto significativo per 'fono' ($F(1, 2) = 105.63$, $p < 0.05$) e per 'modalità di eloquio' ($F(1, 2) = 369.66$, $p < 0.05$)), ma non rivela interazione tra i due ($F(1, 2) = 0.09$, $p > 0.05$). L'effetto significativo delle due variabili indipendenti 'fono' e 'modalità di eloquio' è confermato anche da un test post-hoc: 'fono' $t(235.72) = -14.745$, $p < 0.05$, 'modalità di eloquio' $t(229.36) = -6.081$, $p < 0.05$.

Grafici 3-4 - *Aree di esistenza delle vocali toniche in parlato spontaneo (in alto) e parlato controllato (in basso), parlante GiC*



Il primo e il secondo parlante (GMB e GiC) presentano dei risultati omogenei e simili: una maggiore estensione delle aree di esistenza e il mantenimento dell'opposizione timbrica nel parlato spontaneo. Nel caso di GiC si nota inoltre che le rappresentazioni grafiche delle realizzazioni delle vocali di controllo [a] nel parlato spontaneo sono molto disperse: questo può essere dovuto al fatto che le parole con A tonica non palatalizzata selezionate per il *corpus* sono monosillabiche (ad es. [aŋ] 'anno').

Grafici 5-6 - *Aree di esistenza delle vocali toniche in parlato spontaneo (in alto) e parlato controllato (in basso), parlante MaT*



Per il terzo parlante (MaT) si registra nel parlato spontaneo un'area di esistenza molto estesa lungo l'asse di F2 (anteriorità/posteriorità) per la vocale medio-alta [e]. Vi è inoltre una sovrapposizione minima delle ellissi delle vocali medio-alte e medio-basse. Queste due peculiarità, sommate alla grande dispersione dei dati di [a] nel parlato controllato, possono dipendere dal numero esiguo di occorrenze registrate per MaT (76, cfr. tabella 1).

Appurato che non si registra una confusione timbrica tra le due modalità di parlato si riprende ora la domanda di ricerca esposta al § 1.3, ovvero se la distribuzione complementare del timbro vocalico medio-alto e medio basso differisca, in un contesto comunicativo reale, da quella registrata in contesto di parlato controllato. Da un confronto mirato delle coppie di parole contenute nel *corpus* emerge che essa corrisponde in entrambi i contesti (quindi, si ha l'esito medio-alto nelle parole parossitone e quello medio-basso nelle parole ossitone). Come risulta dalla Tabella 2,

differenze timbriche tra le stesse parole pronunciate nelle due modalità di eloquio si sono verificate soltanto in casi sporadici [22].

Tabella 2 - *Differenze qualitative tra le vocali medie in parlato spontaneo e parlato controllato*

contesto	*parlante*	*parlato spontaneo*	*parlato controllato*
/_ʀ + C	GiC	> [eː]	> [ɛː]
		[ˈbeːrgum] 'boscaiolo'	[lʏˈzɛːɻtɐ] 'lucertola'
	MaT	[eːrb] 'erbe'	[ˈkwɛːɻtɐ] 'coperta'
-ĔLLU	GMB	> [ɛl]	> [ɪl]
		[sɪˈdɛl] 'secchiello'	[vɐˈdɪl] 'vitello'
	GMB	[lʏˈmɛl] 'veratro (erba velenosa)'	[mɐɻˈtɪl] 'martello'
	GMB	[kɐmˈpɛl] (nome proprio di una montagna)	[fræˈdɪl] 'fratello'
monottongazione del dittongo [ɛi̯]	GiC, GMB, MaT	> [ɛ]	> [ɛi̯]
		[trɛ] 'tre'	[trɛi̯] 'tre'

Le vocali medie anteriori in contesto di /_ʀ + C che, come si è visto, nel dialetto di Olivone danno come esito regolare una vocale medio-bassa, sono state realizzate in due casi isolati da due parlanti differenti con la vocale medio-alta. Ulteriori differenze si registrano per gli esiti del suffisso latino -ĔLLU nel parlante GMB: in tre casi, infatti, il parlante non ha pronunciato la vocale alta non tesa peculiare del dialetto della località in questione (che ricorre invece regolarmente nel parlato controllato), bensì la variante con la vocale medio-bassa [ɛl], diffusa in gran parte dei dialetti lombardo-alpini[23]. Un'ultima deviazione rispetto al parlato controllato si registra nel caso del numerale 'tre', che nel parlato controllato in contesto isolato presenta il dittongo [ɛi̯], mentre in un contesto comunicativo reale, in cui solitamente precede un sostantivo e si trova quindi in posizione fonosintatticamente atona, ricorre, nel parlato di tutti gli informatori, in alternanza alla forma non dittongata [trɛ][24].

Come si è visto, queste divergenze sporadiche tra parlato controllato e parlato spontaneo possono essere almeno in parte ricondotte all'influsso che gli altri dialetti della valle, nonché la koinè ticinese e, nel caso del numerale 'tre', anche l'italiano, esercitano sulla varietà locale.

[22] Per gli esiti regolari nei diversi contesti cfr. § 1.3.

[23] Per il dialetto di koinè cfr. *RID*, in cui si registrano regolarmente *fradèll* 'fratello', *vedèll* 'vitello', ecc.

[24] Nel dialetto lombardo-alpino di koinè per il numerale 'tre' è diffusa la forma *trii* (cfr. *RID*: 2, 686), mentre l'esito che emerge nel parlato spontaneo di Olivone sembra essere più vicino all'italiano *trè*.

4. *Conclusione*

Lo scopo di questo studio era quello di verificare se l'analisi di parlato raccolto in un contesto comunicativo reale potesse essere valida, in aggiunta a quella svolta su parlato controllato da laboratorio, per identificare un mutamento linguistico incipiente.

Nella tabella 3 si sono confrontate queste due modalità di eloquio:

Tabella 3 - *Parlato spontaneo e parlato controllato a confronto*

parlato spontaneo (contesto comunicativo reale)	*parlato controllato (inchiesta con questionario)*
maggior naturalezza	minor naturalezza
peggiore qualità del segnale (sovrapposizioni, risate, rumori di sottofondo)	migliore qualità del segnale
sovrapposizione tra le istanze comunicative	nessuna sovrapposizione (un unico parlante)
maggiore velocità di eloquio	minore velocità di eloquio
presenza di fenomeni di coarticolazione	perlopiù assenza di fenomeni di coarticolazione
può permettere di evidenziare un eventuale mutamento linguistico in atto (Cavergno)	non permette di evidenziare un mutamento linguistico in atto (Olivone)

I vantaggi del parlato controllato registrato in condizioni paragonabili a quelle di laboratorio risiedono in una migliore qualità del segnale, nell'assenza di sovrapposizioni tra le istanze comunicative e di fenomeni di coarticolazione. Si tratta di peculiarità che favoriscono senza dubbio un'analisi fonetica sperimentale. D'altro canto, però, il parlato spontaneo, oltre a essere più naturale, ha permesso di ricavare utili informazioni in relazione a potenziali mutamenti linguistici in atto che non sarebbero emerse se si fossero considerati esclusivamente i dati raccolti in un contesto controllato, come mostrato nel caso dell'analisi delle occlusive palatali sorde nel dialetto di Cavergno, dove l'analisi dei dati di parlato spontaneo ha messo in luce un chiaro arretramento del luogo di articolazione dell'occlusiva palatale [c].

Riferimenti bibliografici

Ascoli, G.I. (1873). Saggi ladini. In *Archivio Glottologico Italiano*, 1, 1-537.

Barbato, M. (2008). Prospetto delle trascrizioni fonetiche. In Loporcaro, M., Pescia, L., Broggini, R. & Vecchio, P. (Eds.), *Carlo Salvioni, Scritti linguistici*, vol. 5. Bellinzona: Edizioni dello Stato del Cantone Ticino, 139-141.

Delattre, P.C., Liberman, A.M. & Cooper, F.S. (1955). Acoustic loci and transitional cues for consonants. In *The Journal of the Acoustical Society of America*, 27 (4), 769-773.

Forrest, K., Weismer, G., Milenkovic, P. & Dougall, R. (1988). Statistical analysis of word-initial word-final voiceless obstruent: Preliminary data. In *Journal of the Acoustical Society of America*, 84, 115-123.

Gordon, M., Barthmaier, P. & Sands, K. (2002). A cross-linguistic acoustic study of voiceless fricatives. In *Journal of the International Phonetic Association*, 32, 141-174.

Jannedy, S., Weirich, M. & Helmeke, L. (2015). Acoustic analyses of differences in [ç] and [ʃ] productions in Hood German. In *Proceedings of the 18th International Congress of Phonetic Sciences,* Glasgow UK, 10-14 August 2015 (http://www.icphs2015.info/pdfs/proceedings.html).

Kramer, J. (1981[2]). *Historische Grammatik des Dolomitenladinischen*. *Lautlehre*, vol. 1. Würzburg: Lehmann.

Miotti, R. (2015). Fonetica e fonologia. In Heinemann, S., Melchior, L. (Eds.), *Manuale di linguistica friulana*. Berlino/Boston: De Gruyter, 367-389.

Molino, G., Romano A. (2004). Analisi acustica e articolatoria di alcuni contoidi palatali in un dialetto della Valsesia. In *Bollettino dell'Atlante Linguistico Italiano*, 27 (2003), 203-221.

Negrinelli, S. (2018). Le ostruenti palatali nell'arco alpino: i primi dati dal progetto *AIS Reloaded*, in: Marcato, G. (Ed.), *Dialetto e società. Presentazione di lavori in corso*, CLEUP, Padova, 33-40.

Petrini, D. (1988). *La koinè ticinese. Livellamento dialettale e dinamiche innovative*. Berna: Francke.

RID = Repertorio italiano-dialetti (2013), 2 voll. Bellinzona: Centro di dialettologia e di etnografia.

Romano, A., Molino, G. & Rivoira, M. (2005). Caratteristiche acustiche e articolatorie delle occlusive palatali: alcuni esempi da dialetti del Piemonte e di altre aree italo-romanze. In P. Cosi (Ed.), *Atti del 1° Convegno AISV. Misura dei parametri. Aspetti tecnologici ed implicazioni nei modelli linnguistici*, Padova, Italia, 2-4 dicembre 2004, 389-428.

Romano, A. (2007). La fonetica sperimentale e gli atlanti linguistici: la sintesi romanza di 'pidocchio' e lo studio degli esiti palatali. In Dorta, J. (Ed.), *Temas de dialectología*, La Laguna-Tenerife, Spagna, 179-204.

Salvioni, C. (1898). La risoluzione palatina di k e ĝ nelle alpi lombarde. In *Studi di Filologia Romanza*, VIII (1901), 1-33. Poi in Loporcaro, M., Pescia, L., Broggini, R. & Vecchio, P. (Eds.), *Carlo Salvioni, Scritti linguistici*, vol. 1. Bellinzona: Edizioni dello Stato del Cantone Ticino, 93-125.

Salvioni, C. (1905). Poesie in dialetto di Cavergno (Valmaggia). In *Archivio Glottologico Italiano*, XVI (1902-1905), 549-590. Poi in Loporcaro, M., Pescia, L., Broggini, R. & Vecchio, P. (Eds.), *Carlo Salvioni, Scritti linguistici*, vol. 1. Bellinzona: Edizioni dello Stato del Cantone Ticino, 375-416.

Salvioni, C. (1907). Lingua e dialetti della Svizzera italiana. In *RIL* (*Rendiconti del reale Istituto Lombardo di scienze e lettere*), XL s. II, 719-736. Poi in Loporcaro, M., Pescia, L., Broggini, R. & Vecchio, P. (Eds.), *Carlo Salvioni, Scritti linguistici*, vol. 1. Bellinzona: Edizioni dello Stato del Cantone Ticino, 151-168.

Salvioni, C. (1935). Illustrazioni dei testi di Cavergno (valle Maggia). I. Fonetica. In *Italia Dialettale*, XI (1935), 1-31. Poi in Loporcaro, M., Pescia, L., Broggini, R. &

Vecchio, P. (Eds.), *Carlo Salvioni, Scritti linguistici*, vol. 1. Bellinzona: Edizioni dello Stato del Cantone Ticino, 417-447.

Salvioni, C. (1936). Illustrazioni dei testi di Cavergno (valle Maggia). II. Annotazioni morfologiche. In *Italia Dialettale*, XII (1936), 1-17. Poi in Loporcaro, M., Pescia, L., Broggini, R. & Vecchio, P. (Eds.), *Carlo Salvioni, Scritti linguistici*, vol. 1. Bellinzona: Edizioni dello Stato del Cantone Ticino, 448-464.

Salvioni, C. (1937). Illustrazioni dei testi di Cavergno (valle Maggia). II. Annotazioni morfologiche (continuazione). In *Italia Dialettale*, XIII (1937), 1-55. Poi in Loporcaro, M., Pescia, L., Broggini, R. & Vecchio, P. (Eds.), *Carlo Salvioni, Scritti linguistici*, vol. 1. Bellinzona: Edizioni dello Stato del Cantone Ticino, 465-519.

Schmid, S. (2010). Les occlusives palatales du Vallader. In Iliescu, M., Siller-Runggaldier, H., Danler, P. (Eds.), *Actes du XXVe Congrès International de Linguistique et de Philologie Romanes*, vol. II. Tübingen: Niemeyer, 185-194.

Schmid, S., Negrinelli, S. (2015). Palatal Obstruents in two Rhaeto-Romance Varieties: Acoustic Analysis of a Sound Change in Progress. In *18th International Congress of Phonetic Sciences*, Glasgow UK, 10 - 14 August 2015 (https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0647.pdf).

Sganzini, S. (1928). Osservazioni sul vocalismo dei dialetti della valle di Blenio (Canton Ticino). In *L'Italia dialettale*, 4, 150-167.

Sussman, H., Hoemeke, K. & Farhan, S.A. (1993). A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation. In *Journal of the Acoustical Society of America*, 94, 1256-1268.

Vicari, M. (1992, 1995). *Valle di Blenio. Documenti orali della Svizzera italiana*, 2 voll. Bellinzona: Centro di dialettologia e di etnografia.

Watson, C.I., Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. In *Journal of the Acoustical Society of America*, 106 (1), 458-68.

DOI: 10.17469/O2104AISV000007

GIANCARLO SCHIRRU

Tensione laringea e consonantismo. Il dialetto armeno di Gavar (Nor Bayazet)

This article aims to illustrate the phonetic and phonological properties displayed by the stop and affricate consonants of the Armenian dialect of Gavar, a small town settled on the lake Sevan, in the Republic of Armenia, which is a secondary derivation of the Eastern Anatolia dialect of Bayazet, completely deleted during the World War I. Such a consonant system is based on three series, traditionally described as (I) plain voiceless, (II) aspirated voiceless and (III) voiced stops. Acoustic analysis reveals that the relevant property distinguishing between series II and series III is represented by the laryngeal tenseness, formally accountable either by the phonological feature [±stiff vocal folds] or by the feature [±slack vocal folds]. Such a representation is consistent with comparative data within Indo-European language family.

Keywords: Armenian dialects, tense consonants, vocal tenseness, [±stiff vocal folds], [±slack vocal folds].

1. *Consonantismo e dialettologia in armeno*

Fin dalla nascita della dialettologia armena, avvenuta tra Otto e Novecento, i diversi coefficienti laringei di realizzazione dei foni consonantici hanno rappresentato uno dei principali criteri di classificazione delle singole varietà armene. Il fatto non è motivo di stupore, in un sistema fonologico caratterizzato da una grande ricchezza del consonantismo occlusivo, nel quale si possono distinguere cinque diverse classi per i coefficienti di luogo diaframmatico, e per la presenza o no di affricazione, ciascuno rappresentato da tre serie distinte per i coefficienti laringei, definite tradizionalmente sorde aspirate, sorde semplici e sonore. Ognuno dei quindici suoni occlusivi così descritti è rappresentato nella scrittura tradizionale locale da un segno grafico che, per le proprietà segmentali ora descritte, ha un rapporto biunivoco tra unità fonologiche e segni grafici, oggetto più volte di commento dalla linguistica moderna. Nella seguente tabella riassuntiva (1) ognuno dei quindici elementi occlusivi è classificato, trascritto in IPA, rappresentato con il segno della scrittura nazionale e con la traslitterazione in caratteri latini in uso nella glottologia. La tabella riflette i valori attribuiti allo standard orientale moderno, quello parlato nell'attuale Repubblica d'Armenia, e alla varietà classica di V sec. d.C.

Tabella 1 - *Valori fonologici dell'armeno orientale moderno e attribuiti all'armeno classico (Xaċatryan, 1988; Vaux, 1998)*

	sorde semplici	*sorde aspirate*	*sonore*
bilabiali	/p/ ա *p*	/pʰ/ փ *p̣*	/b/ բ *b*
alveolari	/t/ ո *t*	/tʰ/ թ *ṭ*	/d/ դ *d*
alveolari affricate	/t͡s/ ծ *c*	/t͡sʰ/ ց *ċ*	/d͡z/ ձ *j*
postalveolari affricate	/t͡ʃ/ ճ *č*	/t͡ʃʰ/ չ *č̣*	/d͡ʒ/ ջ *ǰ*
velari	/k/ կ *k*	/kʰ/ ք *ḳ*	/g/ գ *g*

Proprio la realizzazione fonetica delle tre serie rappresenta un forte elemento di distinzione tra le diverse varietà locali della lingua, che, è bene ricordarlo, ha nel mondo contemporaneo una diffusione di tipo diasporico. Il comportamento della laringe distingue per prima cosa, e in modo radicale, i due principali standard di riferimento dell'armeno moderno, quello occidentale e quello orientale, dal momento che i valori laringei delle consonanti sono addirittura invertiti nelle due varietà. Per cui uno stesso segno consonantico della scrittura ha valore di sonoro nella varietà occidentale, e di sordo in quella orientale, e viceversa, ciò che vale come sonoro a Oriente, è realizzato come sordo aspirato a Occidente. Ad esempio un'importante testata di rivista che si stampa a Venezia fin dal 1843 (ed è quindi tra i più antichi periodici attivi nel nostro paese), scritta in alfabeto nazionale Բազմավեպ, è pronunciata [pʰazmaˈveb] in armeno occidentale, e [bazmaˈvep] in quello orientale.

In un suo articolo fondativo pubblicato all'inizio del Novecento, lo studioso armeno Hračya Ačaryan, allora solo un giovane allievo alla scuola parigina di Antoine Meillet, giunse a una classificazione piuttosto articolata delle varietà dialettali armene sulla base di due grandi fenomeni: i coefficienti laringei del consonantismo occlusivo da un lato, e la coniugazione verbale dall'altro (Adjarean, 1909).

Lo stesso autore, pochi anni prima, aveva pubblicato su «La parole: revue internationale de rhinologie, otologie et laryngologie» un saggio contenente osservazioni sperimentali relative a sei parlanti ciascuno proveniente da una diversa località situata tra la Tracia e il Caucaso: Costantinopoli, Izmit, sulla sponda asiatica del Mar di Marmara, Tiblisi, Suša, Muš, Sivas (Adjarean, 1899). Le osservazioni furono da lui ottenute presso il laboratorio di fonetica diretto a Parigi dall'abate Jean-Pierre Rousselot, che era nato soprattutto come centro di supporto per la documentazione dei dialetti francesi: Rousselot aveva animato negli anni precedenti, con Jules Gilliéron, la «Revue des patois gallo-romanes», il nucleo da cui si svilupperà poi l'*Atlas linguistique de la France*. È utile sottolineare un banale dato cronologico: i dati raccolti da Ačaryan nel passaggio di secolo hanno un'eccezionale rilevanza storica, dal momento che sono tra le poche osservazioni dirette sul fonetismo di varietà linguistiche che,

da lì a poco, furono semplicemente cancellate dalla storia per il massacro dei loro parlanti avvenuto dopo il 1915, il tragico evento che fece scomparire una grande parte dei dialetti armeni dalla carta geografica. Oggi non abbiamo più parlanti armeni nella parte anatolica della Repubblica di Turchia.

Già Ačaryan, nel secondo degli articoli menzionati, avanza una serie di osservazioni sulla particolare natura delle sonore di molti dialetti orientali, definendole sonore aspirate, e trascrivendole quindi con un piccolo <h> in esponente. Più o meno negli stessi anni Eduard Sievers, nel suo trattato di fonetica, nel capitolo dedicato alle sonore aspirate, afferma di aver udito nel dialetto armeno di Aštarak, una cittadina oggi situata nella repubblica di Armenia, poco a nord di Erevan, e allora nella Transcaucasia dell'Impero russo, al posto delle attese sonore, consonanti da lui definite *mediae aspiratae*, e descritte in termini assolutamente paralleli a quelli impiegati per le più note sonore aspirate delle lingue indo-arie (Sievers, 1901: 171-72).

Successivamente, il linguista britannico Sidney Allen (1951), pubblicò un ampio articolo dedicato al dialetto armeno di Nuova Giulfa (Nor J̌owła in armeno), la colonia armena storicamente sobborgo della città persiana di Esfahan, ma oggi inglobata senz'altro come quartiere dalla città. La colonia è nata nel XVI secolo dalla deportazione, da parte dei sovrani persiani abassidi, dell'intera popolazione dell'antica città armena di J̌owła, situata in corrispondenza dell'attuale Naxičevan, la piccola enclave azera situata tra il territorio della Repubblica d'Armenia e quello della Turchia. Anche Allen, che afferma di trarre i suoi dati da una dozzina di sessioni di ascolto effettuate a Londra, nel locale laboratorio di fonetica con un solo parlante originario della cittadina persiana, giunge alla conclusione, suffragata da osservazioni sperimentali, che le occlusive sonore della varietà di Nuova Giulfa siano da considerarsi largamente desonorizzate, e in qualche modo aspirate. Proprio questo articolo innescò una discussione piuttosto ampia nella linguistica dell'epoca, su cui torneremo in chiusura.

La varietà linguistica di Nuova Giulfa è stata indagata in anni recenti dallo studioso americano Bert Vaux, che ha annunciato uno studio organico sull'argomento, ma che ha anticipato alcune sue osservazioni fonologiche in più sedi: Vaux conferma la presenza di suoni sonori aspirati nella varietà in questione, che sono da lui classificati, sotto il profilo fonologico, come [-stiff vocal folds, +spread glottis] ([-pliche vocali rigide, +glottide allargata]) (Vaux, 1998: 212-15).

C'è però una terza varietà locale per cui già Adjarian aveva segnalato una chiara presenza di occlusive da lui descritte come sonore aspirate: si trattava, allora, del dialetto di Bayazet, situato nella parte più orientale del grande altipiano del lago di Van, ai piedi del monte Ararat. La presenza armena nella cittadina, oggi denominata Doğubayazıt, in Turchia, presso il confine con l'Iran, non è più vitale, come si diceva, dopo la prima guerra mondiale. C'è però una colonia secondaria della città di Bayazet, situata nell'attuale repub-

blica di Armenia: l'insediamento è costituito dalla cittadina oggi denominata Gavar (fino al 1991 Kamo, dal nome dell'eroe della rivoluzione bolscevica in Transcaucasia), che si distende sulle sponde del lago Sevan, e da una decina di villaggi rurali che le sono intorno[1]. Questa varietà è stata da noi indagata nel maggio del 2005 con una breve inchiesta effettuata sul posto nell'ambito di uno studio sul consonantismo di alcune varietà armene: i dati raccolti in quell'occasione non sono stati però utilizzati nello studio per cui erano tati pensati, presentato in anteprima nel convegno Aisv del 2010, dal momento che essi avevano una certa autonomia rispetto all'argomentazione svolta in quella sede (vd. Schirru, 2010; Schirru, 2012). Essi costituiscono l'oggetto della presente comunicazione.

2. *L'inchiesta*

L'indagine è stata svolta sulla base di una lista composta da 29 coppie minime, differenziate per la presenza, sempre in posizione iniziale, di una consonante avente il medesimo luogo diaframmatico, ma un diverso coefficiente laringeo.

I dati sono stati raccolti all'interno della piccola università locale[2]. In particolare l'inchiesta ha coinvolto cinque parlanti tutti locali, riassunti nella tabella (2), che in questa sede indichiamo come parlante 1, 2, 3, 4, 5:

Tabella 2 - *Parlanti*

parlante	*sesso*	*età*	*note*
1	F	60	Diplomata alla scuola tecnica, impiegata nella locale università.
2	M	39	Diplomato alla scuola tecnica, impiegato nella locale università.
3	F	19	Studentessa nella locale facoltà di lingue.
4	F	20	Studentessa nella locale facoltà di lingue.
5	F	22	Laureata in lingue; nata in città ma da genitori entrambi di Dilĭjan (una cittadina a nord di Erevan).

I segnali acustici sono stati acquisiti per mezzo di un microfono unidirezionale dinamico montato sulla testa, situato quindi a distanza grosso modo costante per tutti i parlanti, e di un registratore digitale a stato solido. Tutte le analisi sono state effettuate con il programma Praat[3].

[1] L'area dialettale è ora oggetto di un'ampia monografia di Viktor Katvalyan (2016) che si sofferma anche sulle distinzioni laringee nel consonantismo (vd. Katvalyan, 2016: 70-80, 98-103).

[2] Vogliamo ringraziare il prof. Viktor Katvalyan, allora di ruolo presso l'università di Gavar, per l'accoglienza e l'aiuto offerto a questa ricerca.

[3] Registratore digitale Marantz pmd 670 (campionamento 24 KHz, 768 kb); microfono unidirezionale dinamico montato sulla testa Shure WH20; analisi acustica condotta con il software *Praat: Doing Phonetics by Computer*, di Paul Boersma e David Weenink (http://www.fon.hum.uva.nl/praat/).

L'analisi si è concentrata solo su cinque delle coppie minime comprese nella lista originaria usata per l'inchiesta, costituita da un primo gruppo di forme funzionali alla distinzione tra prima e seconda serie (sorde semplici / sorde aspirate); un secondo gruppo di coppie minime invece, tutte aventi dopo la consonante oggetto di analisi la medesima vocale centrale, è servito a indagare la differenza tra prima e terza serie (quindi tra sorde semplici e sonore).

Tabella 3 - *Coppie minime contenute nella lista d'inchiesta*

a. Serie I e II

պայտ *payt* /pajt/ 'ferro di cavallo'	~	փայտ *pայt* /pʰajt/ 'legno'
տանկ *tank* /tank/ 'carro armato'	~	թանկ *tank* /tʰank/ 'costoso'
տարմ *tarm* /taɾm/ 'stormo'	~	թարմ *tarm* /tʰaɾm/ 'fresco, nuovo'
տոռ *toł* /toʁ/ 'linea'	~	թոռ *toł* /tʰoʁ/ 'orsù!'
տուրք *towrķ* /tuɾkʰ/ 'tasse'	~	թուրք *towrķ* /tʰuɾkʰ/ 'turco'
տուփ *towp* /tupʰ/ 'scatola'	~	թուփ *towp* /tʰupʰ/ 'arbusto'
ծեռ *ceł* /t͡seʁ/ 'stelo, paglia'	~	ցեռ *ceł* /t͡sʰeʁ/ 'tribù, popolazione'
ճաղ *čał* /t͡ʃaʁ/ 'bacchetta'	~	չաղ *čał* /t͡ʃʰaʁ/ 'grasso'
կոր *kor* /koɾ/ 'curvo'	~	քոր *kor* /kʰoɾ/ 'prurito'

b. Serie I e III

բահ *bah* /bah/ 'vanga, zappa'	~	պահ *pah* /pah/ 'momento, secondo'
բան *ban* /ban/ 'cosa'	~	պան *pan* /pan/ 'padrone agrario'
դաշտ *dal* /dal/ 'colostro'	~	տալ *tal* /tal/ 'dare'
դատ *dat* /dat/ 'giudizio'	~	տատ *tat* /tat/ 'nonna'
գալ *gal* /gal/ 'venire'	~	կալ *kal* /kal/ 'rimanere'
գավ *gav* /gav/ 'caraffa'	~	կավ *kav* /kav/ 'argilla'

Non c'è molto da dire riguardo alla distinzione tra sorde semplici e sorde aspirate, dal momento che essa è affidata sempre alla differenza di VOT: non se ne tratta oltre in questa sede, in cui vorremmo invece concentrarci sull'altra opposizione rilevante, quella tra sorde e sonore, perché, come si mostrerà, quest'ultima presenta molti aspetti inattesi.

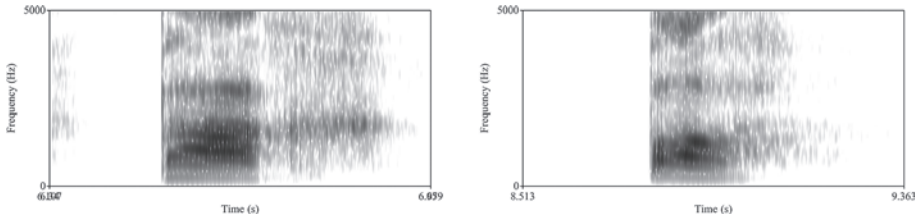
Già Sidney Allen, nell'articolo citato, si rende conto che la differenza specifica tra queste due serie di suoni non va cercata nella porzione di segnale relativa al segmento consonantico, ma si riflette invece sulla successiva vocale: da studioso formatosi alla scuola di John Firth, riesce a individuare una prosodia, piuttosto che andare inutilmente in cerca di un'invarianza:

> The most notable feature differentiating them [*ie*. the voiced stops] from the ejectives, however, is to be found in a following vowel, which is articulated with marked-

> ly stronger breath-force and or a lower pitch than is general in other but comparable contexts (Allen, 1951: 200).

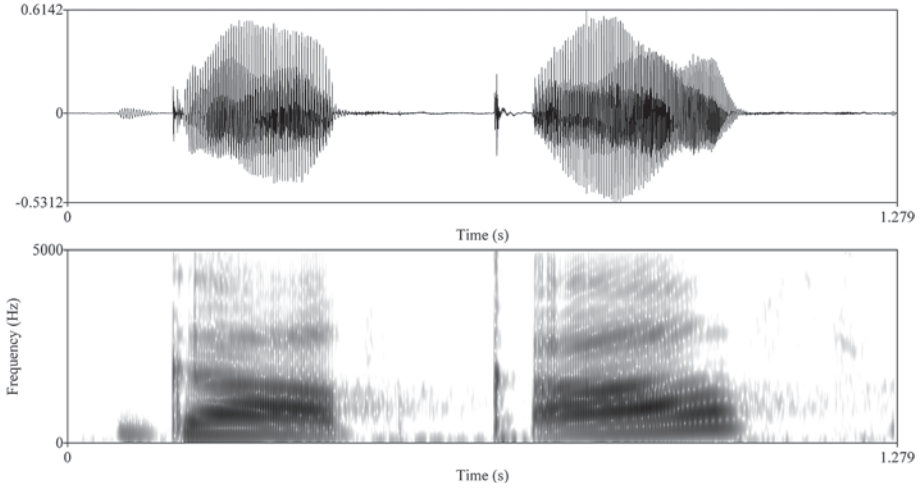
Effettivamente, anche nel nostro caso, come si può notare dai due spettrogrammi relativi alla coppia /bah/ /pah/ realizzata dal parlante 1, la sonora, l'elemento di sinistra, non ha alcuna barra di sonorità: si tratta tecnicamente di una sorda. Ciò che la distingue dalla sorda è nella vocale successiva.

Figura 1 - *Spettrogrammi: [bah] 'vanga' [pah] 'momento', parlante 1*



Le sorde, come si è visto, sono definite senz'altro come eiettive da Allen per il dialetto di Nuova Giulfa: nel nostro caso invece non sono state riscontrate realizzazioni chiaramente eiettive. L'uso di eiettive al posto delle sorde semplici è comunque ben attestato in armeno, ed è stato anche da noi osservato; i dati illustrati nella figura 2 sono relativi a una parlante di 25 anni registrata a Erevan durante la medesima campagna di inchiesta: lo spettrogramma situato a destra illustra una consonante eiettiva, caratterizzata dalla doppia soluzione consonantica, quella orale e la successiva glottidale.

Figura 2 - *Eravan, F, 25 anni: /ɡav/ 'caraffa' (a sinistra) e /kav/ 'argilla' (a destra, eiettiva)*



Per procedere nell'analisi è bene notare che, per varie ragioni (per esempio in alcuni casi è stata omessa inavvertitamente una coppia nell'elicitazione, oppure per una delle due forme il parlante ha usato un lessema diverso da quello indi-

cato nella lista, o il segnale ottenuto non è comunque risultato utilizzabile per l'analisi), non tutti i parlanti hanno dato forme utili per tutte le coppie minime della lista: come si può notare nella Tabella 4, l'ultima coppia manca al parlante 2, le prime due al parlante 3, la terza al parlante 5. C'è comunque un materiale appena sufficiente per avanzare alcune osservazioni.

Tabella 4 - *Coppie di forme disponibili per ogni parlante (serie I e III)*

	parlanti				
	1	2	3	4	5
/bah/ ~ /pah/	x	x		x	x
/ban/ ~ /pan/	x	x		x	x
/dal/ ~ /tal/	x	x	x	x	
/dat/ ~ /dal/	x	x	x	x	x
/gal/ ~ /kal/	x	x	x	x	x
/gav/ ~ /kav/	x		x	x	x

L'analisi è stata condotta con la seguente procedura: nel segnale sonoro sono stati isolati i primi 50 ms. seguenti alla soluzione della consonante; è stato fatto uno spettro di questa porzione di segnale vocalico e sono state misurate le seguenti grandezze, già indicate in letteratura come correlati della diversa tensione laringea: l'ampiezza della prima armonica, l'ampiezza dell'armonica più vicina alla prima formante, l'ampiezza dell'armonica più vicina alla terza formante. La procedura adottata ha alle spalle una buona tradizione di studi: è stata messa a punto da Eli Fischer-Jørgensen nel suo memorabile studio sul consonantismo della gujarati, e più di recente ampiamente utilizzata da Ian Maddieson e Peter Ladefoged nello studio dei sistemi consonantici a tre serie della Cina e dell'Indocina[4]:

a. H2-H1
b. A1-H1
c. A3-H1

dove:

H1 = ampiezza della prima armonica
H2 = ampiezza della seconda armonica
A1 = ampiezza della prima formante
A3 = ampiezza della terza formante

[4] Vd. Fischer, Jørgensen, (1967: 103-15); Maddison, Ladefoged, (1985); Cao, Maddieson, (1993); Gordon, Ladefoged, (2001); Hanson et al., (2001); per l'applicazione ad altre varietà armene cfr. Schirru, (2010); Schirru, (2012).

Come si può osservare nella Tabella 5, l'analisi produce effettivamente un risultato che ha però necessità di una discussione per alcuni aspetti problematici evidenziati in grigio.

Tabella 5 - *Misure*

	sonore	*sorde*	*sonore*	*sorde*	*sonore*	*sorde*
	H2-H1	H2-H1	A1-H1	A1-H1	A3-H1	A3-H1
parlante 1	-2.9	3.0	2.3	8.7	-16.9	-17.1
parlante 2	3.2	1.9	10.0	10.7	-13.3	-11.0
parlante 3	1.8	9.5	2.2	12.7	-25.5	-17.1
parlante 4	0.5	5.8	5.8	13,4	-16.3	-11.0
parlante 5	-0.5	3.6	6.0	10.7	-20.0	-10.0
media	**2.5**	**4.9**	**4.6**	**11.1**	**-18.2**	**-13.4**
p-value		0.00015174		0.00000199		0.01136950

Per prima cosa l'intera terza grandezza presa in esame, consistente nella differenza tra l'ampiezza della terza formante e quella della prima armonica, è risultata poco significativa: il valore di *p* è appena rilevante.

Inoltre, ben due parlanti su cinque reagiscono negativamente al test: si tratta di quelli indicati come parlante 2 e parlante 5.

Tabella 6 - *Misure del parlante del parlante 2 e del parlante 5*

parlante 2

	sonore	sorde	sonore	sorde	sonore	sorde
	H2-H1	H2-H1	A1-H1	A1-H1	A3-H1	A3-H1
	3.2	1.9	10.0	10.7	-13.3	-11.0
p-value	0.39230		0.72542		0.26869	

parlante 5

	sonore	sorde	sonore	sorde	sonore	sorde
	H2-H1	H2-H1	A1-H1	A1-H1	A3-H1	A3-H1
	-0.5	3.6	6.0	10.7	-20.0	-10.0
p-value	0.09996		0.15347		0.07909	

Il comportamento del parlante 5 si spiega tenendo conto che la giovane in questione è nata e cresciuta a Gavar, ma rivelò che entrambi i genitori sono invece di Dilĭjan, un piccolo centro situato a Nord di Erevan, che appartiene a tutt'altra area dialettale. L'altro parlante, l'unico di sesso maschile nel gruppo, ha invece affermato di essere del posto come i suoi genitori: non sappiamo per quale motivo egli presenti un consonantismo diverso da quello atteso. Il fatto ci sembra però di un certo inte-

resse sotto il profilo metodologico: un soggetto che risponde inequivocabilmente in modo negativo a un test, pur avendo tutte le condizioni di contesto che ci porterebbero ad associarlo al comportamento positivo, in realtà conferma la capacità del test di discriminare effettivamente tra due diverse realtà sulla base di dati interni.

Il gruppo, privato di questi due parlanti, presenta dati con altissima significatività per le prime due grandezze analizzate, come si può vedere nella Tabella 7.

Tabella 7 - *Misure del gruppo con esito positivo*

	sonore	*sorde*	*sonore*	*sorde*	*sonore*	*sorde*
	H2-H1	H2-H1	A1-H1	A1-H1	A3-H1	A3-H1
parlante 1	-2.9	3.0	2.3	8.7	-16.9	-17.1
parlante 3	1.8	9.5	2.2	12.7	-25.5	-17.1
parlante 4	0.5	5.8	5.8	13,4	-16.3	-11.0
media	**2.5**	**4.9**	**3.5**	**11.2**	**-19.0**	**15.3**
p-value	0.00013745		0.00000003		0.10926452	

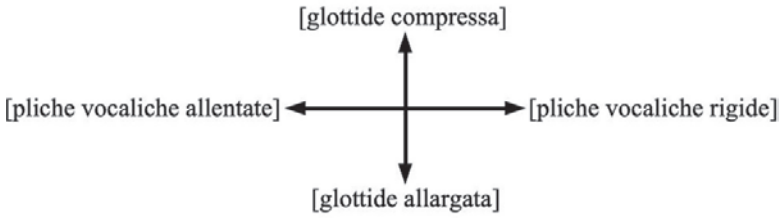
3. *Osservazioni fonologiche*

I risultati raggiunti dall'analisi possono essere proiettati su una riflessione che è stata condotta soprattutto nella fonologia generativa americana, dagli anni settanta in avanti, secondo cui l'insieme dei coefficienti laringei può essere ordinato in uno spazio segnato da due diverse dimensioni: la prima di queste è data dall'apertura o compressione della glottide, ed è responsabile dei fenomeni di aspirazione da un lato, e di glottidalizzazione dall'altro. Una seconda dimensione ortogonale rispetto alla precedente, è relativa alla tensione della glottide, che può spostarsi, rispetto a un valore medio, sia nel senso di una maggiore rilassatezza, sia in quello di una maggiore tensione o rigidità. Da questo punto di vista, la tradizionale distinzione tra sorde e sonore perde di centralità, e può essere vista come un semplice fatto di realizzazione fonetica risultato di configurazioni fonologiche diverse.

Lo schema proposto nella figura 3 fa riferimento solo metaforicamente a uno spazio fisico, come proiezione sul piano cartesiano di organi fisiologici: come è noto, la laringe ottiene i risultati indicati mediante un complesso di molti movimenti compensativi diversi. Non siamo di fronte, tanto per essere chiari, a un dato fisiologico osservabile empiricamente, ma a uno spazio di possibilità che è linguisticamente significativo[5].

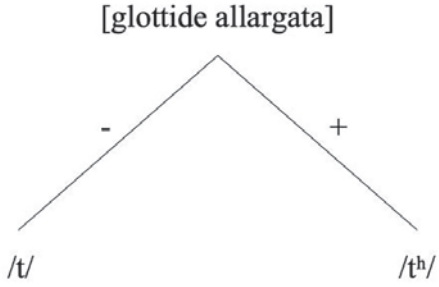
[5] Cfr. Halle, Stevens, (1971); Stevens, (1977); Stevens, (2000); Vaux, (1998).

Figura 3 - *Bidimensionalità dei tratti laringei*



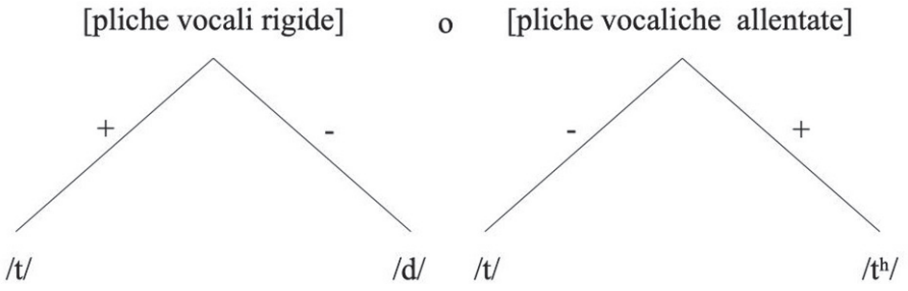
Possiamo senz'altro affidare al tratto [glottide allargata] la differenza tra sorde semplici e sorde aspirate, da collocare quindi sull'asse verticale della distinzione illustrata dalla Figura 3.

Figura 4 - *Distinzione tra serie I e serie II espressa in tratti fonologici*



La differenza tra sorde semplici e sonore sembra piuttosto materia, come abbiamo mostrato, della tensione laringea, e può essere resa sia con il tratto [pliche vocali rigide], come è stato già proposto da Bert Vaux (1998) o mediante il suo complementare [pliche vocali allentate], ovviamente invertendo i valori positivo e negativo. Solo l'analisi dei processi fonologici interni alla lingua può consentire di scegliere tra queste due soluzioni (per esempio, l'emersione dell'elemento non marcato in un processo di neutralizzazione), e non una qualche ragione fisiologica.

Figura 5 - *Distinzione tra serie II e serie III espressa in tratti fonologici*



4. *Osservazioni comparative*

Un altro ordine di osservazioni può essere avanzato sul terreno comparativo: già Hans Vogt (1959) ed Emile Benveniste (1959) reagirono contemporaneamente al citato articolo di Sidney Allen, ricordando che le sonore armene discendono in realtà da sonore aspirate indoeuropee. Ad esempio, la parola armena *dal* 'primo latte, colostro' (presente nella colonna di sinistra della lista di coppie minime della tabella 3), è formata da una base verbale, il cui corradicale sanscrito è il verbo *dhayati* 'succhiare'; quella stessa base è visibile nel greco è τιθήνη (*tithênê*) 'nutrice', nell'elemento θή (*thê*); in latino è la base da cui derivano *fēmina*, *felix*, *fecundus*. L'etimo viene ricostruito come **dʰe(y)-*, con una sonora aspirata iniziale. Secondo alcuni, l'etimologia di *bah* 'vanga', anch'esso un termine presente nella Tabella 3, sarebbe da connettersi col verbo armeno *berem* 'portare'; in ogni caso quest'ultimo è in relazione con il sanscrito *bharati* 'portare', con il greco φέρω (*phérō*) e con il latino *ferō*: abbiamo quindi una sonora aspirata iniziale nell'etimo.

Pertanto, la documentazione di sonore mormorate in vari aree marginali e conservative della dialettologia armena fa pensare a un fenomeno di conservazione, e non di innovazione: che si tratti cioè del riflesso di quei suoni che per l'indoeuropeo ricostruiamo come sonori aspirati[6].

Se così stanno le cose, l'armeno sarebbe l'unica lingua a condividere con l'indo-ario non solo il fatto che la serie sonora aspirata presenta esiti distinti rispetto alle sonore semplici (fatto condiviso anche dal greco, nel germanico e da tutto il ramo italico), ma anche la presenza di altre due proprietà cruciali: (1) la manifestazione di un qualche grado di sonorità; (2) la realizzazione con voce lene e mormorata, come generalmente avviene per la serie che continua la sonora aspirata in tutte le lingue indoarie: per esempio nella hindi, nella gujarati, nella bengali[7].

Inoltre, in molte lingue indo-arie la serie delle sonore aspirate estende le sue proprietà prosodiche sulle vocali adiacenti, dando luogo molto spesso, per esempio, a fenomeni di tonogenesi. Sono osservati casi in cui la consonante esercita un influsso "depressivo" sul tono delle vocali adiacenti[8]: in panjabi, ad esempio, la serie perde la sua indipendenza fonologica (si defonologizza con le sorde), ma trasferisce le sue proprietà sotto forma di un tono ascendente nella vocale successiva[9]. Un fenomeno simile è attestato nell'estremità nord-orientale del continuum indo-ario, nei dialetti bengalesi orientali e in sylheti, dove si ha la perdita delle due serie di occlusive aspirate che confluiscono nelle corrispettive sorde e sonore non aspirate, ma la vocale successiva sviluppa un tono ascendente (o alto)[10].

---

[6] Sul problema, cfr. le osservazioni già avanzate in Pisowicz, (1976); Pisowicz, 1998; Kachaturian, (1983). Prove comparative indipendenti sono addotte in Bolognesi, (1960); Garret, (1991) e Belardi, (2006).

[7] Sulla situazione indo-aria in generale vd. Pandit, (1957); Elizarenkova, (1990: 151-54); Masica, (1991: 102).

[8] Vd. Hombert, Ohala, Ewan, (1979: 47-48); Yip, (2002: 36-38).

[9] Bloch, (1965); Benveniste, (1959); Bahl, (1957); Bahl, (1969: 160-61); Wells, Roach, (1980); Shackle, (1980); Shackle, (1994), Shackle, (2003). Per le lingue dardiche vd. Baart, (1999); Bashir, (2003: 827, 865, 894).

[10] Per i dialetti orientali della bengali, vd. Pal, (1965); per la sylheti, vd. Gope, Mahanta, (2014). Più in generale, sui fenomeni di tonogenesi nelle lingue indo-arie provocati dalle antiche aspirate, vd. Bhascararao, (2016, 384-85), e la bibliografia ivi indicata.

Riferimenti bibliografici

Adjarian, H. (1899). Les explosives de l'ancien arménien. In *La parole. Revue international de rhinologie, otologie et phonétique expérimentale*», 120-27.

Adjarian, H. (1909). *Classification des dialectes arméniens*. Paris: Champion.

Allen, W.S. (1951). Notes on the phonetics of an eastern Armenian speaker. In *Transactions of the Philological Society*, 1950, 180-206.

Baart, J.L.G. (1999). Tone rules in Kalam Kohistani (Garwi, Bashkarik). In *Bulletin of the School of Oriental and African Studies*, 62, 88-104.

Bahl, K.C. (1957). Tones in Punjabi. In *Indian Linguistics*, 17, 1955-1956, 139-47.

Bahl, K.C. (1969). Punjabi. In Emeneau, M.B., Ferguson, C.A. (Eds.), *Linguistics in South Asia*. The Hague: Mouton, 153-200.

Bashir, E. (2003). Dardic. In Cardona, Jain, 2003: 818-94.

Belardi, W. (2006). *Elementi di armeno aureo. II. Le origini indoeuropee del sistema fonologico dell'armeno aureo*. Roma, Il Calamo.

Benveniste, E. (1959). Sur la phonétique et la syntaxe de l'arménien classique. In *Bulletin de la Société de linguistique de Paris*, 44, 46-68.

Bhascararao, P. (2016). Phonetics. In Hock, H.H., Bashir, E. (Eds.), *The Languages and Linguistics of South Asia: A Comprehensive Guide*. Berlin: de Gruyter, 376-87.

Bloch, J. (1965). *Indo-Aryan: From the Vedas to Modern Times*. Paris: Adrien - Maisonneuve.

Bolognesi, G. (1960). *Le fonti dialettali degli imprestiti iranici in armeno*. Milano: Vita e Pensiero.

Cao, J., Maddieson, I. (1992). An exploration of phonation types in Wu dialects of Chinese. In *Journal of Phonetics*, 20, 77-92.

Cardona, G., Jain, D. (Eds.) (2003). *The Indo-Aryan Languages*. London: Routledge.

Elizarenkova, T. Ja. (1990). *Fonologia diacronica delle lingue indoarie*. Napoli: Istituto Universitario Orientale; tr. it. di *Issledovanie po diachroničeskoj fonologii indoarijskich jazykov*. Moskva: Nauka, 1974.

Fischer-Jørgensen, E. (1967). Phonetic analysis of breathy (murmured) vowels in Gujarati. In *Indian Linguistics*, 28, 71-139.

Garrett, A. (1991). Indo-European reconstruction and historical methodologies. In *Language*, 67, 790-804.

Gope, A., Mahanta, S. (2014). Lexical tones in Sylheti. In Gussenhoven, C., Chen, Y. & Dediu, D. (Eds.), *4th International Symposium on Tonal Aspects of Languages (TAL-2014)* (Nijmegen, May 13-14 2014), 10-14 (disponibile in ISCA Archive, http://www.isca-speech.org/archive/tal_2014).

Gordon, M, Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. In *Journal of Phonetics*, 29, 383-406.

Halle, M., Stevens, K.N. (1971). A note on laryngeal features. In *Mit Research Laboratory Quarterly Progress Report*, 101, 198-213; ora in Halle, M. *From Memory to Speech and Back: Papers in Phonetics and Phonology 1954-2002*. Berlin: De Gruyter, 2002, 45-61.

Hanson, H.M., Stevens, K.N., Kuo, H.K-J., Chen, M.Y. & Slifka, J. (2001). Towards models of phonation. In *Journal of Phonetics*, 29, 451-80.

Hombert, J.-M., Ohala, J.J. & Ewan, W.G. (1979). Phonetic explanations for the development of tones. In *Language*, 55, 37-58.

Kachaturian, A. 1983 [= Xačatryan, A.H.]. The nature of voiced aspirated stops and affricates in Armenian dialects. In *Annual of Armenian Linguistics*, 4, 57-62.

Katvalyan, V. (2016). *Bayazeti barbaṙə ev nra lesvakan aṙnčowtyownnerə šṙjaka barbaṙneri het*. Erevan: Hasołik ['Il dialetto di Bayazet e le sue relazioni linguistiche con i dialetti circostanti'].

Maddieson, I., Ladefoged, P. (1985). "Tense" and "lax" in four minority languages of China. In *Journal of Phonetics*, 13, 433-54.

Masica, C.P. (1993). *The Indo-Aryan Languages*. Cambridge: Cambridge University Press.

Pal, A.K. (1965). Phonemes of a Dacca dialect of Eastern Bengali and the importance of tone. In *Journal of the Asiatic Society* (Calcutta), 7, 39-48.

Pandit, P.B. (1957). Nasalisation, aspiration and murmur in Gujarati. In *Indian Linguistics*, 17, 1955-1956, 165-72.

Pisowicz, A. (1976). *Le développement du consonantisme arménien*, Wrocław: Polska akademia nauk.

Pisowicz, A. (1998). What did Hratchia Adjarian mean by 'voiced aspirates' in Armenian dialects?. In *Annual of Armenian Linguistics*, 19, 43-55.

Schirru, G. (2010). La pendenza spettrale come indice acustico della tensione laringea. Osservazioni su alcune varietà armene. In Cutugno, F., Maturi, P., Savy, R., Abete, G. & Alfano, I. (Eds.), *Parlare con le persone, parlare alle macchine: la dimensione interazionale della comunicazione verbale.* Atti del VI Convegno Nazionale AISV - Associazione Italiana di Scienze della Voce (Università di Napoli, 3-5-febbraio 2010). Torriana (Rn): EDK Editore, 339-58.

Schirru, G. (2012). Laryngeal features of armenian dialects. In Whitehead, B.N., Olander, T., Olsen, B.A. & Rasmussen, J.E. (Eds.), *The Sound of Indo-European: Phonetics, Phonemics and Mophophonemics*, Copenhagen: Museum Tuscolanum Press, 435-57.

Shackle, C. (1980). Indko in Kohat and Peshawar. In *Bulletin of the School of Oriental and African Studies*, 43, 482-510.

Shackle, C. (1994). Lahnda. In Asher, R.E., Simpson, J.M.Y. (Eds.), *The Encyclopedia of Language and Linguistics*. Oxford: Pergamon.

Shackle, C. (2003). Panjabi. In Cardona, Jain, 2003: 581-621.

Sievers, E (1901[5]). *Grundzüge der Phonetik zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen*. Leipzig: Breitkopf & Härtel.

Stevens, K.N. (1977). Physics and laryngeal behaviour and larynx modes. In *Phonetica*, 34, 264-79.

Stevens, K.N. (2000). *Acoustic Phonetics*, Cambridge (Ma): The MIT Press.

Vaux, B. (1988). *The Phonology of Armenian*. Oxford: Oxford University Press.

Vogt, H. (1959). Les occlusives de l'arménien. In *Norsk Tidsskrift for Sprogvidenskap*, 18, 143-59.

Xaçatryan, A.H. (1988) [= Kachaturyan, A.]. *Žamankakiç hayereni hnčowyṭababowṭyown*. Erevan: Haykakan SSH GA Hratarakčowṭyown ['Fonologia dell'armeno contemporaneo'].

Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

Wells, C., Roach, P. (1980), An experimental investigation of some aspects of tone in Punjabi. In *Journal of Phonetics*, 8, 85-89.

DOI: 10.17469/O2104AISV000008

BARBARA GILI FIVELA, FRANCESCA NICORA

Intonation in Liguria and Tuscany: checking for similarities across a traditional isogloss boundary

The present work investigates the intonation systems of two varieties of Italian spoken in Liguria, namely in La Spezia and Imperia, with the aims of 1) extending the existing knowledge on the intonation systems of varieties of Italian, and 2) checking if it is possible to detect similarities/differences in systems found in relatively close areas which belong to either the same (e.g., in Liguria itself) or different isoglosses (e.g., in Tuscany). The second goal, in particular, corresponds to a different perspective in comparison to the more frequent attempt to identify patterns that characterise specific areas. The analysis of La Spezia and Imperia Italian and the comparison of their intonation systems with those of the varieties of Italian spoken in Genova, Pisa and Florence allow us to extend the geographical reach of phonological analyses of Italian intonation; furthermore, as for the second goal, results show that, at least as far as yes/no questions are concerned, even though varieties show different "preferred" melodic contours, it is possible to identify similar patterns that occur with different frequency in towns across different isoglosses.

Keywords: Intonation, La Spezia Italian, Imperia Italian, variation, isoglosses.

1. *Introduction*

As shown by Pellegrini's (1977) cartographic representation of the distribution and differentiation of vernaculars (*dialetti*) spoken in Italy, a distinction is traditionally made between the Romance vernaculars spoken on either side of an imaginary line connecting La Spezia and Rimini. Nevertheless, more recent studies on the intonation of Italian varieties have shown that it is not possible to identify homogeneous macro-areas similar to those found when investigating vernaculars as for, e.g., segmental characteristics (Gili Fivela, Avesani, Barone, Bacci, Crocco, D'Imperio, Giordano, Marotta, Savino, Sorianello, 2015; Savino, 2012)[1].

Besides showing the amount of phonetic variability (e.g., Magno-Caldognetto, Ferreno, Lavagnoli & Vagges 1978; Endo, Bertinetto, 1997; Romano, 2003; for an overview, Gili Fivela, 2008), investigations adopting a phonological perspective offer a very interesting insight into the topic of variation, given the effort in identifying the units of a grammar of intonation and to relate them to func-

[1] The varieties described in Gili Fivela et al. (2015), and first analysed and discussed during the Romance Tones and Break Indices (ToBI) workshop, in June 2011 in Tarragona (Spain), are those spoken in Turin, Milan, Florence, Siena, Pisa, Lucca, Rome, Pescara, Naples, Salerno, Bari, Cosenza and Lecce, while Iraci, Gili Fivela (2017) add Palermo Italian to the set of analysed varieties. Savino (2012) focusses on yes-no questions only, as produced in Turin, Bergamo/Brescia, Milan, Venice, Genoa, Parma, Florence, Perugia, Rome, Cagliari, Naples, Bari, Lecce, Catanzaro, and Palermo.

tional categories. A phonological approach, such as that followed within the Autosegmental-Metrical framework (Bruce, 1977; Pierrehumbert, 1980, for an overview, see Ladd, 1996; for an outline on Italian, see Grice, D'Imperio, Savino & Avesani, 2005 and Gili Fivela et al., 2015) is characterized by the effort in finding linguistic categories out of the phonetic continuum, keeping in mind linguistically-induced variation such as that related to sociolinguistic factors. To briefly sum up, in Autosegmental-Metrical analyses high and low tones are phonological events corresponding to target tone levels, belonging to units called pitch accents and edge tones. The former may be monotonal or bitonal accents, which participate in realizing sentence level prominence and are associated to stressed syllables (e.g., H*, L+H*, where the '*' indicate association to the tone bearing unit); the latter are phrase accents and boundary tones (e.g., H-, H%, LH%), which represent relevant cues to prosodic boundaries of different levels and are associated to the edge of prosodic constituents.

The analysis of phonological patterns found in a number of varieties of Italian and in various sentence types produced in different communicative contexts (that is, statements, exclamations, yes/no questions, wh questions, imperatives and vocatives; see Gili Fivela et al. 2015 and Gili Fivela, Iraci 2017) showed that variation through Italian varieties regards the phonological inventory available to speakers, as well as the association of phonological events and functions (besides the expected phonetic variation). Thus, even though in some cases it is possible to find one intonation pattern that can be used by speakers of different varieties of Italian (e.g., in broad-focus statements, lists, wh questions, counterexpectational wh questions, disjunctive questions, and vocatives), in other cases, a high variation is found in relation to both the intonation inventory selected within a variety (e.g., L+H* vs. H*+L to express narrow-correction focus in Florence and Pisa Italian, respectively) and in relation to the specific functions associated with nuclear configurations (e.g., L+H* L% that signals yes-no questions in Cosenza and contrastive correction focus in Florence Italian). Importantly, Gili Fivela et al. (2015) showed that such variation is not bound to isoglosses traditionally identified on the basis of the analysis of Italian *dialetti*. Along similar line, but for yes/no questions only, Savino (2012) highlighted that "the contour type is not geographically related".

Phonological analyses have warrantied the identification of systemic, rather than phonetic, differences or similarities. Consistently, the main goal of previous investigations on variation in Italian intonation was characterising specific varieties, by finding their attested patterns, but with specific attention to their characterizing ones. Nevertheless, a slightly different perspective may be interesting, that is observing if, besides the main features of phonological systems, it is possible to highlight influences due to (internal) contact situations. The idea behind this work is that, by comparing systems found in different varieties and by considering frequency of occurrence of single patterns, it may be possible to

observe influences due to the geographical position of towns in which varieties are spoken, within homogenous or even heterogeneous linguistic areas.

2. *Goals and hypotheses*

Goals of the investigation described here are 1) achieving a better and wider knowledge of intonation variation in Italian, by adding the analysis of two extra varieties, i.e. that spoken in La Spezia and Imperia, and 2) verifying if it is possible to detect similarities/differences between such varieties, and between them and those found in nearby towns which belong to either the same, e.g., in Liguria itself[2], or different isoglosses, e.g., in Tuscany. Importantly, focus of attention is the transition area surrounding La Spezia. Thus, rather than trying to identify, or differentiate, patterns which characterise specific areas, we aim to find out whether the geographical position within a relatively small area, may affect the presence or the frequency of occurrence of attested patterns.

As for the first goal, our view is that a better knowledge on intonation may be reached by focusing on a wide set of communicative contexts and by considering various speech styles; moreover, cross variety variation may be better pointed out by adopting the very same methods used to investigate other varieties. As for the second goal, our hypothesis is that, besides the identity of single varieties, there are reciprocal influences due, for instance, to the amount of contact induced by the geographical position. In particular, we hypothesize that, at least when more than one pattern is available for a specific function, it may be possible to identify similar patterns in different towns, though such patterns may occur with variable frequency.

To verify the hypotheses and reach the intended goals, we collected and analysed speech material for the varieties spoken in La Spezia and Imperia, by adopting the same methods and transcription conventions that have been used to realize the widest investigation available so far on intonation in Italian varieties and other Romance languages too (Gili Fivela et al., 2015; Frota, Prieto, 2015). Moreover, we explicitly compared the intonation systems of the abovementioned varieties with those of the varieties of Italian spoken in Genova, Pisa and Florence. As for the former, we referred to analyses reported in the scientific literature on the topic; as for the others, we looked at analyses of comparable data sets (see references above).

[2] Reference is made to Pellegrini's (1977) proposal of Liguria as part of the northern isoglosses. However, at least four areas can be singled out with reference to the Ligurian dialects (Forner, 1975, 1988, 1995; Loporcaro, 2009, 2015; Benincà et al., 2016). By taking into account La Spezia, Imperia and, as discussed in the next sections, Genoa, we focus on quite distinct dialectal areas within the Liguria region. Therefore, apart from the clear specificity of La Spezia, a deep comparison of intonational features in Genoa and Imperia would also be interesting. This is, though, out of the scope of this paper and will be a goal of future investigations.

3. *La Spezia and Imperia Italian*

3.1 Methods

Along the lines of Gili Fivela et al. (2015), data were collected by audio recording 5 La Spezia and 6[3] Imperia (San Bartolomeo al mare - IM) Italian speakers (respectively, 3F, 2M, aged 20-30 years and 2F, 4M, aged 20-50). All speakers had been continuously exposed to their native variety of Italian, used it for everyday conversation, and had a similar educational level, that is high-school to university degree. Speakers were asked to perform a Discourse Completion Task (Blum-Kulka et al., 1989), including 60[4] situations/contexts presented in pseudo-randomized order. The analysis reported below regards 33 of the 60 contexts. The situation/context favored the speaker immersion in the intended pragmatic condition and induced to produce specific lexical words. Speakers had to spontaneously react to the given context/situation first and, later, to read aloud a sample sentence, in which both lexical entries and sentence structure were controlled and kept unvaried.

On a general basis, two target sentences/contexts were included for each sentence type, mainly to facilitate the collection of patterns realized in the case of nuclear words showing different stress positions; however, in few cases only one situation was included in the corpus. Examples of the former situations are *Mangia i mandarini* 's/he eats tangerins' and *Beve una bibita* 's/he drinks a soft drink', that were included to elicit broad focus statements (by showing someone eating a tangerine/drinking a soft drink and asking the subjects to state what s/he is doing); an example of the latter situation is the counter-expectation polar question *Loredana un ingegnere?!* 'Loredana an engineer?!', elicited by means of a context asking subjects to think to be informed that a friend became an engineer, even though she has never been good at math; the subject is induced to ask for confirmation while explicitly communicating s/he disbelieves it.

In particular, each time a situation/context and an example of response were proposed, speakers were asked to:

1) read carefully and understand a written text describing a context/situation, presented over the PC screen;
2) produce a spontaneous utterance which would fit with the situational context presented;
3) read as spontaneously as possible the target sentence proposed by the experimenters as suitable for the same context.

[3] One extra speaker was added in order to reach the total amount of instances per sentence type, as one of the other speakers produced only one repetition of the corpus.

[4] In comparison to the questionnaire used by Gili Fivela et al. (2015), three contexts/target were added, in order to get data on yes-no questions in which the nuclear pattern is realized on both oxitone and proparoxitone words in final position, and in which the contrastive-corrective focus is realized on nuclear proparoxitone words.

The whole set of target utterances was presented twice, that is we collected 4 target utterances for each situation/context. Interviews were carried out by both the first and the second author[5].

The analysis was carried out within the Autosegmental-Metrical framework (Bruce, 1977; Pierrehumbert, 1980; Ladd, 1996). Auditory analysis and inspection of fundamental frequency tracks were performed by devoting specific attention to spontaneous renditions, though alignment characteristics were confirmed by means of read speech productions. Importantly, if not specified, the analyses are considered not to depend on speech style. In line with annotation convention established by Gili Fivela et al. (2015: 149), "combinations of equal tones are collapsed and represented by one symbol only (e.g., L-L% becomes L%) and sequences of different edge tones are reported with no intermediate hyphen".

3.2 Results

3.2.1 Statements

In both La Spezia and Imperia varieties, broad focus statements are realized by means of a H+L* L% pattern, in line with what reported in the literature on other varieties of Italian. A high variability in H+ scaling is found, due, for instance, to inter-subject or stylistic variability, as well as to differences in illocutionary force (e.g. Gili Fivela, 2006, Gili Fivela et al., 2015; Gili Fivela, Iraci, 2017). This seems to be the case of La Spezia Italian, where most of the time the H+L* pitch accent is implemented as a fall from a gradually falling stretch (66,6%), while 1/3 of cases is characterized by a fall from a plateau (33,3%) – see Fig.1; in Imperia, the abovementioned implementations are found in about 50% of cases each one.

In line with other varieties, lists in La Spezia and Imperia Italian are usually characterized by either H+L* or L+H* on all the items but the penultimate and the last one: the penultimate usually bears a L+H*, and the last one carries a H+L*. In some cases, a delayed peak for the L+H* accent seems to be realized and even possible L*+Hs seem to be realized (but mainly by one speaker). Intermediate edge tones may be high or low, while the final edge tone is low.

[5] The latter is a mother tongue speaker of the variety spoken in La Spezia; the former has been spending in Imperia (San Bartolomeo al mare - IM) about one month a year since the childhood.

Figure 1 - *Broad focus statements* Maria mangia il mandarino *'Maria has a mandarin' by La Spezia speakers (left, spk. 1) and* Maria beve una limonata *'Maria is drinking a lemonade' (right, spk. 2)*

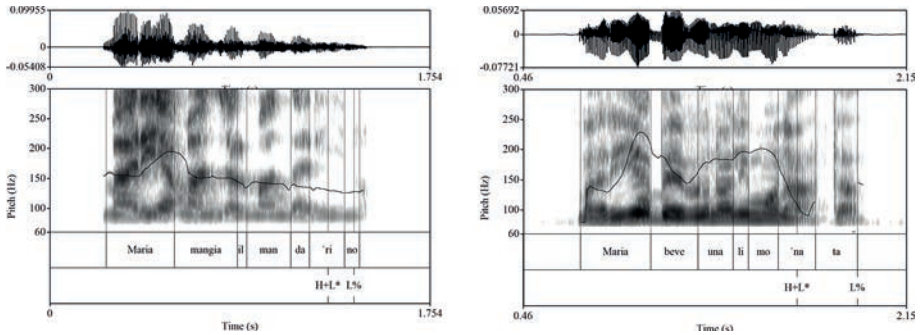
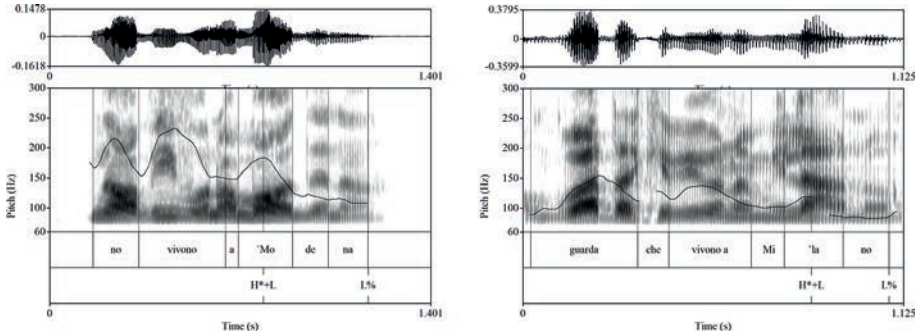
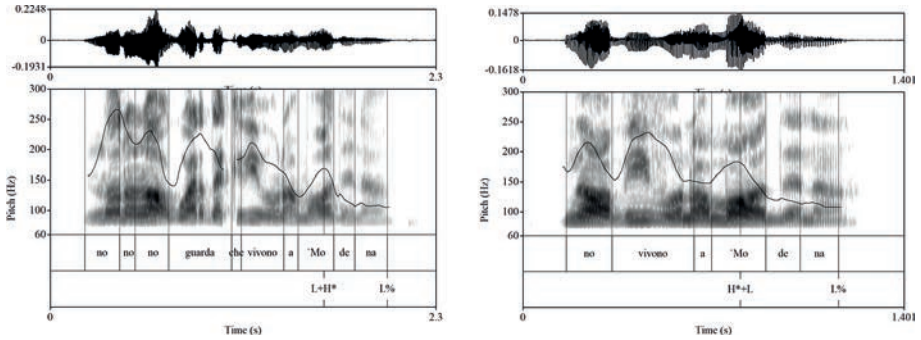


Figure 2 - *Narrow contrastive-correction focus* No vivono a Modena *'No, they live in Modena' and* Guarda che vivono a Milano *'They live in Milan': use of H*+L L% pattern (speaker from La Spezia – left - and Imperia - right)*



In both La Spezia and Imperia Italian, narrow contrastive-corrective focus is realized with an H*+L pitch accent (La Spezia: 62%; Imperia: 70%) – see Fig. 2. Thus, given the two main pitch accents used to express corrective focus in Italian varieties, La Spezia and Imperia Italian show the same phonological categories found in Rome, Pisa, Pescara, Cosenza, Bari, Lecce and Palermo (and different from the L+H* found in Milan, Turin, Florence, Siena, Lucca, Naples and Salerno).

Figure 3 - *Narrow contrastive-correction focus* Guarda che vivono a Modena
'They live in Modena': example of L+H and H*+L pitch accent by La Spezia Speaker 2*



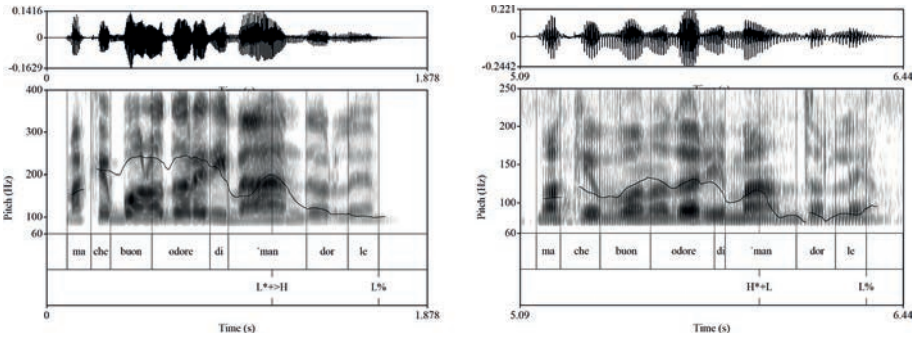
However, as already observed for other varieties, more than one pitch accent may be found in the expression of focus. In particular, in La Spezia Italian a L+H* accent – showing a rise, rather than a rise-fall, through the nuclear syllable – is found (30% of cases) – see Fig. 3, left vs. right. The presence of both H*+L and L+H* was already observed in Palermo Italian, where the choice between patterns appeared to be related to a reduced novelty of the context (as it was observed always in the second repetitions of the task) or to variation in politeness (Gili Fivela, Iraci, 2017). However, in La Spezia Italian the difference between these two patterns seems more clearly related to the speakers' intention to suggest an option, almost with an interlocutory nuance, rather than to correct in a peremptory manner (as in the case of H*+L). In both La Spezia and Imperia Italian, the H+L* pattern often represents a less-peremptory alternative (La Spezia: 8%; Imperia: 30%).

3.2.2 Exclamatives

In both La Spezia and Imperia Italian, in line with other varieties (Gili Fivela et al., 2015), exclamatives are expressed by means of a L*+>H pitch accent (La Spezia: 80%[6] – see Fig. 4, left; Imperia: 60%), followed by a L% boundary tone. Such pitch accent has shown to be characterized by a peak which is aligned earlier than that found in L*+H, independently of the structure and segmental make-up of the target syllable (see also Gili Fivela et al., 2015b).

[6] In some cases (15%), speakers produce shorter nuclear syllables (and seem to speak faster), realizing a steeper rise which could be consistent with a L+H* analysis. However, the difference is taken to be a phonetic one and not to be due to nuances in meaning, eventually requiring a different category.

Figure 4 - *Exclamatives* Ma che buon odore di mandorle! *'What a good smell of almonds!': examples of L*+>H L% by a La Spezia (left) and of H*+L L% by an Imperia speaker (right)*



In both La Spezia and Imperia Italian, alternative patterns are H*+L L% (La Spezia 20%, Imperia 30% – see Fig. 4, right). Moreover, Imperia Italian also shows the nuclear combination H+L* L% (10%). Both H*+L L% and H+L* L% configurations are usually preceded by high F0 values, which are due to a (sequence of) L+H*, especially in the case of H*+L L%, and a wide range in general.

3.2.3 Yes/no questions

3.2.3.1 Information seeking yes/no questions

In both La Spezia and Imperia Italian a pattern used to express information seeking yes/no questions is found to be H*+L LH% (La Spezia: 38.5%, but mainly in two out of the five speakers: 32% in proparoxitone and 45% in paroxitone; Imperia: 81% in both stress positions) – see Fig. 5. Such pattern alternates with L*+H L% in La Spezia (54%: 58% in proparoxitone and 50% in paroxitone). Even though a high variability is observed in the peak alignment (e.g., the peak is aligned before the end of the nuclear syllable by speaker 1 and at the beginning of the post-nucler syllable by speaker 2), the high target does not seem to move far in the post-nuclear syllable and is therefore taken as part of the pitch accent. In La Spezia, the other patterns attested seem sort of idiosyncratic realizations by one speaker[7], while in Imperia the other contour attested in the corpus is H+L* LH% (19%), where the leading high tone is characterized by very high scaling.

[7] However, in case a phonological analysis will turn out to be needed, it could be L*+H H% in 10% of proparoxitone and 5% of paroxitone words.

Figure 5 - *Information seeking yes/no questions* Avete delle mandorle?
'Do you have almonds?' (La Spezia speaker, left), and Avete dei mandarini?
'Do you have tangerins?' (Imperia speaker, right)

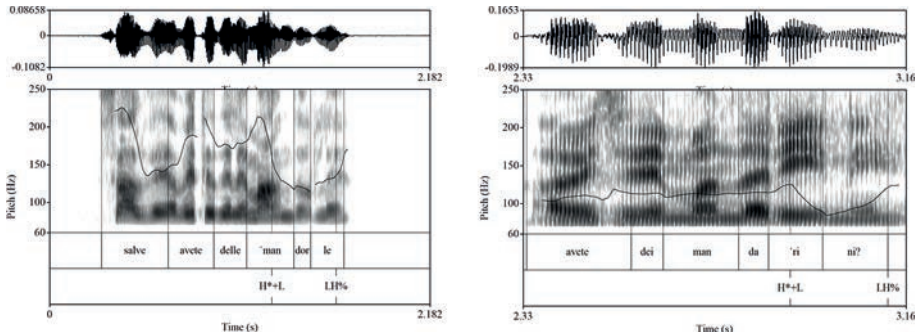
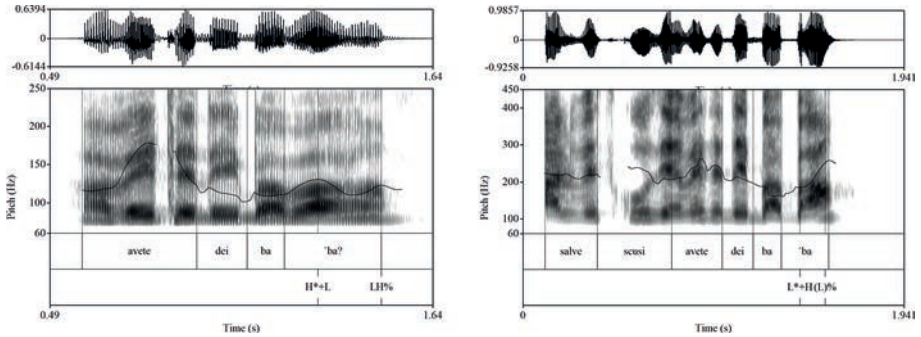


Figure 6 - *Information seeking yes/no questions* Avete dei Babà?
'Do you have Babà?' by La Spezia Italian speakers: no truncation (left panel) and truncation of the final tone (right panel)



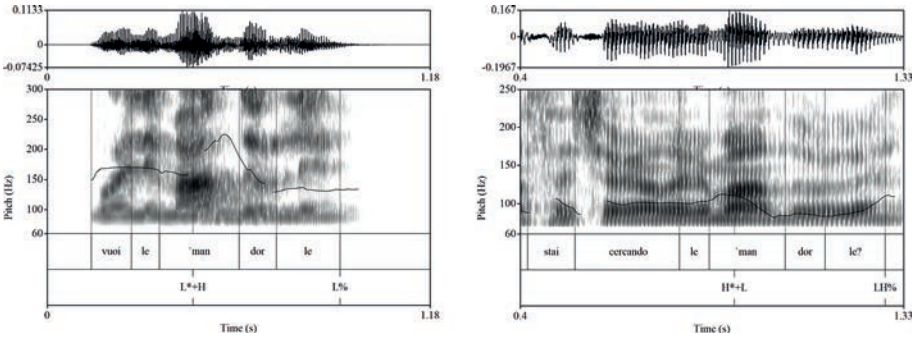
Truncation in oxitone target words is optional in both La Spezia and Imperia. In the former, the H*+L LH% pattern is observed in 55,5% of cases, and only in 38,8% of them truncation of the final H% tone takes place (no truncation in 16,7% – Fig. 6, left). The other option is L*+H L% pattern with truncation of the final L% tone (44,5% of cases, see Fig. 6, right). In Imperia Italian, a quite different situation is observed, in that the H*+L LH% pattern is observed in 20% of cases, while truncation is the most frequent strategy, with total truncation of the H% tone in 40% of cases and a sort of partial truncation, involving a very reduced final rising, in the remaining 40% of items.

3.2.3.2 Echo, counterexpectational and confirmation seeking yes/no questions

In both varieties, confirmation seeking yes/no question may be expressed by means of the H*+L LH% pattern (La Spezia: 33,3%; Imperia: 91%, with a quite early peak). However, in La Spezia the most common pattern is L*+H L% (44,5%), while in Imperia a H+L* LH% may also be found (33.3) – see Fig. 7. In both La Spezia and Imperia, a number of 'statement-like' realizations are also found, in

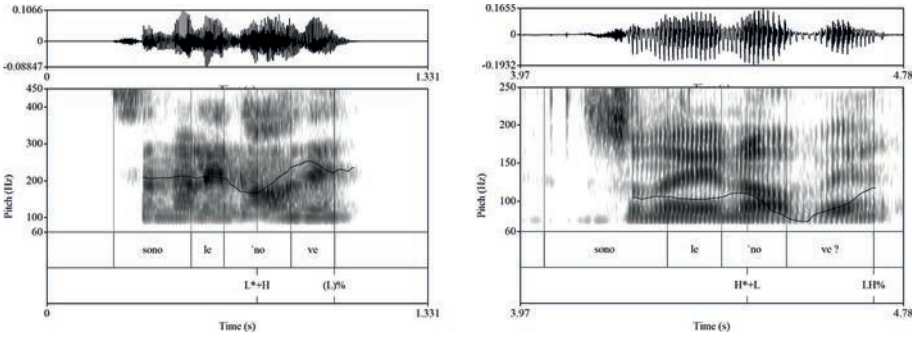
line with the option reported as for other varieties (Gili Fivela et al. 2015), which consists in asking for confirmation by means of the contrastive-corrective focus pattern H*+L L% (La Spezia: 22,2%, by 2 out of the 5 speakers; Imperia: 9%, by one speaker; in any case, speakers are clearly confident that the information is owned by the interlocutor).

Figure 7 - *Information seeking yes/no questions* Vuoi/stai cercando le mandorle? *'Do you want/do you look for almonds?' in La Spezia (left) and Imperia (right)*



Echoes in La Spezia are realized by means of a L*+H L% pattern (65%), though in some cases with partial truncation (38% of the L*+H L%, which are realized on paroxitone words only; see Fig. 8). The H*+L LH% pattern is used to express echos too (mainly by two speakers, 35%). In Imperia Italian, both the H*+L LH% and the H+L* LH% pattern are found (67% and 33% of cases, respectively; the latter shows a possibly upstepped leading tone) – Fig. 8.

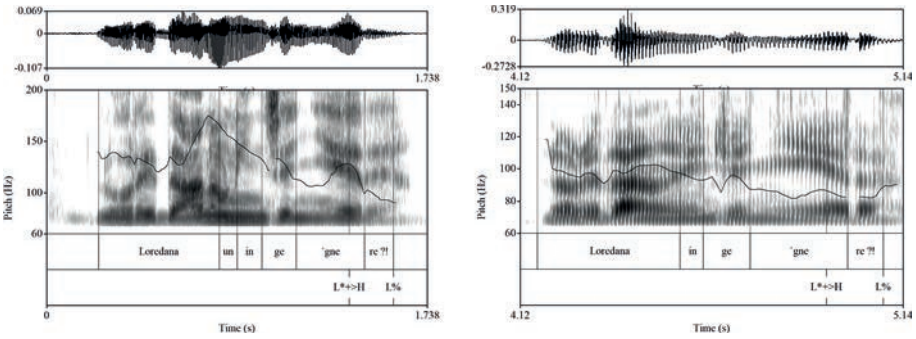
Figure 8 - *Echo yes/no questions* Sono le nove? *'Is It nine o'clock?' in La Spezia (L*+H L% with partial truncation, left) and Imperia (right)*



As in other varieties, counterexpectational yes/no questions may be expressed by means of the same phonological pattern found in echo yes/no questions, although the phonetic implementation may imply differences in syllable lengthening, tonal alignment, and scaling. Thus, such questions are realized by means of an H*+L pitch accent followed by a low and rising pitch track (La Spezia: 6%; Imperia: 30%)

or just a final low (La Spezia: 27%; Imperia: 20%). However, in Imperia Italian the edge tone label seems to correspond a L!H%, as the final rise usually does not reach a very high frequency value.

Figure 9 - *Counterexpectational yes/no questions* Loredana un ingegnere!?
'Loredana an engineer!?' by a La Spezia (left) and Imperia (right) speaker



Rather, counterexpectational yes/no questions may also be expressed by means of a L*+>H pitch accent followed by a L% boundary tone (La Spezia: 47%; Imperia: 50% of cases) – see Fig. 9, or by a L*+>H LH% (rather L!H% usually characterized by a very slight final rise; La Spezia: 20%). Interestingly, the L*+>H L% pattern is the same found in exclamatives. However, by listening to the audio examples and inspecting their F0 tracks, we observed that the pitch range is often more compressed in counterexpectational questions than in exclamatives. This seems to be enough to conveying the question function within a dialogic exchange, while, if the question function has to be explicitly communicated, a final rise is realized (i.e., a final LH% or L!H% edge tone combination which shows a generally compressed, but quite variable scaling is found). Of course these observations need to be confirmed by extensive acoustic measurements[8] and, ideally, by perceptual tests.

Thus, data on La Spezia and Imperia Italian confirm what has been observed for other varieties of Italian, that is one pattern may be common to many (sub) functions (e.g. information seeking, confirmation seeking and echo), though, on the other hand, more than one pattern may play a specific function, possibly conveying different stylistic choices.

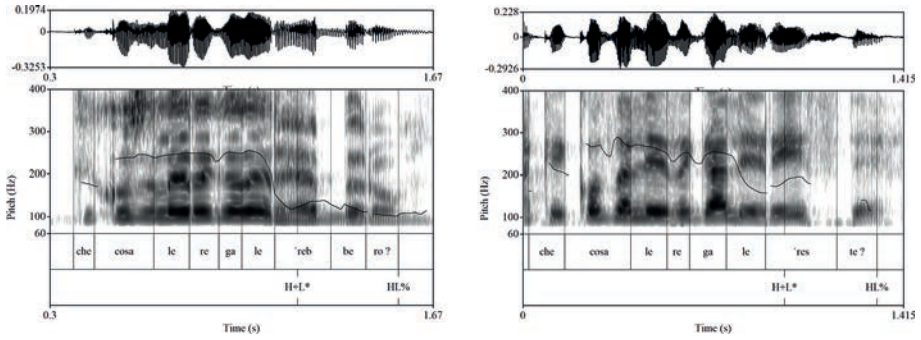
[8] Acoustic measurements performed on contours produced in exclamatives and in counterexpectational questions by the La Spezia speaker who realized higly comparable sentence structures and patterns confirm that, on average, both pitch range (higher prenuclear peak to final low edge tone) and pitch excursion (low to high pitch accent target) are larger in exclamation than in questions.

3.2.4 Wh questions

3.2.4.1 Information seeking wh questions

The most frequent pattern found in the corpus to express information seeking wh questions in la Spezia Italian is H+L* HL% (La Spezia 86% – Fig. 10, left). A very typical feature of La Spezia, but of other Ligurian varieties too, is the rising in the final part of the nuclear syllable. In La Spezia it is a very slight rising, while instances of a more clear rising are found in Imperia Italian, where, though, the pattern is less frequently attested (Imperia 18% – Fig. 10, right)[9]. The analysis proposed, accounting for the different realizations (risings) in the two varieties analysed here and the common nature we hypothesize, is H+L* HL%, where the high edge tone is secondarily associated to the nuclear syllable. Few instances involving the H+L* L% pattern, attested in many varieties of Italian, are also found in both La Spezia and Imperia Italian (La Spezia: 7% Imperia: 27%). In Imperia, the H+L* LH% is the most frequent pattern (55%)[10].

Figure 10 - *Information seeking wh questions* cosa le regalerebbero? *'What would they gift her?' (La Spezia, left) and* cosa le regalereste? *'What would you gift her?' (Imperia, right)*



3.2.4.2 Echo, disjunctive and counterexpectational wh questions

In line with other varieties of Italian, echo questions are realized by means of the same pattern found in information seeking yes/no questions, that is H*+L LH% (La Spezia: 84.2%; Imperia: 91%)[11], and disjunctive questions are realized by means of a L+H* pitch accent on the first item (eventually followed by either a high or a

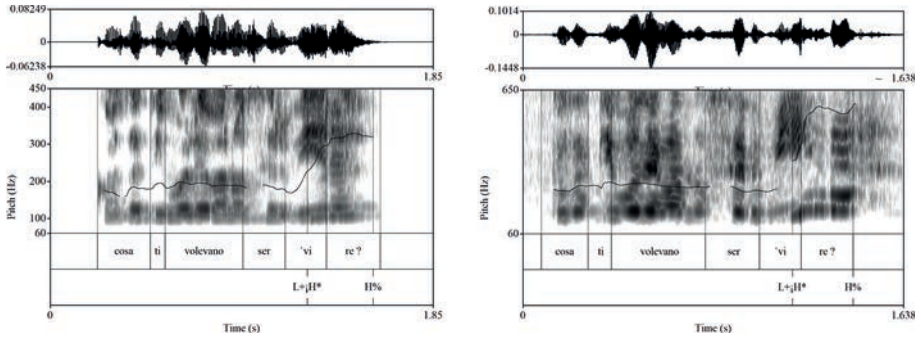
[9] Actually the phonetic shape of the patter in Imperia Italian corresponds to a rising in the nuclear syllable, suggesting even an analysis involving a L+H* L%, preceded though by a high target two syllables before the nuclear one, on an unstressed syllable, where a H+ or even an H* would be needed. Informants suggest that this pattern may be due to the influence of the variety spoken in Genoa.

[10] In the La Spezia corpus, the finally rising pattern attested in many varieties in wh questions is very rare (7%, by one female speaker). Rather, in one case only (spk5, 32 L1) the usual rising by the end of the nuclear syllable is followed by a slight falling and a reduced final rising (no phonological analysis is proposed, but interestingly, from a theoretical point of view, an H+L* HLH% would be needed to account for the pattern).

[11] In La Spezia, one speaker produces a L+H* L% nuclear combination, showing a quite wide pitch range excursion and resembling a L+¡H* L% pattern (15,8% of cases); in Imperia, 9% of contours correspond to H+L* LH%, which is also attested in information seeking yes/no questions.

low phrase accent) and a H+L* on the final item followed by either a low or a high boundary tone (in the Imperia corpus, the final H% is attested in 64% of cases). However, La Spezia Italian speakers also use a L*+H H% pattern (15% of cases)

Figure 11 - *Counterexpectational wh questions* cosa ti volevano servire?
'What did they want to serve you?': La Spezia (left) and Imperia (right)



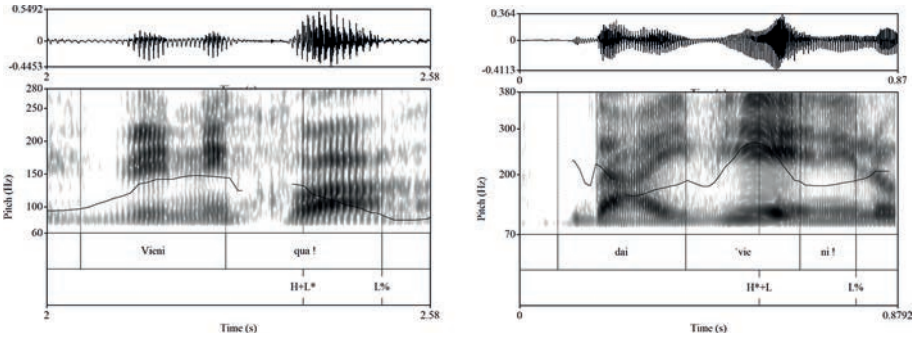
As already observed for many other varieties of Italian, in both the varieties considered here counterexpectational wh questions are expressed by means of a rising pitch accent which is characterised by a wide pitch excursion and is therefore labelled as L+¡H*, followed by a H% boundary tone (La Spezia: 100%; Imperia: 75%) – Fig. 11. In Imperia Italian, in 25% of occurrences an H+L* L% pattern, characterized by a very high leading tone target is also found.

3.2.5 Imperatives: commands and requests

In La Spezia and Imperia Italian commands are usually realized by an H*+L L% (La Spezia: 25%; Imperia: 67%) or H+L* L% (La Spezia: 50%; Imperia: 16.5%)[12] – see Fig. 12. Moreover, in Imperia H+L* LH% may be found (16.5%), and in La Spezia L+H* L% is attested (25% of cases).

[12] Some of the productions which are analysed here as H+L* L% are actually good examples of child directed speech, given the context used in the DCT to elicit the target utterance. In these examples, the L% tone is not really realized as such, but rather as a level sustained tone in the lower portion of the pitch range, and the voice volume is usually raised, with clear changes in the pattern phonetic shape.

Figure 12 - *Imperative command* Vieni qua! *'Come here', left panel, and imperative request* Dai, vieni! *'come on, join us', right panel, by Imperia Italian speakers*



As far as requests are concerned, in both varieties they are realized by means of an H*+L L% (La Spezia: 80%; Imperia; 45%) – see Fig. 12. Further, in Imperia, an H*+L H% pattern is often found (33% of cases, while often the H% is not very high)[13], while in La Spezia Italian they are often realized by a L+H* L% (20% of cases). However, in La Spezia the L+H* L% is especially, but not exclusively, found in non-final position (e.g., *e dai vieni che finisci dopo* 'come on, come here and you'll finish later', by speaker 2 in a spontaneous rendition).

3.2.6 Vocatives

3.2.6.1 Initial call

The pattern found in vocative initial calls is mainly L+H* H!H% (La Spezia: 60%; Imperia 64%), that is the same found in the other varieties investigated so far (i.e., Milan, Turin, Pisa, Lucca, Rome, Pescara, Naples, Salerno, Cosenza, Lecce, Palermo, Florence, Siena) – Fig. 13 (left panel). The L+H* pitch accent may also be followed by a low edge tone L% (La Spezia: 10%; Imperia: 27%[14]) – Fig. 14, left panel. In La Spezia, the H*+L L% pattern is also found (9%).

3.2.6.2 Insistent call

The insistent call may be realized by repeating the same phonological pattern used for the initial call, that is L+H* H!H%, though usually produced with a wider pitch range and H targets on a higher fundamental frequency (La Spezia: 45%; Imperia 36%) – see Fig. 13. However, in la Spezia insistent calls the H*+L L% pattern is even more frequent (55%) – Fig. 14, while in Imperia it is less frequently attested (28%), as L+H* L% – or rather HL%, see n. 14 – is quite often found (36%). Basically, low boundary tones seem to be more often found in insistent calls.

[13] Other attested, though rare, patterns are H+L* H% and H*+L HL%, in 11% of cases each one.

[14] In Imperia Italian, in most cases, an analysis such as HL% would also be appropriate (moreover, such analysis is supported by some extra productions involving a truncated form, i.e. *Seba* instead of *Sebastiano* 'Sebastian', in which an HL% final sequence seems to be realized).

Figure 13 - *Vocative initial and insistent call* Domenico!
'Domenico': example for Imperia Italian

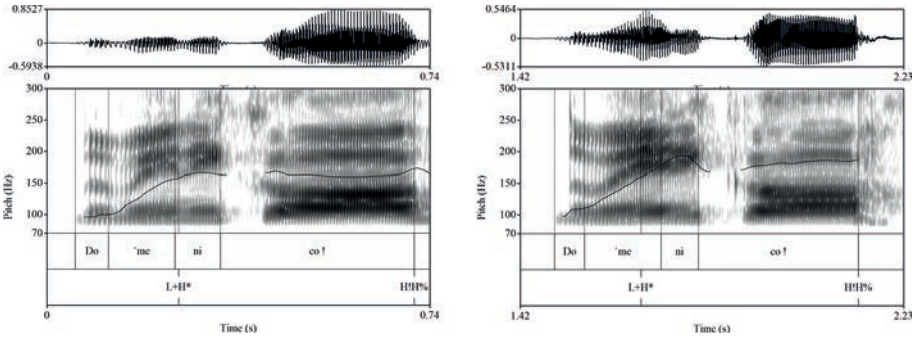
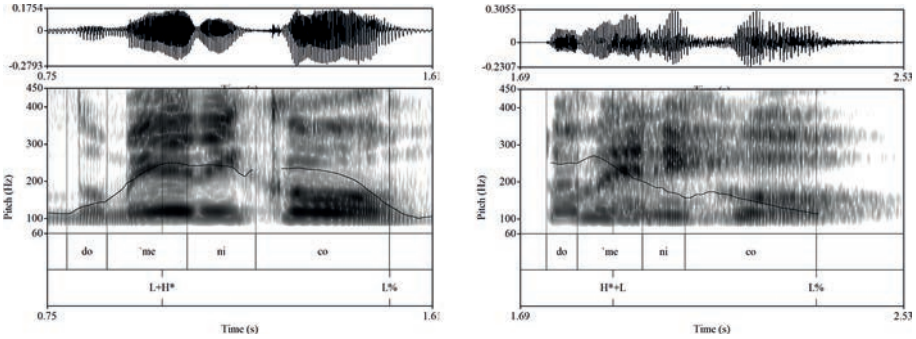


Figure 14 - *Vocative initial and insistent call* Domenico! *'Domenico': alternative patterns produced by a La Spezia speaker*



4. *La Spezia at the «boundary» or within a continuum?*

As already mentioned, our second research goal was to verify if it is possible to detect similarities/differences between systems found in nearby towns which belong to either the same, e.g., Imperia and Genoa (but see n. 2), or different isoglosses, e.g., Genoa vs. Pisa. In order investigate the issue, attention was focussed on information seeking yes/no. Such questions are, in fact, those usually showing the highest cross-variety differences as well as a number of possible intra-variety intonation patterns (though some patterns may also be used for various sub-functions, e.g., the same pattern may express information seeking and confirmation seeking questions).

Table 1 shows patterns attested in the area under investigation, as well as the percentage of occurrence of the attested patterns in the corpora considered, if available. As the table shows, in the area under investigation the patterns used for expressing information seeking yes/no include H*+L LH%, described in this paper for the varieties spoken in La Spezia and Imperia (§3.2.4.1). The same pattern was described for Pisa in Gili Fivela et al. (2015), where the only pattern reported for Florence

was H* LH%[15]. As for Genoa, analyses described in the literature are reported, in particular with reference to Savino (2012) and Crocco (2011), who respectively proposed a L+H* LH% and a (L+)H* LH% analysis[16]. Pitch tracks of yes-no questions by Pisa, Florence and Genoa speakers are given in Fig. 15. The table shows that all analyses include a H* tone in the (monotonal or bitonal) pitch accent, and a following low-high edge tone sequence; moreover, as Fig. 15 shows, independently of the phonological analysis, the phonetic shape of the contours is very similar. This suggests a possible common origin of the pattern, independently of the phonologization (or on the phonological analysis given) in the considered varieties. Apart from this "shared pattern", other contours are usually associated to the function of information seeking yes/no.

Table 1 - *Patterns attested in information seeking yes/no in the varieties considered*

	Pitch accents	*Edge tones*	*Occurrences*
Imperia	**H*+L**	**LH%**	**81%**
	H+L*	LH%	19%
Genoa	**L+H*[18]**	**LH%** LL%	**32.7%** 46.4%
	H+L*	LH%	20.9%
La Spezia[17]	L*+H	L%	54%
	H*+L	**LH%**	**38.5%**
Pisa	H+L*	HL%	86.7%
	H*+L	**LH%**	**13.3%**
Florence	**H***	**LH%**	–

Even though data regard only few towns in the area, they show an interesting trend concerning the pattern under investigation, that is the one resembling the H*+L LH% found in La Spezia and Imperia. In fact, patterns corresponding to similar phonetic shape and similar phonological analyses are reported to occur more or less frequently in the corpora related to the places considered in this investigations. Even though the "picture" is surely more complex than it seems, percentages show a gradual change in the frequency of occurrence of the pattern when moving from western Ligurian varieties to Tuscan ones.

[15] In the case of Florence, there is only one pattern attested in the literature. However, we cannot exclude that others can be found, given that no percentages of occurrence had ever been published as for this variety and that the usual goal pursued so far was identifying the prototypical pattern (more than focussing on intra-variety variation). Analyses concern corpora collected within and along the lines of the Atlas of Romance Intonation (DCT for Pisa, Firenze, Imperia, La Spezia).

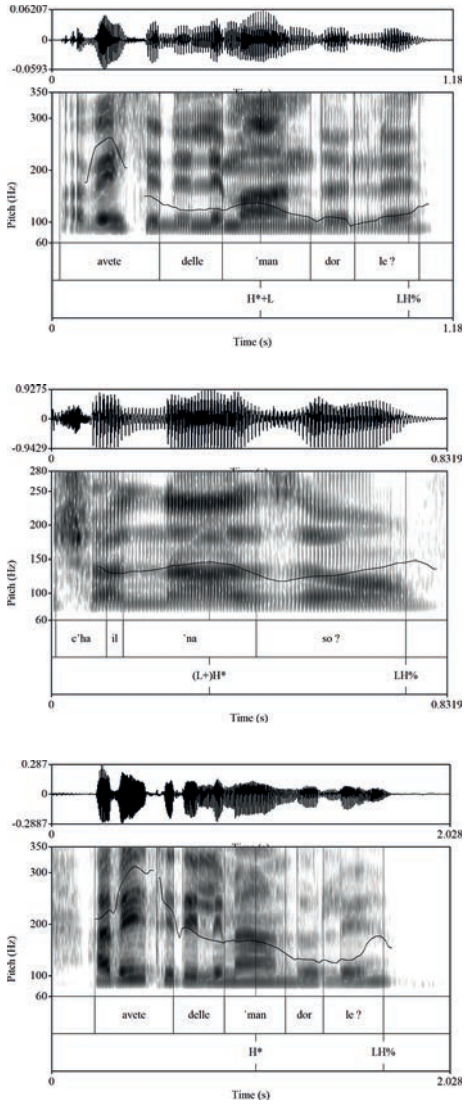
[16] Analyses on Genoa Italian were performed on yes-no information seeking questions collected in the corpus CLIPS (Corpora e Lessici di Italiano Parlato e Scritto – Corpora and Lexicons of Spoken and Written Italian – www.clips.unina.it).

[17] See footnote 7.

[18] The pitch accent is labelled as (L+)H* by Crocco (2011).

Our interpretation of these data is that, when considering the whole set of patterns attested within a variety, it is possible to observe differences in the frequency of occurrence of each pattern and those differences may give hints on the reciprocal influence of intonation systems of varieties spoken in nearby towns.

Figure 15 - *Example of yes-no information seeking questions by Pisa (top), Florence (middle) and Genoa speakers (bottom panel); the former are taken from the corpus collected for the Atlas of Romance Intonation, the latter from the CLIPS corpus and referred to by Crocco (2011)*



5. *Discussion*

The analysis of the intonation systems of varieties spoken in Imperia and La Spezia allowed us to point out peculiar, local features as well as shared ones – see Table 2 for an overview – and to reach our first goal, that is, extending the geographical reach of phonological analyses of Italian intonation.

Details on the patterns were given in section 2, while here it may be worth to highlight the most characterizing aspects of the two varieties. As for Imperia Italian, the H*+L pitch accent is very frequent, in both statements and questions, and it really characterizes the language variety. On the other hand,one of the patterns found in Imperia wh questions is quite typical of the intonation of the investigated area, and some informants consider it to be influenced by the pattern found in Genoa Italian. In particular, wh questions in both Imperia and La Spezia may be characterized by a falling nuclear pattern, ending with a rising pitch (which in Imperia may also be aligned to most part of the nuclear syllable) and being followed by a very final fall (H+L* HL%). In such cases, a fall originates from a prenuclear high on an unstressed syllable which is one or even two syllables away from the nucleus; after the low target, the F0 in the nucleus rises by the end of the nuclear syllable or even through the whole syllable (e.g., in the examples found in Imperia Italian). The analysis proposed, accounting for the different realizations (risings) in the two varieties considered here and the common nature we hypothesize, is H+L* HL%, where the high edge tone is secondarily associated to the nuclear syllable.

Another peculiar characteristic highlighted by the analysis is that in both Imperia and La Spezia Italian counterexpectational yes/no questions may be expressed by means of a L*+>H pitch accent followed by a L% boundary tone. It is a matter of further investigation understating if the difference between such questions and exclamatives, which may be expressed by L*+>H L% too, is related, for instance, to the higher f0 values or the wider range observed in the case of exclamatives.

Besides the abovementioned specific characteristics, data on La Spezia and Imperia Italian confirm that it is not possible to draw clear isogloss boundaries on the basis of intonation, and definitely no isogloss boundaries in the positions proposed by means of the analysis of Italian vernaculars (see Table 3 and 4 for patterns observed through Italy in, respectively, yes/no and wh information seeking questions).

In order to focus on the second goal of the paper, that is to verify if it is possible to detect similarities/differences in systems attested in nearby towns across traditional isogloss boundaries – e.g., Genoa, La Spezia and Pisa –, we focused attention on information seeking yes/no. They usually show the highest cross-variety differences as well as a number of possible intra-variety intonation patterns. A comparison of the main patterns attested in information seeking yes/no in the area under investigation showed that the patterns attested include a H* tone in the (monotonal or bitonal) pitch accent and a following LH% tone sequence. In particular, the patterns are H*+L LH% and H* LH% (the former, in La Spezia and Imperia, see §3.2.3.1 and Pisa, see Gili Fivela et al., 2015; the latter in Florence, see Gili Fivela et al., 2015 again), as well as L+H* or (L+)H* LH% (in Genoa, see Savino, 2012 and Crocco, 2011). Moreover, the phonetic similarity also suggests a possible common origin of the pattern, independently of phonologization processes.

Even though data regard only few towns in the area, they show the presence of highly similar patterns in the area under investigation; moreover, percentages show a gradual change in the frequency of occurrence the H*+L LH% pattern found in La Spezia and Imperia when moving from western Ligurian varieties to Tuscan ones. Thus, even though few locales have been considered in the area, data analysed so far allows us to give a positive answer to the second research question of this paper. In fact, besides expected cross-variety differences, a transition area such as the one considered here may be characterized by the presence of similar patterns that are though used with a different frequency (at least if we take the frequency in the collected corpora as representative of frequency of occurrence in real, complex communication). Further analyses will hopefully confirm these first observations.

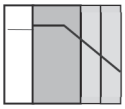
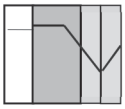
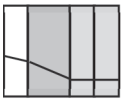
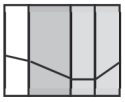
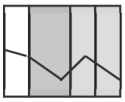
6. *Conclusions*

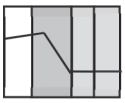
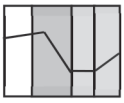
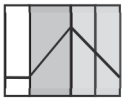
The paper describes an investigation on the intonation system of the varieties of Italian spoken in La Spezia and Imperia, with the aim of extending the existing knowledge on the intonation of varieties of Italian. Another goal instantiates a different perceptive in comparison to the more frequent attempt to identify patterns which characterise specific varieties. A second goal is indeed to detect similarities/differences between the varieties spoken in relatively closed towns, which belong to either the same or different isoglosses.

To reach the intended goals, we collected and analysed speech material for the varieties spoken in La Spezia and Imperia, following the same methods adopted for investigating other varieties of Italian, and other Romance languages too (within the Autosegmental-Metrical framework). Finally, we compared the intonation systems of the abovementioned varieties with those of the varieties of Italian spoken in Genoa, Pisa and Florence, by referring to analyses reported in the scientific literature on the topic.

Results allow to extend the geographical reach of phonological analyses of Italian intonation and to show that, at least as far as yes/no questions are concerned, it is possible to identify similar patterns even in towns across different isoglosses, though the patterns occur with a different frequency.

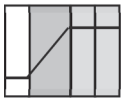
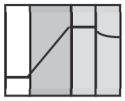
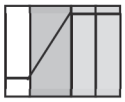
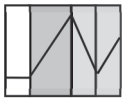
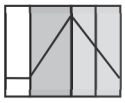
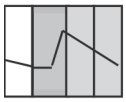
Table 2 - *Inventory of nuclear configurations found in sixteen varieties of Italian, their schematic representations and their use in main sentence types (adapted from Gili Fivela* et al. *2015 and Gili Fivela, Iraci, 2017: updates regarding La Spezia and Imperia Italian data are underlined)*

Nuclear Configuration		*Sentence types where it is used*
	H* L%	Exclamatives (Cosenza).
	H* LH%	Yes/no questions (Florence and Siena).
	H+L* L%	Broad focus statements, intermediate and final item in lists, narrow informational focus (e.g., Firenze and Siena); contrastive-corrective narrow focus statements (in Pescara Italian, when realized as a high pretonic pitch accent that in long constituents corresponds to a high plateau, described as L+H* H+L*; as a second option in some varieties); exclamatives (Milan, Imperia, Lucca, Salerno, Lecce); wh questions (Imperia, La Spezia, Milan, Turin, Pisa, Lucca, Rome, Pescara, Siena, Naples, Cosenza, Salerno, Bari, Lecce, Palermo); final item in disjuctive questions, commands (Imperia, La Spezia, Milan, Turin, Florence, Siena, Lucca, Pisa, Rome, Salerno, Pescara, Lecce, Palermo); imperative requests (Lucca, Rome, Naples, Pescara; in the latter two, the high pretonic pitch accent is found); vocative initial call (Naples and Pescara, where the high pretonic pitch accent is found).
	H+L* LH%	Yes/no questions (Imperia, Milan, Turin, Lucca, Salerno, Cosenza, Lecce); wh questions (Milan, Turin, Rome, Florence, Siena, Lucca, Salerno, Bari, Cosenza, Palermo); possibile in lists.
	H+L* HL%	Yes/no questions (Pisa and Lucca): wh questions in Imperia and La Spezia (with secondary association of H-).

	H*+L L%[19]	Contrastive-corrective narrow focus statements (Imperia, La Spezia, Pisa, Rome, Pescara, Bari, Cosenza, Lecce, Palermo); yes/no questions (Milan, Pisa, Rome, Pescara, Salerno, Lecce); counterexpectational yes/no questions, exclamatives (Imperia, La Spezia, Pisa, Lucca, Rome, Pescara, Salerno, Lecce); commands (La Spezia, Imperia, Cosenza, Lecce, Pescara); imperative requests (La Spezia, Imperia, Pisa, Cosenza, Pescara where the high pretonic variance is found, and Palermo), vocative initial call (La Spezia), vocative insistent call (Imperia, La Spezia, Pisa, Pescara, Lecce).
	H*+L LH%	Yes/no questions (Imperia, La Spezia, Milan, Pisa, Rome, Pescara, Salerno, Lecce)[20], wh echo questions (Imperia, La Spezia).
	L+H* L%	Not final item in lists, early narrow focus (Pisa, Lecce); wh questions (Cosenza);
	L+H* L%	Contrastive-corrective narrow focus statements (Milan, Turin, Florence, Siena, Lucca, Naples, Salerno); exclamatives (La Spezia, Turin, Florence, Siena, Palermo); yes/no questions (Salerno, Cosenza, Bari); counterexpectational yes/no questions; commands (Turin, La Spezia); imperative requests (Milan, Turin, La Spezia, Florence, Siena, Salerno); vocative initial call (Imperia, La Spezia, Pisa, Lucca, Salerno, Cosenza); vocative insistent call (Milan, Turin, Florence, Pisa, Siena, Cosenza, Palermo). Alternative pattern for narrow (non-peremptory) correction-focus (Palermo, La Spezia).
	L+H* LH%	Yes/no questions (Turin, Salerno, Cosenza, Bari).
	L+H* L!H%	Counterexpectational yes/no questions (Lecce).

[19] See Gili Fivela et al. (2015) for a possible analysis in terms of H* secondary association of the shared feature of the early peak alignment in both L+H* L% and H*+L L% as found in contrastive-correction focus.

[20] Possible L!H% in counterexpectational yes/no questions in Imperia.

	L+H* H%	Wh questions (Rome, Cosenza), possible on intermediate item in lists.
	L+H* H!H%	Vocative initial call ((Imperia, La Spezia, Milan, Turin, Florence, Siena, Pisa, Lucca, Rome, Pescara, Naples, Salerno, Cosenza, Lecce, Palermo) and insistent call (e.g., (Imperia, La Spezia, Pisa, Pescara, Salerno, Cosenza, Palermo).
	L+¡H* H%	Counterexpectational wh questions (Imperia, La Spezia, Milan, Turin, Florence, Siena, Pisa, Lucca, Rome, Salerno, Pescara, Cosenza, Palermo).
	L+¡H* LH%	Echo yes/no questions (Lucca).
	L+¡H* L%	Counterexpectational yes/no questions (Bari)[21]; counterexpectational wh questions (Lecce, Salerno, Pescara, Palermo).
	L*+H L%	wh questions in Pescara; yes/no questions, confirmation-seeking and echo yes/no questions in Palermo (where the peak seems to be slightly delayed in comparison to the schema offered to the left); information seking yes-no questions in La Spezia; couterexpectational yes/no questions in Palermo and La Spezia (though the peak is as anticipated to resemble a L*+>H L% pitch accent).
	L*+H LH%	Alternative patterm in couterexpectational yes/no questions (Palermo, Imperia)

[21] On the basis of Savino and Grice (2007, 2011).

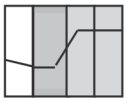
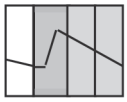
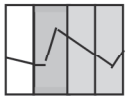
	L*+H H%	wh questions in Pescara and Salerno; disjunctive questions in La Spezia.
	L*+H HL% L*+H HL-L%	Yes/no questions (Turin and Naples, although in the latter the low target in the pitch accent is aligned earlier and a bitonal phrase accent is found; see discussion in Gili Fivela et al. 2015).
	L*+>H L%	Exclamatives (La Spezia, Imperia, Turin. Milan, Lucca, Rome, Lecce, Palermo), Counterexpectational yes/no question (La Spezia, Imperia).
	L*+>H LH%	Counterexpectational yes/no question (La Spezia, Imperia).

Table 3 - *Information-seeking yes/no-questions: transcription of nuclear patterns found in the varieties of Italian (left table) and their stylization (right schemes); motives indicate possible groupings on the basis of nuclear tones; varieties are represented by abbreviations: Milan (MI), Turin (TO), Imperia (IM), La Spezia (SP), Florence (FI), Siena (SI), Pisa (PI), Lucca (LU), Rome (RO), Pescara (PE), Neaples (NA), Salerno (SA), Cosenza (CS), Bari (BA), Lecce (LE) and Palermo (PA) – adapted and updated from Gili Fivela* et al. *(2015)*

	LH%	H%	HL%	L%
H+L*	IM MI TO LU SA CS LE		PI LU	
H*+L	IM SP MI PI RO PE SA LE			
L+H*	TO SA CS BA			SA CS BA
H*	SI FI			
L*+H			TO NA	SP PA

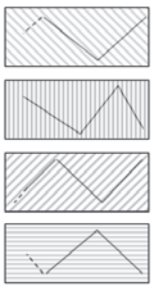
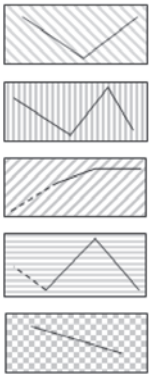


Table 4 - *Information-seeking wh questions: transcription of nuclear patterns found in the varieties of Italian (left table) and their stylization (right schemes); motives represent possible groupings on the basis of nuclear tones; varieties are represented by abbreviations: Milan (MI), Turin (TO), Imperia (IM), La Spezia (SP), Florence (FI), Siena (SI), Pisa (PI), Lucca (LU), Rome (RO), Pescara (PE), Neaples (NA), Cosenza (CS), Salerno (SA), Bari (BA), Lecce (LE) and Palermo (PA) – adapted and updated from Gili Fivela* et al. *(2015)*

	LH%	H%	HL%	L%
H+L*	MI TO IM LU FI SI RO SA BA CS PA		IM SP	IM MI TO SP PI LU SI RO NA PE CS SA BA LE PA
H*+L				
L+H*		RO CS		CS
H*				
L*+H		PE SA		SP PE



Bibliography

Benincà, P., Parry M. & Pescarini D. (2016). The Dialects of Northern Italy. In Maiden, M. & Ledgeway, A. (Eds.), *The Oxford Guide to the Romance Languages*. New York - Oxford: Oxford University Press, 185-205.

Blum-Kulka, S., House, J. & Kasper, G. (1989). Investigating crosscultural pragmatics: an introductory overview. In Blum-Kulka, S., House, J. & Kasper, G. (Eds.), *Cross-Cultural Pragmatics: Requests and Apologies.* Norwood, NJ: Ablex, 1-34.

Bruce, G. (1977). *Swedish Word Accents in Sentence Perspective*. CWK, Gleerup.

Crocco, C. (2011). Profili melodici della varietà genovese. *Contesto Comunicativo e variabilità nella produzione e percezione della lingua: Atti del VII Convegno AISV.*

Endo, R., Bertinetto, P.M. (1997). Aspetti dell'intonazione in alcune varietà dell'italiano. In Cutugno, F. (Ed.), *Atti delle VII Giornate di studio del Gruppo di fonetica sperimentale* (Neaples, 1996), 27-49.

Forner, W. (1975). *Generative Phonologie des Dialekts von Genua*. Hamburg: Buske.

Forner, W. (1988). Areallinguistik I. Ligurien/Aree linguistiche I. Liguria. *LRL* IV: 453-469.

Forner, W. (1995). The Ligurian Dialects. In Parry, M., Maiden, M. (Eds.), *The Dialects of Italy*. London: Routledge, 245-252.

Frota, S., Prieto P. (2015). *Intonation in Romance.* Oxford: Oxford University Press.

Gili Fivela, B. (2006). 'Scaling' e allineamento dei bersagli tonali: l'identificazione di due accenti discendenti. In *Atti del Convegno AISV*, Salerno, 30-2 dicembre 2005, Torriana (RN): EDK, 214-232.

Gili Fivela, B. (2008). *Intonation in Production and Perception: The Case of Pisa Italian.* Alessandria: Edizioni dell'Orso.

Gili Fivela, B., Avesani, C., Barone, M., Bocci, G., Crocco, C., D'Imperio, M., Giordano, R., Marotta, G., Savino, M. & Sorianello, P. (2015). Intonational phonology of the regional varieties of Italian. In Frota, S., Prieto, P. (Eds.) *Intonation in Romance*, Oxford: Oxford University Press, 140-197.

Gili Fivela, B., Interlandi, G. & Romano, A. (2015b). On the importance of fine alignment and scaling differences in perception: the case of Turin Italian. In Romano, A., Rivoira, M. & Meandri, I. (Eds.) *Aspetti prosodici e testuali del raccontare: dalla letteratura orale al parlato dei media*, Atti del 10° convegno AISV, 22-24 gennaio 2014, Università di Torino, Torino: Edizioni dell'Orso, 229-254.

Gili Fivela, B., Iraci, M. (2017). Variation in intonation across Italy: the case of Palermo Italian. In Bertini, C., Celata, C., Lenoci, G., Meluzzi, C. & Ricci, I. (Eds.), *Fattori Sociali e Biologici nella Variazione Fonetica – Social and Biological Factors in Speech Variation*, Collana Studi AISV 3, Milano: Officinaventuno, 169-190.

Grice, M., D'Imperio, M., Savino, M. & Avesani, C. (2005). Strategies for intonation labelling across varieties of Italian. In Jun, S.-A. (Ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing.* Oxford: Oxford University Press, 362-89.

Ladd, D.R. (1996). *Intonational Phonology*. Cambridge: Cambridge University Press.

Loporcaro, M. (2009 [2013]). *Profilo linguistico dei dialetti italiani*. Roma-Bari: Laterza.

Loporcaro, M. (2015). *Vowel Length from Latin to Romance*. Oxford: OUP.

Magno-Caldognetto, E., Ferrero, F., Lavagnoli, C. & Vagges, K. (1978). F0 contours of statements, yes/no questions, and wh-questions of two regional varieties of Italian. In *Journal of Italian Linguistics*, 3, 57-68.

Pellegrini, G.B. (1977). *Carta dei dialetti d'Italia.* Pisa: Pacini.

Pierrehumbert, J. (1980). *The Phonetics and Phonology of English Intonation.* Ph.D thesis, Massachusetts Institute of Technology.

Romano, A. (2003). Accento e intonazione in un'area di transizione del Salento centromeridionale. In Radici Colace, P., Falcone, G. & Zumbo, A. (eds), Storia politica e storia linguistica dell'Italia meridionale: Atti del Convegno internazionale di studi parlangeliani (Messina, 22-3 May 2000). Messina: Scientifiche italiane, 169-81.

Savino, M. (2012). The intonation of polar questions in Italian: where is the rise? In *Journal of the International Phonetic Association*, 42, 1, 23-48.

DOI: 10.17469/O2104AISV000009

MARIA CRISTINA PINELLI, CINZIA AVESANI, CECILIA POLETTO

Is it prosody that settles the syntactic issue? An analysis of italian cleft sentences[1]

The present study aims at investigating the prosodic realization of Italian cleft sentences, in order to provide some new cues for their still debated syntactic interpretation. A monoclausal approach to the analysis of cleft sentences (a.o. Meinunger, 1998) parallels them to left focalization constructions, while a biclausal approach (a.o. Belletti, 2008) considers them composed of a main copular clause and an embedded pseudo-relative clause. A systematic comparison between cleft sentences and left focalizations – carried out through an experimental study and an analysis of pitch accent distribution, scaling, and prosodic phrasing – leads to conclude that their prosodic realization is very similar, thus suggesting that a monoclausal analysis for cleft sentences is supported by prosodic data.

Keywords: Cleft sentences, Left focalization, Prosody-syntax interface.

1. *A syntactic puzzle*

Cleft sentences have been subject of discussion among syntacticians since Jespersen's (1927) work on English, and no unitary theory has yet been developed. One of the reasons for this is that several different languages display different kinds of cleft structures, which often have slightly different syntactic/semantic/discourse properties (Hartmann, Veenstra, 2013).

However, most scholars agree on the assumption that cleft sentences at least can be used with a highlighting function: given the canonical clause in (1), the corresponding cleft clause in (1b) marks focus on the clefted constituent and treats the relative clause as presupposed backgrounded material.

(1) a. Ho visto Andrea in cucina
See-PST.1SG Andrea in the kitchen
"I saw Andrea in the kitchen"

b. È Andrea che ho visto in cucina
be-PRS.3SG Andrea that see-PST.1SG in the kitchen
"It is Andrea that I saw in the kitchen"

In line with this analysis of clefts' information structure, the cleft sentence in (1b) can easily be compared with another syntactic structure that operates the same fo-

[1] Authorship note: this study has been jointly designed by the three authors. Main responsibility for this paper is divided as follows: §1: Pinelli, Poletto; §2: Avesani; §3: Pinelli; §4: Pinelli, Avesani; §5: Avesani.

cus-background partition through a similar word order change: left focalization. Example (2) shows that the focalized constituent (in capitals) of left focalization structures has the same informational status and nearly the same position of the clefted constituent in (1b):

(2) ANDREA ho visto in cucina
Andrea see-PST.1SG in the kitchen
"ANDREA I saw in the kitchen"

These considerations have led scholars to hypothesize that cleft sentences and left focalizations can also have similarities at a deeper level in the syntactic structure.

The most widespread theories on cleft sentences consider them biclausal structures, composed of a main copular clause and an embedded (pseudo-)relative clause (Declerck, 1988; Den Dikken, 2013), as showed in (3a). However, some other authors like Meinunger (1998) and Frascarelli, Ramaglia (2013) proposed a monoclausal analysis of cleft sentences based on the similarities with left focalizations mentioned above (3b).

(3) a. [È Andrea]$_{\text{COPULAR}}$ [che ho visto in cucina]$_{\text{(PSEUDO-)RELATIVE}}$
b. [È Andrea che ho visto in cucina]
"It is Andrea that I saw in the chichen"

In Frascarelli & Ramaglia's view, the free relative clause (*che ho visto in cucina)* is directly merged in the left periphery of the main clause, in a TopP projection (Familiarity Topic), while the clefted constituent moves from its base position to a FocP projection in the same left periphery. This analysis accounts for the similarities with focalizations, since the clefted constituent undergoes the same A'-movement of the focalized constituent and ends up in the same FocP projection in the left periphery, giving as a result a monoclausal structure (4).

(4) [$_{GP}$[$_{IP}$ t'$_{DP}$ è [$_{SC}$ t$_{DP}$ t$_{NP}$]] [$_{\textbf{FocP}}$ [$_{NP}$ Andrea] [$_{FamP}$ [$_{DP}$[$_{SC}$ [$_{NP}$*pro*] [$_{CP}$ che *pro* ho visto]]] t$_{IP}$]]][2]

A third proposal for the analysis of cleft sentences was made by Belletti (2008). The author claims that in Italian there are two different types of cleft sentences, corrective/contrastive clefts and new information clefts, which have different syntactic structures[3]. Corrective/contrastive clefts are used when the clefted constituent is

[2] GP = Ground Phrase; IP = Inflectional Phrase; DP = Determiner Phrase; SC = Small Clause; NP = Nominal Phrase; FocP = Focus Phrase; FamP = Familiarity Topic Phrase; CP = Complementizer Phrase.

[3] Actually, other non-corrective/contrastive clefts can be found in the literature on Italian clefts: they are the so-called non-prototypical temporal cleft sentences – an example of which is reported in (i) – first analysed by Benincà (1978).

i) Sono due ore che ti aspetto
be-PRS-3PL two hours that you-ACC wait-PRS-1SG
"it is two hours that I'm waiting for you"

Whether they can be considered fully-fledged new information clefts is not clear yet. We aim to further investigate this topic from a syntax-prosody-discourse interface perspective.

explicitly opposed to an antecedent in the preceding context or in the common ground: only in this type of sentences the clefted constituent can be a NP, a PP, an AdvP or an AdjP, and if it is an NP it can function both as a subject or as an object[4]. The syntactic structure of corrective/contrastive clefts is said to be still biclausal – with a main copular clause and an embedded pseudo-relative clause – but slightly different from the one in (3a). In fact, the clefted constituent moves to the left periphery of the lexical verb – that is, of the relative clause – and does not reach the copular clause, as the schematic representation in (5) shows.

(5) a. $[\text{È}]_{\text{COPULAR}}$ [Andrea che ho visto in cucina]$_{\text{(PSEUDO-)RELATIVE}}$

b. $[_{\text{CP}} \ldots [_{\text{TP}} \ldots [_{\text{vP}}$ *be* $[_{\text{CP/}\mathbf{FocPcorr}}$ Andrea$_1$ $[_{\text{FinP}}$ che $[_{\text{TP}}$ ho visto t$_1$ ...

The three structures in (3a), (4) and (5) represent the starting point for this prosodic study. It is important to note that in all three structures the clefted constituent ends up to be hosted in a FocP projection, i.e. it is treated as a focus from a syntactic and information-structural point of view.

The similarity between corrective focalizations and clefts, which has been detected by the supporters of the monoclausal analysis, intuitively seems to hold for their prosodic realization as well, but a systematic comparison of these two structures has not been carried out yet: investigating the prosodic phrasing of clefts and left focalizations, as well as their pitch accent selection, can add new elements for the comprehension of their underlying syntactic structures from a prosody-syntax interface perspective.

2. *The prosody of corrective/contrastive Focus in Italian*

In the literature on syntax-prosody interface it has been proposed that when a focus phrase is moved to the left periphery of the sentence – as it is the case for corrective/contrastive Focus – it is followed by an intonational phrase (*IP*) boundary (Nespor, Guasti, 2002): it separates the phrase with the main intonational prominence from the postfocal material, which is argued to be extrametrical (i.e. not included in the "core IP" as proposed by Szendröi, 2001) or to be right-dislocated (Samek-Lodovici, 2006). Frascarelli (2000) proposes instead that an *IP* boundary follows a contrastive focus phrase only if it is not adjacent to the verb, while an intermediate/phonological phrase occurs in the other cases.

Contrary to the claims of the first authors and more in line with Frascarelli, in two production experiments Bocci, Avesani (2005) and Bocci (2013) show that an initial focus is followed by the right boundary of an intermediate/phonological phrase (*ip*). The evidence resides in the acoustic lengthening of the segments preceding the *ip* right boundary: the final syllable and vowel of a fronted focus phrase were not significantly longer than the corresponding elements at the end

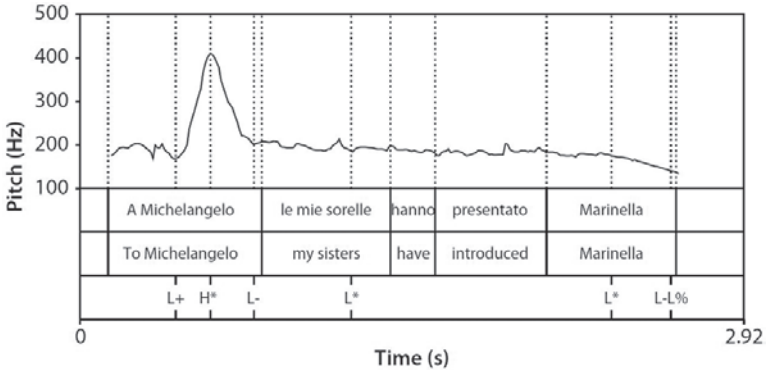
[4] In new information clefts the clefted constituent can only be a subject NP, which means that in Italian new information clefts can only be subject clefts (Belletti, 2008).

of a preverbal subject in broad focus (Bocci, Avesani, 2005) and were shorter than the corresponding syllable and vowel at the end of contrastive and partial topics (Bocci, 2013). They conclude that an utterance divided in left peripheral focus and background, as the one presented in Fig. 1, is metrically phrased as in (6), in which the focus phrase is mapped into an *ip* which occurs in the same *IP* along with the rest of the clause.

The focus phrase attracts the main prominence of the *IP*, leading to an Intonational Phrase whose head is *not* assigned to its rightmost element. The backgrounded material in postfocal position is intonationally realized as a low and flat pitch contour but, despite the apparent absence of intonational cues, it cannot be considered as extrametrical (i.e., dephrased and deaccented). Bocci, Avesani (2011), Bocci (2013) and Bocci, Avesani (2015) show that the postfocal material is prosodically "visible": it is phrased, and L*[5] pitch accents are associated with the metrical head of any postfocal phrase.

(6) $[[[\text{A MICHELANGELO}]_{ip}\,[\text{le mie sorelle hanno presentato Marinella}]_{ip}\,]_{IP}]_{U}$
"To MICHELANGELO my sisters presented Marinella"

Figure 1- *From Bocci (2013:145). Left focalization: "A MICHELANGELO le mie sorelle hanno presentato Marinella" (To Michelangelo my sisters introduced Marinella). U = utterance, IP = Intonational Phrase, ip = intermediate Phrase*



3. *Hypotheses and predictions*

The syntax-prosody mapping rules as formulated by Selkirk (e.g., Selkirk 2005), require that a syntactic clause is mapped into a prosodic constituent of *IP* level. According to this view, if corrective/contrastive cleft sentences are biclausal, we expect that each clause is phrased in a separate Intonational Phrase, i.e. that an *IP* boundary occurs between the main copular clause and the pseudo-relative clause.

[5] In Bocci analysis, L* pitch accents are non fully-fledged accents: they are inserted in the metrical structure only in fulfilment of phonological requirements, and have the function to mark the right side of the focus phrase.

Specifically, in Declerck (1988) and Den Dikken (2013) syntactic analysis (3a) we expect that an *IP* boundary occurs *after* the focused element, while in Belletti (2008) analysis (5) we expect that the *IP* boundary occurs after the copula and *before* the focused element. Conversely, if the corrective/contrastive clefts are monoclausal (Meinunger, 1998 and Frascarelli, Ramaglia, 2013), we expect cleft sentences to be prosodically phrased in one *IP* and, assuming Bocci's analysis for sentences with a left peripheral contrastive/corrective focus, that a prosodic boundary of *ip* level occurs after the focused element. No boundary is expected to occur between the copula and the focused element, as they are both wrapped in the same Intermediate/Phonological Phrase.

Accordingly, if corrective/contrastive clefts are biclausal as in (3a), we expect that the prosodic boundary after the focused element in the copular phrase will be stronger than in left-focalized sentences. Specifically, that the boundary will be cued by a pre-boundary lengthening longer than in focused sentences and by the presence of a pause. If corrective/contrastive clefts are biclausal and have the syntactic structure proposed in (5), we expect the copula "è" to be endowed with a nuclear pitch accent and to be followed by a pause, differently from left-focalized sentences where we expect the copula to have at most a prenuclear accent and no following pause.

Moreover, we expect different prosodic phrasings for biclausal clefts and left-focalized sentences independently of their structural position, i.e. whether or not they are embedded in a superordinate main sentence.

In order to test these predictions, we run a production experiment on minimal pairs of cleft sentences and left focalizations in which we examined their prosodic phrasing and tonal structure with an analysis cast within the framework of the Autosegmental-Metrical Theory of Intonation (e.g. Beckman, Pierrehumbert, 1986; Ladd, 1996).

The paper is structured as follows: after a Method section (§4) in which we present the corpus chosen, the speakers and the prosodic measurements, in section 5 we report the results on phrasing (§5.1) and on type and distribution of pitch accents across the sentence pairs (§5.2). In §5.3 we examine in detail the alignment and scaling properties of the most widely used focal pitch accent in both cleft and focus sentences. In section 6 we discuss how the prosodic analysis impinges on the syntactic interpretation of cleft sentences.

4. *Method*

In order to systematically compare corrective cleft sentences and left focalizations, a corpus of minimal pairs was created, taking into account three syntactic conditions. The first set of minimal pairs includes 8 main clauses with a singular clefted/focalized constituent; the second set includes 4 main clauses with plural clefted/focalized constituents, while the third set includes 4 embedded cleft sentences and focalizations with a singular clefted/focalized constituent. For all three conditions,

subject and object clefts and focalizations were equally distributed within each set of minimal pairs.

In the clefted/focalized constituents, target words are stressed on the penultimate or on the antepenultimate syllable (paroxytones *vs* proparoxytones) and the stressed syllable can be open or close (CV *vs* CVC). All segments of the target words are sonorants, in order to minimize microprosodic effects. The corpus included indeed 5 paroxytone words with an open stressed syllable, and 2 proparoxytone words, 1 with an open stressed syllable and 1 with a closed stressed syllable.

Each target sentence has been inserted in a short conversational context, with the aim of suggesting the desired interpretation and making the reading as natural as possible; the resulting short paragraphs have been pseudo-randomized and presented to the subjects as Power Point slides. 40% of similar extra texts have been added to the experimental set, in order to serve as fillers.

Four female speakers of the Italian variety of Rome, aged 20-28, were asked to read out the small texts three times. They have been recorded with a Shure WH20QTR microphone and a Zoom H2 digital recorder. Two out of three repetitions were segmented and analysed. Out of 144 sentences, was 144, 7 were discarded because reading errors had occurred.

The resulting 137 target sentences (61 focalizations, 76 corrective clefts) have been then extracted from the contexts, segmented and transcribed at phone and syllable level in Praat, and ToBI transcribed. The transcriptions have been carried out separately by the first two authors, and the diverging cases have been discussed untill an agreement was reached[6].

In order to find out whether Italian corrective/contrastive cleft sentences and left focalizations have the same prosodic realization, we measured and compared the following parameters: i) duration of the last vowel[7] of the clefted/focalized constituents; ii) distribution of focal and postfocal pitch accents types in the two structures; iii) alignment of the tonal targets of focal pitch accents with the stressed syllables of the clefted/focalized constituent, obtained by calculating the latency of L and H targets from the stressed syllable onset; iv) scaling of the L and H targets of the focal pitch accents in the clefted/focalized constituents (absolute height, Δraising and Δfalling, in Hz).

Measures of pre-boundary lengthening (i) should inform us about presence and level of prosodic phrasing between the clefted/focal constituents and the rest of the sentence; distribution of focal and postfocal pitch accents (ii) should reveal if the

[6] Out of 137 sentences, in 11 cases the two authors diverged in the trascription of the focal pitch accent (agreement: 80%). An agreed upon transcription for those cases was reached after a discussion of each one. As for prosodic phrasing, a 100% agreement was reached.

[7] Even if desirable, in analyzing the prosodic phrasing we did not take into consideration the duration of the syllable, as the last unstressed syllable of the focal word varies across sentences both in structure (CV, CjV, V) and in segmental composition. The rather limited size of the corpus would not balance out the significant variation in the duration of the syllable as induced by the presence/absence of an onset consonant and by its class (nasal, liquid, stop). The variation induced by the quality of the vowel has been taken care of by setting the target words as a random effect ("Item") in the statistical analyses.

intonation contours are comparable in the two structures; alignment and scaling measures of the focal pitch accents (iii, iv) should provide phonetic evidence for the phonological categorization of the pitch accents.

5. *Results*

5.1 Phrasing

In order to verify if a prosodic boundary occurs after the focal word in the sentence pairs (an example is shown in (7a, b)), we first inspected the F0 contours and then compared the duration of the last unstressed vowel of the focal word in focus and cleft sentences. The focal word may occur in the main clause (7a,b) or in an embedded position (7c, d):

(7) a. $[\text{MARINA}]_{\text{FOC}}$ regala gioielli di valore
b. $[\text{è MARINA}]_{\text{COPULAR}}$ $[\text{che regala gioielli di valore}]_{\text{(PSEUDO-)RELATIVE}}$
c. Ho sentito dire che $[\text{MARINA}]_{\text{FOC}}$ regala gioielli di valore
d. Ho sentito dire che $[\text{è MARINA}]_{\text{COP.}}$ $[\text{che regala gioielli di valore}]_{\text{P.-RELATIVE}}$

Out of 137 sentences analysed, a pause occurred after the focused word only in 7 cases (1 main and 2 embedded left-focus sentences; 2 main and 2 embedded clefts). The pause was easily identified in the set of left-focalized sentences (n=3, average duration = 123.6 ms; $\sigma = 20.8$); while in cleft sentences, as the focused elements is followed by the complementizer "che" ([ke]), the pause, if present, cannot be easily distinguished from the closure duration of the velar stop. Therefore, we calculated the average velar closure duration of the pre-focal "che" in the declarative main clause of embedded clefts (e.g.: 7d) and set that duration as the threshold for distinguishing a pure closure from a closure-plus-pause.

On average, the [k] closure duration of the prefocal "che" is 33.67 ms (n =14; $\sigma = 5.86$); durations of post-focal [k] closures above that threshold were considered the combined outcome of the stop closure and a pause. Four cases exceeded that value, with an average duration of closure-plus-pause of 122.5 ms ($\sigma = 26.5$).

Given that a pause was detected after the focused element only in 5% of the left-focalized sentences (3 cases out of 61) and in 5% of the corrective/contrastive clefts (4 cases out of 76), we excluded the systematic presence of an IP boundary after the focal accent in both sentence types.

In order to ascertain the presence of a boundary of a lower hierarchical level, we fit a Generalized Linear Model (JMP platform) to the duration of the last unstressed vowel with "Type of Sentence" (Focus *vs* Cleft) and "Sentence Position" (main *vs* embedded clause), as fixed factors and "Subject" and "Item" as random effects. Our corpus does not include sentences in which the target word occurs also as a subject in a broad focus sentence, condition that would allow a direct comparison with cases of absence of boundary after the target word. Therefore, we choose to compare the final vowel of the focal word with the final vowel of the post-focal following verb that we can safely assume is phrase-internal, as a verb is usually

phrased with its complement in a same prosodic constituent (see Table 1). As a consequence, a third fixed factor "Position in ip" (ip-internal vs ip-final) and the interaction "Type of Sentence* Position in ip" were included in the model.

Table 1 - *Boundary vowels and phrase-internal vowels compared for final lengthening in main and embedded clauses. Target vowels are bold and underlined.*

	ip-final	*ip-internal*
Main	È Debor**a**	che ved**o** bene in un'azienda a Milano
Embedded	Ho sentito dire che è Marin**a**	che regal**a** gioielli di valore

Results show that the factors "Type of Sentence" and "Sentence Position" are not significant, indicating that the unstressed final vowel of the focal word is not significantly different in focus and cleft sentences, nor is it different if the focal word occurs in a main or in an embedded clause. Instead, "Position in ip" is a significant factor, indicating that the duration of the unstressed final vowel is significantly longer when the word is focal (mean: 127.79 ms) compared to when it is post-focal and ip-internal (mean: 49,30 ms), $F(1,9.73) = 24.063$, $p = .0007$. The interaction between "Type of Sentence" and "Position in ip" is non-significant.

We can infer from these data that both the clefted phrase and the focus phrase are followed by a boundary of *ip* level and consequently we can conclude that the presence of an *ip* boundary points to a monoclausal interpretation of the cleft sentence, in which the copular clause and the pseudo-relative clause are wrapped in two *ips* included in the same *IP*.

However, such a phrasing would not rule out the biclausal interpretation proposed by Belletti, as it shows the presence of an *ip* boundary on the right of the focal word but it does not exclude a boundary on its left. That is, it does not exclude a phrasing in which the copula is mapped into an autonomous *IP* and the rest of the sentence is mapped into two *ips* included in the same *IP* as in (8):

(8) $[[[\text{è}]_{ip}]_{IP}[[\text{MARINA}]_{ip}\,[\text{che regala gioielli di valore}]_{ip}]_{IP}]_{U}$

The prosodic analysis shows that no pause follows the copula in any of the cleft sentences, thus excluding the presence of an *IP* boundary between the copula and the focal word. Moreover, in the vast majority of cases, the copula is not realized with any pitch accent, thus resulting included with the following focal NP in the same intermediate phrase. In the few cases in which a pitch accent is associated with the copula (10% of the total cases), its height is lower than the following focal accent, resulting in a prenuclear PA prosodically subordinated to focal accent within the same *ip*.

Summarizing, results on final lengthening and pitch accent distribution indicate that clefted and focalized constituents are phrased similarly, and that the boundary right-flanking the focus word is of *ip* level. Since there is no evidence that the copula is independently phrased nor headed, the data suggest for cleft sentences the phrasing in (10) that points to a monoclausal interpretation:

(9) $[[[\text{È MARINA}]_{ip}\ [\text{che regala gioielli di valore}]_{ip}]_{IP}]_{U}$

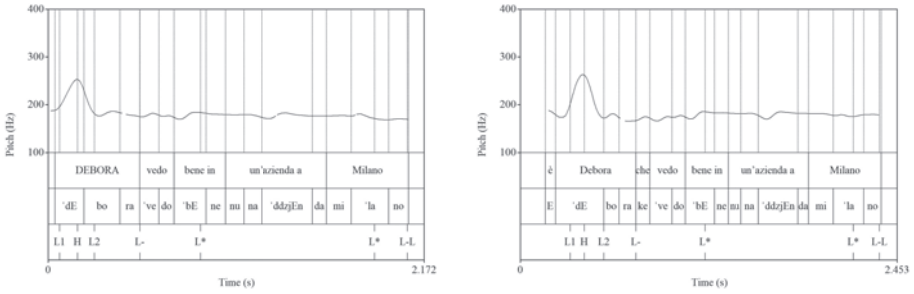
5.2 Clefts and left-focalized sentences: types of pitch accents and their distribution

All cleft sentences and left focalizations collected for this study have been clearly realized as focus-background structures, with a first prosodic constituent bearing a focal pitch accent, and a second constituent realized with a low and flat F0 contour.

In both sentence types the pitch contour starts at a medium-low pitch level and continues roughly flat until the onset of the stressed syllable is reached. There, the pitch sharply rises to a peak in the accented syllable and sharply falls at the end of it. From the offset of the accented syllable on, the F0 contour is slightly falling or flat till the end of the focalized word and continues on a low level until the end of the sentence (Fig. 2). The same pattern is found when the clefted and focalized constituents are embedded within the main clause and no intervening intermediate phrase boundary separates the two clauses (Fig. 6, top). If the main clause is wrapped into an *ip* marked by L- at its right boundary, at the start of the embedded clause the F0 contour continues low and flat until the onset of the accented syllable (Fig. 6, bottom).

In the postfocal constituent, the great majority of prominences that can be found in the corpus are of the L* type (120 out of 137) comparably to what reported by Bocci, Avesani (2006) for left focalization in Tuscan Italian; in cleft sentences, L* amounts to 87% of the postfocal accents and in left-focalized sentences to 96%. In all other few cases (17 out of 137), a compressed !H+L* pitch accent can be detected, comparable to the post-focal pitch accent in the southern varieties of Italian (D'Imperio, 2000; Grice, D'Imperio, Savino & Avesani, 2005) and in European Portuguese (Frota, 2000). The distribution of such compressed post-focal accent is not uniform across speakers but appears to be speaker-specific: one speaker (CC) is responsible for most cases of compressed !H+L* pitch accents (11/32); another speaker produces 4/32 cases and the remaining two speakers one case each.

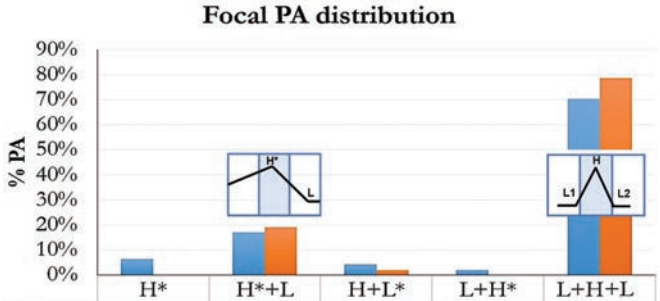
Figure 2 - *Left: Corrective left focalization: "DEBORA vedo bene in un'azienda a Milano" (DEBORA I can imagine in a company in Milan); right: Corrective cleft sentence "è DEBORA che vedo bene in un'azienda a Milano" (it is Debora that I can imagine in a company in Milan)*



The pitch accent selection for the focal constituent is more variable, but no significant difference in distribution between clefts and focalizations can be observed. As showed in Fig. 3, the most widespread pitch accent in the corpus is a rising-falling accent both in clefts (70%) and focalizations (79%). It is phonetically characterized by a sharp rise aligned at the onset of the accented syllable and by a sharp fall that ends at the offset of the same syllable or slightly later, soon after the onset of the postonic syllable (see Fig. 3). We label this pitch accent LHL and will discuss it further in §4.3.

17% of remaining cleft sentences and 19% of remaining focalizations bear an H*+L pitch accent: in contrast with the preceding one, this early falling pitch accent is preceded by a plateau or by a gradual rising movement which spreads from the sentence beginning and reaches a peak in the focal syllable, as shown by the contour in Fig. 4. Rare realizations of H*, H+L* and L+H* pitch accents are found, with a slightly higher frequency in cleft sentences than in focalizations. Fig. 3 shows percentages of occurrence of the focal pitch accents in corrective/contrastive clefts and left-focalized sentences.

Figure 3 - *Distribution of focal pitch accents in corrective clefts and focalizations. Schematic representation of the most widespread pitch accents in the F0 contours: LHL and H*+L*



	H*	H*+L	H+L*	L+H*	L+H+L
CleftC	6%	17%	4%	2%	70%
FocC	0%	19%	2%	0%	79%

In a subset of the corpus (see Appendix "Main clause: plural", 27 items) the clefted/focalized constituent is a complex NP in which two NPs are coordinated as subjects of the clause, as in the example in (8):

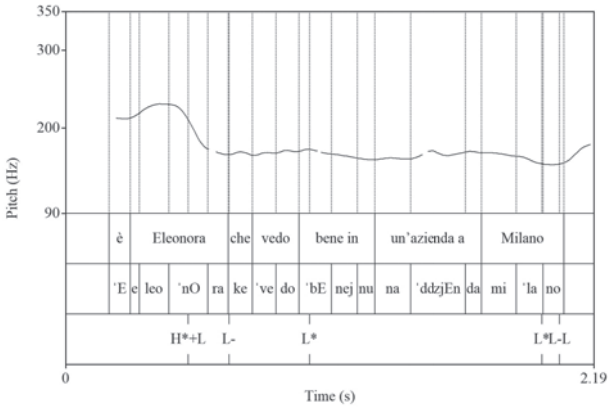
(8) Sono Andrea ed Angelo che vivranno due anni a Londra
It is Andrea and Angelo that will be living in London for two years

In 81% of those cases, a prefocal pitch accent has been realized on the stressed syllable of the first NP. The prefocal pitch accent type is variable, with a great majority of L+H* (59%) and some occurrences of H* (27%) and LHL (14%). The variability of prefocal pitch accents is predicted by the AM theory: their presence is not compulsory, and they do not convey any relevant pragmatic information – as opposed to focal pitch accents. Moreover, all but 2 prefocal pitch accents have been realized with a lower pitch span than the focal pitch

accent of the clause, which in all other cases reaches the highest F_0 of the whole contour.

Overall, the most frequent focal pitch accent in cleft sentences and in early focussed sentences has a rising-falling pattern in which three tonal targets LHL can be detected, all of which appear to align with the stressed syllable.

Figure 4 - *Corrective cleft with a H*+L pitch accent: "è Eleonora che vedo bene in un'azienda a Milano" (it is Eleonora that I can imagine in a company in Milan)*



5.3 Alignment and scaling of the LHL focal pitch accent

To evaluate whether the focal accent displays the same phonetic features in clefts an in left focalizations we analysed the alignment and the scaling properties of each tonal target in both sentence types.

Each tonal target was manually tagged by visual inspection of the F0 curve. The location of the H tone was defined as the point in time where the rise reaches the F0 peak and the peak was automatically detected by Praat as the F0 maximum within the pitch accent. No cases of high plateau are present in the set of the LHL focal accents. The particular nature of our corpus, in which each target sentence has one focal accent which is followed by a long stretch of low and flat pitch contour and generally preceded by no prenuclear pitch accent[8], highly facilitates the identification of the low turning points that precede and follow the F0 peak in the focal accent (see Figures 2 and 4). The first low target L_1 was identified as the last F0 minimum right before the rise (the "elbow" or inflection point of the pitch curve). The second low target L_2 was identified as the first F0 minimum after the peak from which the F0 curve continues low and flat till the end of the utterance. Cases of proparoxytonic focal words (e.g. "Debora", Figure 2 right) are particularly clear in showing that the minimum of the fall is reached at the offest of the accented syllable (or soon after it), and that the postonic syllables lay on the baseline of the contour as the rest of

[8] Prenuclear accents occur only in the subset of sentences in which the focal word is a plural NP.

the sentence. As the speakers of the Roman variety of Italian we analysed are highly consistent in the production of the focal accents, we felt no need to use a line-fitting procedure to automatically calculate the position of the L tones (e.g. D'Imperio, 2000; Frota, 2002; Welby, 2006), in line with Lickley, Schepman, Ladd (2005).

5.3.1 Alignment

For each L and H target in the tritonal L_1HL_2 accent of clefted/focalized constituent we performed the following measurements: i) distance in ms (expressed as a percentage of syllable duration) of the beginning of the rising movement relative to syllable onset (latency L_1-syll); ii) distance of the F0 peak relative to syllable onset (latency H-syll); iii) distance of the end of the falling movement relative to syllable onset (latency L_2-syll).

As different alignments could result from different syllabic and accentual structures (for a review see Prieto, 2011; D'Imperio, 2012), the results for proparoxytone and paroxytone focal words have been kept separate, as well as those for open and close stressed syllables[9]. In Table 2 latency percentages are reported for the alignment of every target of the L_1HL_2 pitch accent, by syllable type (open *vs* close) and word accentual structure (proparoxytones *vs* paroxytones).

Results show that the high target H is always aligned in the middle of the stressed syllable, both in clefted and in focalized constituents, independently of syllable type and word accentual structure. The timing of H is roughly central in paroxytones with open syllable (Focus: 46%, Cleft: 51%); it is slightly later in proparoxytones with open syllable (slightly more than 60%); while it is the earliest in proparoxytones with closed syllable (about 45%).

Table 2 - *Alignment of the three targets of the LHL pitch accents with respect to the stress syllable onset, organized by syllabic and accentual structure*

word	*#items*	*type*	*L1*		*H**	*L2*	
syllable			*pretonic*	*tonic*	*tonic*	*tonic*	*postonic*
*CV.'**CV**.CV*	35	Focus	91%		51%		22%
	23	Clefts	91%		46%	99%	
*'**CV**.CV.CV*	7	Focus		21%	62%		22%
	7	Clefts		31%	65%		21%
*'**CVC**.CV.CV*	7	Focus		0%	42%	86%	
	11	Clefts		2%	48%	93%	

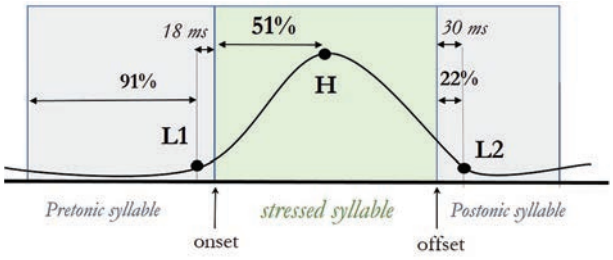
As for the low targets, their alignment appears to depend on the position of the stressed syllable in the word and on the syllable structure. The leading low tone L_1 alignes slightly before the syllable onset in paroxytones (at 91% of the pretonic

[9] Note that, as already stated in §3, the number of items per category is unbalanced, since the CV.'CV.CV words represent the great majority of the corpus.

syllable), both in clefted and in focalized constituents; while it aligns at stressed syllable onset or slightly later in proparoxytones, in a comparable number of cases in focalized and clefted constituents: earlier if the syllable is open, (respectively, 0-2%), later if the syllable is closed (respectively, 21-31%). The trailing tone L_2 aligns with the postonic syllable if the stressed syllable is open, both in focalized paroxytones and proparoxytones (22%); while it aligns differently in clefted constituents: in proparoxtitones, on the postonic syllable (21%) and in paroxytones at the end of the tonic syllable (99%).

Considering the whole LHL pattern, results show that the three tonal targets stably align with the stressed syllable, with small variations induced by the syllabic structure and by the position of the accent in the word. When the trisyllabic word is stressed initially and the syllable is closed, the rising movement is tightly aligned with the syllable onset and the falling movement completed before the syllable offset. When the trisyllabic word has the same accentual structure but the syllable is open, the rising and falling movements are shifted forward in the syllable, with the rising movent starting within the stressed syllable and the falling movement ending in the following unstressed syllable, roughly with the same percentages. If the stressed syllable is word-medial in the trysillable and it is separated by the following *ip* boundary by only one untressed syllable, the LHL pattern is shifted back in the syllable, with the rising movement starting at the end of the preceding unstressed syllable (91%) and the falling movement ending at the offset of the stressed syllable (99%), but only for the clefted constituent. The leftward push of the tonal targets relative to segments in the proximity of a phrase boundary is compatible with what observed in the literature (e.g., Mücke, Hermes, 2007). In the focus cases, the fall takes longer to complete and reaches the low target soon after the onset of the following unstressed syllable (22%, see Fig. 5).

Figure 5 - *Schematic representation of the alignment of the LHL targets in paroxytone words with open syllable in focalizations*



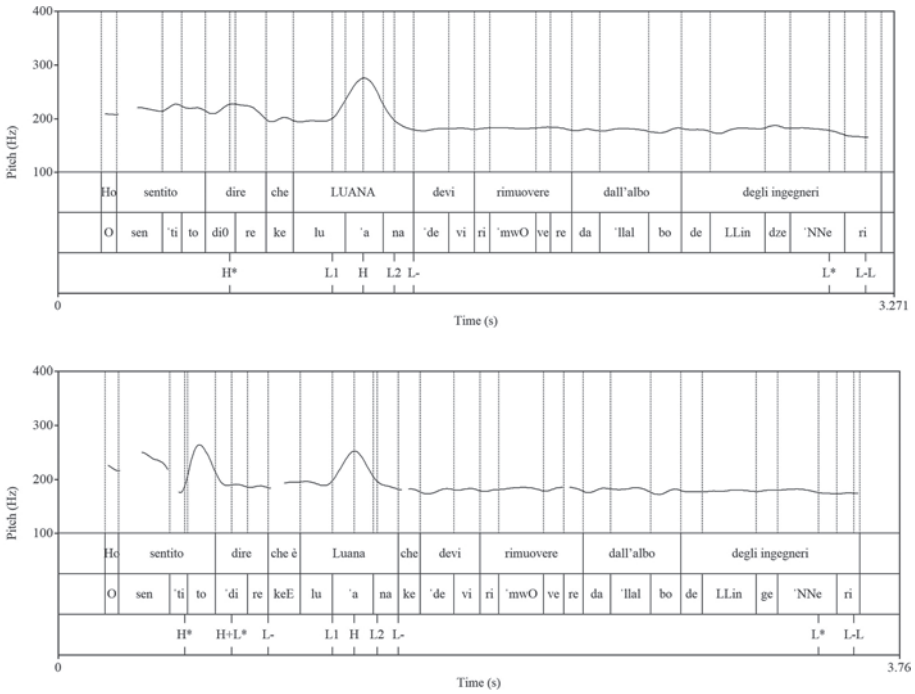
FOCUS – PAROXYTONE – OPEN SYLLABLE

Overall, results show that the tonal targets of the focal pitch accent maintain the same alignment both in cleft and left-focalixed sentences.

5.3.2 Scaling

The focal pitch accent in focalizations and cleft sentences appears to have a symmetrical shape, as can be appreciated in the contours of the following figure.

Figure 6 - *top: Corrective left focalization: "Ho sentito dire che LUANA devi rimuovere dall'albo degli ingegneri" ("I heard that LUANA you should remove from the Register of Engineers"); bottom: Corrective cleft sentence "Ho sentito dire che è LUANA che devi rimuovere dall'albo degli ingegneri" ("I heard that it is Luana that you should remove from the Register of Engineers")*



The height of the peak is comparable in both sentence types. A Generalized Linear Model (JMP platform) applied to the height (Hz) of the H target with "Sentence Type" (focus *vs* cleft), "Sentence Position" (main *vs* embedded), and their interaction ("Sentence Type*Sentence Position") set as fixed factors and "Subject" and "Item" set as random effects indicates that the peak height is not statistically different in clefted and focalized words, and that it does not vary according to of the position of the focalized word in the sentence (mean height: 262.8 Hz, $\sigma = 26.5$).

The low targets of the focal pitch accent, L_1 and L_2, are scaled in the low pitch range of the speakers, with a difference in height compared to the following low phrase accent L- and low boundary tone L% that is compatible with their lying on the declining baseline of the F0 contour[10].

[10] A Generalized Linear Model applied to the F0 of the low targets with "L Type" (L1, L2, L-, L%),

Table 3 - *Mean F0 in Hz (and standard deviation) of the low tones in focus and cleft sentence, in main and embedded position*

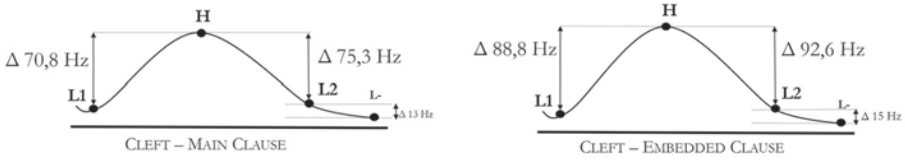
	L1	*L2*	*L-*	*L%*
Main	187.4 (σ 26.4)	179.2 (σ 14.7)	165.4 (σ 12.6)	159.1 (σ 12.9)
Embedded	182.6 (σ 18.3)	175.8 (σ 18.5)	166.1 (σ 9.8)	156.9 (σ 13.6)

In order to verify the symmetry of the focal accent and whether it is kept constant in both sentence types, a Generalized Linear Model was fit on the pitch span of the the rising and in the falling movement as dependent measure, i.e. on the difference in Hz between H and L_1 (Δ raising) and H and L_2 (Δ falling). "Type of Movement" (rising *vs* falling), "Sentence Type" (focus *vs* cleft), "Sentence Position" (main *vs* embedded), interaction "Type of Movement*Sentence" position were set as fixed factors; "Subject" and "Item" were included as random effects. Results show that the pitch span of the raising movement is not significantly different from the falling movement, while Sentence Type and Sentence Position are significant (Sentence Type: $F(1,172.5)= 73.257$ $p <0.0001$; Sentence Position: $F(1,172.3)= 13.595$ $p =0.0003$). The span of the pitch accent is higher in focalized constituents relative to clefted ones (LSM-Hz focus: 86.5 vs cleft: 78.4) and it is higher when it occurs in embedded relative to main clauses (LSM-Hz embedded: 88.9 *vs* main: 76.1; see Table 4). The interaction between type of pitch movement and its position within the sentence is not significant. A summary of the results is presented in Fig. 7 and 8.

Table 4 - *Δ raising and Δ falling values in cleft and left focalized sentences*

		Δ raising	*Standard Error*	*Δ falling*	*Standard Error*
Clefts	Main	71 Hz	6.235	75 Hz	6.235
	Embedded	89 Hz	12.665	93 Hz	12.665
Focus	Main	76 Hz	6.143	88 Hz	6.143
	Embedded	85 Hz	9.249	94 Hz	9.249

Figure 7 - *Scaling in main and embedded clauses: cleft*



"Sentence Type" (focus vs cleft), set as fixed factors and "Subjects" and "Item" set as random effects, shows that "Type of L target" is the only significant factor ($F(3,353)= 61.892$ $p< 0.001$). A Tuckey HSD post-hoc test indicates that each low target is significantly different from all other targets, with a pitch height that steadily decreases from L_1 to L%.

Figure 8 - *Scaling in main and embedded clauses: Focus*

Summarizing, the results presented so far clearly show that the characteristics of the focal pitch accents of clefts and focalizations do not differ: their distribution, alignment and scaling are very similar in all syntactic conditions and for all syllabic structures considered.

6. *Discussion and conclusion*

At the start of this paper we asked whether an analysis of the prosody of cleft sentences would help in disentangling the syntactic issue related to their internal constituency, namely whether they can be considered monoclausal or biclausal structures. Both positions are represented in the literature, with some scholars considering them composed of a copular clause followed by a pseudo-relative clause (see (3a)), while other considering them as monoclausal structures (see (3b)) on the basis of their similarity with left-focalized sentences. Crucially, the similarity is established not only on an informational ground, but also on a prosodic ground: both clefts and left focalizations are sentences partitioned in a focus/background structure and share what appears to be a common prosodic structure, in which the clefted and focalized constituents attract the main prominence of the intonation contour and the following backgrounded material is prosodically subordinated.

Prompted by such reported similarity, our analysis centred on the prosodic properties of a set of minimal pairs of corrective clefts and corrective left focalizations, by examining their phrasing and their accentual structure. Results of the comparative analysis reveal that in both sentence types the focalized/clefted constituent is wrapped in an autonomous prosodic phrase that separates it from the rest of the sentence. The nature of such prosodic constituent is indicated by tonal and metrical evidence: we showed that in both cases i) a L target marks the right boundary of the target constituent, ii) no pause occurs after it, iii) the final vowel of the focalized/clefted word has a comparable duration and it is longer than in a constituent internal position. All prosodic cues concur to indicate the metrical nature of such constituent: in both clefts and left focalizations, the prosodic boundary right-flanking the clefted and focalized constituent is the boundary of an intermediate phrase. The metrical structure of cleft sentences is therefore equivalent to that of sentences in which a corrective focus appears in the left periphery (Bocci, 2013). Namely, the sentence is prosodically phrased in one intonational phrase divided in two intermediate phrases coextensive respectively with the focalized/clefted constituent and with the background. Moreover, no

evidence is found that could indicate a phrasing of the copular clause in an autonomous prosodic constituent: the copula "è" is not associated with any nuclear pitch accent and it is not followed by a pause. Therefore, we argued that the copula is phrased with the clefted constituent in the same *ip* and that the third possible syntactic structure proposed in the literature (see (5)) can be safely ruled out.

That same phrasing holds true also if the linear and structural position of the cleft or left-focalized sentence is manipulated. We increased the syntactic complexity of the sentence by embedding the cleft/focalized sentence in a main one as in (10), and we created the condition in which the prosodic realization of the sentence could be changed: the clefted or the focalized word could become nuclear in a longer prosodic phrase, inclusive of the main clause (Fig. 6, top) or it could remain nuclear in the clefted/focalized constituent but necessarily shifted rightward in the utterance if a prosodic boundary is inserted after the main clause (Fig. 6, bottom).

(10) Ho sentito dire che (è) LUANA (che) devi rimuovere dall'albo degli ingegneri
I heard that it is LUANA you should remove from the Register of Engineers

Results show that the syntactic inclusion in a sovraordinate structure does not change the prosodic phrasing: an *ip* boundary is inserted after the clefted/focalized phrase, while the copula itself has no pitch accent nor is it followed by any boundary that separates it from the clefted noun. Therefore, we can conclude that all metrical evidence supports a monoclausal analysis for cleft sentences.

Clefted and left-focalized sentences are equivalent also on intonational (melodic) ground. They share the same accentual structure: a focal accent on the clefted or left-focalized constituent and postfocal L* accents on the backgrounded material. If, in a minory of cases, a compressed !H+L* pitch accent is used instead, it is equally distributed across sentence types. Clefted and left-focalized sentences also share type and phonetic properties of the most used focal accent: LHL, in which both the rising and the falling movements are tightly aligned with the accented syllable. We did not take any stance on the phonological categorization of such an accent, as we think a more thorough investigation of the intonational system of the Rome variety of Italian is needed before we can define it as truely tritonal pitch accent[11].

In conclusion, our study has shown that corrective clefts and left focalizations share a common prosodic realization in terms of phrasing, accent placement and accent type, and that a detailed analysis of their phrasing suggests a monoclausal interpretation of clefts' syntactic structure. More generally, our study confirms that a thorough prosodic analysis can help to disentangle syntac-

[11] Note that the pitch accent of contrastive/corrective focus in the Rome variety of Italian has been categorized as H*+L in Gili Fivela et al. (2015). No tritonal pitch accents are used in describing the intonational system of the Rome variety of Italian in other previous works (Sardelli , 2006; Sardelli, Marotta, 2009; Giordano, 2006). Contrastive accents with three tonal targets have been attested also in Pisa Italian (Gili Fivela, 2002), and the LHL sequence was categorized as [L]H*+L, i.e. as a bitonal falling accent preceded by a L tone deemed to be a structural property of the peak accent.

tic issues, adding to our comprehension of underlying syntactic structures from a prosody-syntax interface perspective.

Acknowledments

Our special thanks go to one anonimous reviewer for his precious comments and suggestions.

Appendix

Corrective clefts vs corrective focalizations

Main clause, singular

È Andrea che rimane due anni a Londra *it is Andrea that stays two years in London*	ANDREA rimane due anni a Londra *ANDREA stays two years in London*
È Angelo che rimane due anni a Londra *it is Angelo that stays two years in London*	ANGELO rimane due anni a Londra *ANGELO stays two years in London*
È Eleonora che vedo bene in un'azienda a Milano *it is Eleonora that I can imagine in a company in Milan*	ELEONORA vedo bene in un'azienda a Milano *ELEONORA I can imagine in a company in Milan*
È Debora che vedo bene in un'azienda a Milano *it is Debora that I can imagine in a company in Milan*	DEBORA vedo bene in un'azienda a Milano *DEBORA I can imagine in a company in Milan*

Main clause, plural

Sono Andrea ed Angelo che vivranno due anni a Londra *it is Andrea and Angelo that will be living in London for two years*	ANDREA ED ANGELO vivranno due anni a Londra *ANDREA AND ANGELO will be living in London for two years*
Sono Marianna e Valeria che vedo bene in un'azienda a Milano *it is Marianna and Valeria that I can imagine in a company in Milan*	MARIANNA E VALERIA vedo bene in un'azienda a Milano *MARIANNA AND VALERIA I can imagine in a company in Milan*

Embedded clause, singular

Ho sentito dire che è Luana che devi rimuovere dall'albo degli ingegneri *I heard that it is Luana that you should remove from the Register of Engineers*	Ho sentito dire che LUANA devi rimuovere dall'albo degli ingegneri *I heard that LUANA you should remove from the Register of Engineers*
Ho sentito dire che è Marina che regala gioielli di valore *I heard that it is Marina that gives valuable jewelry as a gift*	Ho sentito dire che MARINA regala gioielli di valore *I heard that MARINA gives valuable jewelry as a gift*

Bibliography

Beckman, M.E., Pierrehumbert, J.B. (1986). Intonational structure in Japanese and English. In *Phonology*, 3, 255-309.

Belletti, A. (2004). Aspects of the low IP area. In Rizzi, L. (Ed.), *The structure of CP and IP. The Cartography of Syntactic Structures*, Vol. 2, New York: OUP, 16-51.

Belletti, A. (2008). *Structures and Strategies*. New York: Routledge.

Benincà, P. (1978). Sono tre ore che ti aspetto. In *Rivista di Grammatica Generativa*, 3 (2), 231-245.

Bocci, G. (2013). *The Syntax-Prosody Interface: A Cartographic Perspective with Evidence from Italian*. Linguistik Aktuell 204, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Bocci, G., Avesani, C. (2005). Focus Contrastivo nella periferia sinistra: un accento ma non solo un accento. In Savy, R., Crocco, C. (Eds.), *Analisi Prosodica. Teorie, modelli e sistemi di annotazione. Atti del secondo convegno AISV- Associazione Italiana di Scienze della Voce*, 1-30.

Bocci, G., Avesani, C. (2011). Phrasal prominences do not need pitch movements post-focal phrasal heads in Italian. In Cosi, P., De Mori, R., Di Fabbrizio, G. & Pieraccini, R. *Proceedings of Interspeech 2011*, Firenze, 27-31 August 2011, International Speech Communication Association, 1357-1360.

Bocci, G., Avesani, C. (2015) Can the metrical structure of Italian motivate focus fronting?. In Shlonsky, U. (Ed.), *Beyond Functional Sequence. The cartography of syntactic structures*, vol 10. Oxford: Oxford University Press, 23-41.

Declerck, R. (1988). *Studies on Copular Sentences, Clefts, and Pseudo-Clefts*. Leuven & Dordrecht: Leuven University Press & Foris.

Den Dikken, M. (2013). Predication and specification in the syntax of cleft sentences. In Hartmann, K., Veenstra, T. (Eds.), *Cleft Structures*. Linguistik Aktuell 208, Amsterdam/Philadelphia: John Benjamins Publishing Company, 35-70.

D'Imperio, M.P. (2000). *The Role of Perception in Tonal Targets and Their Alignment*. PhD dissertation, Ohio State University.

D'Imperio, M.P. (2012). Tonal alignment. In Cohn, A.C., Fougeron, C., Huffman, M.K. (Eds.), *The Oxford Handbook of Laboratory Phonology*, Oxford-New York: Oxford University Press, 275-287.

Frascarelli, M. (2000). *The Syntax-Phonology Interface in Focus and Topic Constructions in Italian*. Dordrecht: Kluwer.

Frascarelli, M., Ramaglia, F. (2013). (Pseudo) Clefts at the Syntax-Prosody-Discourse Interface. In Hartmann, K., Veenstra, T. (Eds.), *Cleft Structures*. Linguistik Aktuell 208, Amsterdam / Philadelphia: John Benjamins Publishing, 97-138.

Frota, S. (2000). *Prosody and Focus in European Portuguese: Phonological Phrasing and Intonation*. New York NY: Garland.

Frota, S. (2002). Tonal association and target alignment in European Portuguese nuclear falls. In Gussenhoven, C., Warner N. (Eds.), *Papers in Laboratory Phonology VII*, Berlin: Mouton de Gruyter, 387-418.

Gili Fivela, B. (2002). Tonal alignment in two Pisa Italian peak accents. In Bell, B., Marlien, I. (Eds.), *Proceedings of the 1st International Conference on Speech Prosody*. Aix-en-Provence: Laboratoire Parole et Langage, 339-342.

Gili Fivela, B., Avesani, C., Barone, M., Bocci, G., Crocco, C., D'Imperio, M., Giordano, R., Marotta, G., Savino & M., Sorianello, P. (2015). Varieties of Italian and their intonational phonology. In Frota, S., Prieto, P. (Eds.), *Intonation in Romance.* Oxford: Oxford University Press, 140-197.

Giordano, R. (2006). The intonation of polar questions in two central varieties of Italian. In Hoffmann, R., Mixdorff, H. (Eds.), *Proceedings of the Third International Conference on Speech Prosody (Dresden)*. CD-ROM: PS8-10-155.

Haegeman, L., Meinunger, A., Vercauteren, A. (2013). The syntax of it-clefts and the left periphery of the clause. In Shlonsky, U. (Ed.), *Beyond functional sequences: The cartography of syntactic structures*, vol. 10. Oxford: Oxford University Press, 73-90.

Hartmann, K., Veenstra, T. (2013). *Cleft Structures*. Linguistik Aktuell/Linguistics Today 208, Amsterdam / Philadelphia: John Benjamins Publishing Company.

Jespersen, O. (1927). *A Modern English Grammar on Historical Principles*. Heidelberg: Winter.

Ladd, D.R. (1996). *Intonational Phonology*. Cambridge: Cambridge University Press.

Lickley, R., Schepman, A. & Ladd, R. (2005), "Alignment of "Phrase Accent" lows in Dutch falling-rising questions: Theoretical and methodological implications". In *Language and Speech*, 2005, 48 (*2*), 157-183.

Meinunger, A. (1998). A monoclausal structure for (pseudo-)cleft sentences. In Tamanji, P.N., Kusumoto, K. (Eds.), *Proceedings of NELS* 28. *GLSA,* University of *Massachusetts,* 283-298.

Mücke, D. Hermes, A. (2007). Phrase Boundaries and Peak Alignment: An Acoustic and Articulatory Study. *Proceedings 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany.

Nespor, M., Guasti, M.T. (2002). Focus-stress alignment and its consequences for acquisition. In *Lingue e linguaggio*, 1 (1), 79-196.

Samek-Lodovici, V. (2006). When right dislocation meets the left-periphery. A unified analysis of Italian non-final focus. In *Lingua,* 116 (12), 836-873.

Sardelli, E. (2006). *Prosodiatopia: alcuni parametri acustici per il riconoscimento del parlante.* PhD dissertation, Università di Pisa.

Sardelli, E., Marotta, G. (2009). Prosodiatopia: parametri prosodici per un modello di riconoscimento diatopico. In Ferrari, G., Benatti, R., Mosca, M. (Eds.), *Linguistica e modelli tecnologici di ricerca: Atti del XL Congresso della SLI.* Rome: Bulzoni, 411-435.

Selkirk, E. (2005). Comments on Intonational Phrasing in English. In Frota, S., Vigario, M. (Eds.), *Prosodies.* Berlin: Mouton de Gruyter, 11-58.

Szendröi, K. (2001). *Focus and the Syntax-Phonology Interface*. PhD dissertation, University College, London.

Welby, P. (2006). French intonational structure: Evidence from tonal alignment. In *Journal of Phonetics*, 34(*3*), 343-371.

DOI: 10.17469/O2104AISV000010

SIMON WEHRLE, FRANCESCO CANGEMI, MARTINA KRÜGER, MARTINE GRICE

Somewhere over the spectrum: Between robotic and singsongy intonation

The impressionistic characterisation of intonation as "robotic" or "singsongy" is frequent in many phonetics-related fields, such as forensic linguistics, clinical linguistics, perceptual dialectology and language acquisition. Despite its potential for linguistics, however, the characterisation of intonation as flat or sing-songy remains ill-defined. With this contribution, we propose a dynamic characterisation of intonation, focussing on trajectories of fundamental frequency (F0) across time. We apply this method to the issue of intonation in adults with autism spectrum disorders (ASD), which has variously been reported to be both more singsongy and more robotic than the intonation of neurotypically developed speakers. Our results point to the impossibility of characterising the speech of adults with ASD as a single group, thereby offering an explanation for previous contradictory results and highlighting the importance of individual variability.

Keywords: intonation, sing-song, robotic, pitch range, autism spectrum disorder.

1. *Introduction*

1.1 Overview

At first glance, judging a speaker's intonation style seems to be a relatively straightforward task. Listeners intuitively form impressions based on intonation, amongst other things, in many different contexts and without conscious effort. Putting such impressions into words with any degree of accuracy and confidence is a much more difficult task, however, often resulting in the use of a very limited range of terms, such as "robotic" (i.e. monotonous) or "singsongy" (i.e. lively and repeatedly spanning a large range), at the two ends of the scale. An even greater challenge lies in the formation of scientifically testable operationalisations and the choice of appropriate measurements in an effort to uncover the underlying mechanisms and parameters of intonation styles.

In this contribution, we present a new method of measurement which is shown to be capable of reliably quantifying intonation styles. We exemplify this approach using data from the speech of subjects diagnosed with autism spectrum disorders (ASD). Speech in ASD has in the past been described using *both* extremes of characterisation mentioned above (robotic and singsongy), reflecting the problems inherent in relying on such vaguely defined and technically underspecified terms.

We suggest that a further reason for the contradictory claims on intonation in ASD lies in the infelicitous practice of relying on averaged values across groups of

subjects without due consideration of variation at the level of the individual. The importance of considering scientific data at this level is not specific to this study (cf. Cangemi, Krüger & Grice, 2015). It is, however, made all the more critical when trying to understand the behaviour of a group of speakers as heterogeneous as that of individuals diagnosed with ASD. Considering speaker-specific data becomes nothing less than a necessity when we are additionally dealing with (very) small sample sizes of a population, as has been the case in the vast majority of studies dealing with speech in ASD. Our data seem to show that an inappropriate reliance on mean values across speakers has to be considered as an underlying reason for the conflicting findings describing the intonation style of speakers with ASD as either robotic or singsongy.

1.2 The linguistic interest of intonation styles

Intonation styles in general are of interest to linguists for a variety of reasons. First, they are a property of individual speakers. Besides the character attributions formed in everyday spoken interaction, this facet of individual specificity is of interest from both a more practical and a more theoretical standpoint. Practical applications include forensic phonetics and emotion profiling (Ladd, Silverman, Tolkmitt & Scherer, 1985; Mohammadi, Origlia, Filipponi & Vinciarelli, 2012). Regarding theory, the issue is pertinent both to the long-standing debate on idiolects (Paul, 1880) and to the more recent debate about the concept of individual grammar networks (Cangemi et al., 2015).

Intonation styles are relevant, too, for describing the behaviour of groups of individuals. Intonation has featured particularly prominently in research on the speech of one such group, people with autism spectrum disorders. Speakers with ASD are generally said to have "atypical" intonation. Quite what this means and how it can be measured is less clear and what emerges from the limited number of studies investigating this phenomenon is far from conclusive. It has been suggested since Simmons and Baltaxe (1975) that people with ASD often use a singsongy intonation with excessive pitch variation. However, flat and robotic-sounding behaviour has been documented since Kanner (1943; see also Green & Tobin, 2009). Both of these findings are consistent with Baltaxe (1984), who, intriguingly, showed that autistic children had either a very narrow F0 range or a very large F0 range.

Moving beyond groups of individuals, intonation styles are no less relevant for the description of language varieties. There is abundant evidence for the importance of intonation styles for impressionistic judgements of different dialects: Data from Kuiper (1999: 258) shows that Parisians consider the southern Provencal variety to be "singsongy" and "singing", while they consider the eastern Alsatian variety to be "choppy" and "jerky" (see also Nolan, 2006). It is hard to disentangle such attributions from the wide range of cultural factors and stereotypes that may play a role, but speech styles in themselves are sure to be one crucial factor underlying such descriptions. Such speech styles in turn are influenced by the phonological properties of the regional variety spoken. For instance, the varieties of French spoken in the

South are characterised by final schwas (Coquillon, Durand, 2010). This extension of segmental material available for the production of intonation contours provides an opportunity for more pitch movement (cf. Torreira, Grice, 2018), which, while it does not necessarily have to lead to a more lively intonation, certainly could be one factor underlying the impressions of sing-songiness cited above.

The use of singsongy vs. flat intonation also seems to be related to choices in register, as, for instance, sing-songiness is characteristic of infant-directed speech (IDS) (e.g. Holmes, 2013). A more exaggerated, "motherese" speech style has been shown to lead not only to better mother-infant bonding, but also higher intelligibility and, consequently, better later language development in infants (Liu, Kuhl & Tsao 2003; Kuhl, Conboy, Coffey-Corina, Padden, Rivera-Gaxiola & Nelson, 2008). Lively F0 use furthermore characterises speech by adults talking to attractive conversation partners (Leongómez, Binter, Kubicová, Stolarová, Klapilová, Havlíček & Roberts, 2014). Why a more singsongy intonation is used with interlocutors of a greater attractiveness is not entirely clear, but while probably not being orthogonal to experiences of, and positive associations with, IDS, it might also reflect evolutionally desirable traits such as liveliness and lack of threat. Decreased F0 variability has, conversely, been reported as characteristic of competitive contexts with high aggressiveness (Hodges-Simeon, Gaulin & Puts, 2010).

Finally, intonation styles are relevant for bilingual and second language speech. It has been suggested that different languages can have narrower or larger F0 ranges overall. For instance, Dutch and Japanese have been described as having a narrower F0 range than English; Swiss German and Norwegian have been described to have a wider F0 range than English (Celce-Murcia, Brinton & Goodwin, 1996; Graham, 2014). Celce-Murcia et al. (1996) project data describing F0 range in several native languages onto English as a second language. The described differences of F0 range in the L1 are said to be reflected in the L2, with e.g. Dutch-accented English being described as sounding "somehow flat" and Swiss German-accented English described as having "a somewhat sing-songy quality" (ibid: 193).

Despite the relevance of intonation styles to manifold aspects of language and their being a phenomenon of interest at various levels of linguistic inquiry, the associated methods of measurement have been far from uniform. Perhaps surprisingly, no dedicated attempt that we are aware of has been made to tackle the issue of how intonation styles can be quantified appropriately. The work presented here aims to remedy this situation.

1.3 Intonation styles and pitch range

The precise nature of intonation styles beyond subjective characterisation has remained ill-defined, but there is a long tradition of studies investigating the closely related concept of pitch range (see Lehiste, 1975; Ladd et al., 1985). As a complement to the long-established measurement of mean F0 in the description of prosody, various approaches have been made in order to capture what are essentially the levels and variations of a speaker's minimum and maximum pitch. The most recent and widespread characterisation of

pitch range can be found in the work of Mennen, Schaeffler & Docherty (2012) and subsequent work by e.g. Urbani (2013) and Graham (2014). This approach is based on the assumption that pitch range is best described through a combination of linguistic and distributional parameters. In the following, we will further examine this method and point out how we think it may be complemented and refined in order to better capture different styles of intonation.

Ladd, Terken (1995) and Patterson (2000) first suggested using what they call "linguistic measures" in order to determine pitch range. This entails identifying "linguistically relevant landmarks" (Mennen et al., 2012) in the F0 contour and using them, rather than global minima or maxima, to characterise a speaker's F0 range. In practice, the F0 contour is reduced to a series of high or low turning points, which are then labelled and averaged (within equivalent labels). This approach has shown convincing results in its application to a number of languages, but parts of it still leave room for improvement. For instance, the basis for the chosen operationalization is not driven by theoretical deliberations, but rather by pragmatic reasons, as pointed out by the authors themselves:

> Our decision to assume a direct relationship between turning points and phonological tones was driven by practical reasons so as to ensure consistency in our labelling. However, tones and turning points may not necessarily map in a one-to-one fashion, so that some tones may not be realized as turning points and some turning points may not constitute an underlying phonological tone (ibid., footnote 3).

More importantly, the meaning of intonational labels in itself has come under increasing scrutiny and critical re-examination in recent years (see the contributions in D'Imperio, Grice & Cangemi, 2016). In the method for measuring pitch range outlined above, intonational labels are taken as the *starting point* for further analyses, providing a symbolic reduction of the phonetic signal. This is consistent with a widespread approach in intonation research (from Hirst, Di Cristo, 1998, to Hualde, Prieto, 2016). However, recent developments suggest that it could be more fruitful to take the opposite approach and use intonational labels only as the *outcome* of phonological analysis (Cangemi, Grice, 2016; Frota, 2016). In this perspective, the use of labels requires an evaluation of intonational meaning and of prosodic structure, rather than a discretisation of the phonetic signal.

Besides turning points based on symbolic labels, the second pillar of the approach by Mennen et al. takes the form of so-called "Long-Term Distributional" (LTD) measures. These measures essentially comprise the range, mean, skewness and kurtosis of the distribution of F0 values. Although useful for descriptions of pitch range, the reason why LTDs are nevertheless not ideal for exploring intonation styles can be illustrated with the following example. Consider the F0 contour in Fig. 1. Whilst this contour may not be something that will ever be found in human speech data, it is a useful idealisation of the shape an imagined F0 contour truly worth of the description "robotic" might take. To show why LTDs are problematic for the present purpose, compare the contour in Fig. 1, which is relatively monotonous (but mainly monotonic in the mathematical sense, i.e. entirely non-decreas-

ing), with the one in Fig. 2, which represents the other end of the scale: an extreme version of a thoroughly lively, singsong intonation style. The problem is that these two very different contours yield exactly the same result in an analysis of LTD measures (see Fig. 3), thereby completely obscuring the essential difference between the two styles of intonation – at least in the hypothetical, stylised versions considered here. For this reason, LTDs along with linguistic measures based on phonological labels cannot, to our minds, be considered an entirely satisfactory measurement for the characterisation of intonation styles that is the aim of the present study.

Figure 1 - *Hypothetical F0 contour of a monotonic intonation style*

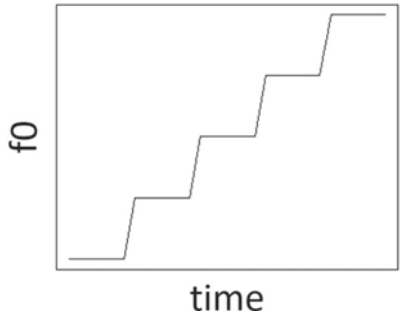


Figure 2 - *Hypothetical F0 contour of a singsongy intonation style*

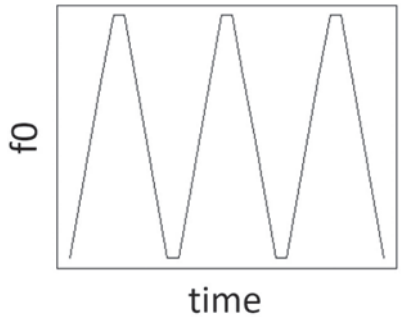
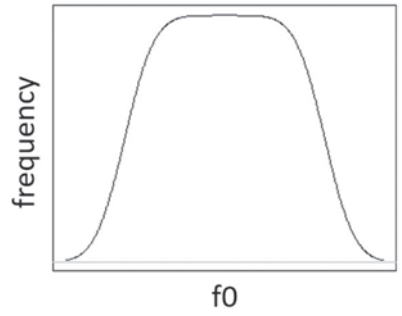


Figure 3 - *Frequency distribution (LTD) of both the monotonic F0 contour shown in Fig. 1 and the lively F0 contour shown in Fig. 2*



2. *Method: A dynamic characterisation of intonation styles*

The novel approach of characterising intonation styles presented in this paper aims to avoid the pitfalls inherent to an approach relying on linguistic and Long Term Distributional measures by instead focussing on the dynamics of F0 contours, represented in the time course of F0 trajectories. Two parameters that capture this aspect are presented in the following: Wiggliness and Spaciousness.

Wiggliness is operationalised as the amount of times an F0 contour "changes direction" over a given stretch of time, i.e. how many different rises and falls are contained within the portion of speech under investigation (based here on a stylisation of the F0 contour with a resolution of 2 semitones).

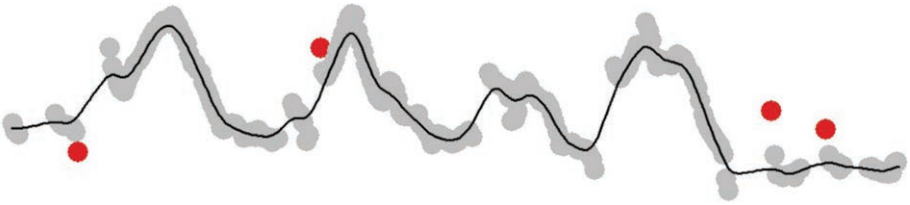
Spaciousness is operationalised as the extent of the slopes of these individual rises and falls, i.e. the maximum F0 excursions.

The more wiggly and spacious the contour, the more singsong we expect it to be and the less wiggly and spacious the contour, the more robotic we expect it to be. As F0 contours can be both more or less wiggly and more or less spacious, the two measures are at least partly independent and are thus chosen to provide a dynamic account of intonation styles.

In a demonstration of how to put this concept into practice, we first choose an excerpt of speech. The length of the excerpt is not fixed and can consist of e.g. one intonation phrase or one interpausal unit.

Next, the F0 contour contained within this excerpt is extracted in Praat (Boersma, Wennink, 2018) and semi-automatically corrected and smoothed using *mausmooth* (Cangemi, 2015). The *mausmooth* procedure is used to first identify any mistakes or artefacts in the Pitch object created by Praat. After correction or deletion of relevant cases, all remaining points are then transformed into a single smooth, continuous contour (see Fig. 4).

Figure 4 - *Correction and smoothing of an extracted F0 contour from an excerpt of speech using mausmooth. Grey dots represent the original pitch contour extracted in Praat, red dots represent points from this original extraction that have been manually corrected or deleted and the black line represents the final smoothed contour.*



In a next step, *Praat*'s Manipulation function is used to stylise the smoothed curve with a 2 semitone resolution (see Fig. 5). This smoothed and stylised curve is the input for further processing which will then yield the characterisation of intonation styles along the dimensions of Wiggliness and Spaciousness introduced above.

The threshold of 2 semitones for smoothing is used here as a first approximation of how the intonation contour might be perceived. By applying smoothing before stylisation, turning points are only located where an actual tonal movement is likely to be perceived. For this reason, we exclude from further analysis certain turning points which are visible in the F0 contour but which are not retained after both the smoothing and stylising procedures (such as the one indicated by the arrow in Fig. 5).

Figure 5 - *Stylisation of the smoothed F0 contour in Praat with a 2 semitone resolution. The arrow indicates an apparent turning point in the F0 contour which is not retained after smoothing and stylising*

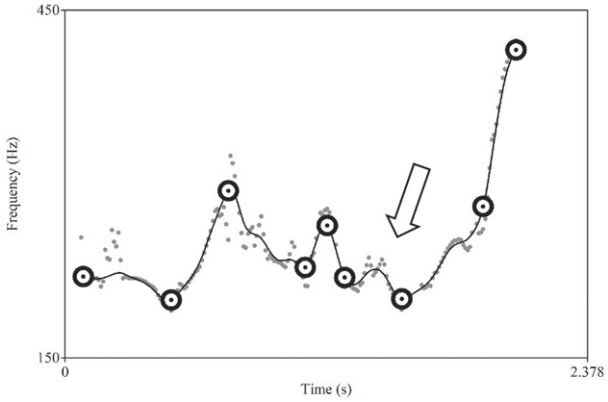
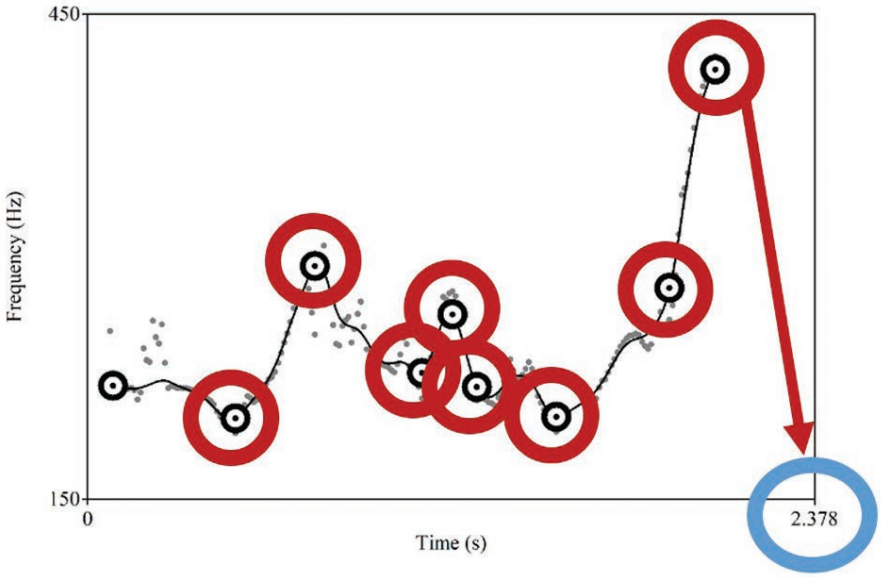


Figure 6 - *The measure of Wiggliness, or Slope Change, is obtained by counting the number of turning points in the stylised F0 contour and dividing it through the length of the excerpt in seconds. In this example, we have 8 turning points after the first one and a total duration of 2.378 seconds, yielding a Wiggliness measure of 3.364*

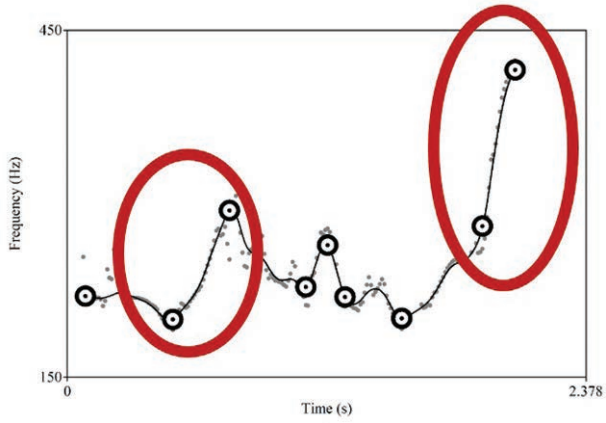


In order to obtain the measure of Wiggliness, or Slope Change, we simply count the number of turning points in the stylised curve and divide this number by the duration of the chosen excerpt in seconds (see Fig. 6).

In order to obtain the measure of Spaciousness, or Maximum Excursions, we simply identify the two largest F0 movements between two turning points and then calculate their average (see Fig. 7).

It is worth pointing out that neither the choice of a 2 semitone resolution for stylisation nor the choice of precisely the 2 largest F0 movements to obtain the averaged value for Spaciousness are extrinsically or theoretically motivated, but simply reflect a starting point for exploration that has proven successful for our data so far. The exact values of these parameters can be adapted and fine-tuned in future work depending on the speech material under investigation. Furthermore, results gained from perception studies designed to test for the perceptual relevance of the measures proposed here will either corroborate or refute their usefulness and guide subsequent refinement of these values.

Figure 7 - *The measure of Spaciousness, or Maximum Excursions, is obtained by identifying the two largest F0 movements between two turning points and then calculating their average. In this example, the two largest excursions have a value of approximately 80 Hz and 135 Hz, respectively. This yields a Spaciousness measure of approximately 105 Hz*



3. *Application*

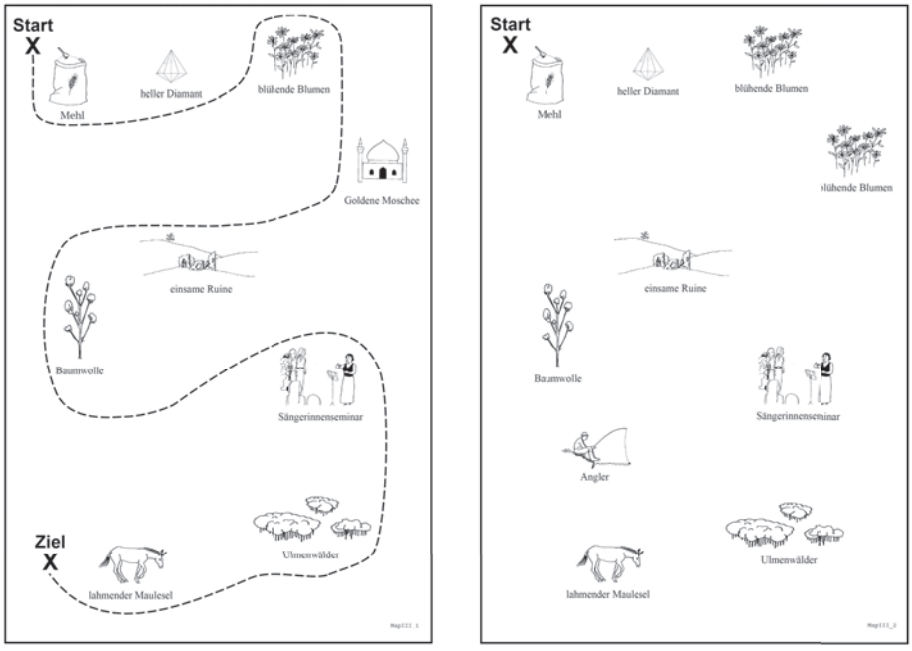
As a test case for this procedure designed to characterise different styles of intonation, we return to one of the issues mentioned in the introduction: the speech of individuals diagnosed with autism spectrum disorders (ASD). This strikes us as a particularly good such test case due to the contradictory claims in the literature about speakers with ASD as having either robotic or singsong intonation. Although these claims have in part been made several decades ago, the issue has not been resolved in any way since. In the following section, we hope to shed some light on why this

might be the case and to demonstrate why our new approach to the characterisation of intonation styles can be helpful in this and other cases.

3.1 Subjects and materials

As part of an ongoing collaboration with the psychiatry department of the University Hospital of Cologne (see e.g. Krüger, Cangemi, Vogeley & Grice, 2018), we have been collecting Map Task recordings (Anderson et al., 1991) between dyads of subjects diagnosed with ASD and dyads of neurotypical (NT) control speakers (all native speakers of German). The materials used in the task are shown in Fig. 8. For the present purpose, we will evaluate data from one female ASD dyad (subjects aged 25 and 46) and one female NT dyad (subjects aged 23 and 26). For each speaker, we extracted 20 excerpts with an average length of 2.5 seconds for further analysis.

Figure 8 - *Map Task materials from the production task. The map for the instruction giver is on the left, the map for the instruction follower is on the right*



3.2 Results

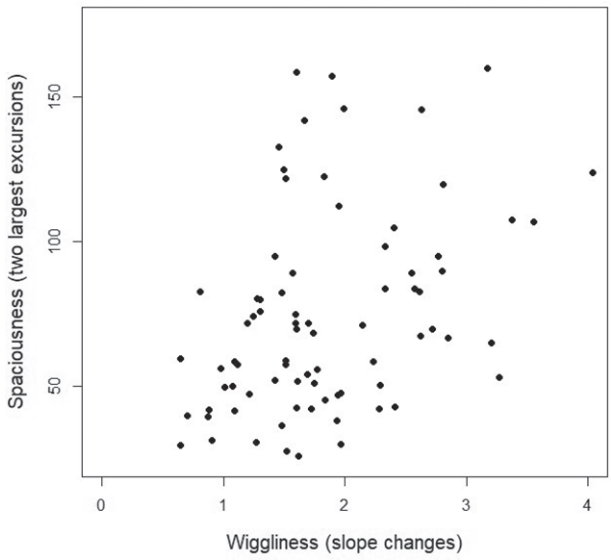
The results are plotted in Fig. 9. The measure of Wiggliness (Slope Changes) is plotted along the x-axis, the measure of Spaciousness (Maximum Excursions) is plotted along the y-axis.

From a general point of view, the plot seems to show that there is some amount of correlation between the two measures. This is not entirely surprising in itself, but will have to be tested with more data in order to quantify the exact strength of the correlation. For the time being, the pattern nevertheless appears to be clearly in

line with the assumption that the two dimensions we have chosen are at least partly independent from each other.

Although the plot in Fig. 9 contains data from two speakers each for the ASD group and for the NT group, it is evident that the data does not cluster into two distinct parts, as would be the case if ASD speakers' intonation was simply either clearly more robotic or clearly more singsongy than that of NT speakers. To make sense of the data, we therefore need to investigate the data at the level of individual speakers.

Figure 9 - *Aggregated data from all 80 excerpts of all 4 speakers. Wigglines is on the x-axis, Spaciousness is on the y-axis*



In Fig. 10 datapoints are colour-coded by speaker. The two ASD speakers are represented by blue and cyan dots, while the two NT speakers are represented by red and orange dots. The NT speaker in red is shown to have a wide range of Wiggliness values and a slightly more limited range of Spaciousness values. This shows that there is a lot of variability in different (parts of) utterances for this speaker. The other NT speaker (in orange) seems to have an intonation style somewhat more towards what could be described as the robotic end. Most values are concentrated in the bottom left quadrant of the plot, representing lower values for both Spaciousness and Wiggliness. Nevertheless, there is variability here, too, with some data points gradually spaced out towards the higher end of both the Wiggliness and Spaciousness scales.

Two examples for the pitch contours represented by the datapoints in Fig. 10 are given in Fig. 11.

Figure 10 - *Wiggliness and Spaciousness values for two NT speakers (red and orange) and two ASD speakers (blue and cyan). The circles in the top right and bottom left of the graph mark the pitch contours shown as examples in Figure 11*

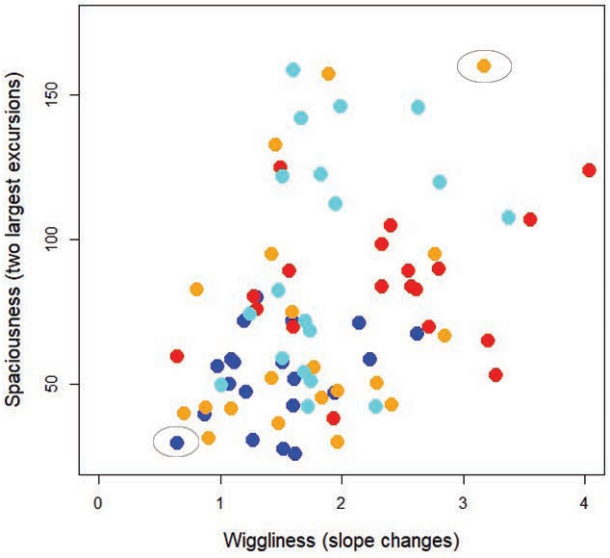
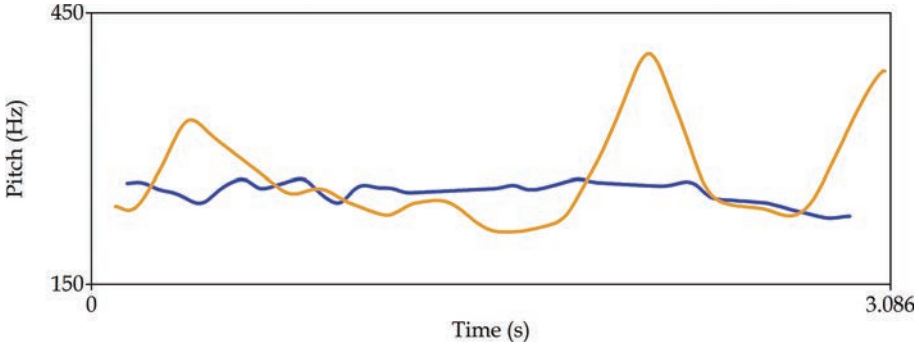


Figure 11 - *Examples of pitch contours represented by dots in Figure 10: The contour in orange is the one marked by a circle in the top right of Fig. 10, the contour in blue is the one in the bottom left of Fig. 10*



Considering the two ASD speakers, the productions of the speaker represented with the blue dots are similar to those of the NT speaker in orange, in being concentrated in the lower regions of both Wiggliness and Spaciousness. The crucial difference between the ASD speaker and the NT speaker is that the productions of the ASD speaker in blue seem to be less variable and therefore much closer to a *uniformly* robotic intonation style, with very few values towards the higher end of Wiggliness and none in the top half of the Spaciousness scale. The second ASD speaker (in cyan) produces a different pattern altogether. Values are spread out along the full range of both dimensions. However, the values are not evenly spread out. There are

very few values in the middle of the graph, around the midpoints of Wiggliness and Spaciousness. Instead, values almost seem to be split into one singsong half and one robotic half. The bottom half overlaps with the rather more robotic productions of the ASD speaker in blue, while the top half, taken on its own, might be considered as a typical representation of a singsong style.

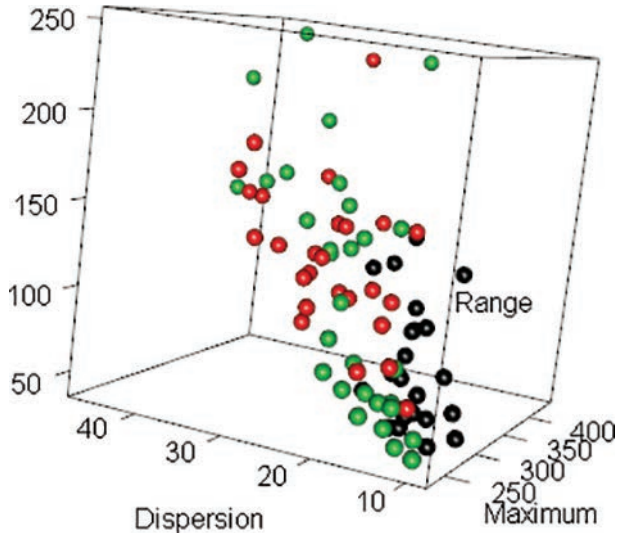
Taken together, the broadest and at the same time most urgent message to be taken from this analysis is that it confirms the absence of a hard dividing line between subjects with a diagnosis of ASD and those without such a diagnosis. Just as autism spectrum disorders *within* themselves cover a range of phenotypical expressions of atypicality that range from low-functioning to high-functioning (amongst other things), there is an *overlap between* the portion of the general population with more autistic-like traits and the portion of the ASD-diagnosed population with outwardly less conspicuous expressions of ASD. This holds true both for general behaviour and for the specific data on intonation styles presented here.

4. *Discussion*

In this contribution, we have pointed out that intonation styles are important in a variety of ways in a number of areas of linguistic inquiry, from the applied to the theoretical. Despite this, accurate and reliable methods for measurement and analysis of what lies behind descriptions of singsong and robotic intonation styles (and what lies in between) have been lacking to date. We have demonstrated the application of a novel, largely automated procedure that fills this gap by reliably quantifying intonation styles, using the example of data from speakers with ASD. These data have also highlighted the necessity of taking into account speaker-specific strategies in the analysis of intonation styles.

Due to the presence of massive individual variability and the absence of clear differences between groups, it is impossible for us to run a conclusive comparison between the metrics employed in this paper and the Long Term Distributional metrics employed in previous research. However, to support the claim that Wiggliness and Spaciousness do indeed provide a better characterisation of speakers' productions, we have plotted utterances from three of the speakers in our dataset into a multidimensional space. The cube in Fig. 12 shows utterance as datapoints, colour-coded per speaker. Points are scattered along the main dimensions of LTD metrics, notably F0 maximum, F0 range and F0 dispersion (calculated as standard deviation of F0 over each individual utterance). The plot indicates that the three LTD metrics are highly correlated, and that they only allow the separation of speakers on the basis of physiological characteristics: Having higher F0 maxima also entails larger values for F0 range and F0 dispersion.

Figure 12 - *LTD measures for three of the speakers in the dataset. F0 dispersion is plotted on the x-axis, F0 range on the y-axis and F0 maximum on the z-axis*



With the approach demonstrated in this contribution we have shown that the picture that emerges from an analysis that is indeed able to capture different dimensions of intonation styles, whilst at the same time giving appropriate consideration to individual differences, confirms the impression that it is inaccurate to describe speakers with ASD as having one particular intonation style. Instead, these speakers seem to show behaviour that goes more towards either end of the spectrum lying between the two poles of "singsongy" and "robotic", both within and across individuals. This reflects similar recent results regarding the prosodic encoding of givenness in ASD (Krüger et al., 2018). Furthermore, our analysis demonstrates that the simplistic labels previously used to describe intonation styles in ASD do not in themselves stand up to thorough investigation.

Although an understanding of the true nature of the data at hand cannot be gained without giving due consideration to individual variability, we submit that this is not the case merely because we are dealing with the somewhat elusive topic of intonation styles in conjunction with the somewhat broad range within the ASD spectrum. In fact, this particular example serves as a useful illustration for understanding the nature and the import of individual variability more generally. Moreover, across different domains of language and different fields of scientific endeavour, not to give due consideration to individual-specific differences is to allow ourselves to be misled by an only apparent simplicity of explanation.

Acknowledgements

We would like to thank Prof. Kai Vogeley of University Hospital Cologne, whose support and expertise were essential to the research presented here. We would also like to thank Harriet Hanekamp and Anika Müller for their invaluable help with data processing. Further, we would like to thank two anonymous reviewers for their extremely helpful comments, as well as the participants of AISV 2018 in Bolzano for insightful and stimulating discussion.

The research for this paper has been funded by the German Research Foundation (DFG) as part of the SFB 1252 "Prominence in Language" in project A02 "Individual behaviour in encoding and decoding prosodic prominence" at the University of Cologne. The first author is partly funded through a doctoral scholarship by the German Academic Scholarship Foundation (Studienstiftung des deutschen Volkes).

Bibliography

Anderson, A.H., Bader, M., Bard, E.G., Boyle, E.H., Doherty, G.M., Garrod, S.C., (...) & Weinert, R. (1991). The HCRC map task corpus. In *Language and speech*, 34(4), 351-366.

Baltaxe, C. (1984). Use of contrastive stress in normal, aphasic, and autistic children. In *Journal of Speech, Language, and Hearing Research*, 27(1), 97-105.

Boersma, P., Weenink, D. (2018). *Praat: doing phonetics by computer* [computer program]. Version, 6.0.39, retrieved 7 April 2018 from http://www.praat.org/.

Cangemi, F. (2015). *mausmooth* [computer program]. Version 1.0, retrieved 7 April 2018 from http://phonetik.phil-fak.uni-koeln.de/fcangemi.html.

Cangemi, F., Grice, M. (2016). The importance of a distributional approach to categoriality in autosegmental-metrical accounts of intonation. In *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 7(1)9, 1-20.

Cangemi, F., Krüger, M. & Grice, M. (2015). Listener-specific perception of speaker-specific production in intonation. In Fuchs, S., Pape, D., Petrone, C., Perrier, P. (Eds.), *Individual Differences in Speech Production and Perception.* Frankfurt am Main: Peter Lang, 123-145.

Celce-Murcia, M., Brinton, D. & Goodwin, J. (1996). *Teaching Pronunciation: A Reference for Teachers of English to Speakers of Other Languages*. Cambridge: Cambridge University Press.

Coquillon, A.-L., Durand, J. (2010). Le français méridional: éléments de synthèse. In Detey, S., Durand, J., Laks, B. & Lyche, C. (Eds.), *Les variétés du français parlé dans l'espace francophone: ressources pour l'enseignement*. Paris: Ophrys, 243-262.

D'imperio, M., Grice, M. & Cangemi, F. (Eds.) (2016). Advancing prosodic transcription. Special collection in *Journal of the Association for Laboratory Phonology*.

Frota, S. (2016). Surface and structure: transcribing intonation within and across languages. In *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 7(1)7, 1-19.

Graham, C. (2014). Fundamental frequency range in Japanese and English: The case of simultaneous bilinguals. In *Phonetica*, 71(4), 271-295.

Green, H. & Tobin, Y. (2009). Prosodic analysis is difficult... but worth it: A study in high functioning autism. In *International Journal of Speech-Language Pathology*, 11(4), 308-315.

Hirst, D., Di Cristo, A. (Eds.) (1998). *Intonation Systems: A Survey of Twenty Languages*. Cambridge: Cambridge University Press.

Hodges-Simeon, C., Gaulin, S. & Puts, D. (2010). Different vocal parameters predict perceptions of dominance and attractiveness. In *Human Nature*, 21(4), 406-427.

Holmes, J. (2013). *An Introduction to Sociolinguistics*, 4th edition. London-New York: Routledge.

Hualde, J., Prieto, P. (2016). Towards an International Prosodic Alphabet (IPrA). In *Journal of the Association for Laboratory Phonology*, 7(1), 5.

Kanner, L. (1943). Autistic disturbances of affective contact. In *Nervous Child*, 2(3), 217-250.

Kuhl, P.K., Conboy, B.T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M. & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). In *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493), 979-1000.

Kuiper, L. (1999). Variation and the norm: parisian Perceptions of regional French. In Preston, D. (Ed.), *Handbook of Perceptual Dialectology Vol. I,* Amsterdam-Philadelphia: John Benjamins, 243-262.

Krüger, M., Cangemi, F., Vogeley, K. & Grice, M. (2018). Prosodic marking of information status in adults with Autism Spectrum Disorders. In *Proceedings of Speech Prosody*. Poznan, 182-186.

Ladd, D.R., Terken, J. (1995). Modelling intra- and inter-speaker pitch range variation. In *Proceedings of XIIIth International Congress of Phonetic Sciences*. Stockholm, 386-389.

Ladd, D.R., Silverman, K.E., Tolkmitt, F., Bergmann, G. & Scherer, K.R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signalling speaker affect. In *The Journal of the Acoustical Society of America,* 78(2), 435-444.

Lehiste, I. (1975). The phonetic structure of paragraphs. In Cohen, A., Noteboom, S. (Eds.), *Structure and Process in Speech Perception*. Berlin, Springer-Verlag, 195-206.

Leongómez, J.D., Binter, J., Kubicová, L., Stolarová, P., Klapilová, K., Havlíček, J. & Roberts, C. (2014). Vocal modulation during courtship increases proceptivity even in naive listeners. In *Evolution and Human Behavior*, 35, 489-496.

Liu, H.-M., Kuhl, P.K. & Tsao, F.-M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. In *Developmental Science*, 6, F1-F10.

Mennen, I., Schaeffler, F. & Docherty, G. (2012). Cross-language differences in F0 range: a comparative study of English and German. In *Journal of the Acoustical Society of America*, 131 (3), 2249-2260.

Mohammadi, G., Origlia, A., Filippone, M. & Vinciarelli, A. (2012). From speech to personality: Mapping voice quality and intonation into personality differences. In *Proceedings of the 20th ACM international conference on Multimedia*, 789-792.

Nolan, F. (2006). Intonation. In Aarts, B., Mcmahon, A. (Eds.), *Handbook of English Linguistics*. Oxford: Blackwell, 433-457.

Patterson, D.J. (2000). *A linguistic approach to pitch range modelling.* PhD dissertation, University of Edinburgh.

Paul, H. (1880). *Principien der Sprachgeschichte.* Tübingen: Niemeyer.

Simmons, J.Q., Baltaxe, C. (1975). Language patterns of adolescent autistics. In *Journal of Autism and Childhood Schizophrenia*, 5(4), 333-351.

Torreira, F., Grice, M. (2018). Melodic constructions in Spanish: Metrical structure determines the association properties of intonational tones. In *Journal of the International Phonetic Association*, 48(1), 9-32.

Urbani, M. (2013). *The pitch range of Italians and Americans. A comparative study.* PhD dissertation, University of Padova.

DOI: 10.17469/O2104AISV000011

ANTONIO ORIGLIA, ANTONIO RODÀ, CLAUDIO ZMARICH,
PIERO COSI, STEFANIA NIGRIS, BENEDETTA COLAVOLPE,
ILARIA BRAI, CRISTIAN LEORIN

Gamified discrimination tests for speech therapy applications

The integrity of phonetic perception abilities is necessary for a normal functioning future speech development. Since the ability to discriminate linguistic sounds is typically associated to the correct acquisition and production of the same sounds, an alteration of this ability could contribute to the onset of speech and language disorders. Support for presenting discrimination tests to young children (5- and 6-years-old), however, is provided when gamified settings are put in place. Moreover, moving beyond static tests in favour of dynamically generated ones may help personalise the test. In this work, we propose an acoustic discrimination test as the first step for the creation of a renovated *Italian Literacy Tutor*. Presented results show promising indications concerning the application of the proposed approach both from the user experience and from the reporting point of view.

Keywords: speech therapy, gamification, discrimination tests.

1. *Introduction*

1.1 The raising of intersubjectivity and cultural learning in infancy

Typical contact with language, in the first years of life, consists of a playful activity where parents and infants engage protoconversations made of rhythmical and musical content. This manifests the emotional regulation of primary intersubjectivity (Trevarthen, 1979), where interaction with the caregiver, either reciprocally directed or mediating access to objects of interest for the infant, manifests the typical playfulness often observed in mammals. At 9 months, secondary intersubjectivity arises (Trevarthen, 1978) and the baby's interest moves onto sharing the ways companions use objects as she starts to interact with the material world in a more informed way. The caregivers' language also shifts, in this phase, from questions and rhetorical comments to instructions and informative comments to support the baby's interest in participating to a task (Halliday, 1975). This is "[...] the start of cultural information transfer between generations" (Trevarthen, 2009: 74). Playful behaviour adapts to new roles as the child grows older but always stays in the background, motivating access to cultural information, reinforcing memory and supporting the creation of meaning (Trevarthen, Aitken, 2001; Reddy, 2008). Language development strongly depends on intersubjective experiences: from the effective engagement of minds and bodies depends cultural learning (Donald, 2001). The naturalistic and social context is also facilitating phonetic learning, because

nine-month-old children can easily learn a new language only if they are involved in a real communicative exchange, but not when they are exposed to the acoustic signal or integrated acoustic-visual signals (like a movie) (Meltzoff, Kuhl, Movellan & Sejnowski, 2009). As a matter of fact, it has been shown that ecological learning is faster, more effective and more lasting than learning from non-naturalistic setting (Kuhl, Conboy, Coffey-Corina, Padden, Rivera-Gaxiola & Nelson, 2008).

In this paper, we will describe a software architecture designed to present discrimination tests in a playful setup depicting a social situation with different kinds of virtual agents. In fact, although humans appear to be born with a natural disposition towards cultural learning (Trevarthen, Aitken, 2003), successful acquisition of cultural skills depends on the interaction quality, especially considering social feedback. Perceived affection and playfulness by the infant towards the parents helps to establish a mutually teasing situation (Reddy, 2008) that focuses attention on rituals that may later become skills (Eckerdal, Merker, 2009). When interaction is insensitive, coercive or qualitatively poor in general, however, it elicits avoidance and protest (Gratier, Trevarthen, 2008), highlighting how "[...] infants are equipped with defensive emotions that repel unsympathetic communication" (Trevarthen, 2009: 80). Digital games, in the modern context, can be a powerful mean to channel literacy contents towards children and modern, engaging technology can be used to design a well-rounded intervention spanning different aspects of the problem at hand. It is therefore necessary to carefully consider what games are, what they are made for, and how they can provide both entertainment and tutoring. The present, ongoing work builds upon the experience of the Colorado Literacy Tutor (Cole, 2003) and of the Italian Literacy Tutor (Cosi, Delmonte, Biscetti, Cole, Pellom & Van Vuren, 2004). We decided to begin the building of such a tutor with a phonetic/phonological module, for two main reasons:

- a correct perception (and reproduction) of the sound system of a language is the *sine qua non* condition to be able to access the other levels of the spoken language;
- speech disorders and language disorders with a phonetic-phonological component are an important, if not the main, portion of the caseloads of speech therapists who deal with voice-speech-language disorders in childhood (Law, Boyle, Harris, Harkness & Nye, 2000).

1.2 Phonological discrimination in childhood

Generally speaking, phonological discrimination means that process of categorical perception through which differences that unfold along a physical continuum (of frequency, intensity, duration) are traced to discrete categories.

Phonological discrimination skills are an essential part of a normal speech perception development, and they are systematically improving up till 10 years of age (Edwards, Fox & Rogers, 2002; Hazan, Barrett, 2000; Nittrouer, 1992), although the cornerstones for a correct discrimination are already laid down by 5 years of age (Weber, Cutler, 2004; Tamashige, Nishizawa, Itoda, Kasai, Igawa & Fukuda, 2009).

Their normal development fortunately can be tested by 4-5 years of age onwards by using the same tests used for adults (Polka, Jusczyk & Rvachew, 1995).

Phonological discrimination tests are an important procedure for assessing proficiency in speech acquisition. In fact, the integrity of phonetic perception abilities is necessary, albeit not sufficient, for a future normal functioning speech development and an alteration of the ability to discriminate "similar" sounds could contribute to the onset of speech and language disorders (Brancalioni, Bertagnolli, Bonini, Gubiani & Keske-Soares, 2012; Freitas, Mezzomo & Vidor, 2015; Nithart, Demont, Majerus, Leybaert, Poncelet, & Metz-Lutz, 2009; Rvachew, Jamieson, 1989; Rvachew, Ohberg, Grawburg & Heyding, 2003; Tallal, 1976). Phonological discrimination tests could vary as to both the form and the content:

- regarding the form, i.e. the procedural paradigm used to test the phonological discrimination skill, the AX or "same/different" paradigm, is to prefer when testing young children, because of less taxing the work-memory in comparison to more sophisticated design (Polka et al., 1995).
- regarding the content, i.e. the verbal material composing the stimuli, the choice is between words and non-words stimuli. While the first are normally easier to administer even to ages earlier than five years, nonwords stimuli are to be preferred because of "the potential usefulness of processing-based measures generally in providing culturally nonbiased assessments of linguistic abilities" (Weismer, Tomblin, Zhang, Buckwalter, Chynoweth & Jones, 2000: 874).

In fact, the non-words discrimination is a task of speech perception less dependent from previous lexical knowledge, thus engaging only the perceptual system and/or the phonological memory, but not the lexical/semantic system. Phonological memory achieves the capacity to face with this task by 4-5 years of age (Polka et al., 1995), and children are successful in comparing short speech sequences out of the context, as it happens in a discrimination task, even because their vocabulary is more than 6000 words great and promotes phonological awareness (Carroll, Snowling, Stevenson & Hulme, 2003). Regarding the perceptual system, a short introduction about the perceptual skills in relationships with phonetic/phonological proficiency of children at the end of pre-school years is needed here. Because of shortage of space, we prefer not even try to resume the huge literature about the development of speech perception abilities from womb up till five years of age, and the interested reader is referred to Choi, Black & Werker (2018), Kuhl et al. (2008), Saffran, Werker & Werner (2006), Walley (2005) for some recent surveys. We focus on the period at the end of the pre-school years because it is the age-range of our sample (the younger age, as written before, at which is possible to apply the same methodologies used with the adult population). We will consider first the ability to process the acoustic dimensions of speech. Jensen, Neff (1993) demonstrated that children tested at four-years-of-age and re-tested 12-18 months later, improved speech discrimination skills beginning with variations in intensity, followed by frequency changes and finally by duration changes, but at the final assessment, for

many of them, frequency and duration discrimination were still poorer than adults' discrimination. The delay in sensitivity maturation of the temporal information is due to both the central level of processing and working memory capacity.

Regarding phonological categories, the perception of the consonants is, generally speaking, less categorical and more influenced by the context than adults' perception until 5-6 years-of-age (Walley, 2005). At this age, stops are better recognized than fricatives (Tamashige et al., 2008), whose recognition is still obscured by vocalic transitions. At the same time, vowels' identification is favoured more by their relative durations than by the contextual consonants. According to Walley (2005) all these results are compatible with the hypothesis that the 4-to 5-years-old children are more dependent from a global, syllabic representation than a segmental one (Tamashige et al., 2008; Nittrouer, 1996; Bijeljac-Babic, Bertoncini & Mehler, 1993), and they still need to increase and consolidate their lexicon in order to extract all the relevant phonetic information as adults do (Walley, 2005). In fact, children aged five are less sensitive towards the position of errors within the recently acquired words, than they are towards errors within familiar, earlier-acquired words. Furthermore, the identification of consonants is more disturbed by noise in 5-years old children than in adults, especially for place identification, while, as for sonority, voiced are identified better than voiceless consonants (VOT contrasts are perceived in adults-like manner by 4-to 6 years of age, Tamashige et al. 2008). Regarding the influence of the relative position within the utterance of the consonants to be compared, previous works found that children facing with a non-word discrimination test were found to be more successful for consonants in initial rather than in final position (see McAllister-Byun, 2015). As for manner, nasals, liquids and stops are identified better than fricatives and affricates (Walley, 2005), and the identification is more facilitated if the contextual vowel is [a] rather than [u] or [i]. Generally speaking, the more two consonants share distinctive features, the more they will be confused, especially if they are voiceless, but, according to McAllister-Byun (2015) which compared adults' (Weber, Cutler, 2004) with children's phonological discrimination, the perceived distance between pairs of speech stimuli follows the same trend in the two populations, thus demonstrating that by five years children's discrimination skills are essentially adult-like.

1.3 Distinctive features in infancy

At this point a critical discussion about the concept of "distinctive feature" in relationships to phonological acquisition is needed: the term "distinctive feature" in phonology refers to a particular property of a phone/phoneme; according to the traditional theory, we can imagine the distinctive features as abstract cognitive entities that characterize a certain sound in the mind of the speaker/listener (Chomsky, Halle, 1968). In particular, Cristià, Seidl & Francis (2011) identify two main purposes in using distinctive features:

- distinctive function: they are used to distinguish sounds in contrast with each other (an acoustic difference can lead to a change in meaning, as shown by the minimum pairs, e.g. it. /'pane/ - /'kane/);
- classifying function: they determine the classes of sounds based on common characteristics, which may be subject to the same phonological rule.

It is important to underline that both of the functions described fall within the definition of distinctive features in the adult phonological system, but there is no a priori reason to think that if a child is able to make a distinction between two sounds that we describe as [+FEATURE] and [–FEATURE] (e.g. sound [+ continuous] and sound [–continuous]), then she is also able to group and classify all the sounds belonging to the [+FEATURE] category (e.g. [+continuous]) in opposition to all those with the characteristic [–FEATURE] (e.g. [–continuous]) (Cristià et al., 2011; Menn, Vihman, 2011). This is to say that we need to emphasize the importance of distinguishing the ability to discriminate two sounds (minimum pairs) from the ability to use this contrast in a phonologically relevant way (to learn new sounds), skills that can have different time courses.

Some contrasts are initially difficult to discriminate – for example, /f/ from /θ/ – and errors in the production of these consonants may have their basis in perceptual abilities (Vihman, 1996). Similarly, production errors in older children who have a speech disorder may reflect either motor problems, or an inadequate phonemic representation (Rvachew et al., 2003; Gierut, 1998). At present, the question of how perception and linguistic production are interrelated is still unresolved. According to some hypotheses there would be an integration between the two abilities from the beginning; according to others, instead, they would follow two different development paths, at least at the beginning. However, there is ample evidence that highlights the relationship between production and perception in the child's phonological development, and a large number of studies show that children have a specific difficulty in discriminating the same contrasts that neutralize in their productions (eg, McAllister-Byun, 2012; Vance, Rosen & Coleman, 2009; Whitehill, Francis, & Ching, 2003; Rvachew, Jamieson, 1989; Velleman, 1988; Hoffman, Daniloff, Bengoa & Schuckers, 1985; Locke, 1983).

In recent decades, numerous studies have shown that individual variability in linguistic production is related to individual differences in discrimination and categorical perception of linguistic sounds. According to Perkell, Guenther, Lane, Matthies, Stockmann, Tiede & Zandipour, (2004) adults who exhibited greater sensitivity in discriminating intermediate signals along the continuum /s/-/ʃ/, showed at the same time a greater acoustic contrast in the production of the same consonants. The correlation between perception and production is confirmed by other studies (Newman, 2003; Villacorta, Perkell, & Guenther, 2007; Perkell et al., 2004). The links observed between perceptual acuity and the robustness of contrast in production find a reason in theoretical and computational models such as DIVA - Directions Into Velocities of Articulators (Tourville, Guenther, 2011;

Guenther, 1995). According to this model, speakers who identify a narrower region of auditory space as the target of a certain sound are also more precise in the phonetic realization of that sound in contrast to other phonemes. In line with this model, many studies have shown that children who make mistakes in producing a given contrast also have a lower perception of the same contrast than children who produce it correctly (McAllister-Byun, 2012; Whitehill, Francis & Ching, 2003; Rvachew, Jamieson, 1989; Hoffman et al., 1985; Locke, 1983). In a recent study by Terband, Van Brenk & van Doornik-van der Zee (2014) two groups of Dutch-speaking children were compared: the first was formed by children aged 4 to 8 years with a typical language development, the second by children of the same age with language disorders. The audio recordings of the children's productions of the vowel /e/ in words with CVC structure were played back to them with changes in the height of F1 and F2. The authors observed that children with typical development successfully compensated for changes in both F1 and F2; the children with language disorders, on the other hand, did not compensate in either case, especially for F1, where they even exaggerated the perturbation instead of compensating it. This leads the authors to think that the children of the second group have an adequate auditory-perceptual ability to perceive the perturbation, but lack of the ability to modify their production to compensate for this change. In other words, it seems that children with speech impairment have difficulty integrating auditory feedback with motor planning.

These studies highlight an interesting dissociation. In adults, where the motor and perceptual systems of language are well established, there seems to be a close relationship between the two abilities (perception and production). In children a similar situation is observed, but more complex: in fact, the motor system, as well as the perceptual one, are still in the maturation phase and are developing through experience. During development, it is possible that the child perceives speech at almost the adult level, but does not yet has the motor skills to achieve a certain target (McAllister-Byun, Tiede, 2017). On the other hand, it is possible that the child has adequate motor skills, but still an auditory-perceptual representation too wide of the target, with the consequence of not being able to receive the error feedback that would lead him to modify the motor planning (Shiller, Rochon, 2014). The latter statement suggests that understanding more deeply the relationships between perception and production in the course of development would better clarify the factors that underlie the enormous variability of production capacity observed in children.

Understanding when a dissociation of development between perception and production is likely to occur, and whether one or the other will be a limiting factor, would be useful in the clinical management of language delay/disturbance. However, it is difficult to make good predictions in this regard. In fact, in the perception and in the production of speech the children undergo profound changes in the first stages, followed by a period of time in which the abilities gradually mature and become refined. The development of production compared to perceptual skills can show a variation across the different linguistic targets, speakers and develop-

mental stages (McAllister-Byun & Tiede, 2017). Discrimination tests are usually administered following scripted approaches (as to Italian, see for instance Tressoldi, Vio, Gugliotta, Bisiacchi & Cendron, 2005). These, however, cannot take advantage of information collected during the test itself as a human therapist cannot track the child performance and find the most appropriate stimuli in real time. An Artificial Intelligence provided with fundamental knowledge about language and about the structure of a discrimination test can, instead, continuously track the child's performance and adjust the test dynamically generating new stimuli by estimating their *usefulness*.

2. *System architecture*

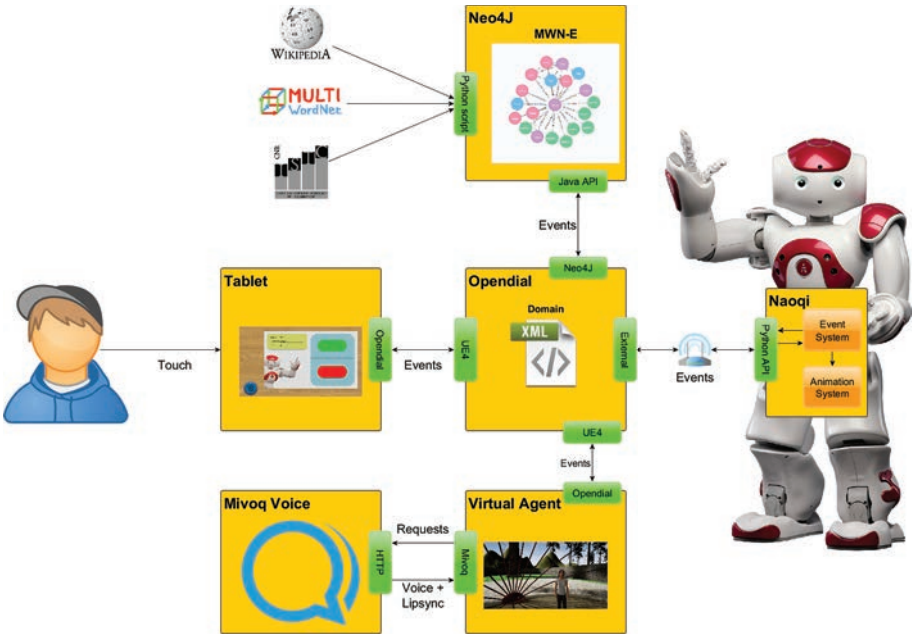
In our work, we represent the discrimination test as a dialogue model where each stimulus, once paired with the child's answer, generates a new stimulus as a system response. This stimulus is selected depending on a utility function taking into account linguistic knowledge and the child's performance. From an architectural point of view, this reflects in a dialogue manager acting as the system's controller and in linguistic knowledge being distributed between the dialogue manager and a database of Italian words. The dialogue manager, implemented using the Opendial framework (Lison, Kennington, 2016) is provided with the capability to establish which kind of information can be obtained by presenting each available stimulus and with a non-words generator which make use of phonotactic rules to avoid structures not belonging to the Italian language. The database contains morpho-syntactic, phonological and frequency data to improve the quality of the selected stimuli. In order to present the discrimination test in a social, gamified, setup, the dialogue manager controls a set of virtual agents with different characteristics. In our case, a virtual avatar is presented on a computer screen and acts as the game's guide while a social robot is used to implement a learning-by-teaching approach. The virtual avatar is controlled using the Unreal Engine 4[1] and its voice is dynamically generated using the Mivoq Voice Synthesis Engine[2], which represents the state of the art of Italian synthesis (Tesser, Sommavilla, Paci & Cosi, 2016). The synthetic voice has a number of advantages: it allows the system to be easily updated as the proposed stimuli are not pre-recorded, it allows the 3D characters to address the child by calling her by name, thus establishing a closer relationship, and it can be adapted to different kinds of characters. In the specific case of Mivoq, personalised voices and specific prosodic styles can also be synthesised, opening to a number of applications for game-like software artefacts. A tablet interface, also controlled using the Unreal Engine 4, is provided to the child to evaluate the proposed stimuli. Since the ability to adequately use a tablet interface appears to be reliable for 5 years old and onwards children (Vatavu, Cramariuc & Schipor, 2015), this is the minimum age recom-

[1] www.unrealengine.com.

[2] www.mivoq.it.

mended to apply this technology. The robot used in our implementation is Nao, which is a well established robotic platform to work with children. An overview of the system is shown in Figure 1. The technical details of the software architecture are presented in Origlia, Cosi, Rodà & Zmarich (2017).

Figure 1 - *System Architecture*



For the task of finding phonological neighbours presenting specific phones in opposition given a syllabic structure, it is possible to exploit the MWN-E database (Origlia, Paci & Cutugno, 2017), implemented as a graph in the Neo4J module (Webber, 2012). Since the phonological transcriptions are included among the properties of words, it is possible to extract phonological neighbours sharing the same syllabic structure to isolate phonological neighbours obtained through a substitution operation. Also, it is possible to specify which phonemes should be involved in the substitution in order to obtain the stimuli needed for the test. Word pairs that include words present in the *Primo Vocabolario del Bambino* (Caselli, Casadio, 1995), the Italian version of the MacArthur-Bates Communicative Developmental Inventories), are given precedence. As an alternative, word pairs with the highest average Wikipedia frequency are selected. The list of consonantal phonemes is shortened for reasons of space. A word pair represents a stimulus in the test. For each possible phoneme opposition, the system checks if a stimulus with the considered syllabic structure exists. If this is the case, the stimulus is a candidate to be presented during the test and its utility is computed, at each turn, using utility functions, computed as described in the next Section. By making the 3D avatar pronounce the target word and the Nao robot pronounce the second word, the

system asks the child to decide if Nao repeated the first word correctly or not[3]. This is the same as asking if the two words are the same or not and it has the additional advantage of putting the child in an advantage position with respect to one of the involved agents, thus letting them playing the role of *teachers*.

3. *Utility functions*

In this section, we summarise the principles of the statistical modelling technique used to dynamically choose the best stimulus, given the subject's observed performance among the ones obtained by querying the MWN-E database. First of all, on the basis of the study presented in Zmarich, Bonifacio (2005), we consider the acquisition age of each phoneme. For the sake of simplicity, in this version of the model we assume that the complexity of the phones substitution is the same whether it consists of substituting a phoneme acquired later with a phoneme acquire earlier or vice-versa. For our experiments, we refer to Schmid (1999) for the distinctive features of standard Italian (see Table 1 for the reader's convenience).

Table 1 - *Distinctive features for standard Italian (Schmid, 1999, translated)*

	p	b	t	d	k	g	f	v	s	z	ʃ	ts	dz	tʃ	dʒ	m	n	ɲ	l	ʎ	r	j	w
Consonantal	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-
Sonorant	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+
Continuous	-	-	-	-	-	-	+	+	+	+	+	-	-	-	-	-	-	-	+	+	+	+	+
Del. release	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-
Voiced	-	+	-	+	-	+	-	+	-	+	-	-	+	-	+	+	+	+	+	+	+	+	+
Nasal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-
Lateral	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-
Coronal	-	-	+	+	-	-	-	-	+	+	+	+	+	+	+	-	+	-	+	-	+	-	-
Anterior	+	+	+	+	-	-	+	+	+	+	-	+	+	-	-	+	+	-	+	-	+	-	-
Posterior	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+

In this work, we do not consider the Consonantal feature in our experiments in order to concentrate on more subtle oppositions. On the other hand, we introduce the *length* feature in order to allow the system to distinguish between Italian words that are only differentiated by the phonetic realisation of a phonological geminate, as in palla /'palla/(*ball*) versus pala /'pala/ (*shovel*). With this decision we don't want to take a theoretical position about the phonological status of Italian geminates (Bertinetto, 1981): this choice is dictated by the SAMPA representation of Italian words that is provided by the pronunciations database. The probability of a subject to assign a label to the presented opposition is a binomial distribution (Equal/Different). Therefore, to represent a priori probabilities built using previous feedback, the conjugate prior of the binomial distribution, the Beta distribution, is used. Following the Opendial implementation, a two dimensional Dirichlet

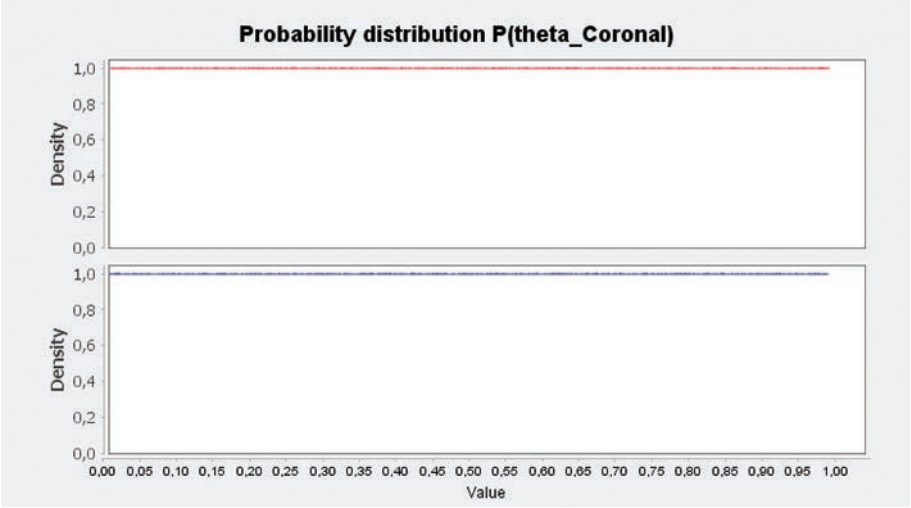
[3] The presented experiments also highlighted technological and theoretical problems due to the highly pioneering nature of the proposed system. First of all, the voice synthesizers have shown, in rare cases, problems in providing non-words. The performance of the Mivoq engine seemed superior to the one provided by the synthesizer built in the Nao robot so this problem has been addressed by integrating the Mivoq engine in the robotic platform.

probability density function with parameters is used to model the conjugate prior. An entropy-based utility function for a given opposition is computed, assigning higher utility values to stimuli presenting the opposing features for which the associated probability density functions present the higher uncertainty. An opposition presenting more than one highly entropic feature is not an optimal choice as it is not possible to evaluate which feature influenced the outcome. This is the reason why, for scripted tests, it is not possible to use phoneme pairs opposing more than one feature, which becomes a problem in tests opposing words as there may not be phonological neighbours with the specified structure opposing exactly the phonemes involved in the feature of interest. In a dynamically generated test, however, the estimated incapability of the subject to perceive oppositions on a given feature may allow to investigate other features that are never opposed in isolation, as in the case of the *posterior* feature always being opposed with the *anterior* feature. For this reason, we compute a utility function based on the mean probability value that each feature has not to be discriminated. This function assigns a higher utility value to oppositions presenting a single, highly entropic, feature together with features that have been found not to be discriminated by the subject. The higher the likelihood of other features not to be discriminated, the higher the utility. Since the task complexity can be influenced by the age acquisition difference in the involved phonemes, we model a substitution-based utility function assigning a higher value to phoneme oppositions that are closer to each other in the acquisition sequence. As this is a relative measure of phoneme-based complexity for the opposition, we also need an absolute measure to prefer phonemes acquired earlier. We therefore define an acquisition-based utility function. Since all utility functions are different measures of the same object (the phoneme opposition) sharing the same range (0, 1), the final utility function for the opposition is computed as the harmonic mean of these four measures. We use the harmonic mean because, in the considered case, different measures sharing the same range (0-1) are performed on the same subject. In this situation, the harmonic mean is the averaging method to be used. This function lets the dialogue manager select the optimal stimulus for the next turn. The algorithm for dialogue management, implemented in Opendial and exploiting the MWN-E data, proposes a stimulus at each step and updates the probability distributions according to the feedback given by the subject using Bayesian inference.

4. *Example*

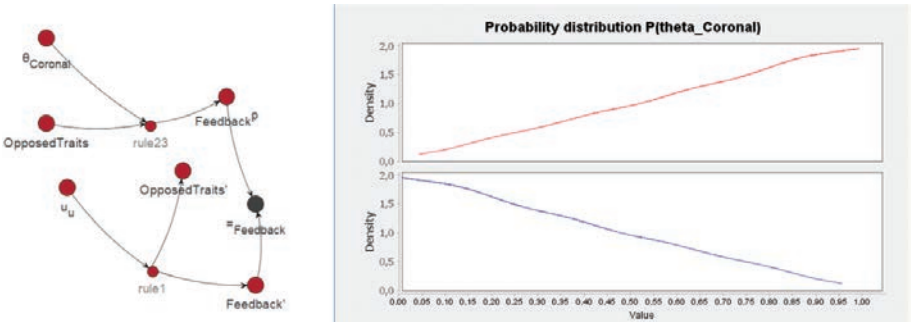
As an exemplification of the way utility functions are included in the dialogue manager, consider the case of an opposition involving the *coronal* feature that is being proposed for the first time. The Bayesian network in Opendial predicts that the probability of the child detecting or missing the opposition, given current information, is maximally entropic as no data are present. In this situation, the probability distribution is uniform on both dimensions and the probability is 50% for both cases. This situation is summarised in Figure 2.

Figure 2 - *The initial distribution of the coronal trait is uniform and maximally entropic for the probability of the child missing/detecting the opposition*



After presenting the stimulus and comparing the actual feedback from the child and the prediction, the dialogue manager uses Bayesian inference to update the probability distributions involved in the previous prediction. In this case, the probability distribution associated with the *coronal* feature is updated: as the child answers correctly, high probabilities are assigned with the first dimension of the distribution while the second is updated symmetrically. This leads to a less entropic distribution, as shown in Figure 3.

Figure 3 - *The Bayesian network compares the actual answer with the previous prediction and updates the appropriate distributions to improve future predictions*



The same strategy used for the considered features is also used to collect feedback on control stimuli, consisting of presenting two times the same stimulus or presenting two very different stimuli. A t-test on the final distributions is used to validate or reject the test.

5. *Experiments*

To investigate the validity of the proposed approach, we performed a pilot test by recruiting a group of 5-years old children (3 males, 2 females) to participate in a series of test sessions administered in different days and masked behind the game narrative. For each recruited child, we programmed six sessions distributed over three weeks. In order to establish a baseline for the prototype system, a group of Italian native speaking children with no reported speech and hearing and/or cognitive problems were recruited. As a reference for the children's capabilities, an entry test consisting of a) the non-word phonological discrimination subtest of BVN 5-12 (Tressoldi et al., 2005) and the b) word phonological discrimination and c) word and d) non-word repetition subtests of BVL 4-12 (Marini, Marotta, Bulgheroni & Fabbro, 2015) was also administered.

The first session (Intro) lasted approximately 6 minutes and served as an introduction to let the children familiarise with the experimental setting and with the narrative situation. The experimenter described the problem of the Nao robot, learning to speak, and introduced the virtual character, Ellie, as Nao's teacher and friend. After administering the first battery of standard tests, (BVN), the child was guided through a tutorial session to demonstrate the use of the tablet interface. During the tutorial, Nao performs a small set of funny behaviours following requests from the virtual character and the child was asked to evaluate Nao's performance using the tablet interface. At this stage, in accordance with the narrative, Nao only communicates using a set of non-verbal digital sounds. After the tutorial session, the second battery of standard tests (BVL) is administered and the child is asked to give his consent to continue helping Nao to solve its problem in the following sessions. The second session (NW1) lasts approximately 10 minutes and consists of discrimination and repetition tasks using non-words. Concerning the second phase, the system is not able, at present time, to have Nao repeat the word stated by the child and has to follow the same strategy used for the discrimination test. Currently, the system collects the audio recording to be analysed subsequently by the expert but otherwise keeps using the strategy used during the discrimination test. Future work will consist of integrating a Speech Recognition Engine specialised on children voice to address this current limitation (Cosi, Paci, Sommavilla & Tesser, 2015). The starting situation consists of assigning all features an a-priori probability corresponding to the uniform distribution. This corresponds to an initial situation in which entropy is maximised. The dialogue manager selects the most appropriate stimulus using the utility function described in Section 5.2 and coordinates the two agents so that one presents the first non-word and the second presents the second. The child's feedback, collected through the tablet interface, is used to update the statistical model and select the next stimulus. The discrimination session lasts 3 minutes and is interrupted by a cutscene in which Nao acts sadly and the virtual character explains that it is representing discouragement. This has the effect of providing emotional reactions to Nao so that the child can more easily relate. It also allows the child to release his attention from the task. Nao is ready to start the repetition test after the

child, having been instructed by the virtual character, caressed Nao on the head. The repetition test lasts as much as the discrimination test and makes use of Nao's voice activity detection (VAD) system to establish when the child has repeated the non-word pronounced by the virtual character. The system has been set to a slightly higher sensitivity to stimulate the child to repeat the stimuli loud and clear. The NW1 session ends with Nao starting to produce vocalisations together with digital sounds before going to sleep and the virtual character highlights the change. In the proposed approach, Nao's evolution represents the reward for the child's effort as no feedback can be provided during the tests as part of the protocol. The third session (NW2) is identical to the first one. At the end of it, Nao stops producing digital sounds and uses vocalisations only. The virtual character informs the child that, starting from the next sessions, Nao will start to learn real words. The fourth session (W1) is identical to the first two sessions but it makes use of real words instead of non-words. The cutscene is also different, as in this case, Nao will stand up and assume an opposing pose, with its head looking away from the child. The virtual character explains that Nao does not want to study anymore and it has to be scolded. Once again, the VAD system is used to detect the child's voice and have Nao get back to work. At the end of this session, Nao starts mixing real words with non-words. The fifth session (W2) has the dialogue manager initialised with the statistical model obtained at the end of the W1 session, in order to check if the collected information improves by providing more stimuli or if one session is sufficient. The control distribution is the only one reset in this session. The cutscene is also slightly altered as this time Nao will protest to the child scolding him and will need to be scolded again before apologising (by saying sorry instead of a non-word) and going back to work. At the end of this session, Nao only uses isolated real words. The last session (Story) was designed to let us check to what extent the child's attention can be retained by the system and if the same architecture can be used to implement discrimination tests with reaction times. In this session, the virtual character reads, phrase by phrase, the Little Red Riding Hood story and Nao repeats the phrases while trying to introduce funny errors. The AI uses the MWN-E database to substitute words with phonological neighbours having the same grammatical role of the word being substituted. In such cases, for example, *Cappuccetto Rosso* (Little Red Riding Hood) becomes *Cappuccetto Rotto* (Little Broken Riding Hood) and *Nonna debole e malata* (Old and sick granny) becomes *Nonna debole e salata* (Old and salty granny). The child clicks a red button on the tablet increasing a counter to mark Nao's errors. The duration of this session is doubled with respect to the preceding ones in order to check whether it is possible, for the system, to keep the child engaged in a repetitive task for a prolonged period of time. At the end of this session, Nao becomes able to speak correctly and thanks the child for the help while the virtual character congratulates her for completing the task. The child is then administered a set of questions inspired by the USE questionnaire (Lund, 2001) to collect a subjective impression for the experience. The original 7-points Likert scale was substituted with a 3-points scale (Yes, No, So-so) in order to simplify the task for the children. At the end of the recording period,

more than 4 hours of video recordings were collected for the subsequent analyses. The sessions logs were also saved to allow offline analyses. In order to objectively evaluate interest and emotional feedback, the collected video material was manually annotated by two human judges using the ELAN software (Wittenburg, Brugman, Russel, Klassmann & Sloetjes, 2006). The judges annotated data on two tiers: on the first one, they marked the children's gaze targets (Ellie, Experimenter, Nao, Tablet) while, on the second one, they marked positive and negative emotional expressions. The annotation directives for the first tier were to mark the frame containing the fixation instant of an object belonging to the experimental setup as the starting instant and the frame preceding the one where the gaze leaves the object as the ending time. Transitions and gazes that were directed away from the experimental setup were automatically marked as Other. Given the strict directives for the first tier, the two annotators produced practically identical results and it was not necessary to merge the two annotations. The annotation directives for the second tier were more subjective. The judges were asked to mark positive and negative expressions. As in this case the annotations were influenced by subjective judgement, the final segments considered for the analysis of the results are obtained by considering the annotations overlaps only. Due to video files corruption, the F1/Intro, M2/Story and M3/W1 sessions could not be analysed for the objective evaluation.

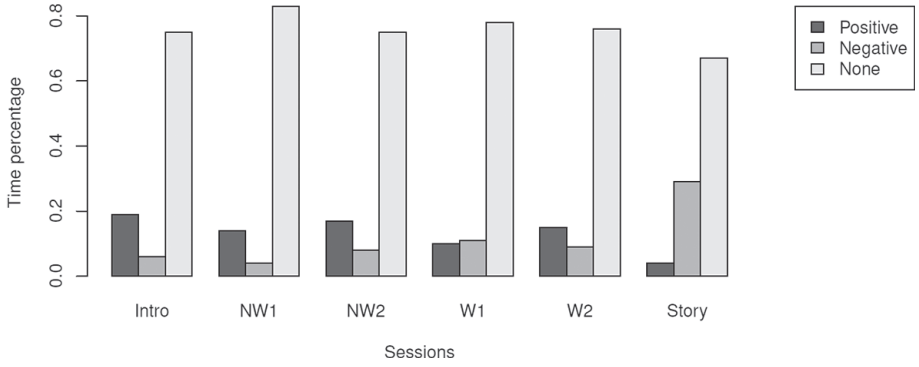
5.1 User experience

From the manual annotations of eye gaze targets, we obtain an estimate of the degree of attention children gave to the system's actors. From the general view presented in Figure 4, we observe that the Nao robot receives high attention during most of the sessions, particularly during the Intro session, highlighting the novelty effect. Children were looking at an element of the experimental setup (Ellie, Nao, Tablet) 64% of the recording time. The experimenter was not often looked at, indicating that the children had limited need to obtain support during the test and were engaged in performing the given tasks.

Figure 4 - *Overall gaze distribution during the NW1, NW2, W1 and W2 sessions*

Concerning the display of emotional feedback from the children, obtained results are shown in Figure 5. The amount of positive feedback is generally higher than the amount of negative feedback.

Figure 5 - *Emotional displays during the different sessions*



Lastly, we consider the scores collected by administering the modified USE questionnaire to the children. Given the limited sample and the reduced size of the scale, we consider the median values of the scores, represented in Figure 6. The children appear to have perceived the task as difficult and not so easy to learn but they unanimously considered it to be pleasant and fun. This is consistent with the goal of task gamification: no benefits are to be expected from the point of view of the perceived difficulty of the task, which is cognitively challenging for 5 years old children, but a good disposition of the subjects towards doing it was observed in this pilot study.

Figure 6 - *Median values for the subjective scores*

5.2 Linguistic report

Examples of the final reports obtained by the proposed system are presented in Table 2 (M1/W2) and in Table 3 (F2/W2). The M1 subject was borderline in the entry tests and the system appears to correctly detect the apparent difficulties on

this subject. The F2 subject, on the contrary, scored highest in the entry test and the system also identifies her as the best performing subject. The only problem detected on F2 is on the Voiced feature. Given the age of the subject, however, this is consistent with her expected capabilities, thus supporting a positive view of the feedback provided by the proposed system.

Table 2 - *Example summary (M1/W2)*

Feature	$\mu(Ok)$	$\mu(Wrong)$	σ	N	Evaluation
Control	0,865	0,135	0,079	19	Reliable
Sonorant	0,672	0,328	0,245	2	Weak Ok
Continuous	0,418	0,582	0,123	13	Strong problem
Delayed Release	0,273	0,727	0,196	4	Strong problem
Voiced	0,265	0,735	0,166	8	Strong problem
Nasal	0,500	0,500	0,289	0	Unknown
Lateral	0,702	0,298	0,236	2	Weak Ok
Coronal	0,218	0,782	0,140	15	Strong problem
Anterior	0,288	0,712	0,118	17	Strong problem
Posterior	0,626	0,374	0,267	6	Weak Ok
Length	0,227	0,773	0,111	12	Strong problem

A potential theoretical issue coming with the use of distinctive features can be observed in the two Tables: in the model we adopted for our experiments, the considered features are not necessarily opposed in a single pair of phonemes. As a consequence, the AI may not be able to test some features if an opposition on the other features they are opposed with is perceived by the considered subject. As an example, the *posterior* feature is never opposed in isolation in any possible pair of phonemes and, in the best case, a pair presenting an opposition on the *posterior* feature also opposes the *anterior* feature.

Table 3 - *Example summary (F2/W2)*

Feature	$\mu(Ok)$	$\mu(Wrong)$	σ	N	Evaluation
Control	0,735	0,265	0,101	19	Reliable
Sonorant	0,604	0,396	0,266	1	Unknown
Continuous	0,642	0,358	0,142	11	Strong Ok
Delayed Release	0,698	0,302	0,195	3	Weak Ok
Voiced	0,302	0,698	0,113	13	Strong problem
Nasal	0,500	0,500	0,289	0	Unknown
Lateral	0,500	0,500	0,289	0	Unknown
Coronal	0,776	0,224	0,138	7	Strong Ok
Anterior	0,658	0,342	0,142	9	Strong Ok
Posterior	0,500	0,500	0,289	0	Unknown
Length	0,791	0,209	0,166	3	Strong Ok

The *anterior* feature, on the other hand, can be tested in isolation (i.e. by opposing the /s/ and /S/ phones, SAMPA coding). Given the stimuli choice model, the AI therefore tests the *anterior* feature and, if the child answers correctly, estimates that it is not useful to attempt to test the *posterior* feature (i.e. by opposing the /p/ and

/k/ phonemes) as a correct answer by the child can be attributed to the opposition on the *anterior* feature. For this reason, the system checked the *posterior* feature when examining the M1 child after observing that the probability of this particular child to provide a correct answer on a stimulus opposing the *anterior* feature was low. For the F2 child, an opposition on the *anterior* feature was found to be perceived by the child, so the system did not propose stimuli involving the *posterior* feature, which is maximally uncertain (Unknown) in the report. Another problem is represented by the *lateral* feature, which was often not investigated by the system although a pair presenting an opposition on this feature in isolation exists (/l/ and /r/). This is because the system appears to overestimate the importance of the acquisition age in the utility computation (/l/ and /r/ are very distant from each other in the phonological acquisition natural history.

6. *Conclusions*

We have presented a technological system designed to administer discrimination tests to evaluate the linguistic competence of young children using a gamified setup. The system dynamically adapts the test to the children's performance using a Bayesian dialogue manager that combines linguistic knowledge with utility functions to iteratively select the most informative stimuli to be presented. Our experiments, although limited to a small sample, indicate that the chosen gamification style is able to keep the children engaged over multiple sessions distributed in a time window of three weeks, when the novelty effect introduced by the Nao robot, in particular, has worn off. The linguistic competence report obtained at the end of the administered sessions provides a detailed view of the test results, as opposed to standard tests, which provide a more general view. While the clinical validity of the approach cannot be stated at present, the ranking obtained by considering the system's report is compatible with the one obtained by administering standard tests. We consider this result to be very encouraging for our future work. The presented experiments also highlighted technological and theoretical problems due to the highly experimental nature of the proposed system. Once the problems highlighted in this first test will be solved, a larger sample of children with no reported linguistic impairments will be recruited to confirm the indications we obtained. Also, the effectiveness of the system to detect existing problems will be tested by considering children affected from dyslexia (Joanisse, Manis, Keating & Seidenberg, 2000) and/or Phonological Disorder (Brancalioni et al., 2012). The same system can also be extended to support speech treatment.

Acknowledgements

Antonio Origlia's work is supported by Veneto Region and European Social Fund (grant C92C16000250006).

Bibliography

Bertinetto, P.M. (1981). *Strutture prosodiche dell'italiano.* Accademia della Crusca.

Bijeljac-babic, R, Bertoncini, J. & Mehler, J. (1993). How do 4-day-old Infants Categorize Multisyllabic Utterances?. *Developmental Psychology,* 29(4), 711-721.

Brancalioni, A.R., Bertagnolli, A.P., Bonini, J.B., Gubiani, M.B., Keske-Soares, M. (2012). The Relation between Auditory Discrimination and Phonological Disorder. *Jornal da Sociedade Brasileira de Fonoaudiologia* 24(2), 157-161.

Carroll, J.M., Snowling, M.J., Stevenson, J. & Hulme, C. (2003). The Development of Phonological Awareness in Preschool Children. *Developmental Psychology, 39*(5), 913.

Caselli, M.C., Casadio, P. (1995). *Il primo vocabolario del bambino.* Milano: Franco Angeli.

Choi, D., Black, A.K. & Werker, J.F. (2018). Cascading and Multisensory Influences on Speech Perception Development. *Mind, Brain, and Education*, 12, 212-223.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. Harper & Row.

Cole, R.A. (2003). Roadmaps, Journeys and Destinations Speculations on the Future of Speech Technology Research, in *Eighth European Conference on Speech Communication and Technology, Eurospeech 2003,* 2905-2908.

Cosi, P., Delmonte, R., Biscetti, S., Cole, R., Pellom, B. & Van Vuren, S. (2004). Italian Literacy Tutor Tools and Technologies for Individuals with Cognitive Disabilities. In *Proceedings Of Instil/Icall2004-NLP And Speech Technologies In Advanced Language Learning Systems-Venice* (Vol. 17), 19-27.

Cosi, P., Paci, G., Sommavilla, G., Tesser, F. (2015) Kaldi: Yet another ASR toolkit? Experiments on Adult and Children Italian Speech. In Vayra, M., Avesani C. & Tamburini F. (Eds.), *Studi Associazione Italiana Scienze della Voce* (Vol. 1), 429-438.

Cristià, A., Seidl, A. & Francis, A.L. (2011). Phonological Features in Infancy, in G.N. Clements & R. Ridouane (Eds.), *Where do Phonological Contrasts come from? Cognitive, Physical and Developmental Bases of Phonological Features,* John Benjamins Publishing Company, 303-326.

Donald, M. (2001). *A mind so rare: The evolution of human consciousness.* WW Norton & Company.

Eckerdal, P., Merker, B. (2009). Music and the 'Action Song' in Infant Development: An Interpretation. In Malloch S., Trevarthen C. (Eds.), *Communicative Musicality: Exploring the Basis of Human Companionship*, Oxford University Press, 241-262.

Edwards, J., Fox, R.A. & Rogers, C.L. (2002). Final Consonant Discrimination in Children: Effects of Phonological Disorder, Vocabulary Size, and Articulatory Accuracy. *Journal of Speech, Language, and Hearing Research,* 45(2), 231-242.

Freitas, C.R., Mezzomo, C.L. & Vidor, D.C.G.M. (2015). Phonemic Discrimination and the Relationship with Other Linguistic Levels in Children with Typical Phonological Development and Phonological Disorder. In *CoDAS 2015,* 27(3), 236-241.

Gierut, J.A. (1998). Treatment Efficacy: Functional Phonological Disorders in Children. *Journal of Speech, Language, and Hearing Research,* 41, 85-100.

Gratier, M., Trevarthen, C. (2008). Musical Narrative and Motives for Culture in Mother-Infant Vocal Interaction. *Journal of Consciousness Studies,* 15(10), 122.

Guenther, F.H. (1995). Speech Sound Acquisition, Coarticulation, and Rate Effects in a Neural Network Model of Speech Production. *Psychological Review,* 102(3), 594-621.

Halliday, M. (1975). *Learning how to mean*. London: Edwin Arnold.

Hazan, V., Barrett, S. (2000). The Development of Phonemic Categorization in Children Aged 6-12. *Journal of Phonetics,* 28(4), 377-396.

Hoffman, P.R., Daniloff, R.G., Bengoa, D. & Schuckers, G.H. (1985). Misarticulating and Normally Articulating Children's Identification and Discrimination of Synthetic [r] and [w]. *Journal of Speech and Hearing Disorders,* 50(1), 46-53.

Jensen, J.K., Neff, D.L. (1993). Development of Basic Auditory Discrimination in Preschool Children. *Psychological Science, 4*(2), 104-107.

Joanisse, M.F., Manis, F.R., Keating, P. & Seidenberg, M.S. (2000). Language Deficits in Dyslexic Children: Speech Perception, Phonology, and Morphology. *Journal Of Experimental Child Psychology 77*(1), 30-60.

Kuhl, P.K., Conboy, B.T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M. & Nelson, T. (2008). Phonetic Learning as a Pathway to Language: New Data and Native Language Magnet Theory Expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences,* 363(1493), 979-1000.

Law, J., Boyle, J., Harris, F., Harkness, A. & Nye, C. (2000). Prevalence and Natural History of Primary Speech and Language Delay: Findings from a Systematic Review of the Literature. *International Journal of Language and Communication Disorders, 35*, 165-188

Lison, P., Kennington, C. (2016). Opendial: A Toolkit for Developing Spoken Dialogue Systems with Probabilistic Rules, *Proceedings of ACL-2016 System denonstrations*, 67-72.

Locke, J.L. (1983). Clinical Phonology: The Explanation and Treatment of Speech Sound Disorders. *Journal of Speech and Hearing Disorders,* 48, 339-341.

Lund, A.M. (2001). Measuring usability with the USE questionnaire*, Usability Interface* 8(2), 3-6.

Marini, A., Marotta, L., Bulgheroni, S. & Fabbro, F. (2015). BVL_4-12. *Batteria per la valutazione del linguaggio in bambini dai 4 ai 12 anni*. Firenze: Giunti OS.

MacAllister - Byun, T. (2012). Bidirectional Perception-Production relations in Phonological Development: Evidence from Positional Neutralization. *Clinical Linguistics & Phonetics*, 26(5), 397-413.

Mcallister - Byun, T. (2015). Perceptual Discrimination across Contexts and Contrasts in Preschool-Aged Children. *Lingua 160,* 38-53

Mcallister - Byun, T., Tiede, M. (2017). Perception-Production Relations in Later Development of American English Rhotics. *PloS one, 12*(2).

Meltzoff, A.N., Kuhl, P.K., Movellan, J. & Sejnowski, T.J. (2009). Foundations for a New Science of Learning. *Science, 325* (5938), 284-288.

Menn, L. & Vihman, M. (2011). Features in Child Phonology: Inherent, Emergent or Artefacts of Analysis? In Clements G.N. & Ridouane R. (Eds.), *Where do Phonological Contrasts Come From? Cognitive, Physical And Developmental Bases of Phonological Features.* John Benjamins Publishing Company, 261-301.

Newman, R.S. (2003). Using Links between Speech Perception and Speech Production to Evaluate Different Acoustic Metrics: A Preliminary Report. *The Journal of the Acoustical Society of America*, 113(5), 2850-2860.

Nithart, C., Demont, E., Majerus, S., Leybaert, J., Poncelet, M., Metz-Lutz, M. (2009). Reading Disabilities in SLI and Dyslexia Result from Distinct Phonological Impairments. *Developmental Neuropsychology*, *34*(3), 296-311

Nittrouer, S. (1992). Age-Related Differences in Perceptual Effects of Formant Transitions within Syllables and Across Syllable Boundaries. *Journal of Phonetics.*

Origlia, A., Cosi, P., Rodà, A. & Zmarich, C. (2017). A Dialogue-Based Software Architecture For Gamified Discrimination Tests. In *Proc. of the 1st Workshop on Games-Human Interaction (online)*.

Origlia, A., Paci, G. & Cutugno, F. (2017). MWN-E: A Graph Database to Merge Morpho-Syntactic and Phonological Data For Italian. In *Proc. of Subsidia 2017*.

Perkell, J.S., Guenther, F.H., Lane, H., Matthies, M.L., Stockmann, E., Tiede, M. & Zandipour, M. (2004). The Distinctness of Speakers' Productions of Vowel Contrasts is Related to Their Discrimination of the Contrasts. *The Journal of the Acoustical Society of America,* 116(4), 2338-2344.

Polka, L., Jusczyk, P.W., Rvachew, S. (1995). Methods for Studying Speech Perception in Infants and Children. In Strange, W. (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, New York: Press Timonium, 49-89.

Reddy, V. (2008). *How infants know minds.* Harvard University Press.

Rvachew, S. E, Jamieson, D.G. (1989), Perception of Voiceless Fricatives by Children with a Functional Articulation Disorder. *Journal of Speech and Hearing Disorders,* 54, 193-208.

Rvachew, S., Ohberg, A., Grawburg, M. & Heyding, J. (2003). Phonological Awareness and Phonemic Perception in 4-year-old Children with Delayed Expressive Phonology Skills. *American Journal of Speech-Language Pathology, 12,* 463-471.

Saffran, J.R., Werker, J.F. & Werner, L.A. (2006). The Infant's Auditory World: Hearing, Speech, and the Beginnings of Language. In Kuhn, D., Siegler, R. (Eds.), *Handbook of Child Psychology,* 1308-1315.

Schmid, S. (1999). *Fonetica e fonologia dell'italiano.* Paravia scriptorium.

Shiller, D.M. & Rochon, M.L. (2014). Auditory-perceptual learning improves speech motor adaptation in children. *Journal of Experimental Psychology: Human Perception and Performance, 40*(4), 1308.

Tallal, P. (1976). Rapid Auditory Processing in Normal and Disordered Language Development. *Journal of Speech and Hearing Research,* 19, 561-571.

Tamashige, E., Nishizawa, N., Itoda, H., Kasai S., Igawa, H.H. & Fukuda S. (2009). Development of Phonemic Distinction in Japanese Preschool Children, *Folia Phoniatrica et Logopaedica,* 60, 318-322

Terband, H., Van Brenk, F. & Van Doornik-Van Der Zee, A. (2014). Auditory Feedback Perturbation in Children with Developmental Speech Sound Disorders. J*ournal of Communication Disorders,* 51, 64-77.

Tesser, F., Sommavilla, G., Paci, G. & Cosi, P. (2016). Automatic Creation of TTS Intelligibility Tests, in Savy R., Alfano I. (Eds.), *La fonetica sperimentale nell'insegnamento e nell'apprendimento delle lingue straniere*, Studi AISV, Vol. 2, 289-303.

Tourville, J.A., Guenther, F.H. (2011). The DIVA Model: A Neural Theory of Speech Acquisition and Production. *Language and Cognitive Processes,* 26(7), 952-981.

Tressoldi, P., Vio, M., Gugliotta, M., Bisiacchi, P. & Cendron, M. (2005). BVN 5-11. *Batteria di valutazione neuropsicologica per l'età evolutiva*. Trento: Erickson.

Trevarthen, C. (1978). Secondary Intersubjectivity: Confidence, Confiding and Acts of Meaning in the First Year. *Action, Gesture, And Symbol: The Emergence of Language.*

Trevarthen, C. (1979). Communication and Cooperation in Early Infancy: A Description of Primary Intersubjectivity. *Before Speech: The Beginning of Interpersonal Communication,* 321-347.

Trevarthen, C. (2009). The functions of emotion in infancy. In: Fosha D., Siegel, S.M.F. & Solomon, M. (Eds.), *The healing power of emotion: Affective neuroscience, development & clinical practice* (Norton Series on Interpersonal Neurobiology), WW Norton & Company, 55-85.

Trevarthen, C., Aitken, K.J. (2001). Infant intersubjectivity: Research, theory, and clinical applications *Journal of child psychology and psychiatry 42*(1), 3-48.

Trevarthen, C., Aitken, K. (2003). Regulation of Brain Development and Age-Related Changes in Infants, Motives: The Developmental Function of "Regressive" Periods. In M. Heimann (Ed.), *Regression Periods in Human Infancy.* Mahwah, NJ: Erlbaum, 107-184.

Vance, M., Rosen, S. & Coleman, M. (2009). Assessing Speech Perception in Young Children and Relationships with Language Skills. I*nternational Journal of Audiology,* 48(10), 708-717.

Vatavu, R.D., Cramariuc, G. & Schipor, D.M. (2015). Touch Interaction for Children Aged 3 to 6 Years: Experimental Findings and Relationship to Motor Skills. *International Journal of Human-Computer Studies*, 74*,* 54-76.

Velleman, S.L. (1988). The Role Of Linguistic Perception in Later Phonological Development. *Applied Psycholinguistics*, *9*(3), 221-236.

Vihman, M.M. (1996). *Phonological Development: The Origins of Language in the Child.* Blackwell Publishing.

Villacorta, V.M., Perkell, J.S. & Guenther, F.H. (2007). Sensorimotor Adaptation to Feedback Perturbations of Vowel Acoustics and Its Relation to Perception. *The Journal of the Acoustical Society of America*, *122*(4), 2306-2319.

Walley, A.C. (2005). Speech Perception in Childhood. In Pisoni D.B., Remez R.E. (Eds.), *The Handbook of Speech Perception.* Blackwell publishing, UK, 449-468.

Webber, J. (2012). A programmatic introduction to neo4j. In *PROCEEDINGS OF THE 3RD ANNUAL CONFERENCE ON SYSTEMS, PROGRAMMING, AND APPLICATIONS: software for humanity*. ACM, 217-218.

Weber, A., Cutler, A. (2004). Lexical Competition in Non-Native Spoken-Word Recognition. *J. Mem. Lang. 50*, 1-25.

Weismer, S.E., Tomblin, J.B., Zhang, X., Buckwalter, P., Chynoweth, J.G. & Jones, M. (2000). Nonword Repetition Performance in School-Age Children with and

without Language Impairment. *Journal of Speech, Language, and Hearing Research,* 43(4), 865-878.

Whitehill, T.L., Francis, A.L. & Ching, C.K. (2003). Perception of Place of Articulation by Children with Cleft Palate and Posterior Placement. *Journal of Speech, Language, and Hearing Research,* 46(2), 451-461.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. & Sloetjes, H. (2006). ELAN: A Professional Framework for Multimodality Research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, 1556-1559.

Zmarich, C., Bonifacio, S. (2005). Phonetic Inventories in Italian Children Aged 18-27 Months: A Longitudinal Study. In *Proceedings of Interspeech*, Lisboa, 757-760.

DOI: 10.17469/O2104AISV000012

CECILIA DI NARDI, ROSANNA TURRISI, ALBERTO INUGGI, NILO RIVA, ILARIA MAURI, LEONARDO BADINO

An automatic speech recognition Android app for ALS patients

This paper describes AllSpeak, an Automatic Speech Recognition (ASR) Android Application developed for Italian-speaking patients with Amyotrophic Lateral Sclerosis (ALS). It allows to recognize a predefined and customizable set of basic utterances that are used by the patient in everyday life (e.g., "I'm thirsty", "I feel pain", etc...). The ASR engine is based on deep learning architectures and it uses a simple decoding strategy to allow offline (i.e., w/o any network connection) and fast decoding. Although deep learning approaches have achieved outstanding results on different speech recognition tasks, recognition of impaired speech is still quite challenging for an ASR system mainly due to a scarce availability of training data and a large variability of impairments. We have addressed these two problems by limiting recognition to a set of key phrases/words corresponding to the patient's primary needs and by strongly adapting the neural networks to the target speaker's voice. Results show that the type of network architecture and the training strategy have both a very significant impact on recognition accuracy of dysarthric speech. Although different architectures and training strategies perform similarly on healthy speakers, recurrent neural networks trained in sequence-to-sequence fashion significantly outperform any other method on most of ALS speakers.

Keywords: automatic speech recognition, amyotrophic lateral sclerosis, smartphone application, deep neural networks.

1. *Introduction*

Amyotrophic lateral sclerosis (ALS) is characterized by progressive muscle paralysis caused by degeneration of motor neurons in the primary motor cortex, corticospinal tracts, brainstem, and spinal cord (Van Es et al., 2017). ALS is relentlessly progressive – 50% of patients die within 30 months of symptom onset and about 20% of patients survive between 5 years and 10 years after symptom onset (Talbot, 2009). Modern technology has allowed people with ALS to compensate to some degree for almost every loss of function, making it possible even for those with almost no muscle function to continue to breathe, communicate, eat, travel and use a computer. In particular, for many people with ALS, the speaking ability may be lost as weakness increases in the muscles of the mouth and throat that control speech and in the muscles that help generate the pressure that moves air over the vocal folds. Dysarthria is indeed the presenting symptom in 30% of patients with ALS and is found in >80% of patients (Hardiman, 2017) and this loss of communication

prevents patients from participating in many activities and may lead to social isolation, reducing the quality of life (QoL).

The goal of management of dysarthria in ALS patients is to optimize communication effectiveness for as long as possible. Speech therapy can delay the progression of dysarthria, and augmentative and alternative communication techniques are the treatments of choice and can enhance QoL in the most advanced phases of ALS. Nevertheless, although there have been several attempts to improve speech recognition for dysarthric speakers as communication techniques based on brain–computer interfaces, these efforts have not until recently converged and their use in the clinical setting is still limited. Moreover, modern automatic speech recognition (ASR) is ineffective at understanding relatively unintelligible speech caused by dysarthria and traditional representations in ASR such as Hidden Markov models (HMMs) trained for speaker independence that achieve 84.8% word-level accuracy for non-dysarthric speakers might achieve less than 4.5% accuracy given severely dysarthric speech on short sentences (Rudzicz, 2010a; Rudzicz, 2010b; Rudzicz, 2012). Recently, more accurate dysarthric speech recognition system has been developed by using deep learning based approaches (España-Bonet, Fonollosa, 2016; Joy, Umesh, Abraham, 2017; Vachhani, Bhat, Das & Kopparapu, 2017). However, in case of severe disability, the ASR performance still remains poor. Causes of poor performance may include slurred speech, weak or imprecise articulatory contacts, weak respiratory support, low volume, incoordination of the respiratory stream, hypernasality, and reduced intelligibility (Kim, Kent & Weismer, 2011). Additionally, dysarthric speech is not sufficiently covered in the training datasets of state-of-the-art commercial ASR systems.

As a result, dysarthria can *have* dramatic consequences for speech intelligibility among artificial listeners – that is, speech recognition systems. In some preliminary experiments we have carried out on the TORGO dataset (Rudzicz, Namasivayam & Wolff, 2012), Google Speech API and IBM speech-to-text could misrecognize more than the 80% of words in single word utterances.

This paper describes AllSpeak, an Automatic Speech Recognition (ASR) Android Application specifically developed for patients with ALS. It allows patients to communicate, through their residual speech abilities, their basic needs to their families and caregivers.

2. *The AllSpeak App*

The AllSpeak App is a hybrid App developed with the Ionic 1.X framework for the applicationAndroid 6.0 platform. All the speech processing and recognition modules are implemented within a custom multi-threaded Cordova Plugin. The latter is composed by the following modules, each running on its own independent thread:

- audio acquisition (INPUT);
- voice activity detection (VAD);
- spectral features extraction (FE);
- speech command recognition, mainly based on Tensorflow neural networks (SR);

Once recognition is activated, these four processes run in parallel.

The INPUT module extracts speech from the smartphone's microphone and sends it to the VAD module.

When the VAD module recognizes speech activity, it sends the extracted speech segments to the FE module that calculates the spectral features and, once completed, sends concatenated feature vectors to the SR module.

The VAD module sends a detected speech segment to the FE module only if its duration is longer than a predefined threshold (500 ms in our case). The sent segment also contains a non-speech "tail", i.e., up to 400 ms long "active samples" after the last speech sample identified as speech. Then the resulting segment is considered as a command and after feature extraction its associated verbal command will be inferred by the SR module. The SR module consists in a simple speech decoder and runs preloaded Tensorflow deep neural networks.

This four-thread approach optimizes the recognition process, since the to-be-inferred features are already present in the SR module when the VAD module decides that a new command has been pronounced by the App user.

3. *The ASR engine*

The ASR engine (the SR module of the previous section) is based on deep neural networks. The spoken command decoding is simply the classification of the input speech segment and depends on the type of neural network used.

Neural networks training have been split in two steps: speaker-independent training on a control data set (i.e. healthy speakers) and speaker-adaptation to the patient of interest. Speaker adaptation has been applied to the deep feedforward neural networks (DNNs) to compensate the mismatch between clean speech-trained model and a small set of impaired speaker's data.

The ASR can use two different types of deep neural networks: deep feedforward neural networks (DNNs) and deep recurrent neural networks (RNNs).

3.1 Feature extraction technique

Feature extraction is the main part of the speech recognition system. The goal of feature extraction is to compute a sequence of feature vectors to have a compact representation of input signal. Because every speech and speaker has different individual characteristics embedded in their speech utterances, it is better to perform feature extraction in short term interval that would reduce these variabilities. Hence, the input voice signal is examined over a short period of time where the characteristics of speech signal become stationary. In general, a speech signal contains some acoustic information which can be represented by these features. There are several feature extraction techniques, however the use of Mel Frequency Cepstral Coefficients (MFCCs) can be considered as one of the standard methods for feature extraction (Motlíček, 2003) and it is also the technique employed in our algorithm. MFCCs are the result of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale.

In our algorithm, the speech signal is first divided into time frames consisting of an arbitrary number of samples. Each time frame is then windowed with 25 ms length Hamming window shifted every 10ms and for each speech frame, a set of MFCCs is computed. The number of spectral features employed in the DNN is listed below.

FBANKS (log mel-filter bank channel outputs):

- 24 FBANKS + temporal delta and acceleration coefficients (72 parameters per frame);
- 24 FBANKS + spectral delta and acceleration coefficients (72 parameters per frame);

MFCCs (mel-frequency cepstral coefficients):

- 13 MFCCs + temporal delta and acceleration coefficients (39 parameters per frame)
- 13 MFCCs + spectral delta and acceleration coefficients (39 parameters per frame)

Note that spectral deltas and acceleration coefficients are heuristic-based features we have proposed to account for the small training dataset. These features have an important impact for feedforward DNN as we will see in the result section.

3.2 ASR based on feedforward DNNs

Built on DNNs, the decoder simply averages the spoken command posterior probabilities that the DNN outputs at each speech frame and selects the command with the highest posterior.

The control dataset consisted of 23 commands spoken by eight healthy subjects, each command repeated from 8 to 10 times and the patient dataset comprised the same 23 commands spoken by eight ALS patients, each command repeated from 4 to 10 times – depending on patient's medical condition.

Regarding the DNN architecture and training, a three hidden layer DNN was implemented for the first training step on controls with an input layer containing 792 nodes (72 features x 11 context frames), the hidden layers with 500 nodes each and the output layer with 23 nodes, as many as the number of commands.

Once the first training step is completed, speaker adaptation comes in. We have experimented with a simple speaker-adapted layer insertion strategy consisting in adding input, output or hidden layers to the original net and then optimizing the parameters of that/those layer(s) only (see for example, Neto, Almeida, Hochberg, Martins, Nunes, Renals & T. Robinson, 1995; Gemello, Mana, Scanzio, Laface & De Mori, 2007; Li, Sim, 2010). For example, adding a first input layer should serve as "normalization" of the input, where the patient's input speech is transformed in order to closely match the input of the control training data.

As mentioned above, the DNN outputs are sentence/command posteriors:

$$(1) \quad y^* = \underset{s}{argmax} \frac{1}{T} \sum_{t=1}^{T} (p(s|x_t))$$

where y^*= selected sentence, s=sentence/command T= number of frames, x_t = concatenated vectors at time t.

This very simple decoding strategy resembles the key phrase recognition strategy proposed for Google small footprint keyword spotting in Chen et al. (2014).

3.2.1 Results

Averaged results for both the acoustic features employed with 23 different recorded commands pronounced by eight ALS patients and by eight healthy controls are shown in Table 1.

Table 1 - *Average Performance (Command Error Rate)*

	Acoustic Features	
	FBANKS	MFCC
Spectral	17.8 %	24.9 %
Temporal	32.7 %	32.7 %

Previously showed results are primarily related to patient's vocabulary size and modality of speech (intelligible or degraded speech) depending on the extent of the disease in each patient at the time of this study. A more detailed per-speaker accuracy is displayed in Table 2, together with a Therapy Outcome Measuring (TOM) tool using a Rating Severity Scale from 0-5 to rate scores of dysarthria (0 = normal, 3 = moderate and 5 = severe).

Table 2 - *Per Speaker Command Error Rates*

Speaker	*Temporal FBANKS*	*Spectral FBANKS*	*Temporal MFCC*	*Spectral MFCC*	*TOM*
BB	60.9 %	21.7 %	47.8 %	39.1 %	1
DAD	4.3 %	0.0 %	0.0 %	0.0 %	4
DG	21.7 %	8.7 %	43.5 %	4.3 %	2
PN	36.4 %	13.6 %	36.4 %	13.6 %	3
RS	39.1 %	26.1 %	30.4 %	34.8 %	1
SE	30.4 %	26.1 %	47.8 %	52.2 %	3
TE	54.5 %	31.8 %	31.8 %	36.4 %	3
VL	14.3 %	14.3 %	23.8 %	19.0 %	3

3.3 ASR based on RNNs

This section refers to the decoding strategy based on recurrent neural networks (RNNs) trained using a sequence-to-sequence approach. The sequence-to-sequence approach is not the only one we have tested (e.g., we also experimented with connectionist temporal classification) but it is the one that turned out to be the most successful. In this approach, the entire variable-length sequence of feature vec-

tors representing the speech segment is fed into the RNN that returns one single vector of posterior probabilities with one element for each command. The decoder simply selects the command with the largest posterior probability.

3.3.1 Architecture

RNNs have recently drawn the attention of researchers as they have proven to be a suitable tool to model temporal sequences. Indeed, it has been shown that RNNs can outperform feed-forward networks on large-scale speech recognition tasks (Sak et al., 2014). A recurrent neural network is a neural network that consists of a hidden state h which operates on a variable-length sequence $x = (x_1, \ldots x_T)$ through a non-linear activation function f. In our system, the input x is a vector that represents the acoustic features and we aim at finding the most likely corresponding command y. At each time step t the hidden state $h_t^{\rightarrow}$ of the RNN is updated by $h_t^{\rightarrow} = f(h_{t-1}^{\rightarrow}, x_t)$. Finally, it estimates the label posterior $p(y_t \mid x_t, h_t^{\rightarrow})$. The power of RNN relies on taking into account temporal dependencies over the input sequence, either unidirectionally or bidirectionally. Unidirectional RNN estimates the label posteriors using only the left (past) context of the recurrent input, while bidirectional RNN uses separate layers for processing the input in the forward (i.e., from left to right) and backward (i.e., from right to left) directions. In the latter case, we will have $p(y_t \mid x_t, h_t^{\rightarrow}, h_t^{\leftarrow})$, where $h_t^{\leftarrow} = g(h_{t+1}^{\leftarrow}, x_t)$ for some nonlinear function g. The limit of RNNs is that they can capture only very short time dependencies. To overcome this problem, we looked at a particular type of recurrent neural networks: the long short-term memory (LSTM) (Hochreiter, Schmidhuber, 1997). In this work, we implemented the bidirectional LSTM (BLSTM) architecture.

Typically, in speech recognition, both recurrent and feed-forward networks are trained as frame-level classifiers. As a consequence, the alignment between audio and transcription sequences has to be determined in order to have a target for every frame. Typically, alignments are provided by a Gaussian Mixture Model – Hidden Markov Model (GMM-HMM) system trained with the Baum-Welch algorithm. However, a good alignment of impaired speech may not be feasible, and that can have catastrophic consequences on the (frame-level) training of neural networks (as labels would be very noisy). To address these issues, we trained the BLSTM as a sequence-to-sequence model (Sutskever, Vinyals & Le, 2014). This method allows to train the network by taking in input a sequence of length T and giving as an output the correspondent sequence of length T', where T and T' are not necessarily the same. In our case, the output sequence is a command and, therefore, $T' = 1$.

The underlying idea is very simple: an encoder (or reader) BLSTM processes the input sequence and emits a fixed-size context variable C, which represents a summary of the input sequence. A decoder (or writer) takes as input the context C and generates the output sequence. Usually, the final hidden state of the encoder is used to compute C. In terms of probability, the sequence-to-sequence architecture maximizes the probability of the command, given the whole acoustic sequence, $p(\mathrm{y} \mid x_1, \ldots z_T)$.

3.3.2 Experimental setup

We evaluated the sequence-to-sequence BLSTM on the AllSpeak dataset. In particular, we tested five patients and two control speakers in order to cover the whole range of dysarthric degrees (on the TOM scale). From the speech of these speakers we extracted the adaptation data and the testing data. We only considered the temporal MFCC feature vectors, as they are the most conventional choice. For all the experiments, we used the BLSTM network with 5 hidden layers and 250 units per layer. We set the initial learning rate to be 0.01, and we exponentially decayed the learning rate by a factor of 0.7, every 3000 steps. Our model was trained to minimize the cross entropy (within the sequence-to-1 paradigm), by using the momentum optimizer with momentum equal to 0.9. We also clipped the gradient to avoid the vanishing/exploding gradient problem. Cross-validation was employed to get the best number of training epochs. To reduce the mismatch between the acoustic model and the testing speaker, we performed the speaker adaptation. More precisely, after training the network, we added a feed-forward layer atop the input. We trained the new layer, freezing the other ones, on the adaptation data. Finally, we used the testing data to measure the level performance of the model.

3.3.3 Results

Table 3 shows the command error rate (CER). As expected, the error is lower on the control speakers testing. Surprisingly, the sequence-to-sequence BLSTM achieves a good performance even in presence of dysarthric speech, with a minimum error of 4% on the speaker SG. In every case, the error is reduced (or remains equal) after speaker adaptation. In the best case, adaptation provides an error reduction from 71.7% to 21.7%. Note that the averaged error rates are not referred to all speakers but only to BB, PN, RS and SE. This is to compare the CERs with the ones coming from Table 2. As we can see, we obtained a CER reduction from 36.0% to 16.6%.

Table 3 - *Sequence-to-sequence BLSTM results*

Speaker	*Patient/Control*	*CER (without adaptation)*	*CER (after adaptation)*	*TOM*
AI	Control	7.0 %	7.0 %	0
CD	Control	8.0 %	1.3 %	0
BB	Patient	25%	18.2 %	1
PN	Patient	34.8 %	13 %	3
RS	Patient	71.7 %	21.7 %	1
SE	Patient	4.0 %	4.0 %	3
SG	Patient	44.4 %	25.9 %	NA
Average	Patients	36.0 %	16.6 %	–

4. *Conclusion*

Despite their growing presence in home computer applications and various telephony services, commercial automatic speech recognition technologies are still not easily employed by everyone, especially individuals with speech disorders. ALLSpeak is an App designed for Android equipped smartphones and tablets that allow ALS patients to go on communicating with the rest of the world, both when speaking becomes an effortful task and when their voice intelligibility almost vanishes. The first version of our algorithm running on the App was based on a DNN trained on non-dysarthric speech. This recognizer had an averaged command error rate ranging from 32.7% to 17.8% using temporal and spectral FBANKS respectively and from 32.7% to 24.9% using temporal and spectral MFCC for dysarthric speech. With the aim of improving the recognizer performance, we explored a further method: the sequence-to-sequence LSTM model. We observed that the best performance is accomplished by applying the speaker adaptation, providing an averaged command error rate of 15.0% over all 7 speakers, and 16.6% over the 5 patients. In order to compare the DNN and LSTM models, we analyzed the results related to the common tested speakers. What we found is an averaged error rate difference of 19.4%, showing that the LSTM model trained in a sequence-to-sequence fashion is a more suitable tool to address the dysarthric speech recognition. Thus, the following step will be the integration of this method to the mobile application. Our belief is that, by using our AllSpeak application, people with speech disorders will have the opportunity to participate in the technology present and experience the benefits of smartphones which are powerful devices able to mitigate their disabilities.

Bibliography

Chen, G., Parada, C., Heigold, G. (2014). *Small footprint keyword spotting using deep neural networks.* In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 4087-4091.

Espana-Bonet, C., Fonollosa, J.A.R. (2016) *Automatic Speech Recognition with Deep Neural Networks for Impaired Speech.* A: International Conference on Advances in Speech and Language Technologies for Iberian Languages. "Advances in Speech and Language Technologies for Iberian Languages: Third International Conference, IberSPEECH 2016: Lisbon, Portugal, November 23-25, 2016: proceedings". Lisbon: Springer, 97-107.

Gemello, R., Mana, F., Scanzio, S., Laface, P. & De Mori, R. (2007). *Linear hidden transformations for adaptation of hybrid ANN/HMM models*. Speech Communication, Elsevier: North-Holland, 49 (10-11), 827.

Hardiman, O. (2017). *Amyotrophic lateral sclerosis*. Nature Reviews Disease Primers, 3, 17071.

Hochreiter, S., Schimidhuber, J. (1997). *Long short-term memory*. Neural computation, 9(8), 1735-1780.

Joy, N.M., Umesh, S., Abraham, B., (2017) *On Improving Acoustic Models for TORGO Dysarthric Speech Database*. Proceedings Interspeech 2017, 2695-2699.

Kim, Y., Kent, R.D., Weismer, G. (2011). *An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria*. In *Journal of Speech, Language, and Hearing Research*, 54(2), 417-429.

Li, B. & Sim, K.C. (2010). *Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems*. Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, 526-529.

Motlicek, P. (2003). *Feature Extraction in Speech Coding and Recognition, Report, Portland, to research, data, and theory*. Technical Report of PhD research internship in ASP Group, OGI-OHSU, < http://www. fit. vutbr. cz/~motlicek/publi/2002/rep ogi. pdf.

Neto, J., Almeida, L., Hochberg, M., Martins, C., Nunes, L., Renals, S. & Robinson, T. (1995). *Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system*. Proceedings from Eurospeech '95: The 4th European Conference on Speech Communication and Technology, 2171-2174.

Rudzicz, F. (2010a). *Toward a noisy-channel model of dysarthria in speech recognition*. Proceeding of the NAACL HTL 2010 Workshop on Speech and Language Processing for Assistive Technologies, Association for Computational Linguistics, Los Angeles, California, 80-88.

Rudzicz, F. (2010b). *Correcting errors in speech recognition with articulatory dynamics*. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10), Association for Computational Linguistics, Stroudsburg, PA, USA, 60-68.

Rudzicz, F. (2012). *Using articulatory likelihoods in the recognition of dysarthric speech*. In Speech Communication, 54, 430-444.

Rudzicz, F., Namasivayam, A.K., Wolff, T. (2012). *The TORGO database of acoustic and articulatory speech from speakers with dysarthria*. In Lang Resources & Evaluation, 46: 523-541.

Sak, H., Senior, A., Beaufays, F. (2014). *Long short-term memory recurrent neural network architectures for large scale acoustic modeling*. In Fifteenth annual conference of the International Speech Communication Association.

Sutskever, I., Vinyals, O., LE, Q.V. (2014). Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems, 3104-3112.

Talbot, K. (2009). *Motor neuron disease: the bare essentials*. Practical neurology, 9, 303-09.

Vachhani, B., Bhat, C., Das, B., Kopparapu, S.K. (2017). *Deep auto encoder based speech features for improved dysarthric speech recognition*. In Proceedings of Interspeech 2017, 1854-1858.

Van ES, M.A., Hardiman, O., Chio, A., Al-Chalabi, A., Pasterkamp, R.J., Veldink, J.H., Van Den Berg, L.H. (2017). Amyotrophic lateral sclerosis. The Lancet, 377(9769), 942-955.

DOI: 10.17469/O2104AISV000013

MARIA DI MARO, SARA FALCONE, FRANCESCO CUTUGNO

Prosodic analysis in human-machine interaction

In this paper, we are going to present some experiments concerning the analysis of prosodic features in the spoken production of requests by human users in human-machine interactions. The main aim of this analysis is to understand if and how much a speaker adapts the spoken production to his/her virtual interlocutor, and to which extent this could be caused by the type of user and his/her representational preconception towards the specific interaction. The collected results are, therefore, considered as an important means for developing spoken dialogue systems whose speech recognition module skills are better suited to the characteristic of the human interlocutor.

Keywords: human-machine interaction, spoken dialogue systems, pitch, speaking rate, vowel space, hyperarticulation.

1. *Introduction*

Human-machine interaction is a field of research that covers several aspects, from text to speech to pragmatic aspects of the verbal and nonverbal interaction. In this paper we focus on speech and, particularly, on the prosodic anomalies analysed in several human-machine dialogues. In this work, the analysis of the prosodic features of questions and commands posed to a domain-dependent spoken dialogue system and to Google Assistant is carried out. The collected results were compared with speech materials produced at a different point on the diaphasic continuum. In particular, spontaneous narrations were recorded, as users were asked to tell the plot of a movie they saw. The main aim of this analysis was to understand if and how much a speaker adapted the spoken production to the virtual interlocutor and if a previous acquaintance with the system had an impact on the interaction.

The conventional assumption towards human speech in conversing with computer systems is that speakers usually tend to simplify their language to avoid not being understood. The reason behind the use of a simplified register takes origin from the perception of the non-expertise of computers in conversing naturally. On the other hand, other empirical observations show how the use of a virtual assistant on a personal smartphone can lead to a more spontaneous language production, as if the user was speaking with another human interlocutor. This difference lies in the representational perception of the other which explains how the language production can be modified according to the users' preconceptions towards a specific channel of communication or context of interaction. With this analysis we would like to start a deepened pragmatic analysis of human-computer interaction, taking prosodic features as a starting point. A theory of mind of user behaviour is

the desired future goal, useful to shape a better model (Soltau, Waibel, 2000), which would improve speech recognition for spoken dialogue systems.

In this paper, we sketched out the steps of a prosodic analysis concerned with the alleged hyperarticulation of sounds in user's utterances addressed to machines. We started from what it is described in other state-of-the-art studies on hyperarticulation and simplified registers. Hyperarticulate speech is intended as a strategy to ease the communicative exchange in situations where misunderstanding or non-understanding of the speaker intention occur. Other than using simpler syntactic structures, common words, pointing gestures and other visual paralinguistic means of communication, the use of a greater articulatory effort in producing sounds, is seen as helping in increasing the success of the informative exchange. This strategy is employed in different compromising situation, where the listener could have problems to catch the message because of a noisy environment, or because he/she is a foreigner, or has hearing impairment or is a child, who is taking his/her first steps in language learning. In fact, researches on hyperarticulation are mainly concerned with infant-directed (McMurray et al., 2013; Hartman et al., 2017; Kalashnikova et al., 2018), foreigner-directed (Scarborough et al., 2007) and hearing-impaired-persons directed speech (Picheny et al., 1986). Computer-directed speech is another field of interest (Oviatt et al., 1998; Stent et al., 2008; Akira et al., 2017) where the hyperarticulation hypothesis is studied, especially as an error resolution strategy (Oviatt et al., 1998). Conversely, with this paper we start considering possible hyperarticulated realizations without paying particular attention to phonetic adaptations triggered by error resolutions, which has already been studied (Oviatt et al., 1998) and which will be however further researched in future studies, considering different classes of errors causing different types of adapting strategies. This was useful to prove if the alleged adaptations were the result of a general phenomenon occurring in human-machine interaction or if they were merely related to error resolution patterns.

The paper is structured as follows: in the first section, the description of tests used to collect data in two subcorpora is provided; the second paragraph explains the parameters taken into account in the conducted analysis; in the third section, results for both subcorpora are shown; finally, conclusions outline interpretation of the previously mentioned results.

2. *Corpus collection*

The corpus used for the analysis consists of annotated audio files recorded in two different test sessions. During the first session, users were asked to pose questions to a task-oriented dialogue system (Di Maro et al., 2017), designed to give information concerning paintings in a virtual museum. To guide users through the test, they were provided with twenty different conceptual classes, such as *Name of the artist*, *Name of the painting*, *Techniques*, *Iconography*, and similar. The combination of these classes and the paintings shown in the 3D scene – developed with the game engine Unreal Engine 4[1] – led users to ask questions

[1] https://www.unrealengine.com.

naturally. For this work, we selected one hundred questions posed by five different users, each of whom was asked to pose two questions per class (a simpler one and a more articulated one in terms of syntactic structures or vocabulary). The interaction was structured here as a question answering system. The same users were asked to recount the last movie they saw or the movie they liked the most, in order to be able to compare the previously mentioned human-machine interactions with spoken utterances collected in human-human conversations, thus in a different situational context for which the interlocutor was human and not virtual.

For the second session, we selected a different dialogue system, which was the well known general-purpose virtual assistant Google Assistant. This choice is motivated by the necessity to understand if different tools, their designed purposes and the quality of ASRs could be important in affecting the language production by human users. This test was divided in four phases. The first one was concerned with collecting users' personal data, useful to later map the extracted language features to their specific characteristics. Those who never used a virtual assistant on their smartphone were introduced to it through a brief explanation and examples. Afterwards, they were asked to complete three different tasks. Specifically, they had to interact with the assistant to find a place (street, route or a specific place) with Google Maps, to memorize an appointment on the calendar, specifying time, place and everything that was important to them, and to send a message to someone using Whatsapp. In this third phase, we collected 67 different recordings, corresponding to conversational turns by the users. The resulting corpus comprises clauses of different type: 65,67% consists of imperative clauses, 16,41% declarative clauses, 7,48% infinitive clauses, 4,47% noun clauses, and 5,97% interrogative clauses. The dialogue act (requesting information, saving an appointment, etc) were fulfilled sometimes in one turn and some other times in more than one turn, according to the expertise or preference of the user. Not-understood or misunderstood user inputs, for which error resolution reformulations were needed, were only the 15,67% of the total turns collected. Finally, they were asked to complete a questionnaire, to evaluate usability and satisfaction of the interaction. The collected evaluations were used to better interpret the results. The ten selected participants, who did not overlap with users tested in the previous experiment, were different for age, gender and attitude towards technology. As for the first test, we also recorded spontaneous narrations given by the same participants.

The recordings were paired with TextGrid files containing graphic and phonetic transcriptions. The phone alignment was automatically processed using WebMAUS (Kisler et al., 2017), whose outputs have been manually corrected. These files were therefore used to extract suprasegmental features, precisely pitch mean, pitch range, speaking rate and stressed vowels' formants, as we are going to illustrate in the next section.

3. *Corpus analysis*

Starting from an empirical observation of our subcorpora, it was noticeable that the collected human-machine interactions were characterized by specific hyperarticulation-related acoustic properties, such as loudness, reduced speaking rate, and the absence of

hypo-clear sounds. Hyperarticulate speech is used with "at-risk" listeners, such as children, hearing-impaired interlocutors, and non-native speakers. In each of these communicative risky situations, different parameters have been noticed: in infant-directed speech elevated pitch, higher pitch rate and stress on new words are used (Ferguson, 1977); with hearing-impaired speakers, amplitude and frequency values are higher and speaking rate is lower (Picheny et al., 1986). Since there are no distinct characteristics in determining hyperarticulation and stated that hyperarticulation in human-computer interaction is still not well defined (Oviatt et al., 1998), our long-term goal here is to find specific traits characterizing this peculiar context of spoken production. As a starting point, prosodic parameters, which are going to be presented in the course of this section, were selected.

As user commands appear to be hyperarticulated, their speaking rate is expected to be lower. As a matter of fact, in hyperarticulate speech, all syllables are pronounced, and many short pauses are used, resulting in an increased speech time. Pauses are here mainly used to pragmatically segment units of meaning, such as phrases, as if avoidance of system information overloading was intended. To get accurate results, we manually counted the perceived syllables (i.e. the ones really produced by the speaker, and not the ones expected to be produced) divided by the seconds of speech production.

Many studies on infant-directed speech (Fernald et al., 1989; Song et al., 2010; Gauthier, Shi, 2011), pointed out that higher pitch values, exaggerated pitch contours, and wider pitch ranges occur when asking for attention, communicating intentions, or even for lexical teaching purposes. Since being clear is fundamental in those situations, we may consider pitch trends to be crucial to also differentiate human-machine interaction utterances compared to spontaneous language productions. In fact, one of the pragmatic difference arising in human-computer interaction is concerned with the necessity of being understood, by means of a clear language, in a situation where the linguistic expertise of the interlocutor is perceived to be lower. For this research, we therefore decided to start from computing pitch mean and pitch range values, where the former is important to collect single utterances pitch trends, and where the latter is useful to compute differences in trends for each utterance. These values are manually calculated for every single file recording in our corpus, using the pitch frequency waves computed in Praat (Boersma, Weenink, 2018). The average values obtained for each user are also noted.

One assumption about hyperarticulate speech is that exaggerating speech sounds production leads to an extension of the vowel space (Story, Bunton, 2017; Wedel et al., 2018). This means that when vowels are hyperarticulated, they tend to be further from the centroid in the vowel space triangle. Conversely, when speakers do not articulate sounds carefully, they tend to produce vowels closer to the centroid, meaning that their articulation differences are less detectable. To compute the vowel position in the triangle, we extracted the formant values (F1 and F2) for each manually annotated stressed vowel in the corpus, using a Praat script. Moreover, since speakers can show their own articulation differences, we calculated the vowel space dispersion using the centroid of

each speaker's vowel triangle. Specifically, centroids were computed by grand means of vowels' average formant values (Koopmans-Van Beinum, 1983).

4. *Results*

In this session, the experimental results are presented. Firstly, we will have a look to the selected parameters (speaking rate, pitch contours and vowel space) within the first test, then the same will be displayed for the second one. Discussions and interpretations follow the outcomes.

4.1 Dialogue system-directed speech vs. spontaneous speech

As far as the speaking rate is concerned, a tendency can be outlined: speakers tend to talk slowly in interacting with a goal-oriented dialogue system (Figure 1). The system was indeed new to them; therefore, they thought that speaking fast could have jeopardized the understanding by the virtual interlocutor. Only for the fourth user the tendency is not applicable. Nevertheless, it must be noticed that this user started the task with a very high speed (7,3) and ended it with a lower value (3,49). In Figure 2, we can observe how the aforementioned speaker continuously changes the speed value, as a gradual attempt to adjust his speech to the interlocutor. The general tendency of a lower speaking rate is a first value in favour of the hypothesis of hyperarticulation in this particular speech use.

Figure 1 - *Speaking rate results – human-human vs. human-machine interaction (Users 1, 2= female; users 3, 5= male)*

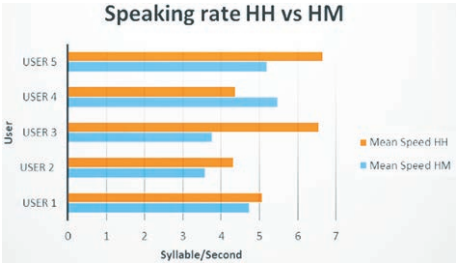


Figure 2 - *Speaking rate values – user 4 (male) – on the x-axis is the id-number of the utterance, whereas on the y-axis is the speed (syllables/second)*

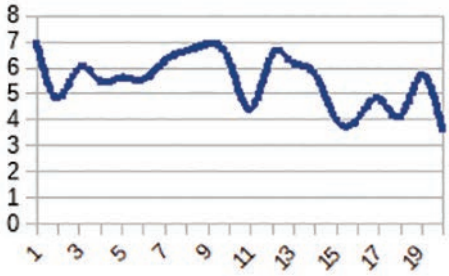


Figure 3 - *Mean pitch results – human-human vs. human-machine interaction (Users 1, 2= female; users 3, 5= male)*

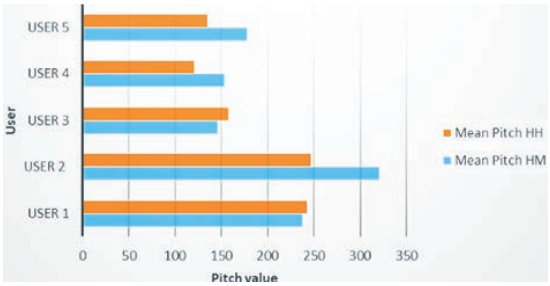
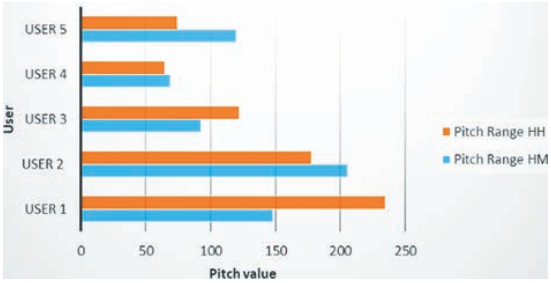


Figure 4 - *Pitch Range Results – Human-Human vs. Human-Machine Interaction [Users 1-2= female; Users 3-5= male]*



Conversely, the hypothesis of an evidence in pitch contours cannot be confirmed. As a matter of fact, pitch mean is higher in human-machine interaction in three out of five users, which is a difference not that significant in defining a tendency. But, although this increasing dynamics, pitch mean values are not so low compared to the ones occurring in narration. In fact, HM and HH values for user 1 and 3 are very closed to each other (Figure 3). Therefore, also pitch range values are not significant to underline the hypothesized tendency (Figure 4). The tendency of increasing pitch contours was empirically selected especially because of the speech performance of some users, such as users 2 and 5, who tended to stress sounds in an unnatural way. The non-confirmation of this prosodic adaptation can be actually motivated by the fact that in human-machine interaction there is no need to focus the computer's attention to specific lexical items via a pitch variation, as it could be useful in infant-directed speech (Oviatt et al., 1998). The pragmatic needs in this diaphasic situation are in fact of different kinds.

Concerning the vowel space based on formants values, we can infer that users tend to articulate sounds differently. For the female users (Figures 5, 6), the dialogue system's vowel triangle appears to be less extended that the one resulting from the spontaneous speech, especially what front vowels' F2 is concerned. Despite this reduced extension, for user 2 (Figure 6) the back closed vowel is further from the centroid compared to its equivalent in narration. On the other hand, for the other three users, vowel spaces are much more extended in interacting with the virtual agent, confirming the hypothesis of hyperarticulation.

Figure 5 - *Vowel chart - human-human vs. human-machine interaction (female)*

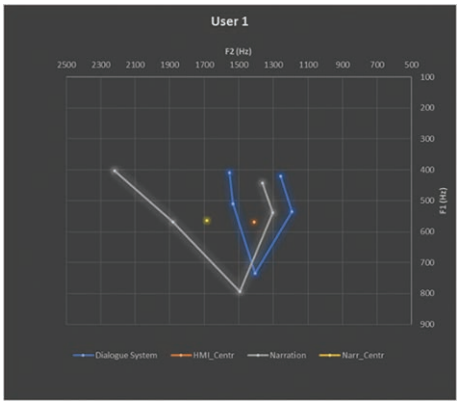


Figure 6 - *Vowel chart - human-human vs. human-machine interaction (female)*

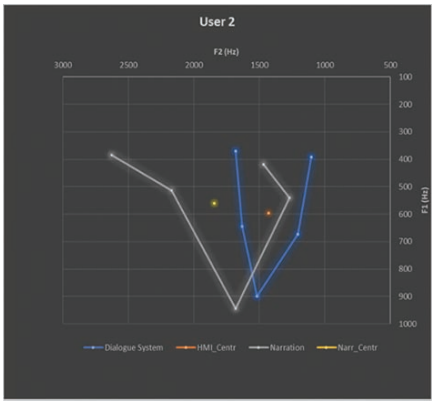


Figure 7 - *Vowel chart - human-human vs. human-machine interaction (male)*

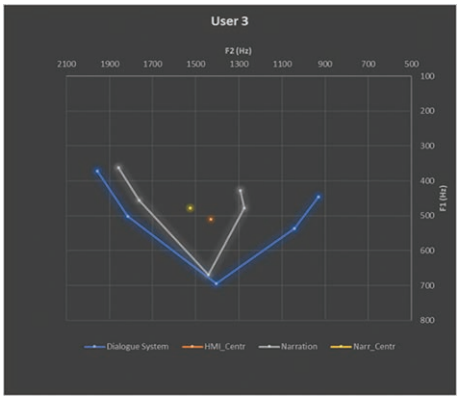


Figure 8 - *Vowel chart - human-human vs. human-machine interaction (male)*

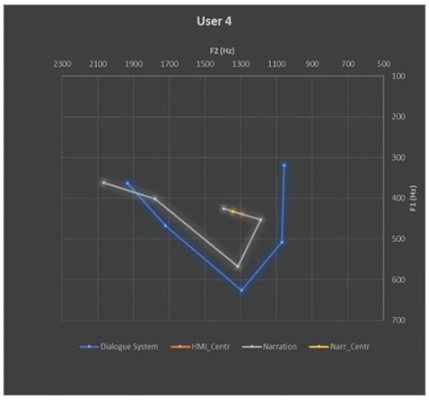


Figure 9 - *Vowel chart - human-human vs. human-machine interaction (male)*

4.2 Google assistant-directed speech vs. spontaneous speech

Even when interacting with Google Assistant, speakers tend to decrease their speaking rate, although the difference registered was not always significantly remarkable (Figure 10). Only for three (user 3, 7, 8) out of ten users this tendency was not observed. In Figure 11, we can notice how both users 3 and 7 start with a medium speed (on average 4-5 syllable/second), increase it up to 6, and eventually drop it to a slower value (2,94 for user 3 and 4,42 for user 7). User 8's behaviour is less stable, since his speaking rate increases and decreases at every utterance. Interestingly, the first utterance of each task (we refer to the points 1, 3, 4 on the x-axis in Figure 11 – user 8) is always slower than the others. The reason may lie in the tendency to be hesitant about the correct speaking rate to use to avoid misunderstanding; therefore, once the first turn is understood, the others uttered to complete the pragmatic act and to get the desired information is faster produced.

Concerning pitch values, six out of ten users employed this prosodic strategy regarding speech adaptation to computers, despite a major difference was evident only in users 1, 5, 7 and 9 (Figure 12). Pitch range values confirm the unreliability of this parameter in this context of use (Figure 13), as explained for the first test.

Contrary to what being observed for the previous test, the hyperarticulation hypothesis cannot be confirmed when analysing vowel spaces as a matter of fact, its extension is wider only in three users (1, 5 and 6), although this difference is not statistically relevant. For users 4 and 7, only specific vowels are over-articulated: the closed back rounded vowel for user 4 and user 7 (Figures 16, 20) and the open unrounded vowel for user 7 (Figure 20). Interestingly, in user 9 (Figure 22), the reduction phenomenon occurring in the interaction with the conversational agent is so strong that the back closed rounded vowel overlaps with the centroid.

The observable reductions can be explained with reference to the perceived experience during the interaction. Only users 2 and 4 stated that they had never conversed with a dialogue system. Consequently, not being an unprecedented experience, it makes them believe that they have a better expertise and can interact more naturally. The users who perceived themselves as experts are also the ones with whom the assistant had understanding problems. As a consequence, they evaluated the system use negatively.

Figure 10 - *Speaking rate results - human-human vs. human-machine interaction (users 1, 5 = females; users 6-10 = males)*

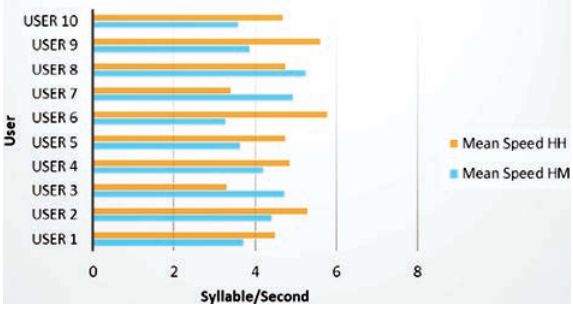


Figure 11 - *Speaking rate values - users 3, 7, 8 On the x-axis is the id-number of the utterance, whereas on the y-axis is the speed (syllables/second)*

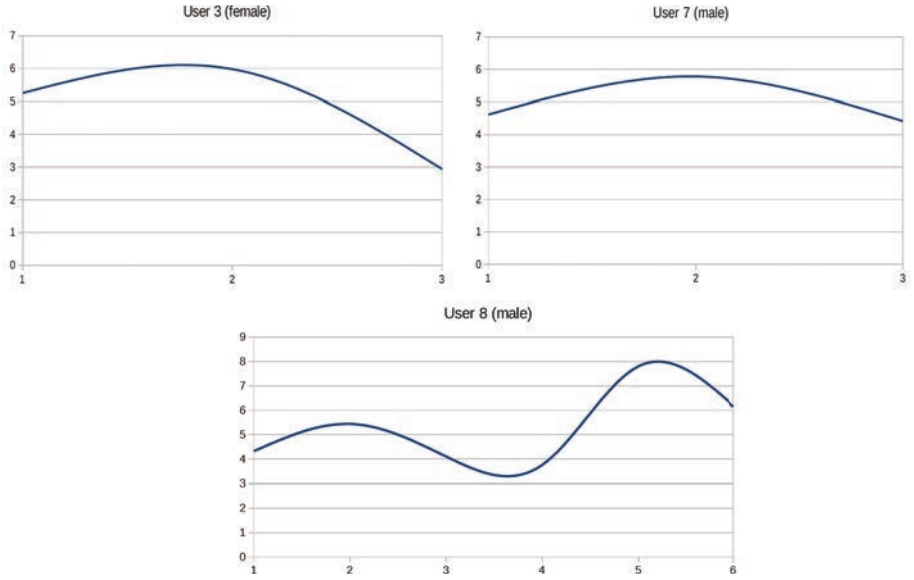


Figure 12 - *Mean pitch results - human-human vs. human-machine interaction (users 1, 5 = females; users 6, 10 = males)*

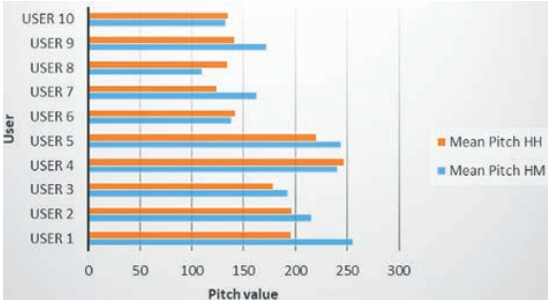


Figure 13 - *Pitch Range Results - human-human vs. human-machine interaction (users 1, 5= females; users 6, 10= males)*

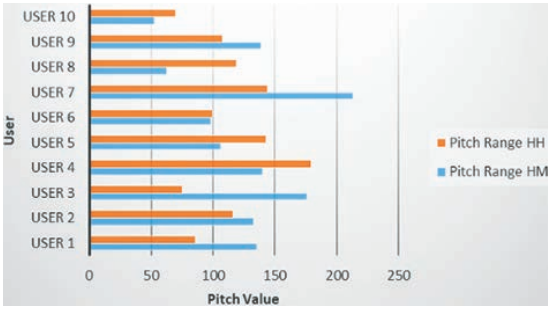


Figure 14 - *Vowel chart - human-human vs. human-machine interaction (female)*

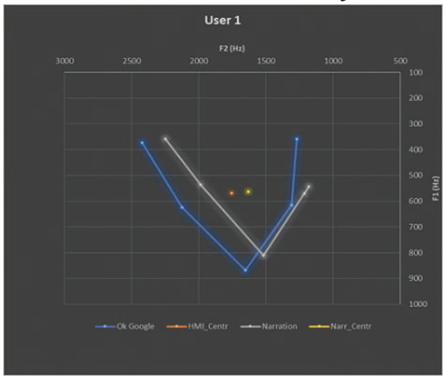


Figure 15 - *Vowel chart - human-human vs. human-machine interaction (female)*

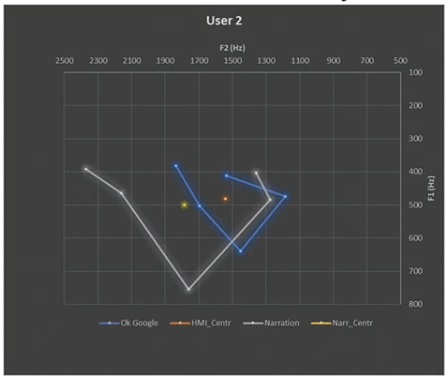


Figure 16 - *Vowel chart - human-human vs. human-machine interaction (female)*

Figure 17 - *Vowel chart - human-human vs. human-machine interaction (female)*

Figure 18 - *Vowel chart - human-human vs. human-machine interaction (female)*

Figure 19 - *Vowel chart - human-human vs. human-machine interaction (male)*

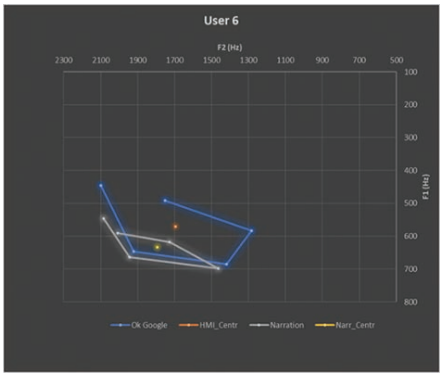


Figure 20 - *Vowel chart - human-human vs. human-machine interaction (male)*

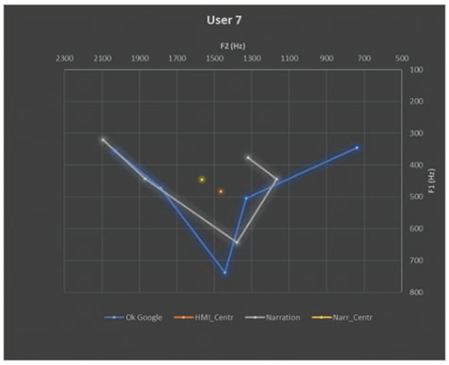
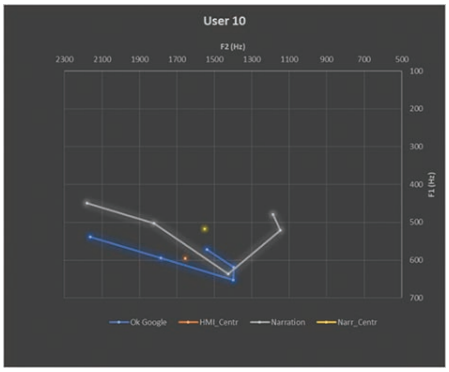


Figure 21 - *Vowel chart - human-human vs. human-machine interaction (male)*

Figure 22 - *Vowel chart - human-human vs. human-machine interaction (male)*

Figure 23 - *Vowel chart - human-human vs. human-machine interaction (male)*



5. *Conclusions*

Although human-computer dialogue could be considered as an "at risk" context of interaction because of the present limited expertise of machines in the field of encyclopaedic knowledge and social signal processing, our results show that, as far as the selected parameters are concerned, no measurable acoustic difference were observed. As a matter of fact, only the speaking rate was slower as expected. Pitch and vowel space expansion did not change significantly, confuting the hypothesis of hyperarticulation. Precisely, pitch values are considered not important in defining hyperarticulation in human-machine interaction, as stated in other studies (Oviatt et al., 1998), whereas F1 and F2 extension is an intriguing approach which did not lead to the expected results, even in studies on infant-directed speech (Miyazawa et al., 2017). Nevertheless, a slightly difference in prosody, due to length and type of pauses or to amplitude and tone, can still be perceivable. For this reason, further phonetic and pragmatic analysis are needed to define how hyperarticulation or clearer speech production is generated in human-machine interaction.

All in all, users, even the ones who were not used to talk with a virtual agent as stated in the questionnaires, appeared to perceive the interaction as customary, especially as far as the interaction with the digital personal assistant was concerned, since speakers tend to identify these technological tools as something which is no longer far from our experiential horizon. In fact, in drafting a theory of mind (Goldman, 2012) applied to this special context of use, we can assert that users tend to interact accordingly to their preconception of the affordances of the tool: since nowadays technology is perceived as having unlimited capacities, speakers start adopting a communicative code which is closer to the one used in interacting with other humans.

The counter-adaptation resulting from the non-hyperarticulation of sounds in human-machine interactions can be explained with the increased error rate occurring when speakers try to speak clearer, as shown in other studies (Oviatt et al., 1998b; Soltau and Waibel, 1998). For this particular reason, the prosodic characteristics triggered in error resolution scenarios or in noisy environments (virtual agents for call-routing or for driver assistance) represent an interesting future investigation. Finally yet importantly, increasing the awareness of the specific pragmatic traits arising from this context of interaction will be advantageous in the development of better-performing acoustic and linguistic models.

Bibliography

Akira, H., Vogel, C., Luz, S. & Campbell, N. (2017). Speech Rate Comparison when Talking to a System and Talking to a Human: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task. *Proceedings of Interspeech 2017*, 3286-3290.

Boersma, P., Weenink, D. (2018). *Praat: Doing Phonetics by Computer*. Version 6.0.37. Accessed 14.03.2018 from http://www.praat.org/

Di Maro, M., Valentino, M., Riccio, A. & Origlia, A. (2017). Graph Databases for Designing High-Performance Speech Recognition Grammars. IWCS 2017 – 12th International Conference on Computational Semantics – Short papers.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., De Boysson-Bardies, B. & Fukui, I. (1989). A crosslanguage study of prosodic modifications in mothers' and fathers' speech to preverbal infants, In *Journal of Child Language*, 16, 477-501.

Ferguson, C.A. (1977). Baby Talk as a Simplified Register. In Snow C.E., Ferguson C.A. (Eds.), *Talking to Children*. Cambridge, Cambridge University Press, 209-235.

Gauthier, B., Shi, R. (2011). A Connectionist Study on the Role of Pitch in Infant-directed Speech, In *The Journal of the Acoustical Society of America*, 130(6), EL380-EL386.

Goldman, A.I. (2012). Theory of Mind. In Margolis, E., Samuels, R. & Stich, S.P. (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science*. New York: Oxford University Press, 402-424.

Hartman, K.M., Ratner, N.B. & Newman, R.S. (2017). Infant-Directed Speech (IDS) vowel clarity and child language outcomes. In *Journal of child language*, 44(5), 1140-1162.

Kalashnikova, M., Goswami, U. & Burnham, D. (2018). Mothers speak differently to infants at-risk for dyslexia." In *Developmental science*, 21(1), 10-15.

Kisler, T., Reichel U.D. & Schiel, F. (2017). Multilingual Processing of Speech via Web Services. In *Computer Speech & Language*, 45, 326-347.

Koopmans-Van Beinum F.J. (1983). Systematics in Vowel System. In van den Broecke M., van Heuven V. & Zonneveld W. (Eds.), *Sound Structures*, FORIS Publications, Dordrecht, 159-171.

McMurray, B., Kovack-Lesh, K.A., Goodwin, D. & McEchron, W. (2013). Infant Directed Speech and the Development of Speech Perception: Enhancing Development or an Unintended Consequence? In *Cognition*, 129(2), 362-378.

Oviatt, S.L., MacEachern, M. & Levow, G. (1998). Predicting Hyperarticulate Speech During Human-Computer Error Resolution. In *Speech Communication*, 24(2), 87-110.

Oviatt, S.L. (1998). The CHAM Model of Hyperarticulate Adaptation During Human-Computer Error Resolution. *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia paper 49.

Picheny, M.A., Durlach, N.I. & Braida L.D. (1986). Speaking Clearly for the hard of hearing II: Acoustic Characteristics of Clear and Conversational Speech. In *Journal of Speech, Language, and Hearing Research*, 29(4), 434-446.

Scarborough, R., Dmitrieva, O., Hall-Lew, L., Zhao, Y. & Brenier, J. (2007). An Acoustic Study of Real and Imagined Foreigner-Directed Speech. In *Journal of the Acoustical Society of America*, 121(5), 3044.

Soltau, H., Waibel, A. (1998). On the Influence of Hyperarticulated Speech on Recognition Performance". *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia paper.

Soltau, H., Waibel, A. (2000). Specialized Acoustic Models for Hyperarticulated Speech. In *Proceedings of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3, 1779-1782.

Song, J.Y., Demuth, K. & Morgan, J. (2010), Effects of the Acoustic Properties of Infant-Directed Speech on Infant Word Recognition. In *The Journal of the Acoustical Society of America,* 128, 389-400.

Stent, A.J., Huffman, M.K. & Brennan, S.E. (2008). Adapting Speaking after Evidence of Misrecognition: Local and Global Hyperarticulation. In *Speech Communication*, 50(3), 163-178.

Story, B.H., Bunton, K. (2017), Vowel Space Density as an Indicator of Speech Performance. In *The Journal of the Acoustical Society of America*, 141(5), EL458-EL464.

Wedel, A., Nelson, N. & Sharp, R. (2018), The phonetic specificity of contrastive hyperarticulation in natural speech. In *Journal of Memory and Language*, 100, 61-88.

DOI: 10.17469/O2104AISV000014

VALENTINA SCHETTINO, ANTONIO ORIGLIA, FRANCESCO CUTUGNO

Dynamic time warping and prosodic prominence

In this study, an investigation on methodological issues linked with prosodic prominence rating is carried out. Our main research goal is the resolution of a particular issue related to a specific rating scale – i.e. the one resulting from the PromDrum method (Samlowski, Wagner, 2016). In this approach, namely, the rater is left free to *drum* as many syllabic units as perceived, thus resulting in rated material in which the number of rated units possibly does not correspond to the number of actually expected syllables. In order to solve this problem, a forced alignment algorithm is developed with the Dynamic Time Warping procedure. In this way, we are able to find the best possible alignment without subjective choices. Moreover, the procedure allows a qualitative evaluation of the rated material.

Keywords: prosodic prominence, rating scales, Dynamic Time Warping.

1. *Prosodic Prominence: An introduction*

In the last years, many scholars have investigated a prosodic phenomenon with an important communicative function (Streefkerk, 2002), known as prosodic prominence. Prominence could be defined as "the property by which linguistic units are perceived as standing out from their environment" (Terken, 1991: 1768). In spoken productions, salient units are produced and perceived in a more detailed way, being fundamental in the understanding of the linguistic message. The way in which this emphasis is achieved, however, is a complex and multi-layered process, that comprehends both acoustic behaviours and linguistic expectancies (Streefkerk, 2002). The disentanglement of the different influences and of their relative importance is particularly complex. Furthermore, every language seems to adopt its own manner of signaling prominence, both through different acoustic correlates and through the linguistic meaning assigned to each; prominence patterns can scarcely be compared in different linguistic and phonetic settings in the same language too: this means that the same prominence pattern could bear two different meanings in two different contexts. Eventually, prominence can be referred in various ways to different prosodic domains: intonation patterns are strictly related to prominence, seen that the shape of the pitch curve is related to both phenomena; stress and prominence mutually influence each other, too. As such, then, prominence is a complex phenomenon, acting on different prosodic levels; its production and perception, indeed, cannot be properly described without observing the various contribution to it from different, although related, fields. Disentangling the different features contributing to this phenomenon on different levels is a big challenge that the academic community is facing (Wagner, Origlia, Avesani, Christodoulides, Cutugno,

D'Imperio, Escudero Mancebo, Gili Fivela, Lacheret, Ludusan, Moniz, Chasaide, Niebuhr, Rousier-Vercruyssen, Simon, Simko, Tesser, Vainio & 2015). One of the most interesting issues related to this topic, however, is the annotation of prosodic corpora regarding prominence. Rating procedures, indeed, are both interesting for cognitive reasons and for methodological interest: how comes that some syllables are perceived as standing out, and which elements are considered most relevant in the perceptual phase? In order to shed some light on these questions, in this paper we deal with a specific rating methodology, explaining its theoretical benefits but also presenting some drawback aspects linked with it. Specifically, we suggest an improvement of the rating scale known as PromDrum method (Samlowski, Wagner, 2016), using Dynamic Time Warping.

In the next sections, the problem will be presented, and our proposal will be depicted. We will present the material used for the investigation and display the results achieved through the suggested implementation.

2. *Prosodic Prominence and rating scales*

One of the first issues to be solved when approaching the annotation of prosodic prominence is the kind of rating scale to use for prominence levels, linked with the more general question of whether prominence should be regarded as a discrete or continuous phenomenon. In the literature, in fact, different kinds of scales have been used: given that the linguistic community still does not agree upon the number of relevant linguistic categories as regards prominence scales, some settle a binary system, in which the distinction between prominent and non-prominent syllables is considered sufficient for the description of the phenomenon; on the other side, however, we find completely opposed approaches, in which prominence categories are discretized in 31 different classes.

In particular, the so-called 31 degrees scale, first introduced by Fant, Kruckenberg (1989), aims at a fine-grained evaluation of perceived prominence levels; this type of scaling can be easily related to a physical perspective and presupposes a gradient functioning of prosodic patterns, both on the level of production and perception. However, it could be difficult for the annotators to discretize between such close categories.

In the other scale type mentioned above, known as binary scale, syllables are expected to be either prominent or non-prominent, no further category is contemplated (Wightman 1993; Streefkerk, Pols, Ten Bosch, 1999); this approach should result in an easier job for the annotator: nevertheless, this simplistic view risks to leave aside important distinctions and consequent interpretation of intonational patterns. Indeed, different degrees of prominence, related to different linguistic levels, are entirely lost in this kind of approach.

Another scale type is a compromise between the other two and is in fact known as intermediate scale: it uses four different levels of perceived prominence and usually distinguishes between non-prominent, almost non-prominent, almost promi-

nent and prominent syllables (Jensen, 2004). However, results are still not satisfying.

Lately, a new methodology has been developed (cf. Samlowski, Wagner, 2016), which exploits the prosody-gesture link and tries to avoid all drawbacks of the other approaches. In this case, prominence perception is related to the beating movement: participants in the experiment are asked to listen to some short sentences and reiterate them by beating on a DrumPad, modulating the intensity of the beat in a directly proportional way. The drumming task permits an easy, intuitive processing of prosodic prominence, allowing drummers to produce a fine-grained annotation without necessarily being experts: discretizing between close categories results in an intuitive task in this case, because drummers are not confronted with the choice between very similar, close categories, but rather reproduce what they hear exploiting the prosody-gesture link; moreover, this procedure proved to be very fast, thus enabling the annotation of very large corpora in a consistent way as regards prosodic prominence. Moreover, perception of prominence patterns remains central all over the annotation task.

All things considered, the PromDrum method is in our opinion the best suited approach for the investigation of prosodic prominence perceptual patterns, and for different reasons: firstly, it does not require a long preparation phase and it allows naïve speakers to take part in the perceptual experiment of rating spoken productions; secondly, it is fast and enables the annotation of very large corpora; lastly, but most importantly, native speakers (and listeners) should be able to evaluate the degree of prominence of the data reflecting the actual functioning of perceptual processes related with prosodic prominence, allowing an examination of the complex dynamics developed along sequences of prominence peaks in relation with non-prominent, contextual units.

Summing up, this methodology seems suitable for many purposes: from our point of view, the major advantage of this approach is that it refers to a definition of prominence that is strongly intuitive and based on acoustics, but that at the same time does not leave aside the mental categorization of the phenomenon: it succeeds, then, in mixing both signal-based and expectation-based factors. For these reasons, we will use this methodology for annotating a corpus of spontaneous speech. However, we will also describe the limitations still inherent in this approach and propose an improvement on that side.

3. *PromDrum method: An issue*

As we have stated, prosodic prominence is a quite complex phenomenon with different elements interplaying at the same time on different levels. As such, a good research angle is on the perceptual side, as prominence is mainly referred to as a perceptual process, with units "*perceived as standing out from their environment*". In analyzing this phenomenon, then, an important methodological choice regards the rating scale used for examination. In the preceding section, we have presented

the most widespread ones and explained why we chose the so-called PromDrum method (Samlowski, Wagner, 2016). In this part of the paper, however, we intend to depict an issue inherent in this approach, in order to propose an improvement on this side towards the end of this contribution.

Indeed, the main problem with this approach is that the drumming procedure – with which annotators rate prominence exploiting the prosody-gesture link through an electric drum pad – leaves raters free to decide how many beats they hear. Traditionally, manually annotated corpora for prosodic prominence have used a strict correspondence of syllables and annotation units: annotators, indeed, were forced to rate the relative salience of each syllable as marked by the researchers (cf. Fant, Kruckenberg, 1989; Wightman, 1993; Streefkerk et al., 1999; Jensen, 2004). The drumming procedure has the advantage of removing this constraint: in fact, drummers are left free to produce the number of beats they consider to be the best representation of what they have heard in the input audio file. This choice, however, implies that, in our data, the drumming associated with a specific file may not contain an amount of beats that is equal to the number of reference syllables[1]. In Samlowski, Wagner (2016), authors examined only drummed sentences in which the number of expected syllables and the number of beats coincide, as their aim consists mainly in the validation of the drumming procedure. In our case, on the contrary, we are interested in exploiting the freedom left to the annotators, because in this way we believe we can further understand the connection between perceptual processes and rating procedures. Nevertheless, we still have to overcome the alignment problem between rating and drumming: in fact, once we have our input audio file and the concerning annotation, we have to be reasonably sure that a given drummed beat can be correctly assigned and aligned with a given spoken chunk. In order to display our proposal to overcome this issue, we have to present the material used for the investigation first. In the next section, we will proceed with that; after that, we will introduce our proposal.

4. *Material*

Prosodic prominence (Terken, 1991) has gained attention over the past decades. Nevertheless, the different elements related to this topic still have to be examined in a detailed way, disentangling notational problems and domain mix-up (Wagner et al., 2015). Due to these investigation issues, large corpora annotated on the prosodic level for prominence are very small in number, especially when dealing with multilingual data and L2 acquisition – with some exceptions[2]. In addition to the lack of comparability of databases, it is also hard to compare annotation methods:

[1] With the concept *reference syllables*, we refer to the syllabic units that can be expected in an utterance on the basis of lexical and/or phonematic constraints.

[2] Cf. Kohler, (1996); Campione, Véronis, (1998); Ostendorf, Price & Shattuck-Hufnagel, (1995); Oostdijk, (2000); Cheng, Greaves & Warren., (2005); Hirst, Bigi, Cho, Ding, Herment & Wang, (2013), among others.

indeed, a variety of methodological approaches as regards annotation schemata and rating scales has resulted in a low level of comparability between different works and a core notational problem (Wagner et al., 2015). Attempting an annotation of prominence can be both a difficult challenge and, at the same time, an important occasion for examining this complicated situation; in order to gain further insights on this investigation, we decided to annotate prominence from a perceptual point of view (Portele, Heuft, 1995; 't Hart, Collier & Cohen, 1990). Moreover, we decide to concentrate this study on the examination of German L1 and Italian L2, mainly because of personal competencies. However, we also find stimulating the idea of investigating perceptual processes in L2 productions: indeed, if the DTW algorithm can be applied to L2 material, too, the implementation scope of this technique would be *a fortiori* wide.

To our knowledge, at present there is only one available corpus structured for prosodic studies taking into consideration German L1 and Italian L2 together (Schettino, 2015). This database consists of 24 German native speakers producing more than nine hours of spoken speech, both in German L1 and Italian L2. Most of the speakers were university or school students, with just one teacher of Italian. Mean age was 25.6 years, with a total amount of nine men and 15 women; different levels of fluency in Italian were examined, and specifically 14 speakers of the level A, eight of the level B and two of the level C (CEFR); the diatopic variation was not taken into consideration. In the following tables, precise quantitative information is reported.

Table 1 - *Quantitative information about the corpus*

	Read speech	Commentaries	Dialogue (German L1)	Dialogue (Italian L2)
Speakers	24	9	24	24
Recording time	1h 36m 40s	50m 04s	2h 52m 18s	3h 49m 18s

Table 2 - *Additional information about the informants' fluency level in Italian L2 (CEFR)*

	A1	A2	B1	B2	C1	C2
Male	4	0	2	2	1	0
Female	7	3	1	3	0	1
Total	11	3	3	5	1	1

The elicited spontaneous productions consisted of two participants (for each session) who were asked to play TicTacToe together. The material is thus characterized for similar syntactic organization, comparable lexicon and congruous duration across files; moreover, this game is a perfect situation for analyzing prominence distribution predictability (Watson, Arnold & Tanenhaus, 2008). Participants played alternatively in German and in Italian, with a randomized sequence of languages. The starting move of the game and the first speaker were randomized, too. At the

beginning of every game, the participants were instructed on the starting move and the language they should have used for that particular game. In total, every pair of participants played eight games in Italian and eight in German, for a total amount of 16 games per couple. All games were recorded in a single session. The files were registered with high quality microphones in an anechoic chamber.

As regards the segmentation procedure, the above described "dialogues" were then segmented in speech turns. For this study, we used a sample of this segmented dialogic turns, both in German and in Italian; in particular, we used nine Italian drummers, each of them evaluating the degree of perceived prominence of 61 different turns, and three German drummers, drumming 51 turns each. Turns were selected trying to locate the files in which intonation and stress patterns differed from the norm: if a stress was put on the "wrong" syllable, or the pitch shape diverged from usual Italian[3] patterns and/or alignment, the file was considered to be a good element of investigation. In total, then, we let twelve annotators drum prominence, listening to about 700 files. In this way, we obtained 700 drummed files in which information about prominence perception and functioning is concealed. Still, we have to overcome the alignment problem with the original audio file. In the next section, we will advance our proposal for solving this issue.

5. *PromDrum and alignment: Our proposal*

As previously mentioned, prosodic prominence has been described as a complex phenomenon, in which bottom-up features and top-down knowledge are intertwined in the perception phase, resulting in a complex dynamic whose different elements and their mutual influences are difficult to be told apart. In this respect, we believe we can further understand the nature of this phenomenon through the examination of the connection between perceptual processes and rating procedures. For this reason, it is important to overcome the alignment problem between ratings and drumming. In order to do that, we develop an objective procedure that is able to both evaluate the quality of the annotations in the first place and to find the optimal alignment of drummed beats and expected syllabic units. We choose to use the Dynamic Time Warping (henceforth DTW, Sakoe, Chiba, 1978), because this algorithm does not assume that the number of units in the sequences to align has to be the same; furthermore, it reports alignment paths that minimize a given distance function. As such, then, it is compatible with both our aims of evaluating the quality of the annotations and to find the best suited alignment with respect to the reference number of syllables.

Concerning the qualitative evaluation of the annotated files, we have to bear in mind that – given the degree of freedom left to the annotators – it is possible that, in some cases, the number of beats does not exactly match the amount of ref-

[3] We do not expect German productions to be mis-produced, as speakers in this corpus are German native speakers.

erence syllables. Little discrepancies between drummed beats and syllabic units are acceptable for our analysis: we do not impose a "right" number of units that have to be recognized, but rather leave freedom of perception to the drummer. In our opinion, indeed, these little differences could be a big help in our investigation, as we regard them as possible expressions of perceptual processes, with which it would be possible to interpret the relationship between signal and perception in a more extensive way. Moreover, differences in the amount of reference syllables and beats do not represent a problem in our approach, because the DTW procedure is able to align beats and signal in a straightforward way. For example, if a drummer drums 10 beats instead of 11, the algorithm would be able to calculate to which units the beats are probably referred, and it is possible to retrace the non-drummed syllable. At this point, it becomes possible to add linguistic interpretation to the missing beat, trying to understand why that particular syllable was not perceived in the rating phase. On the contrary, drumming sequences that greatly differ from the reference ones cannot be used for further analyses and must be discarded: if – for instance – a drummer should have drummed 17 beats, but only 5 beats are found in the file, the alignment cannot be objectively reconstructed.

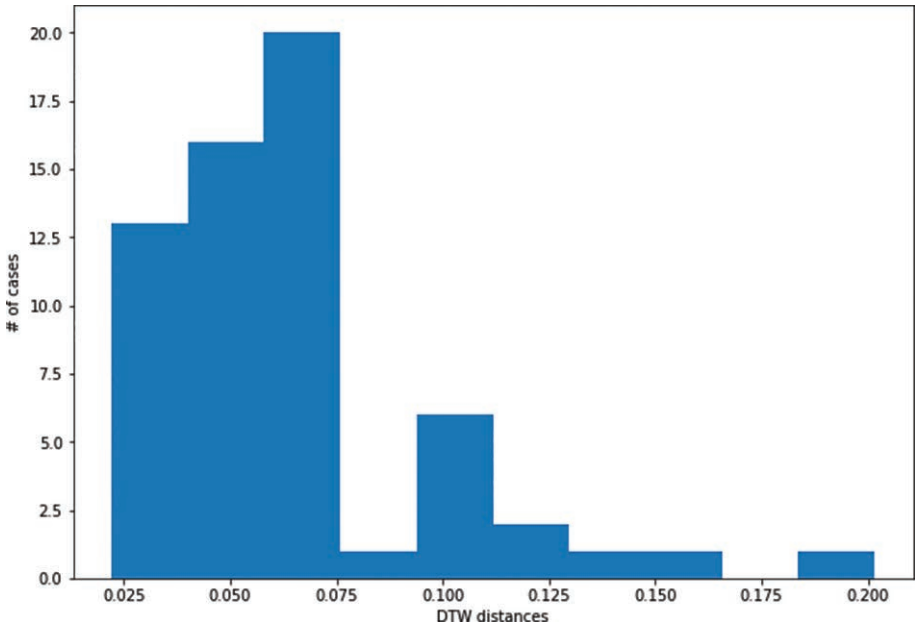
For all these reasons, we reckon that the DTW procedure applied to the PromDrum rating method can successfully improve this technique and help us improving our understanding of the phenomenon known as prosodic prominence. In the next section, specific results corroborating this assumption will be depicted and discussed.

6. *PromDrum method and DTW implementation: Our results*

In this section, results about the application of the DTW algorithm to the PromDrum rating method will be presented.

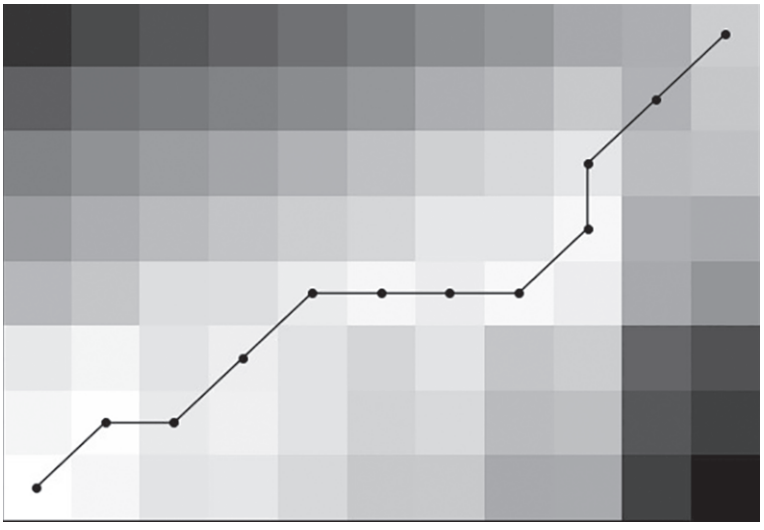
In the first place, the minimum distance provided by the DTW procedure – interpreted as the effort that the algorithm has to do in order to connect beats and reference syllables – gives us indications about the annotation quality: the less effort the system makes, the lower the DTW distances values are, and the surer we can be about the quality of the drummed sentences. In Figure 1 it is possible to observe the minimum distance value plotted along the number of cases: in this specific case, it seems that most of the sentences drummed by this drummer have a minimum distance ≤ 0.075. With this procedure, we can calculate the qualitative threshold for each drummer, leaving aside all the drummed files that diverge too much from the reference, thus being sure that the actually evaluated files have been produced with a good degree of correlation between perceived units and linguistic signal.

Figure 1 - *Minimum distance value plotted along the number of cases. Qualitative evaluation of the drummer*



Along the whole set of our drummers, we calculated that the best threshold is 0.07 in our data: most of the annotators, in fact, have produced the vast majority of acceptable drummed files under this DTW distance value, indicating that it could be considered a good qualitative limit, sufficiently strict but quite fair, too.

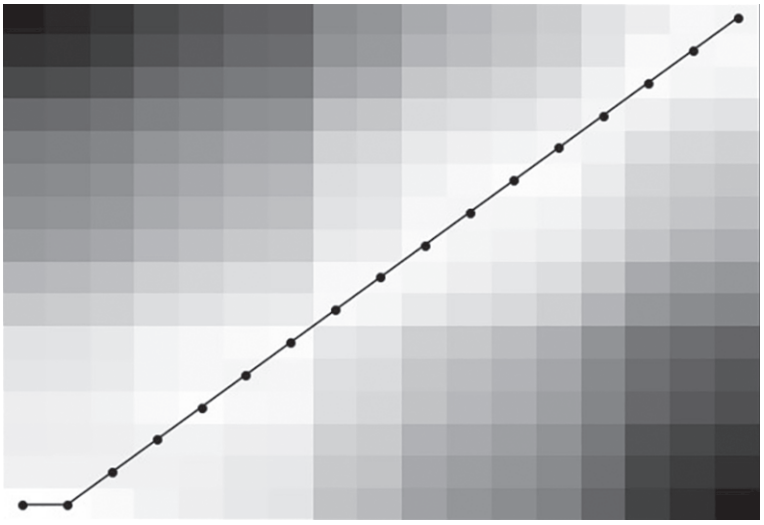
Figure 2 - *Warping path of a non-acceptable drummed file. Too big discrepancies between the reference syllables distribution and the beats one*



The second advantage of the DTW approach is that this procedure indicates what is the best alignment between the two sequences: this is named warping path. In this way, the alignment does not result from a subjective choice: it is the result of an optimization procedure that minimizes the chosen distance function.

In Figures 2 and 3, two different accounts of the warping path are reproduced: in the first case, there are consistent discrepancies between the rating annotations and the number of reference beats: as can be seen, the warping path does not find a one-to-one correspondence and the line appears to be consequently not straight.

Figure 3 - *Warping path of an acceptable drummed file. Tolerable discrepancies between the reference syllables distribution and the beats one*



In the second figure, on the contrary, the relationship between number of reference beats and actual beats is much more uniform: the only drumming hit that prevents a bijective correspondence is the first one. In this specific case, the Italian word *io* "I" – that is expected to contain two syllabic units – is drummed as a single beat: however, for the DTW algorithm it is not much onerous to recognize that two adjacent, very close syllables may be perceived as one unit and drummed accordingly; this is shown in the Figure, too: given that the colour of the cells correlates with the alignment cost – with white signalling a relative smooth and effortless alignment and black cueing almost impossible connections, we can observe that the dots relative to the first two reference syllables, although connected with one single beat, are positioned in white cells, indicating the relative ease with which they are related to the same drumming hit.

As regards the forced alignment of the beats' values and the acoustic cues, we decide to relate drummed beats with the vocalic portion of the syllable. The vowel, indeed, is the portion of speech that carries most of the relevant acoustic features; furthermore, it was proven in many studies that the quality and the distribution of

vocalic phones in an utterance is the only acoustic correlate of rhythmic categorization that seems to be valuable in comparative works (cf. Dauer, 1987; Mehler, Dupoux, Nazzi & Dehaene-Lambertz, 1996; Ramus, Nespor & Mehler, 1999; Ling, Grabe & Nolan, 2000; Grabe, Low, 2002). As the PromDrum method mainly reflects prominence perception on the rhythmic level, i.e. provides a reflection of the rhythmic perception of sequences of "strong" and "weak" units – it seems appropriate to base our examination on vocalic productions.

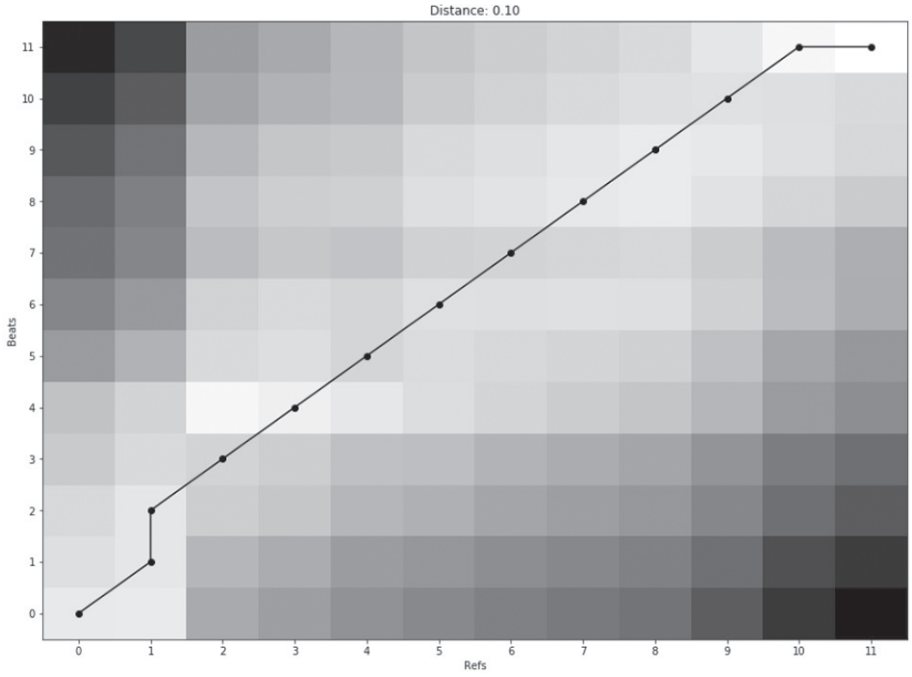
Concerning the temporal alignment of the two units (beat and vowel), we set the onset of the vowel as the point in time to which the beats are connected. The temporal alignment, indeed, may cause problems in our approach: as we mentioned, the DTW procedure reports alignment paths that minimize a given distance function, in our case, the Euclidean distance. In case we have less beaten units than vowels, then, the temporal distance between the beats will be counted as a relevant indication in the DTW algorithm in the alignment process. As a consequence, the temporal alignment of the beat with the vowel is crucial in the calculation of distance and in the following assignment of beats to vowels. However, the alignment can be carried out in a successful way if we make an important consideration: as a beat is realized faster than a vocalic breath emission, aligning the beat with – for example – the intensity peak of the vowel introduces delays due to the different speed with which the intensity curve can reach its peak. Putting the reference on the vowel onset, on the contrary, marks well the moment in which the vowel is perceived. The best way to align beaten units and vowels, then, is using the vowel onset as a fixed reference in the signal.

The obtained successive values relative to the beats sequences are normalized in the time domain: we assume that the first beat and the first vocalic onset are located at time 0, while the last units will be displaced at time 1. In this way, each file has a temporal sequence that can be aligned without too much effort in a considerably successful way. Pauses also play a role, helping in the right alignment process and disambiguating between different alignment paths.

The described procedure allows the optimal alignment without forcing us to make too strong assumptions: unlike the approach found in Samlowski, Wagner (2016), where – in order to validate the methodology – only the drummed file with an equal amount of beats and reference syllables are counted, we do not need to set a number of linguistic units that should be regarded as the right amount. The DTW algorithm, indeed, is able to prevent strict constraints in the empirical phase, avoiding the obligation of enforcing some methodological protocols due to underlying theoretical assumptions. In our case, moreover, the simple fact that reference syllables and beaten units are equal in number does not suffice in assuring a good quality of the alignments: it could well be possible, indeed, that a drummer – in the attempt of reiterating the "correct" number of syllables, concentrates in repeating the exact amount of expected units, without being successful in reproducing the rhythmic contour of the input file. In this case, the Euclidean distances between the beats would not reflect the temporal distribution of vowels in the audio file,

thus resulting in a higher degree of effort in the DTW algorithm and consequently in a badly correlated – perhaps not acceptable rated file. An example of this sort can be observed in Figure 4: although the two sequences have the same number of units in it, the alignment is not straightforward: most of the cells are grey, with the algorithm interpreting the Euclidean distances between the different units as a bad rhythmic reflection of the input file. As a consequence, this drummed sequence is not accepted, because the degree of effort in the alignment procedure exceeds the threshold of acceptability for the given drummer.

Figure 4 - *Warping path of a non-acceptable drummed file in which the number of reference syllables and does number of beats coincide, but their distribution over time do not*



This is not a drawback in our opinion: in this way, in fact, we can be sure that only the drummed files that really reflect rhythmic perceptual processes are accepted.

On the other side, we can find drummed sequences that contain a number of beats that is not at all equivalent to the reference number of syllables, but that still can be aligned in a successful way through the DTW procedure. In Figure 5, an example of this type of drumming is shown: in this case, although we have a number of beats lower than the expected reference syllables, the algorithm is capable of aligning the two sequences with not too much effort: as we can see, the points are mostly distributed in white cells. Evidently, even if the produced drummed beats are less than expected, their sequence succeeds in reflecting the rhythmic contour of the input audio file.

Figure 5 - *Warping path of an acceptable drummed file in which the number of reference syllables and the number of beats do not coincide, but their distribution over time does*

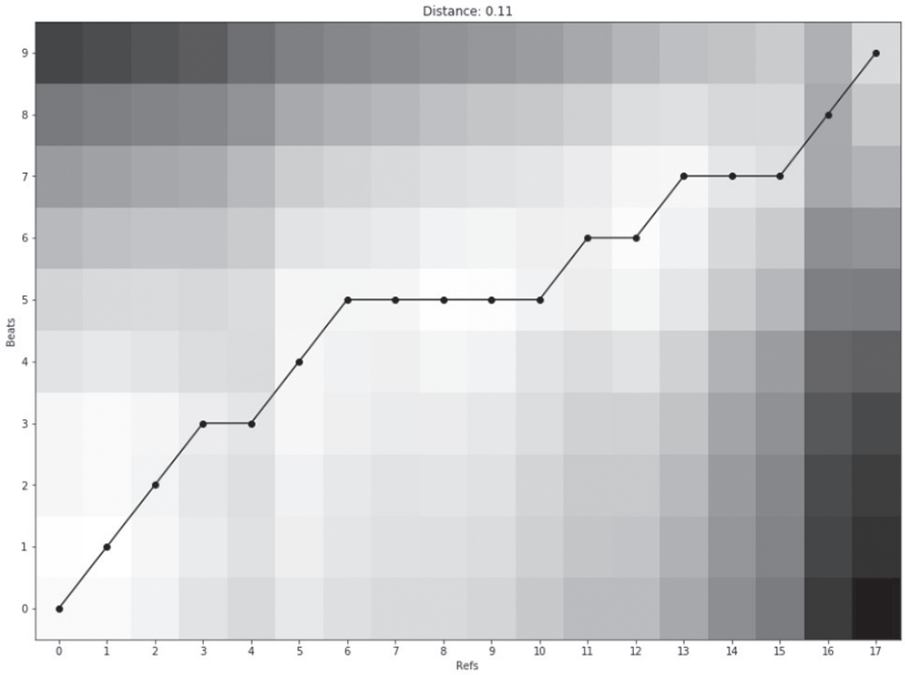
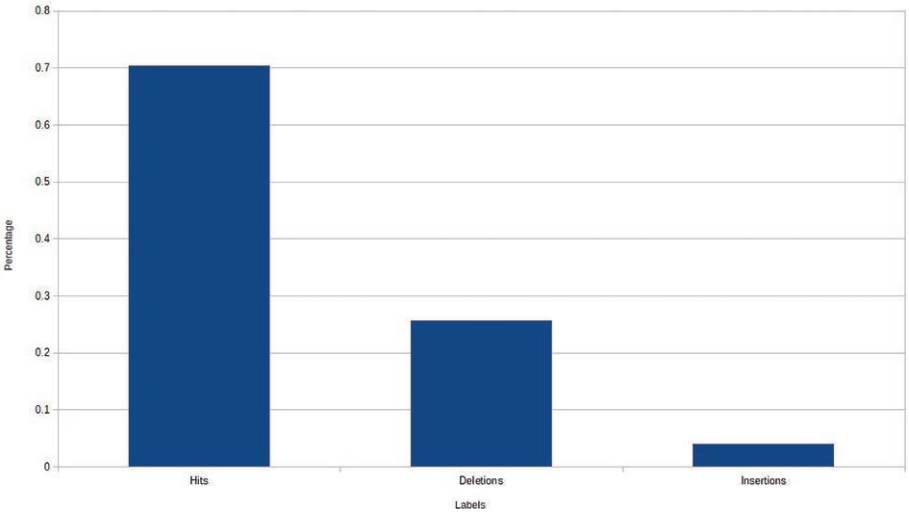


Figure 6 - *% of Hits, Deletions and Insertions in the DTW procedure*

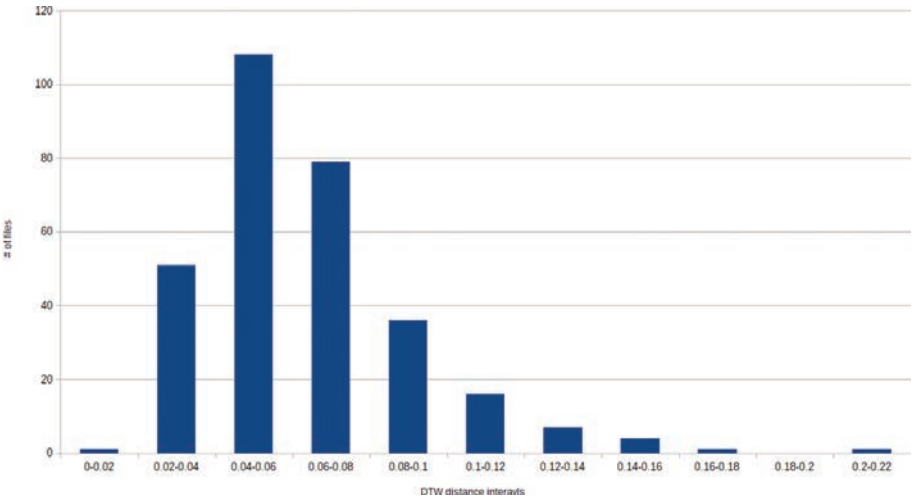


As it can be observed in the last two figures, the methodology developed in this work allows both the insertion of "non-expected" beats and the deletion of "expected" one, with respect to the reference number of syllabic units. In general, we are

expecting a great amount of beats that actually match the reference unit without alignment problems; we are defining these cases as Hits. Deletions – i.e. absences of an expected beat – are also expectable, given the methodological organization of our approach and in particular the freedom left to the drummer to rate the amount of syllables they hear; however, their influence on the analysis is not so big, because the DTW provides a qualitative level of the alignments: as a result, missing beats do not automatically result in a file ineligible for analysis. Insertions, on the contrary, are not expected to take place often: it is not probable that drummers produce more beats than what they listened to, even if it is possible – as shown in Figure 4 – that the warping path assigns two different beats to the same reference syllable.

In Figure 6, the percentage of Hits, Deletions and Insertions for the whole examined corpus is reported. Our expectations are confirmed: Hits represent the 70% of the cases, a value that further legitimates the PromDrum procedure enhanced with our methodological improvements; Deletions instances are about a quarter of the cases, whereas Insertions are really rare (less than 5% of the whole investigated data). As it can be seen, although Hits represents a vast majority of the cases, the number of files in which this correspondence does not take place is statistically significant. As a consequence, developing a procedure that allows the use of most drummed sentences is a worthwhile effort. Therefore, the DTW technique can be considered a fruitful improvement in order to obtain the best possible alignment between the beats sequence and the relative audio file.

Figure 7 - *Relationship between the number of examined files and DTW distance intervals*



A last acknowledgement about the value of this approach relies in the number of acceptable drummed sequences: if with the basic PromDrum method only those annotations could be accepted in which the number of reference syllables and the number of actual beats coincided, with the DTW technique it is possible to accept most of the annotations, disregarding only those qualitatively scarce because not

reflecting the rhythmic nature of the input audio file. As it can be seen in Figure 7, in fact, a histogram about the relationship between the number of examined files and the DTW distance intervals tells us that we are on the right track: indeed, most of the files can be found under the 0.1 distance threshold. A limited amount of files display a distance value higher than 0.12: for those ones, the effort put by the algorithm in finding the best possible alignment is too high, and they must be discarded. But the whole rest can be accepted and used for linguistic analyses and interpretation.

7. *Conclusion and future work*

In Samlowski, Wagner (2016), the prosodic prominence rating procedure admitted the possibility of annotated files having non-equivalent numbers of beats and expected syllables, but solved the issue just discarding the considered files. In this work, we develop a procedure for considering those files, too, as good candidates for perceptual analyses of prosodic prominence. The DTW algorithm suits our aim, and for different reasons: firstly, the best alignment between the two sequences, the so-called warping path, does not result from a subjective choice, but is the manifestation of objective parameters. The second advantage is that it allows us to evaluate the quality of the annotations: the minimum distance provided by the DTW algorithm – interpreted as the effort that the algorithm has to do in order to connect beats and reference syllables – gives indications in this sense. Furthermore, we can calculate the qualitative threshold for each rater, leaving aside all the drummed files that diverge too much from the reference. Little discrepancies between the amount of drummed beats and syllabic units are acceptable for our analysis; indeed, we regard them as possible expressions of perceptual processes. On the contrary, drumming sequences that greatly differ from the reference ones cannot be used for further analyses and must be discarded.

We consider this procedure as a big improvement in the drumming methodology, and in general in the rating of prosodic prominence. With this approach, indeed, it is possible to rate prosodic prominence in a simple and intuitive way, exploiting the existing link between prosody and gestures; in this way, large corpora of spoken speech can be annotated in a rather fast way. Moreover, the great enhancement connected with the DTW procedure consists in the acquittal from background assumptions: we do not need to state how many units have to be rated, nor do we need to set them manually. In this way, we reckon that the perceptual process can be investigated in a more straightforward way, together with the relationship between perceived degree of prominence and acoustic correlates. This concrete understanding can be used in a future phonological description of the phenomenon.

As regards possible improvements of this approach and particular precaution to take when using this methodology, we observed that the best quality of the analysis comes from drummed data registered with a set sound on the DrumPad whose mean pitch is attested between 150 and 250 Hz. Moreover, input audio files with

long pauses in it should be regarded as bad candidate for the examination, because long silences and their relative distance to adjacent vowels are improbable to be reproduced in a satisfying way, and the pertaining files are systematically discarded during the DTW procedure.

Summing up, we think that this procedure can ameliorate the DrumPad method, which in turn could result in a better annotation system, applicable to large corpora of spontaneous speech.

Bibliography

Campione, R., Véronis, J. (1998). A Multilingual Prosodic Database. *Proceedings of the Fifth International Conference on Spoken Language Processing*, 7, 3163-3166.

Cheng, W., Greaves, C. & Warren, M. (2005). The Creation of a Prosodically Transcribed Intercultural Corpus: The Hong Kong Corpus of Spoken English (prosodic). In *International Computer Archive of Modern and Medieval English (ICAME) Journal*, 29, 47-68.

Dauer, R. (1987). Phonetic and Phonological Components of Language Rhythm. *Proceedings of the XIth International Congress of Phonetic Sciences*, 447-450.

Fant, G., Kruckenberg, A. (1989). Preliminaries to the Study of Swedish Prose Reading and Reading Style. In *Speech Transmission Laboratory. Quarterly Progress and Status Reports*, 1-83.

Grabe, E., Low, E.L. (2002). Durational Variability in Speech and the Rhythm Class Hypothesis. In Warner, N., Gussenhoven, C. (Eds.), *Papers in laboratory phonology* 7. Berlin: Mouton de Gruyter, 515-546.

t' Hart, J., Collier, R. & Cohen, A. (1990). *A Perceptual Study of Intonation*. Cambridge: Cambridge University Press.

Hirst, D.J., Bigi, B., Cho, H., Ding, H., Herment, S. & Wang, T. (2013), Building OMProDat: An Open Multilingual Prosodic Database. *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP)*, 11-14.

Jensen, C. (2004). *Stress and Accent*. Ph.D. Dissertation, University of Copenhagen.

Kohler, K.J. (1996). Labeled Data Bank of Spoken Standard German: The Kiel Corpus of read/Spontaneous Speech. *Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP)*, 3, 1938-1941.

Ling, L.E., Grabe, E. & Nolan, F. (2000). Quantitative Characterizations of Speech Rhythm: Syllable-timing in Singapore English. In *Language and speech*, 43(4), 377-401.

Mehler, J., Dupoux, E., Nazzi, T. & Dehaene-Lambertz, G. (1996). Coping with Linguistic Diversity: The Infant's Viewpoint. In Morgan, J.L. & Demuth, K. (Eds.), *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. Mahwah: Lawrence Erlbaum Associates, 101-116.

Oostdijk, N. (2000). The Spoken Dutch Corpus. Overview and First Evaluation. www.lrec-conf.org/proceedings/lrec2000/pdf/110.pdf. Accessed 27.10.2018.

Ostendorf, M., Price, P.J. & Shattuck-Hufnagel S. (1995). The Boston University Radio News Corpus. In *Technical Report ECS-95-001*, 1-19.

Portele, T., Heuft, B. (1995). Two Kinds of Stress Perceptions. In Elenius, K., Branderud, P. (Eds.), *Proceedings of the 13th International Conference of Phonetic Sciences*, 1, 126-129.

Ramus, F., Nespor, M. & Mehler, J. (1999). Correlates of Linguistic Rhythm in the Speech Signal. In *Cognition*, 73(3), 265-292.

Sakoe, H., Chiba, S. (1978). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1), 43-49.

Samlowski, B., Wagner, P. (2016). PromDrum-Exploiting the Prosody-gesture Link for Intuitive, Fast and Fine-grained Prominence Annotation. *Proceedings of Speech Prosody 2016*, Paper 212.

Schettino, V. (2015). Prosogit: A Corpus for Prosodic Studies in German L1 and Italian L2. In *AION*, 25(1/2), 237-255.

Streefkerk, B.M. (2002). *Prominence. Acoustic and lexical/syntactic correlates*. Utrecht: LOT.

Streefkerk, B.M., Pols, L.C.W. & Ten Bosch, L.F. (1999). Towards Finding Optimal Features of Perceived Prominence. In Ohala, J.J., Hasegawa, Y., Ohala, M., Granville, D. & Bailey, A.C. (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences*, 1-7 August 1999, 1769-1772.

Terken, J. (1991). Fundamental Frequency and Perceived Prominence of Accented Syllables. In *The Journal of the Acoustical Society of America*, 89(4), 1768-1776.

Wagner, P., Origlia, A., Avesani, C., Christodoulides, G., Cutugno, F., D'Imperio, M., Escudero Mancebo, D., Gili Fivela, B., Lacheret, A., Ludusan, B., Moniz, H., Chasaide, A.N., Niebuhr, O., Rousier-Vercruyssen, L., Simon, A.C., Simko, J., Tesser, F. & Vainio, M. (2015). Different Parts of the Same Elephant: A Roadmap to Disentangle and Connect Different Perspectives on Prosodic Prominence. *Proceedings of the 18th International Congress of Phonetic Sciences*, Paper 202.

Watson, D.G., Arnold, J.E. & Tanenhaus, M.K. (2008). Tic Tac Toe: Effects of Predictability and Importance on Acoustic Prominence in Language Production. In *Cognition*, 106 (3), 1548-1557.

Wightman, C. (1993). Perception of Multiple Levels of Prominence in Spontaneous Speech. In *The Journal of the Acoustical Society of America*, 94(3), 1881-1881.

DOI: 10.17469/O2104AISV000015

ROBERTO GRETTER, MAURIZIO OMOLOGO,
LUCA CRISTOFORETTI, PIERGIORGIO SVAIZER

A vocal interface to control a mobile robot

A multi-modal interface has been integrated on a moving robotic platform, which allows the user to interact at distance, through voice and gestures. The platform includes a microphone array, whose processing provides speaker localization as well as an enhanced signal acquisition. A multi-modal dialogue management is combined with a traditional HMM-based ASR technology, in order to give the user the possibility to interact with the robot in different possible ways, e.g., for platform navigation purposes. The system is always-listening, it operates in real-time, and has been tested in different environments. A corpus of dialogues was collected while using the resulting platform in an apartment. Experimental results show that performance are quite satisfactory in terms of both recognition and understanding rates, even when the user is at a distance of some meters from the robot.

Keywords: multi-microphone signal processing, multi-modal interfaces, distant-speech recognition, spoken dialogue management, human-robot interaction.

1. *Introduction*

In human-machine communication, one of the most common dreams is about talking with a robot. In spite of recent important advances in automatic speech recognition (ASR), primarily obtained thanks to a corresponding tremendous progress in deep learning (Goodfellow, Bengio & Courville, 2016; Yu, Deng, 2015), most of the existing robotic technologies do not have the capability of conducting a human-like voice interaction (Markovitz, 2015). A fundamental problem is that in most of the cases speech recognition and spoken dialogue are treated as plug-in technologies. Speech processing and related feedback mechanisms are not part of a fully integrated framework, that manages them in a coherent manner, together with other sensing (e.g., vision, touch), sensorimotor channels, knowledge and "world" representation. Moreover, this framework is often disconnected by all the cognitive processes that are needed to obtain a fully autonomous and effective solution.

This paper describes our recent activities towards the development of a robotic platform able to interact by voice thanks to an integrated framework that includes a multi-microphone input device, a Kinect device, an Arduino based GUI interface, and a low-cost moving robotic platform.

The main focus of our work was to realize a spoken dialogue management and, inherently related to it, a scheduler to real-time process in a coherent way the various information captured by the available sensing platforms, and execute different kinds of actions, such as platform movements, navigation, voice feedback, and GUI

output. As outlined in the following, the system can access and modify its internal setup (e.g. speed), can handle low- and mid-level navigation commands, can manage an agenda, learn new information.

At this moment, the platform does not include cameras and related video processing, which would further increase the possible functionalities and skill, for a deeper understanding of the surrounding scene. Nevertheless, the robotic platform performs a lot of possible actions in a highly multimodal fashion, based on the interpretation of the acoustic scene as well as of what can be deduced from Kinect, and other navigation related devices.

An important aspect to highlight is also related to hardware: the system runs on processing units available on the platform itself. It does not exploit any connection to external computers or to remote services (e.g., such as those possible through Amazon Alexa, GoogleHome, Siri, etc.). In other words, it can be classified as a "very-low" cost solution, fully relying on off-the-shelf devices and on-board computing platforms.

Finally, it is worth mentioning that our work also aims to show how effective a robot-control can be just based on a mix of voice and gestures, with few linguistic restrictions, and from a reasonable distance (typically up to 4-5 meters) of the user from the platform. Indeed, an always-listening distant-speech interface (Wolfel, McDonough, 2009) gives a lot of flexibility to the user in real-time controlling the robot. On the other hand, robust speech preprocessing (Vincent, Virtanen & Gannot, 2018) and ASR technologies are necessary, in order to tackle environmental noise, reverberation, as well as the mechanical noise produced by the platform itself, when it moves.

The remainder of this paper is organized as follows. Section 2 provides an overview of the system, including the architecture and the user interface. Section 3 describes the corpus that was collected and the related experimental set-up that was used to develop and test the system. Section 4 provides some experimental results, while Section 5 draws some conclusions and outlines the envisaged future work.

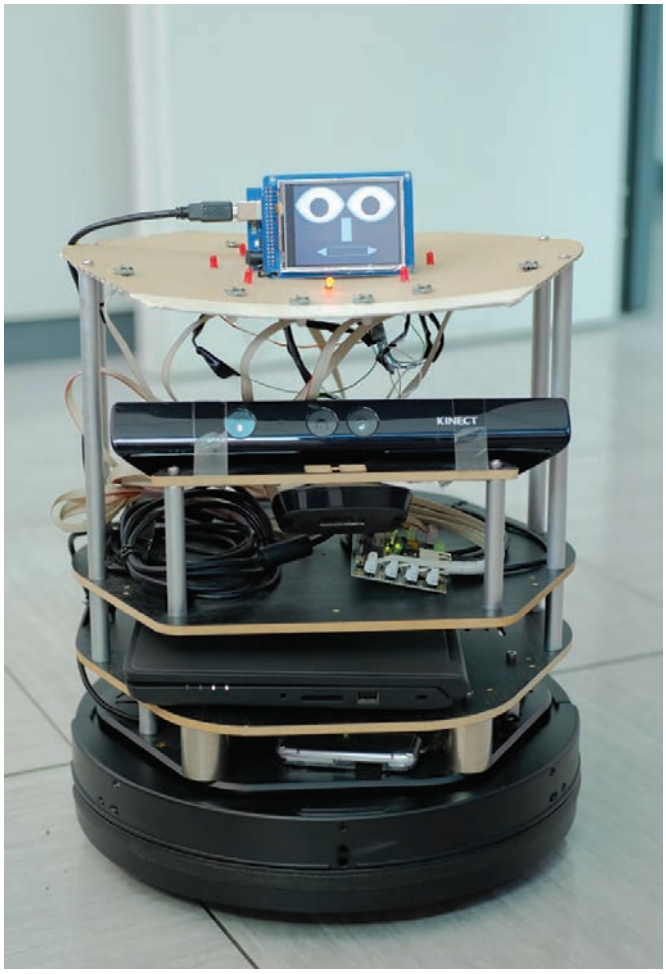
2. *System overview*

2.1 Architecture

Our robot is based on the TurtleBot2[1] platform. The basic structure is a Kobuki moving base, a Microsoft Kinect device and an entry-level laptop (eventually replaced by an Odroid XU4 in the most recent version). The Kobuki base is equipped with two motors, bumpers, encoders on the wheels and a gyroscope. To this standard setup we added eight digital MEMS microphones, some LEDs and an Arduino-based LCD screen. The robot can be seen in Figure 1.

[1] https://www.turtlebot.com/turtlebot2.

Figure 1 - *The robot equipped with all the installed devices*

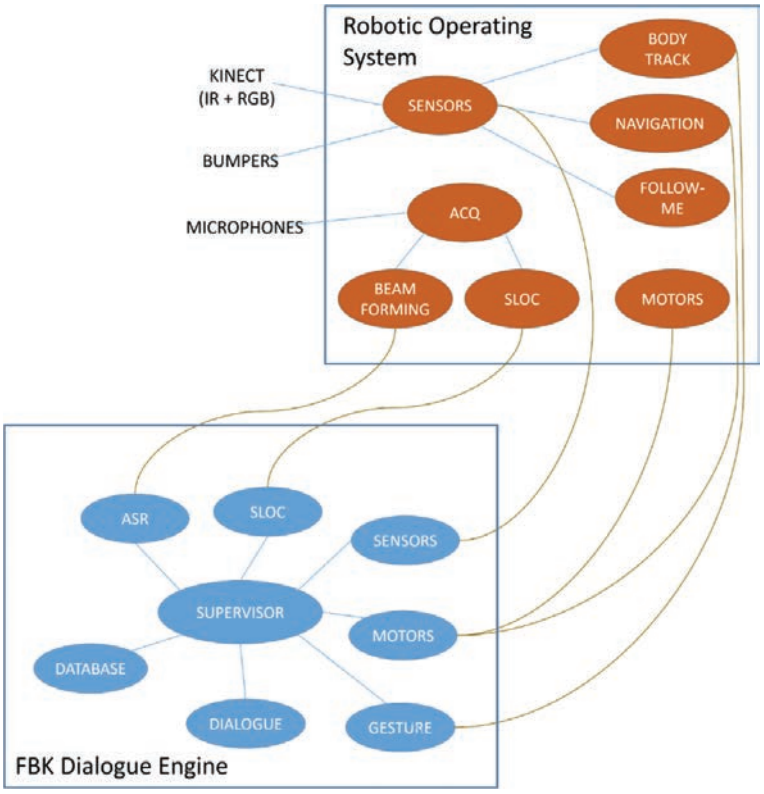


The software architecture is based on ROS[2], an open source platform for Ubuntu that helps in building robot applications. The main concept is based on software nodes that collaborate, supervised by a master. Nodes can also run on different machines to obtain a distributed architecture.

We added new nodes to handle multichannel audio acquisition, beamforming, sound source localization, Arduino LCD. In addition, other specific nodes interact with the speech recognizer and the dialogue framework. Figure 2 depicts the different ROS nodes, connected to physical devices and connected to the dialogue environment modules, described later.

[2] http://www.ros.org.

Figure 2 - *ROS nodes and dialogue modules and their interaction*



2.2 User interaction and feedback

The interaction between the user and the robot is multimodal: it can be based on speech or gestures, or both. Gestures are captured by the Microsoft Kinect installed on the robot: publicly available Kinect libraries provide a skeleton representation of the user in front of the device. This representation is processed to detect the user position and arms movements, a feedback is provided by an LCD display. The display on top of the robot represents a simple face, whose eyes track the user's head position. Nose turns red when the user moves the right hand to stop the robot, while mouth corners turn red to indicate the estimation of the direction pointed by the user. Around the display, eight red LEDs turn on to indicate the direction from which the acoustic wave is impinging the microphone array.

Speech is captured by eight digital MEMS microphones installed on a plate on top of the robot. They are distributed around the borders to have a 360-degree coverage. More details are provided in the following.

2.3 Multi-microphone signal processing

In the ideal scenario, the robot should always be ready to detect, decode and understand questions or commands uttered by a potential user and addressed to it. For this reason,

sound acquisition must always be active, such that the robot can listen continuously at what is happening in its surroundings. In practice, as we did not implement barge-in capabilities yet, the robot is always listening except when it is speaking: this condition is notified by means of a LED that becomes red, while it is green when the user is allowed to speak to the robot.

The availability of multiple microphones (8 sensors distributed in a circular pattern, with higher concentration in the frontal direction) enables the processing of the acquired signals not only by means of a time/frequency analysis, but also taking into account the spatial characteristics of the arriving acoustic waves. In particular, by evaluating global coherence field (GCF)-based acoustic maps related to the surrounding space (De Mori, 1997; Knapp, Carter, 1976), the robot is able to derive sound direction and also to estimate the distance of the source.

Once the mutual delays on the eight channels are compensated and the corresponding signals are summed up (i.e. by means of a delay-and-sum beamforming operation) the robot is also able to perform a sort of spatially-selective listening (Brandstein, Ward, 2001). The main goal is to detect human speech and isolate potentially interesting utterances.

The speech/non-speech classifier currently implemented on the platform is rather basic, as it uses only simple features: signal dynamics and a periodicity index denoting the presence of pitch within the frequency range typical of human speech. In rather quiet conditions this is already sufficient to provide a satisfactory voice activity detection (VAD). Experiments in more adverse acoustic conditions show that the use additional acoustic features such as the Mel-frequency cepstral coefficients (MFCCs) and the inter-channel spatial coherence yield improved performance especially under medium-low SNR conditions (Armani, Matassoni, Omologo & Svaizer, 2003).

An exhaustive evaluation of the VAD is outside the scope of this paper. Nevertheless, the VAD is a delicate part of the whole chain: in particular, it suffers from the noise produced during movement by the electrical motors, when the robot is moving at high speed: in these cases, the system is usually able to discard false alarms due to the constraint that each valid command must begin with a keyword ("robottino"), as will be explained later.

2.4 Acoustic models

The beamformed signal represents the input to the next stage of the speech decoding process.

The acoustic features are the traditional 13-dimensional MFCCs, augmented by first and second order temporal derivatives, and mean-normalized. The front-end processing is applied on 25 ms length windows, with a 10 ms shift. A standard HMM based recognition system is built, based on the FBK's ASR technology developed during the past years, and already used in other application fields (Brutti, Coletti, Cristoforetti, Geutner, Giacomini, Maistrello, Matassoni, Omologo, Steffens & Svaizer, 2005; Sosi, Ravanelli, Matassoni, Cristoforetti, Omologo & Ramella, 2014): fifty Italian monophone models are trained using phone segmentation, obtained automatically and manually checked. Around 1000 tied-state context-dependent triphones are then derived (with ~14000

Gaussians). For state tying, a standard decision tree based on specific phonetic questions for the Italian language is employed. Acoustic models are trained on a dataset composed of Apasci clean (Angelini, Brugnara, Falavigna, Giuliani, Gretter & Omologo, 1994), Apasci reverberated (Ravanelli, Sosi, Svaizer & Omologo, 2012), and part of DIRHA database (Cristoforetti, Ravanelli, Omologo, Sosi, Abad, Hagmüller & Maragos, 2014). The total duration of the speech dataset is about 7 hours and 23 minutes. The final model is obtained after a MAP adaptation stage with a limited amount (25 minutes) of in-domain recordings.

2.5 Finite state grammars

The language model used to perform ASR is a set of Finite State Networks (FSNs), where each transition can be either a word or the link to another FSN. In this way, the resulting grammar is in fact a context free one (De Mori, 1997).

Output labels can be assigned to both words and links, so the output of the ASR results to be a parse tree of the sentence. In order to reduce the risk of recognizing generic sentences as commands, valid commands have to be preceded by the name of the robot, which is *Robottino*.

Each FSN can be designed following different criteria: some of them, representing well defined sublanguages like numbers or time expressions, were designed by hand by means of regular expressions. Some others can be lists of words or phrases, while often N-grams are used to represent for instance the surface of the sentences (Falavigna, Gretter & Orlandi, 2000).

In principle, each dialogue state could activate a specific grammar, but in practice this possibility is used only to favour expected default data, like assigning to the utterance "25" the label *degree* after the question "give me the rotation angle". In this application, a set of about 170 FSNs is used, with a global lexicon composed of about 6000 words.

2.6 Dialogue

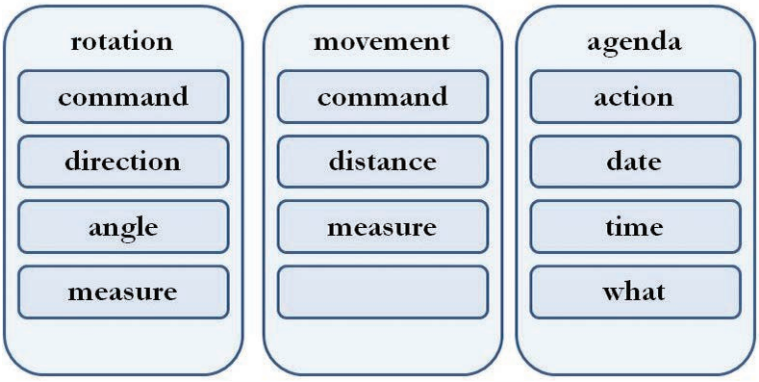
FBK built a dialogue engine in the past (Falavigna, Gretter, 1999; Giorgino, Azzini, Rognoni, Quaglini, Stefanelli, Gretter & Falavigna, 2005), which was used in several European projects (e.g., C-Oral-Rom, Homey, Vico, Dirha)[3] and commercial prototypes and systems. It is able to handle system- and mixed-initiative dialogues, and can cope with relatively complex sentences in natural language in limited domains. The dialogue component implemented in the actual version of the FBK robot is composed of the following main parts:

- engine, written in Perl (also a version implemented in C is available), which is basically an interpreter of a description, written in a Perl-like proprietary language and compiled in a Perl data structure;

[3] C-Oral-Rom (http://cordis.europa.eu/project/rcn/60720_en.html);
Homey (http://cordis.europa.eu/project/rcn/61013_en.html);
Vico (http://cordis.europa.eu/project/rcn/60714_en.html);
DIRHA (http://dirha.fbk.eu).

- description of the dialogue, which consists in:
  - a dialogue strategy – a number of structures and procedures common to many applications, that implement the philosophy of the dialogue – very briefly, there is a main loop in which the status of the dialogue is analyzed in order to decide which is the next move to do. A number of semantic concepts are organized into contexts, each one corresponding to a subdomain: the dialogue strategy has to cope with all of them, being able to change context depending on the user input;
  - a number of tools that can be easily included in the application, to handle common concepts like numbers, dates, confirmations, etc.;
  - an application-dependent part, where the semantic data necessary to model the desired domains are defined, together with their dedicated procedures. In this project, some of the contexts and concepts defined for the navigation and for the agenda are shown in Figure 3.

Figure 3 - *Some concepts related to navigation and agenda, that are normally filled by voice. They are grouped into contexts, that can be considered as subdomains. All subdomains can be activated at any time without the need for an explicit switch*

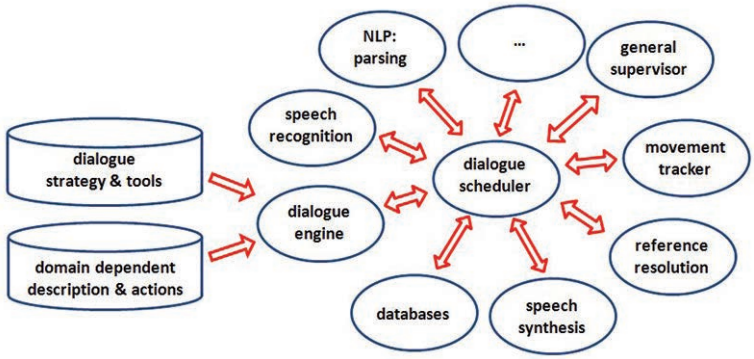


The dialogue module communicates through a scheduler – responsible only to exchange messages among processes – with several external modules, each one dedicated to some precise task. The most important processes for this application, depicted in Figure 4, are:

- speech acquisition: it gets the audio corresponding to a sentence by the user, at present in a synchronous way, i.e. the user can speak only after the prompt given by the robot. This process can be simple – just one audio channel, for instance for phone applications – or complex – in the case of the robot, 8 microphones located on the robot itself: the eight speech signals are properly processed in order to obtain a unique, enhanced, speech signal. Note that also a text input mode is possible;
- ASR + parsing: the former processes the input speech in order to get the word sequence, the latter produces a parse tree of the sentence. There must be an alignment between the labels of the parse tree and the labels of the semantic concepts defined in the dialogue description;

- speech synthesis: has to produce a synthetic speech output, starting from the word sequence produced by the dialogue;
- database: has to collect information from the "external world", for instance labels and positions recorded in past interactions, a personal agenda, etc.
- reference resolution: it keeps track of the objects that are nominated during the dialogue and in case of pronouns, tries to find out the most probable object (how much is your speed? ... augment it by 0.2 ... bring it to 0.7...), taking into account both properties of the objects (speed can be augmented but not cancelled) and distance in time (objects nominated recently are more likely to be selected).
- robot modules: there are several processes which perform speaker localization, return or set the values of some parameters (translation and rotation speed, battery level, etc.), execute navigation commands that could be simple (go forward 1 meter, turn right 30 degrees) or complex (follow me, perform autonomous navigation to reach coordinates XY), etc.
- movement tracker: based on the skeletal tracking capability of the Kinect, the robot can track the movements of the users in front of it. Some gestures with arms are recognized and used as additional directives for the dialogue manager that, besides spoken commands, is able to handle multimodal input.

Figure 4 - *Architecture of the processes connected with the dialogue. New processes can be easily added to the architecture and each component can be easily replaced with an improved version*



Newly added features to the dialogue concern the interaction with the last three modules, namely reference resolution, robot modules and movement tracker. Further, due to the characteristics of the application, the dialogue engine was updated to address the following issues:

- compound commands: the system is now able to get multiple commands (go forward 1 meter, turn left 30 degrees and then go on for 2 meters) and to properly execute them in the right order, waiting for the formers to be completed before starting the next one.
- multimodality: the system is able to get input from speech only, from gesture only (equivalent to the command "*stop*") or from a combination of the two (the command "*go there*", indicating a direction with the arm). It is also capable to detect the posi-

tion of the speaker from his voice, by exploiting the delay with which it arrives to the eight microphones. This feature is essential to be able to properly react to the command *"come here"* even if the speaker is behind the robot. At present, the two channels (speech and gesture) are completely independent and their fusion only requires a synchronization in time (i.e. the speech command and the gesture recognition must be detected within a couple of seconds, otherwise they will be ignored). Gesture processing is always active.

2.7 Language recognized by the robot

The language that the robot can understand covers some subdomains, each modeled by hand-defined grammars that allow to express in natural language the possible commands, with relatively complex sentences in a variety of ways. When a command is incomplete, a mixed-initiative dialogue helps to get the missing parameters. The most important command types are:

- basic navigation commands ("*go forward one meter and a half*"; "*turn right 90 degrees*"; *"go backward 50 centimeters"*; *"stop"*);
- multimodal commands, which merge spoken commands with information coming from various sensors (gesture detection using Kinect, speaker localization via audio processing) to be properly executed:
  - ***come here*** needs to localize the position of the speaker using his voice – he could be behind the robot;
  - ***go there + gesture*** combines speech and gesture, seen by the robot via Kinect – user must be in front of the robot;
  - ***follow me*** combines the two: based on localization via speech, first the robot will move to the speaker and then will use Kinect to follow him;
- information commands and self-awareness: the system can answer questions about the capabilities of the robot; the system can access and change some of its internal parameters (*"which is the value of your speed?"*; *"bring it to 1.0"*; *"set angular speed to 2"*; *"what can you do?"*; *"shut up"*);
- teaching commands: to provide new information in a dynamic way. The user can teach the robot a new position ("*learn, this is the entrance door"*) that will be memorized and that could be immediately used to drive the robot in that position;
- mid level navigation commands: ("*go to the entrance door"*; "*go home"*; *"go to the office of Maurizio"*) that use ROS primitives to perform autonomous navigation in order to reach the coordinate X,Y associated with the given label;
- compound commands: the system can handle a list of commands, for instance up to four navigation commands that will be executed in sequence (turn right 90 degrees, go forward one meter, turn 30 degrees to the left);
- agenda: the user can save in the agenda a new appointment, or ask for the appointments previously saved. Each appointment needs a date, a time slot, a description to be complete (*"set a new appointment for next Tuesday: phone call with John at 3PM"*; "*tell me what I have to do this afternoon"*).

The system is also able to solve pronoun resolution, useful in dialogues like:

User: *robot, how much is your speed?*
Robot: *speed is 0.4.*
User: *raise **it** by 0.2.*
Robot: *changing speed from 0.4 to 0.6.*

3. *Experimental set-up*

3.1 Corpus of dialogues

Data for the evaluation of the system were acquired in the ITEA Apartment, described below. Seven subjects performed interactions with the robot in order to execute some predefined tasks, designed in order to collect data covering most of the operations that *Robottino* could perform. Two of the subjects were involved in the design of the vocal interface of the system, the others were researchers working on other topics. Each subject first read a page describing the main features of the robot and reporting some sample commands, then performed the recordings in three sessions. He/she had to drive the robot to execute 14 tasks in different conditions (close to the robot or not, turned toward the robot or not); about 30 minutes were needed on average to perform all the tasks (min 21, max 40). Some examples of the tasks are reported in Table 1. The entire speech data sequences (total duration 3h 19m) were acquired and processed to get time markers and ASR output.

Manual correction was then performed to obtain a reliable transcription. Also a semantic representation was automatically derived and manually checked from the transcriptions. Only the pure speech commands were considered; we decided to leave evaluation of gesture and mixed commands to future works. Some samples are in Table 2.

Table 1 - *Some examples of the tasks to be executed. Each task was performed in the indicated conditions: speaker close to Robottino or not – respectively, less than 2 meters or more than 2.5 meters; talking towards Robottino or turned on the other side*

Task description	*Conditions*
Teach Robottino where the living room window is located (you must first bring it to the desired position and then tell it to memorize the position).	Robottino starts from the center of the living room. Hand-clap. Speaker close, talks towards Robottino.
Make Robottino run a 8-path between two chairs, with basic movement commands.	Speaker away, talks towards Robottino.
Ask Robottino for Friday's appointments.	Speaker close, talks to the other side.
Ask Robottino for the speed value and then decrease it by 0.1.	Speaker away, talks to the other side.
Set up a new appointment: Saturday morning at 10.00 am, trip to the museum with Paolo and Alessio.	Speaker close, talks towards Robottino.

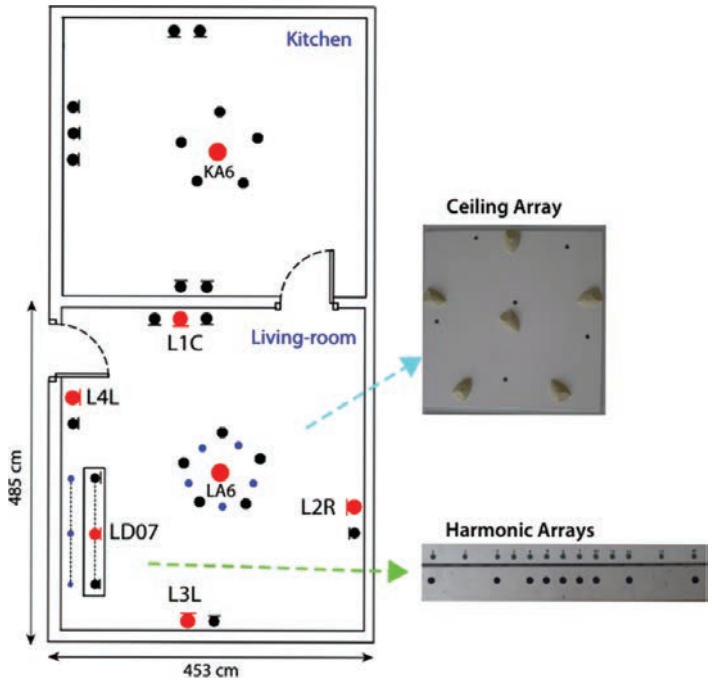
Table 2 - *Some examples of transcriptions of original sentences (in Italian) with the corresponding semantic representations*

robottino vai avanti di ottanta centimetri (FORW_CMD(vaiavanti)FORW_CMD) (FORW_DIST(ottanta)FORW_DIST) (FORW_MIS(centimetri)FORW_MIS)
robottino ruota a sinistra di novanta gradi (ROT_CMD(ruota)ROT_CMD) (ROT_DIR(asinistra)ROT_DIR) (ROT_ANG(novanta)ROT_ANG) (ROT_MIS(gradi)ROT_MIS)
aumenta= =la di zero punto due (PRN_VRB(aumenta=)PRN_VRB) (PRN_CLT(=la)PRN_CLT) (PRN_VAL(di(NUM3(zero)NUM3)punto(NUM3(due)NUM3))PRN_VAL)
robottino dimmi gli appuntamenti di domenica prossima (AG_ACTION(dimmigliappuntamenti)AG_ACTION) (AG_DATE((WEEK(domenica)WEEK)prossima)AG_DATE)

3.2 The microphone network

A data collection took place in an apartment in Trento, called ITEA Apartment. The flat comprises five rooms which are equipped with a network of several microphones. Most of them are high-quality omnidirectional microphones (Shure MX391/O), connected to multichannel clocked pre-amp and A/D boards (RME Octamic II).

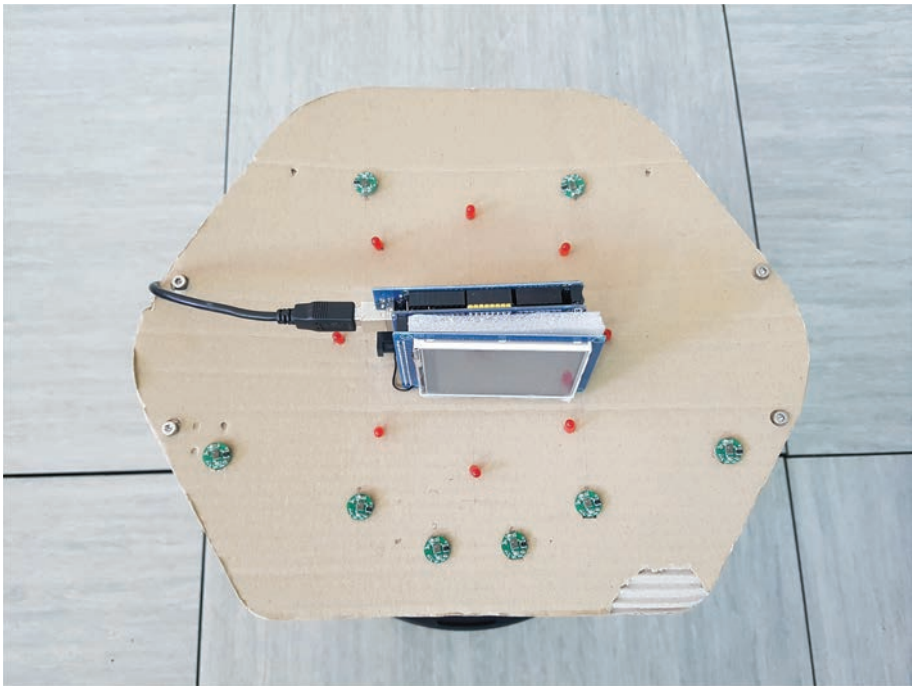
Figure 5 - *Details of the microphone network installed in the kitchen and in the living-room*



The bathroom and two other rooms were equipped with a limited number of microphone pairs and triplets (i.e., overall 12 microphones), while the living-room and the kitchen comprise the largest concentration of sensors and devices. As shown in Figure 5, the living-room includes three microphone pairs, a microphone triplet, two 6-microphone ceiling arrays (one consisting of MEMS digital microphones), two harmonic arrays (consisting of 13 electret microphones and 16 MEMS digital microphones, respectively).

Eight digital MEMS microphones are installed on the top plate of the robot (see Figure 6). Their polar pattern is omnidirectional and they are distributed on both front and back sides to acquire sound coming from all directions. Sampling rate is 16 kHz with 16 bit resolution.

Figure 6 - *Robot top plate, with MEMS microphones and LEDs*



Along with audio recordings, the interactions have been video-recorded with a digital camera. Audio-video synchronization is obtained with a hand-clap.

The experimental dataset was collected under realistic conditions, i.e. with the robotic platform moving inside this apartment. Note that it was not possible to synchronize at sample level the acquisition boards devoted to different sets of microphones, which operated at the same nominal sampling rate but with independent clocks, leading to possible time drift. Therefore, only a rough synchronization was achieved between the robot, the distributed microphones and the video recordings.

This misalignment was then compensated off-line for each utterance by means of a method based on GCC-PHAT (Knapp et al., 1976).

Beamformed signals were used to accomplish the manual segmentation and annotation of the dataset, which served then as reference in order to evaluate the whole speech interaction system.

4. *Evaluation/Results*

To evaluate the system from the ASR point of view we considered the beamformed signal derived from the acquisitions done by the robot's array in the ITEA apartment. From the whole corpus we extracted only the segments, manually checked, corresponding to sentences pronounced by the users, without considering sentences in which also *Robottino* was speaking. In total we have 950 utterances, amounting to 33:07 minutes of speech. This material was divided into development and test set, composed of 254 (09:08 minutes, development) and 696 (23:58 minutes, test) utterances, respectively.

Results are reported both in terms of Word Accuracy (WA) by considering words as units, and in terms of Semantic Accuracy (SA). In this case, and for ASR evaluation only, we consider as a semantic unit all the words and the semantic labels used: even a small error in a functional word will cause the semantic unit to be considered wrong, like the following two examples:

(AG_DATE((WEEK(domenica)WEEK)prossima)AG_DATE)
(AG_DATE((WEEK(domenica)WEEK)prossimo)AG_DATE)

(AG_WHAT(gitainmontagnaconMirco)AG_WHAT)
(AG_WHAT(montagnaconMirco)AG_WHAT)

Table 3 reports results in terms of Word and Semantic Accuracy for development and test set.

Table 3 - *Results both in terms of Word Accuracy (WA) and Semantic Accuracy (SA). Also the total number of units as well as deletions, insertions and substitutions are reported*

	Words					*Semantic Units*				
	WA	#Units	Del	Ins	Sub	SA	#Units	Del	Ins	Sub
dev	77.96%	1384	88	37	180	81.55%	542	55	9	36
test	81.73%	3366	219	97	299	85.28%	1427	71	36	103

To further understand the distribution of the errors, we joined development and test set and rearranged the whole corpus following the acquisition conditions:

- **F** vs **N**: Far (>2.5 meters) vs Near (<2 meters);
- **GO** vs **BO**: GoodOrientation (user speaks turned toward Robottino) vs BadOrientation (user speaks turned opposite);

- **NH** vs **NL**: Noise produced by *Robottino* when moving (High vs Low: the amount of noise depends on the speed of Robottino).

Table 4 reports results for the different acquisition conditions. Despite the fact that in some cases (in particular N-BO and NH) the number of utterances is too low to have statistical significance, a clear trend can be observed. The distance from the robot seems not to be a critical issue for comprehension when the orientation is good: SA drops only from 86.67% (N-GO) to 85.90% (F-GO), despite a bigger drop in WA (84.76% to 79.70%) which involves insertion and deletions of semantically irrelevant words. Much more critical seems to be the drop in orientation: 85.90% drops to 58.90% when far, 86.67% drops to 68.57% – a bit less severe – when near the robot. This is probably due to the fact that the voice impinges the microphones with less energy and there are much more reflections, which also determines lower quality of the resulting beamformed signals. Note that, for the very few sentences recorded in the NH condition, we got a really low WA (0.00%), mainly due to insertion and deletion of semantically irrelevant words; still the SA remained at an acceptable level (62.50%).

Table 4 - *Recognition results under different conditions*

	Words					*Semantic Units*				
	WA	#Units	Del	Ins	Sub	SA	#Units	Del	Ins	Sub
F-BO	57.93%	435	26	20	137	58.90%	163	52	0	15
F-GO	79.70%	2803	228	73	268	85.90%	1156	79	14	70
N-BO	71.23%	73	0	7	14	68.57%	35	6	1	4
N-GO	84.76%	1450	74	45	102	86.67%	615	16	19	47
NH	00.00%	15	0	8	7	62.50%	8	1	1	1
NL	80.05%	4746	325	133	489	84.29%	1961	143	31	134

Finally, it is interesting to report a rough indication of the performance that can be achieved by this system, in terms of speaker localization performance. Given the aforementioned on-board microphone array geometry, and the state-of-the-art techniques that were embedded in our solution, estimating the direction from which the user is speaking is relatively easy, when no other speakers or noise sources are active. In this case, we observe an average error of less than 5 degrees in the azimuth angle estimation, which normally leads to a quite accurate activation of the LED in the direction of the user.

On the other hand, a more challenging task is the depth estimation, i.e., the distance of the user from the robot, which is more difficult to evaluate due to the geometry of the array (all the microphones are concentrated in a rather restricted area on the top of the platform). The average error in depth estimation is typically of about 50 cm - 100 cm, though this strongly depends on some factors, including how

loud the command was uttered by the speaker, and how her/his head was oriented (i.e., facing the array increases the chance of obtaining a satisfactory localization).

5. *Conclusions and future work*

We described recent activities conducted by Fondazione Bruno Kessler for the development of a multi-modal robotic platform, which integrates a multi-microphone input device, a Kinect device, and a Kobuki moving base, and realizes different functionalities relying on audio and gesture input.

Spoken dialogues of variable complexity can be realized, related to robot's self-awareness information, to management of a user's agenda, to execution of commands for platform movements, as well as to navigation in the surrounding space.

A corpus has also been collected, based on user-robot interactions in a real-environment, in order to evaluate system performance.

Experimental results on this corpus and on-field sessions showed that in most of the cases the word accuracy of 70-80% is reached, which corresponds to more than 80% semantic accuracy. This performance allows the user to interact with the robot in a smooth and effective way, even when he/she is rather distant from the robot. Although some performance loss was observed when the user did not speak in the direction of the robot, this situation does not seem to be critical, since interacting with a robot always induces one, if possible, to face it (and its GUI interface).

It is also worth mentioning that the presented solution is characterized by a very low complexity, which allowed a low-cost implementation that does not require any connection with external computing platforms, clusters, or cloud-based services.

Future developments could take different directions.

A first one is based on the integration of a camera and related video processing, in order to take advantages of the complementarity between audio and video modalities, for the introduction of additional functionalities as well as for a more robust user-robot interaction performance. A first example is showed in (Qian, Xompero, Brutti, Lanz, Omologo & Cavallaro, 2018), which investigated on the joint use of audio and video input for 3D person tracking.

Concerning multimodal dialogue management, we also envisage several possible evolutions of this work.

For instance, the system is capable to handle multiple dialogues at a time: this feature was implemented for instance in the DIRHA system where in each room a different dialogue was running, being able to serve different people at the same time. This could be useful in future to allow different users to interact with the robot and, in case of audio/video person recognition, access personal information (agenda, preferences, etc.). Recently, a web interface has been implemented to be able to use the dialogue potentially from any device connected to Internet. This feature, in addition to some webcam, could allow to control the robot remotely, or in general to use the dialogue to access information from anywhere. The dialogue architecture is modular, in the sense that it is quite easy to add new processes to the

system (recently: robot modules, pronoun resolution, multimodality, etc.): this will easily allow to introduce new processes that, maybe exploring the characteristics of the environment, give suggestions to the dialogue about some move to do to improve the interaction with the user (for instance, when detecting a noise source, trying to do something to reduce it – close a window – or proposing to go near the speaker to get a better signal to noise ratio).

It is also worth noting that the collected corpus will give us the chance to explore another possible approach, which is based on integrating all the available microphone signals for speaker localization as well as for speech recognition purposes. This perspective can eventually be related to very attractive and novel application contexts, e.g., when the user is very distant from the platform, or is not facing it, a collaborative processing based on both the on-board array and the distributed microphone network will increase the overall system's robustness, i.e., robot behaviour.

Finally, the adoption of deep learning together with distributed processing will be explored for instance based on the recent paradigm introduced in (Ravanelli, Brakel, Omologo & Bengio, 2017) in order to realize an on-board small footprint deep learning based solution.

Acknowledgment

We would like to thank our colleagues Alessio Brutti, Marco Matassoni, Marco Pellin, Mirco Ravanelli and Alessandro Sosi, for various contributions provided during the realization of the described robotic platform.

Bibliography

Angelini, B., Brugnara, F., Falavigna, D., Giuliani, D., Gretter, R. & Omologo, M. (1994). Speaker Independent Continuous Speech Recognition Using an Acoustic-phonetic Italian Corpus. *Proceedings of the Third International Conference on Spoken Language Processing (ICSLP 94)*, Yokohama, Japan, September 18-22, 1994, 1391-1394.

Armani, L., Matassoni, M., Omologo, M. & Svaizer, P. (2003). Use of a CSP-Based Voice Activity Detector for Distant-Talking ASR. *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003 - INTERSPEECH 2003)*, Geneva, Switzerland, September 1-4, 2003, 501-504.

Brandstein, M., Ward, D. (2001). *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin: Springer.

Brutti, A., Coletti, P., Cristoforetti, L., Geutner, P., Giacomini, A., Maistrello, M., Matassoni, M., Omologo, M., Steffens, F. & Svaizer, P. (2005). Use of Multiple Speech Recognition Units in a In-car Assistance System. In Hüseyin, A. Hansen, J. & Takeda, K. (Eds.), *DSP for In-Vehicle and Mobile Systems*. Berlin: Springer, 97-111.

Cristoforetti, L., Ravanelli, M., Omologo, M., Sosi, A., Abad, A., Hagmüller, M. & Maragos, P. (2014). The DIRHA Simulated Corpus. *Ninth International*

Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, *May 26-31*, 2014, 2629-2634.

De Mori, R. (1997). *Spoken Dialogues with Computers*. Cambridge, MA, Academic Press, Inc.

Falavigna, D., Gretter, R. (1999). Flexible Mixed Initiative Dialogue over the Telephone Network. *Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU99)*, December 12-15, 1999, Keystone, Colorado, USA, 12-15.

Falavigna, D., Gretter, R. & Orlandi, M. (2000). A Mixed Language Model for a Dialogue System over the Telephone. *Proceedings of 6th International Conference on SpokenLanguage Processing (ICSLP 2000)*, October 16, 2000, Beijing, China, 585-588.

Giorgino, T., Azzini, I., Rognoni, C., Quaglini, S., Stefanelli, M., Gretter, R. & Falavigna, D. (2005). Automated Spoken Dialogue System for Hypertensive Patient Home Management. *International Journal of Medical Informatics*, 74(2):159-167.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep learning*. Cambridge, MA, MIT Press. http://www.deeplearningbook.org.

Knapp, C., Carter, G., (1976). The Generalized Correlation Method for Estimation of Time Delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4), 320-327.

Markowitz, J. (2015). *Robots that Talk and Listen*. Berlin: Walter de Gruyter.

Ravanelli, M., Sosi, A., Svaizer, P. & Omologo, M. (2012). Impulse Response Estimation for Robust Speech Recognition in a Reverberant Environment. *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, August 27-31, 2012, Bucharest, Romania.

Ravanelli, M., Brakel, P., Omologo, M. & Bengio, Y. (2017). A Network of Deep Neural networks for Distant Speech Recognition. *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, March 5-9, 2017. New Orleans, USA, 172-176.

Sosi, A., Ravanelli, M., Matassoni, M., Cristoforetti, L., Omologo, M. & Ramella, S. (2014). Interazione vocale a distanza in ambiente domestico. In Romano, A., Rivoira, M. & Meandri I. (Eds.), *Aspetti prosodici e testuali del raccontare: dalla letteratura orale al parlato dei media*. Alessandria: Edizioni dell'Orso, 2015.

Qian, X., Xompero, A., Brutti, A., Lanz, O., Omologo, M. & Cavallaro, A. (2018). 3D Mouth Tracking from a Compact Microphone Array Co-located with a Camera. *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, April 15-20, 2018. Calgary, Canada, 3071-3075.

Vincent, E., Virtanen, T. & Gannot, S., Eds. (2018). *Audio Source Separation and Speech Enhancement*. Hoboken, NJ: Wiley.

Wolfel, M., McDonough, J. (2009). *Distant Speech Recognition*. Hoboken, NJ: Wiley.

Yu, D., Deng, L. (2015). *Automatic Speech Recognition - A Deep Learning Approach*. Berlin: Springer.

DOI: 10.17469/O2104AISV000016

DUCCIO PICCARDI, FEDERICO BECATTINI

Voice Onset Time Enhanced User System (VOTEUS): a web graphic interface for the analysis of plosives' release phases

The paper proposes an up-to-date literature review of the works using AutoVOT, a discriminative large-margin learning algorithm developed for the semi-automatic measurement of voice onset times. In order to expand the accessibility of the tool in linguistic research, we present VOTEUS, a user-friendly graphic interface written in Python. The interface is conceived to assist the researcher throughout the whole process of annotation, from the forced alignment of the corpora to the refinement of the AutoVOT tier and the extraction of the durations. The general aim is to speed up this phase of data analysis, providing a significant improvement on prevalent practice to date.

Keywords: AutoVOT, Graphic User Interface, VOTEUS, annotation, forced alignment.

1. *Voice Onset Time: Tools for a middle-aged feature*

It has been roughly fifty years since the first description of the Voice Onset Time (VOT) as «the interval between the release of the stop and the onset of glottal vibration, that is, voicing» was proposed in Lisker, Abramson (1964: 389). Celebrating this anniversary, Abramson and Whalen (2017) wrote a retrospective essay[1] discussing the evolution of its denotation and some critical points, not without proposing recommendations on Praat (Boersma, 2001) tiers labeling in VOT research. In the last paragraph, the authors give some space to a brief recollection of tools developed for the automatic measurement of VOT, hoping that «these systems will continue to improve in the coming years» (Abramson, Whalen, 2017: 84). Particular attention is dedicated to AutoVOT, described as «the most widely used system» (ibid.: 83). In the following sections, we will describe AutoVOT and provide an up-to-date account of its applications, highlighting the necessity to broaden its audience. We will then introduce Voice Onset Time Enhanced User System (VOTEUS), a web-based interface currently in development that will facilitate the usage of AutoVOT, also integrating other functionalities for VOT annotation.

[1] Abramson, Whalen (2017) leads the way to a special VOT issue of the 2018 *Journal of Phonetics*, dedicated to theoretical and experimental aspects of voicing contrasts (Cho, Docherty & Whalen, 2018).

1.1 AutoVOT: Procedures and performances

AutoVOT is a discriminative large-margin learning algorithm for VOT semi-automatic measurement originally developed by Morgan Sonderegger and Joseph Keshet (2012)[2], and subsequently integrated in a software written in Python and based on declarative programming (Keshet, Sonderegger & Knowles, 2014). While first thought for the annotation of voiceless plosives, AutoVOT expanded its agenda to work with prevoiced plosives (i.e. negative VOT; Henry, Sonderegger & Keshet, 2012) and preaspirated plosives (Sheena, Hejná, Adi & Keshet, 2017). AutoVOT is compatible with *.wav files (16 kHz mono) and Praat textgrids (*.TextGrid). The algorithm can be used to train models providing *.wav files with hand-measured textgrids containing a common label (e.g. *vot*) as input. The trained model can later be applied to new *.wav files matched with textgrids structured with a tier with aligned intervals in order to segment the contained VOTs. The recommended, optimal tiers should not include more than one stop consonant and should begin 50 ms before the stop burst or 30 ms before the entire segment. Eventually, AutoVOT predictions should be checked and adjusted by a human annotator. Among the studies that made use of AutoVOT (see below), few actually reported precise information on its performance. In this aspect, Stuart-Smith, Sonderegger, Rathcke & Macdonald (2015) provides the most comprehensive picture and will be here summarized as an example of the algorithm's potentials and problematics.

1) The aligned tiers were automatically generated, and not subsequently modified;
2) the authors provided two different small training sets (100 VOTs from each of five analyzed speakers per set) to generate a model for voiceless plosives and one for voiced plosives;
3) after the application of the models to the inquired corpus, the phase of evaluation and correction of AutoVOT predictions had an astounding 1:1 ratio between the actual duration of the annotated file and the time of human adjustment;
4) the variable ANNOTATOR in the statistical analysis did not hold significance, hinting to the good quality of the semi-automatic measurement;
5) the miscellaneous quality of the recordings contained in the inquired corpus did not alter the effectiveness of the algorithm;
6) a total 2564 predictions were labeled as "not usable" (21,6%; 1736 voiced stops, i.e. 29,8% and 828 voiceless ones, i.e. 7,9%), 5860 as "correct" (62,6%; 3171 voiced stops, i.e. 54,4% and 2689 voiceless ones, i.e. 76,2%) and 1474 as "corrected" after the phase of human adjustment (15,8%; 916 voiced stops, i.e. 15,7% and 558 voiceless ones, i.e. 15,8%)[3].

[2] A first attempt by the two authors to tackle the issue of automatic VOT measurement can be read in Sonderegger, Keshet (2010).

[3] The reported percentages refer to parts of the total number of analyzed tokens (9898; 5823 voiced and 4075 voiceless stops; see below). A prediction was coded as "not usable" in the case of alignment or transcription errors, sounds overlapping to the token production or variation phenomena hindering

1.2 Literature review

In this section we will provide a review of all the linguistic research and ongoing projects[4] reporting the use of AutoVOT and indexed as such in Google Scholar[5]. Starting with studies on corpora of read speech, Chodroff, Godfrey, Khudanpur & Wilson (2015) applied AutoVOT on a total of 68000 tokens produced by 129 American speakers. The authors searched for /b d g p t k/ VOT variability in a corpus thought to be quantitatively appropriate for observations at both the talker-specific and the population level. Results showed indeed significant differences in individual productions, such as the entities of the effect of stop category or speech rate on VOT lengths. The authors also found that the individual, within-category durational means and standard deviations were consistently connected, and that VOT lengths were strongly correlated across the stop categories elicited in individual productions, pointing to structured variability of VOT patterns[6].

Bang, Sonderegger, Kang, Clayards & Yoon (2018) explored the topic of a sound change regarding Seoul Korean aspirated plosives through the analysis of 6849 intonational phrase-initial stops elicited by 118 speakers and contained in an apparent-time corpus[7]. The study confirms the previously retrieved distribution showing that the female speakers are leading the substitution of VOT length with f0 patterns as primary phonological cue of the aspirated series; moreover, the change has slowed down in recent years, hinting to its near completion. The frequency of a word is positively correlated with both the degree of VOT reduction for aspirated plosives and f0 contrast enhancement; since this last result is contra-

---

the realization of the token as a plosive.

[4] From the moment that the main goal of this section is to create a broader understanding of the potentialities of the tool in actual linguistic inquiries, we will exclude from the review the project papers stating the intent to integrate AutoVOT in other tools for linguistic analysis, such as McAuliffe, Stengel-Eskin, Socolof & Sonderegger (2017). In the review, we will focus our attention on the results of the experiments concerning VOT measures, with the *caveat* that VOT is not always the only, nor the main feature analyzed in the reported research.

[5] We are well aware that the transparency practices concerning the use of software in linguistic research are not homogeneous among the different subfields of the discipline. However, phonetics reportedly values the quotation of the equipment (Berez-Kroeker, Gawne, Kung, Kelly, Heston, Holton, Pulsifer, Beaver, Chelliah, Dubinsky, Meier, Thieberger & Woodbury, 2018: 9-10) so that we hope that our search will result exhaustive.

[6] Chodroff, Wilson (2017) confirmed these results comparing the data with hand-labelled productions in a laboratory setting, while increasing the number of tokens annotated with AutoVOT (88725 tokens, 180 speakers). In this study, the authors present an extensive discussion on the implications of these outcomes for structural constraints on phonetic systems and perceptual adaptation. Finally, Chodroff, Wilson (2018) replicated the findings through the semiautomatic annotation of 96357 VOTs from the same corpus, also describing a similar structured variability for plosives' center of gravity and onset f0 in the following vowel.

[7] Partial results from this study (5888 tokens) can be found in Bang, Sonderegger, Kang, Clayards & Yoon (2015). In Bang (2017) the data is further compared to corpus-retrieved American English (126 speakers, 4208 tokens) and German (118 speakers, 2660 tokens) read speech.

ry to typological expectations, the authors suggest that the f0 enhancement is an adaptive change to the VOT reduction. On the other hand, the presence of a subsequent high vowel inhibits these processes, suggesting that general coarticulatory lengthening mechanisms could have conditioned the modalities of the change. Finally, intrinsic f0 vowel differences after voiceless plosives are dampened as the phonological f0 distinctions arise over time. These results are discussed in relation to a potential Seoul Korean tonogenesis.

The same research topic is investigated in Cheng (2017) from the point of view of South Californian Heritage Korean. 32 speakers were recruited to read a set of 35 words presenting fortis/lenis/aspirated stops and affricates minimal pairs. Participants were classified in three generational levels, corresponding to different times and modalities of exposure to Korean and American English. The set was read in a fixed carrier phrase and in more naturalistic sentences, resulting in a total of 2240 tokens. The tonogenetic shift was well represented by the first-generation speakers, while the second-generation ones seemed to use primarily VOT lengths to express phonological distinctions. Still, a perceptual counterpart focused on language attitudes showed that this difference alone cannot be considered a marker of linguistic proficiency for South Californian Koreans. The results are compared to similar tendencies described in other studies and interpreted in the light of potential attrition with American English.

Together with this last study, Schertz, Kang, & Han (2017) is of particular interest to the aims of this section for successfully applying AutoVOT to different consonantal typologies. The authors gathered 11121 productions of Korean and Mandarin sibilants and affricates from an isolated-words reading task proposed to 107 bilingual speakers from the two Chinese prefecture-cities of Hunchun and Dandong, located at the border with North Korea. After analyzing the VOTs of all the tokens for the phonetic description of the phonological categories of the two coexisting systems, a subset of corresponding sounds is further compared to observe the intertwined participation of the two languages to potential sound changes. In regard of the corresponding affricates, results show that older interviewees equate the Korean and Mandarin VOT values, while peculiar trends can be observed in their younger counterparts. In Dandong, young speakers present a Seoul-like tonogenetic tendency in their Korean production, leaving the Mandarin tokens unaffected. In Hunchun, this demographic group has shorter VOTs in both Korean and Mandarin productions; however, no VOT merger is observable in the Korean phonological categories, being the change probably Mandarin-driven.

Singh, Keshet, Gencaga & Raj (2016) tackled the debated issue of VOT-physical age patterns, grounding their results on unprecedented quantities of tokens[8] and observed speakers (630, American English). The authors make a successful use of AutoVOT in also predicting the Voice Offset Times contained in the corpus, i.e.

[8] It should be noted that the exact number of VOTs is not stated in the paper. The authors report that all the stop plosives (/b d g p t k/) were represented at least once for each speaker in the corpus. We can infer that the study applies AutoVOT *at least* to 7560 tokens.

«the duration between the cessation of voicing in a voiced phoneme, and the onset of the burst of the subsequent plosive sound» (ibid.: 2). Contrarily to previous results, both these features do not show significant correlations with speakers' age. The accuracy of the annotation method is believed to be an important factor in determining the outcome of the research.

Chen, Xiong & Hu (2018) preferred a real-time approach to this same research topic. The authors extracted 1001 voiced and 2297 voiceless plosives from 40 recordings of the Christmas speeches by Queen Elizabeth II ranging from 1953 to 2016. While controlling for potentially conditioning linguistic factors, the researchers observed a declining trend in the amplitude of annual fluctuations in VOT productions, parallel to a similar tendency in VOT mean values. These findings are tentatively interpreted as dependent from physiological factors of vocal aging.

Goldrick, Keshet, Gustafson, Heller & Needle (2016) studied VOT durations of the slips of the tongue occurring during tongue twisters. 34 American English speakers were invited to read the materials in time to a metronome. The twisters were composed by monosyllabic stimuli selected to differ just by the sonority of the first plosive, with four different typologies (ABBA, BAAB, ABAB, BABA). 68000 tokens were segmented with AutoVOT. The slip of the tongue was defined as a deviation from a normal VOT duration in the direction of the other member of the twister (e.g., /b/ with long VOTs and /p/ with short releases). The switching patterns (ABBA, BAAB) showed errors with smaller degree of VOT deviations than those from the alternating ones. Moreover, erroneous productions had a higher degree of variation than the correct ones. The authors discuss their acquisitions in the light of the two proposed explanatory factors for this kind of speech errors, i.e. planning and articulatory processes, finally suggesting an integrated account.

Coming to the analyses of large corpora of spontaneous speech, Stuart-Smith, Sonderegger, Ratchke & Macdonald (2015)[9] studied 9898 voiceless and voiced[10] plosives uttered in Glaswegian vernacular by 23 working-class women. Two clusters of recordings were taken into account, one from the 1970s and the other from the 2000s; three age groups were established per cluster, searching for proofs of a historical process of VOT lengthening in the inquired variety. Elderly speakers from the 1970s had significantly shorter VOTs than their younger counterparts; in the 2000s, the situation is reversed, with the least pronounced aspirations uttered by the youngest speakers. These two different directions underline a sociophonetic potential of VOT related to speakers' age in Glaswegian. Moreover, the fact that middle-aged and old speakers from the 2000s showed longer VOTs than their respective age groups from the 1970s seems to suggest a real-time lengthening. The aberrant results from the youngest group from the 2000s is tentatively explained in

[9] Stuart-Smith, Ratchke, Sonderegger & Macdonald (2015) reported preliminary results from 12 speakers and 6125 tokens, reduced to 3012 reliable measures.

[10] As of 2015, the algorithm for negative VOTs (Henry et al., 2012) did not prove to be reliable; as consequence, all the voiced plosives observed in this study had positive VOTs. This problem seems to be somehow resolved at the time of Solanki (2017) (see below).

reference to a reported tendency of this demographic cluster to follow vernacular patterns of speech: in this case, Scots is known for its short VOTs, whereas Scottish Standard English has longer values, more similar to Anglo-English.

Sonderegger, Bane & Graff (2017)[11] took 25584 VOT durations from the speech of twenty participants to a British reality television show produced over more than fifty consecutive days. The research goal was to describe individual speech dynamics in medium term. Time dependence was pervasive in the productions of all the participants; in particular, by-day variability was the norm, while time trends interested around half of the observations. This result somehow conciliates the concepts of individual dynamicity in short-time and individual stability in long-time, from the moment that not all the daily fluctuations have the potential to become consistent change. Moreover, the study bore little evidence of overall convergence over time between the productions of the participants, challenging the assumptions of change by accommodation theories. However, a consistent convergence between the values of two participants after their romantic engagement was observed, hinting to the fact that strong social bonds represent a determinant factor in such dynamics. Finally, the participants showed different estimates of phonetic plasticity: the authors suggest that this parameter is central in determining the role of a speaker as innovator or early adopter in language change.

Two Ph.D. dissertations cited AutoVOT for the segmentation of semi-spontaneous productions in laboratory tasks. Turnbull (2015) used the algorithm to segment 1748 VOT tokens derived from an experiment with 19 participants, in the framework of listener-oriented accounts of predictability-based phonetic reduction. The participants sat in front of a screen, that showed highlighted words beginning with a plosive. The task was to instruct a confederate to click on the same word on his other screen, without the possibility of directly viewing it but being instructed that the two supports were showing the same elements. The researcher tested for the effect on VOT length of stop place, context (the fact that the word had no minimal pairs, or had minimal pairs without competitors on screen, or had minimal pair with competitors on screen), phonological neighborhood density, log frequency and individual scores to assess the extent of the Theory of Mind of the participants. Surprisingly enough, only the place of articulation had a significant effect, probably due to the choice of the experimental materials.

Solanki (2017) studied speech accommodation in live conversation in a laboratory setting. 12 female participants from Glasgow were recruited and paired to verbally interact in front of two separated screens with the aim of finding a number of small graphic differences between elements placed in three different scenarios. A total of 14494 (negative and positive) VOTs was retrieved during these sessions. Results showed that neither the previous production of the interlocutor, nor the position of the interaction in the course of the experiment had a significant effect

[11] Preliminary VOT results can be found in Bane, Graff & Sonderegger (2010) (circa 800 manually segmented VOTs), Sonderegger (2012) (6494 tokens, from manual annotations and automatic measurements) and Sonderegger (2015) (20822 tokens analyzed with AutoVOT).

causing convergence. However, interaction length proved to be a significant factor in VOT accommodation. The author infers that, while the time of communicative contact does not imply phonetic convergence *per se*, the content of the contact is crucial in assessing the will of cooperating. In this case, longer interactions were a signal of more difficult tasks, triggering convergent behaviors.

Finally, two research projects are planning to implement AutoVOT for the analysis of spontaneous productions. Chen, Kozbur & Yu (2015) transcribed fifteen years of oral arguments (1998-2013, 975 hours of recordings) that took place at the U.S. Supreme Court for the sake of analyzing speech accommodation phenomena. While preliminary results are available for vowel formants, the authors will also check for convergence in VOT values. Singh, Raj & Gencaga (2016) lists the voice onset time among those "stable" sub-phonemic features that could be of help in the field of forensic anthropometry from voice. The idea is to automatically extract VOT values from short audio segments involved in criminal activities, such as hoax calls (Singh, Keshet & Hovy, 2016), to infer physical features of the culprit facilitating the process of profiling.

1.3 Discussion

The proposed review highlights the versatility of the tool, that proved its usefulness in disparate research conditions (from real-time to apparent-time corpora, from lab speech to spontaneous and read speech), for very diverse corpus dimensions (from 1748 to 96357 VOTs) and topics of investigation (individual differences, sociophonetic values, interactional processes etc.). A concept recurring in the summarized studies is that automatic segmentation procedures are a key factor for exploring new nuances of the VOT feature, and grounding previous results on more adequate quantities of observations. However, in six years since Sonderegger, Keshet (2012), the algorithm was adopted in just 13 research projects, including Ph.D. theses and ongoing works. In addition to that, the direct involvement in 6 of these research of one of the original authors of AutoVOT definitely catches the eye[12]. Among the many factors that could explain these numbers, our take is that the level of accessibility of the technology at hand should not be taken lightly, especially in a field that dwells in deeply-rooted interdisciplinarity. Linguists have to master a wide variety of competences, both humanistic and scientific. The lack of expertise in one of its essential components results in being detrimental to the field itself[13], e.g. precluding the access to convenient tools. One possible solution resides in the development of user-friendly graphic interfaces apt to lighten the burden of specific tasks on research projects. The academic community is already working in this direction, with the planning of web-based interfaces such as DARLA (Reddy, Stanford, 2015) for semi-automated forced alignment and vowel extraction. In particular, the project

[12] In line with this fact, Solanki (2017) was written under the supervision of Stuart-Smith at the University of Glasgow.

[13] On this topic, see e.g. the informal sarcasm chosen by Foulkes (2015) to describe sociophonetics at the *ICPhS* dedicated session.

of Visible Vowels (Heeringa, Van de Velde, 2017), an internet tool for vowel plotting, normalization and analysis of dynamic features, puts great emphasis on its user-friendliness and accessibility (ibid.: 4034)[14]. It is from these premises that we present here Voice Onset Time Enhanced User System (VOTEUS), an open-access framework for semi-automatic VOT annotation of speech corpora.

2. *VOTEUS*

The framework Voice Onset Time Enhanced User System (VOTEUS) that we present in this paper is a tool intended to bridge the gap between the linguistic and computer science knowledge domains in the usage of AutoVOT. Our goal is to provide an intuitive interface to configure and run the algorithm on large speech corpora, without requiring any computer programming skills and therefore allowing everyone to exploit all AutoVOT's capabilities. More than that, we provide a set of functionalities that guide the user to easily generate fully annotated VOT datasets starting from raw speech recordings and their transcriptions. In particular we integrate a forced alignment routine to provide initial speech segments on which AutoVOT can be applied and we developed an intuitive interface within VOTEUS to manually refine the predicted VOT tiers, in order to produce high quality annotations. Furthermore, VOTEUS has been developed keeping in mind a modular software structure. This allows it to be extended and integrated with additional methods for detecting VOT and possibly compare them with AutoVOT. In the following we provide an overview of VOTEUS' architecture and organization and explain its main use cases, namely corpus inspection, semi-automatic annotation and training AutoVOT models. VOTEUS is currently under development for Linux and Windows operating systems and is going to be released under the MIT license, therefore allowing users to include and modify its source code within other projects. An alpha release of VOTEUS is scheduled to be released in early 2019. Code and installation guide will be available for download at the following link: https://github.com/fedebecat/VOTEUS.

2.1 Architecture

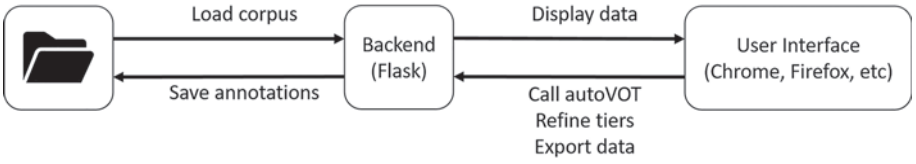
VOTEUS is organized into two software components: a backend that integrates and extends the functionalities of AutoVOT and the actual graphical user interface (GUI) for controlling and invoking these functionalities. This architectural choice keeps the control logic separate from the user interface, which reflects in a better code maintenability and implements the concept of separation of concerns by encapsulating different logical modules into separated software components[15]. We de-

[14] This point was also firmly made during a presentation of Visible Vowels held by Van de Velde at the Scuola Normale Superiore of Pisa (28/04/17).

[15] The concept of Separation of Concerns was initially introduced in Dijikstra (1982) and nowadays is at the basis of the most diffused architectural design patterns such as Model-View-Control (MVC).

veloped our backend framework in Python, by creating wrappers for AutoVOT and integrating them into a Flask[16] webserver which exposes the GUI in the form of an interactive web page, that runs in the browser. We adopted a web-based solution developed in Python more than integrating our system into existing tools or external engines such as Praat (Boersma, 2001) to focus on portability and diffusion among inexperienced users. Moreover, the advantages of this choice are twofold, in the one hand we provide a familiar environment to the user, minimizing the cognitive burden required to learn how to utilize the system, on the other we could exploit the vast resource of available libraries and toolkits for web development available online. The resulting user interface has been developed in JavaScript, largely exploiting jQuery[17] and the wavesurfer.js[18] and Google Chart Libraries[19]. All styling materials have been taken from the resources of materializecss[20]. Whereas our framework is developed as a web-based interface, we propose VOTEUS as a standalone application to be run on personal computers/workstations and not as a remote application accessible online, since users would need to upload and store large amounts of data for their corpora. At the same time it should be noted that remote access to interact with a VOTEUS instance could be easily enabled. To maintain compatibility with others systems we rely on the same data representation formats used by AutoVOT, in particular we store all audio annotations in textgrid files. VOTEUS is thought to handle different speech datasets that can be added to the framework simply including a folder to its search path. Also in this case we refer to AutoVOT specifications for input files (see above). In Figure 1 a schematic representation of the main modules and functionalities of our system is shown, depicting how the interface interacts with the data through the backend.

Figure 1 - *Schematic representation of VOTEUS' architecture. Data is stored on the disk and read by the backend. The interface allows the user to browse the data and call the functions exposed by the backend. The generated annotations are then saved back on the disk by the backend*



[16] Flask is a Python based micro-framework for developing web applications (http://flask.pocoo.org/).
[17] https://jquery.com/.
[18] We used the wavesurfer.js library (https://wavesurfer-js.org/) in combination with the spectrogram plugin (http://wavesurfer-js.org/example/spectrogram/).
[19] https://developers.google.com/chart/.
[20] http://materializecss.com/waves.html.

2.2 Corpus inspection

The simplest functionality offered by VOTEUS is to display large speech corpora in an aggregated fashion, in order to interactively inspect all the available annotations. Figure 2 shows how this is presented to the user through the interface. Once a corpus is loaded in the interface, the user can navigate through all the *.wav files and study their waveforms and spectrograms. A temporal representation of all the available tiers in the textgrid annotation file is shown and the user can highlight the correspondent interval in the audio representation (waveform and spectrogram) by simply clicking on the tier of interest. If multiple types or tiers are present in the textgrid, they are stacked inside the interface and color coded for simple inspection. Figure 3 depicts three different details of the interface, showing how the user can interact with the annotations by clicking on the tiers. The selected audio file can also be reproduced, both in its entirety or focusing on specific tier intervals. For long audio files, the waveform and spectrogram can be zoomed-in and out to examine the details of the recording at a fine-grained level.

Figure 2 - *Main Graphical User Interface of VOTEUS. When a corpus has been loaded, the user can browse its files and display the corresponding waveform and spectrogram. All available annotations are shown in the timelines below. The user can interact with the tiers to highlight or listen specific audio segments. The buttons in the lower part of the GUI can be used to call some of the functionalities of the backend.*

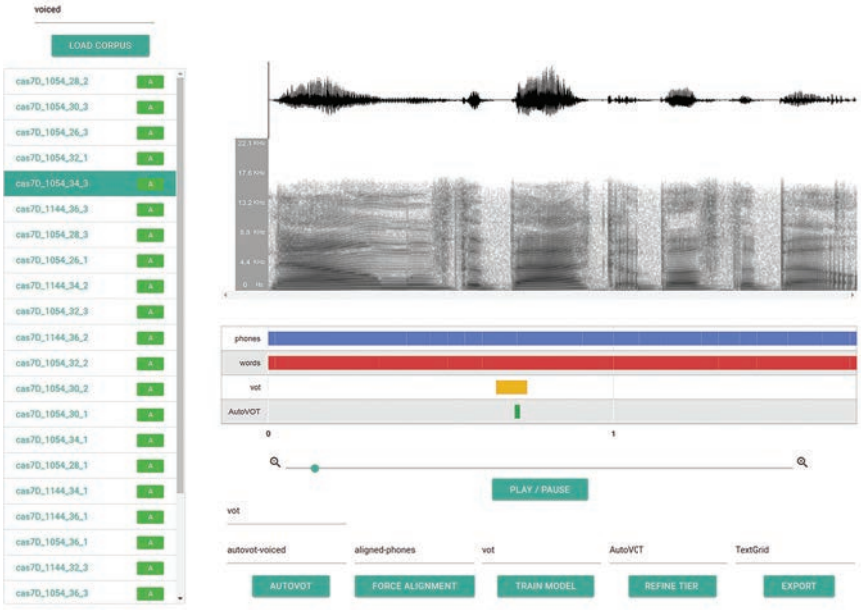
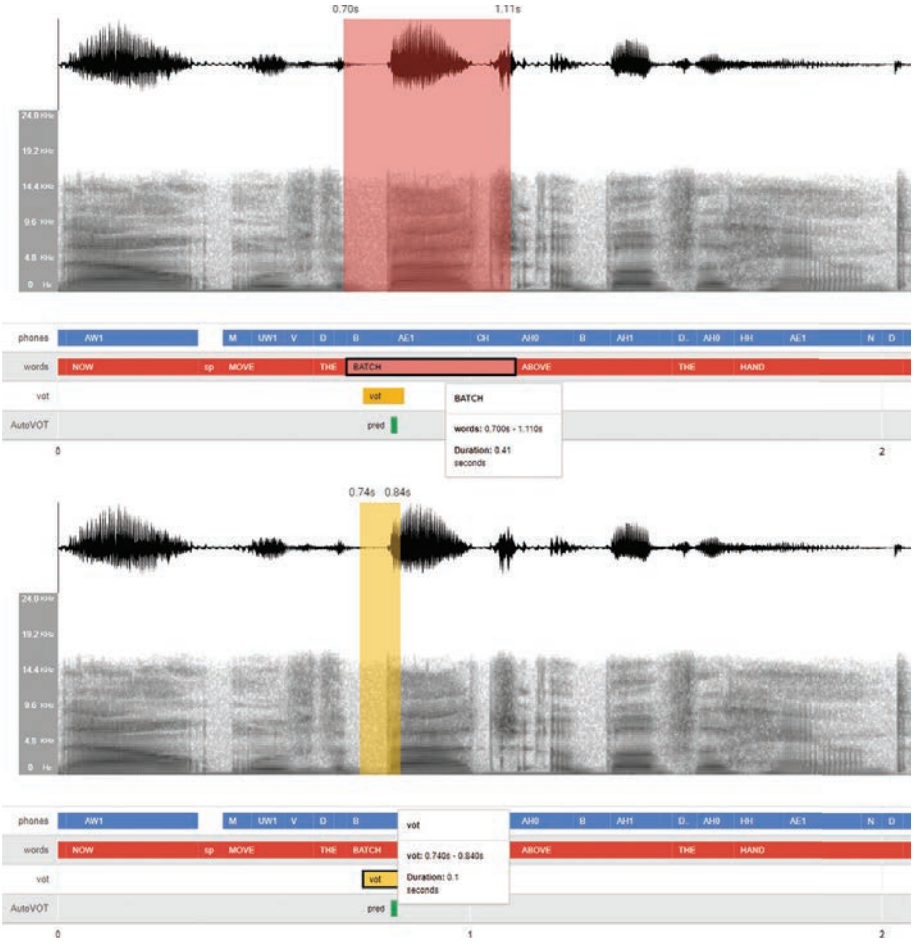


Figure 3 - *By interacting with the annotations, the user can isolate the interested portion of the audio file and reproduce it. Different tiers are highlighted with different colors*



2.3 Semi-automatic annotations

The most important feature provided by VOTEUS is the possibility of generating semi-automatic annotations of speech corpora for VOT intervals. This functionality is articulated into three distinct steps:

a) Automatic forced alignment
b) Fast refinement of textgrid tiers
c) Batch annotation with a pretrained model

To generate the annotations we rely on an AutoVOT model, which can be applied on text segments to generate VOT predictions. Whereas this process is fully automatic, it requires as input a collection of candidate speech intervals that should contain no more than one stop consonant and start 50 ms before the stop burst or 30 msec before the entire segment. To provide such segments we rely on SPPAS (Bigi,

2015; Bigi, Meunier, 2018), an additional tool that performs automatic forced alignment. The term forced alignment denotes a process for determining the time segment of a recording that contains a given portion of a transcription. SPPAS aims at automatizing this process to produce annotations with a granularity that ranges from utterance to phoneme. To this end, SPPAS performs three sub tasks divided into *tokenization*, *phonetization* and *time-alignment*. Tokenization (text-normalization) converts input text into a linguistic representation with standardized and ordinary words, phonetization applies a grapheme-to-phoneme translation and finally time-alignment deals with aligning the sequence of phonemes to the speech signal. SPPAS is provided with resources for multiple languages[21], but the authors state that most of the algorithms have been developed to be as much language-independent as possible and that adding a new language reduces to integrating a few resources such as lexicons and dictionaries. This aspect of SPPAS, in combination with the ready-to-use Python bindings for automatic phonetic segmentation, is what motivated our choice towards this tool. We wrapped SPPAS inside VOTEUS' backend and it can be easily invoked by the interface to obtain candidate intervals on which to apply AutoVOT. Since AutoVOT input requirements are quite strict, to provide better search intervals we implemented a fast refinement procedure for allowing the user to modify existing tiers[22] or adding new ones. By opening this view, VOTEUS shows in a rapid sequence all the annotations for the selected tier for each audio file in the corpus. The user can examine and click directly on the spectrogram to define the precise boundaries of the interval and move to the next annotated entry (Figure 4). If any modification is made, the annotations are automatically updated and saved to disk when the user visualizes the next annotation. This allows the user to rapidly skim through the annotations and adjust them without the need of going through the whole file. Furthermore it eliminates the apparently negligible overhead time for manually loading individual files, displaying them and locating interesting segments before performing the annotation. We argue that this procedure will significantly lower the time needed by an annotator to manually label segments of interest within a big speech corpus.

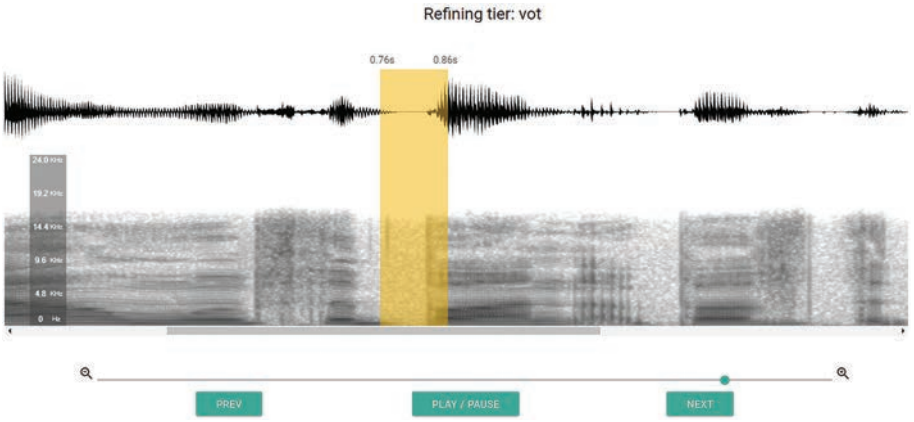
Once a set of sufficiently accurate time intervals is obtained, the user can apply a pretrained AutoVOT model on the whole dataset. This is a fully guided and customizable operation that does not require any programming skill to interact with AutoVOT. AutoVOT's parameters can be configured through VOTEUS and the output is saved directly into the textgrid of each audio file. Among the customizable parameters, the user can select a pretrained model, the dataset on which to apply it and the tier name on which to search for VOTs. All the other parameters that the

[21] Available languages are English, French, Italian, Spanish, Mandarin Chinese, Catalan, Polish, Portuguese, Southern Min, Cantonese, Japanese, Korean and Naija.

[22] Note that we refer to a generic tier present in a textgrid, which includes the output of intermediate steps of our annotation procedure. This procedure in fact will also be used at a final stage to manually refine VOT tiers provided by AutoVOT, adding a layer of human supervision to assess the quality of the predictions and correct them if needed.

original implementation of AutoVOT offers, such as the size of search window and the minimum and maximum length of the detectable VOTs, are fully controllable from the interface. This is sufficient to use VOTEUS as a proxy module to test AutoVOT, but if the final goal is to obtain accurate VOT annotations, the user can rely on the aforementioned fast tier refinement procedure to check and eventually adjust the predicted tiers. The advantages of this semi-automatic annotation pipeline therefore reflect on two important use cases: testing AutoVOT to obtain VOT predictions and precisely annotating a corpus with a head start provided by AutoVOT's predictions.

Figure 4 - *Users can refine tiers in sequence, rapidly skimming through the whole dataset. For precise refinements the annotation can be zoomed in and out within the interface*

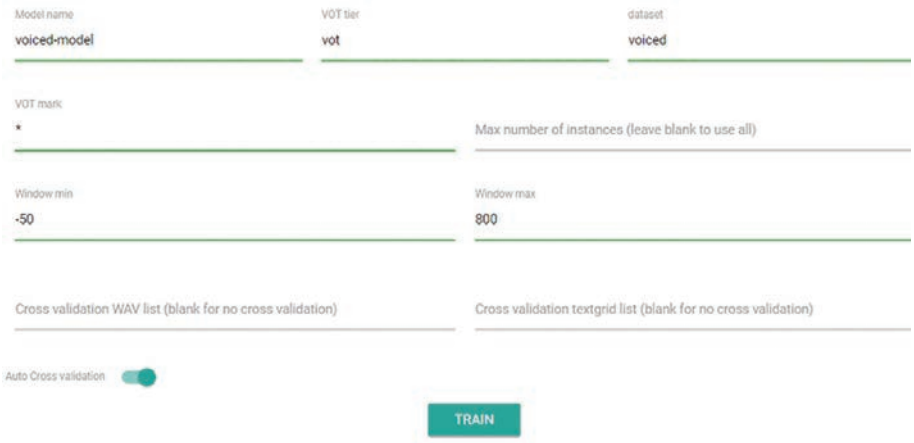


2.4 Training AutoVOT models

To obtain VOT detections it is necessary to use a functional AutoVOT model. Whereas pre-trained models can be downloaded along with AutoVOT and integrated with VOTEUS, one could need to train a model suitable for the data at hand. Through VOTEUS we permit to train new models on a custom speech corpus providing another guided procedure. Similarly to the AutoVOT evaluation functionalites, no programming is required and everything is configurable through our interface. The annotations required to train the model can be selected from an existing tier in the textgrid or manually defined by the user. The user can customize the training procedure setting all the parameters expected by AutoVOT. In Figure 5 the training interface is shown. The required parameters that the user has to set are a name to save the model, the dataset to use for training and the name of the tier with the VOT annotations. In addition there are optional parameters for AutoVOT such as the VOT mark to select a subset of annotations (e.g. "vot", "pos", "neg"), the number of instances to be used and the left and right boundaries of the annotation window in milliseconds relative to the VOT interval. The user can also decide whether to perform cross validation during training and if so which files to use as

validation set. The files for cross validation can be explicitly listed or selected by random through the selection of the "auto cross validation" option. The generated model is then saved into VOTEUS in order to be tested on new data. Again, for the sake of simplicity and compatibility, we store the trained models in the same data format originally used by AutoVOT.

Figure 5 - *Users can customize all the parameters required by AutoVOT and train a model on a selected dataset*



2.5 Exporting results

All the results produced within VOTEUS can be exported and reused with external tools. We offer a choice between different data formats to export the annotations generated with VOTEUS. Users can directly save from the interface the textgrid files to be inspected with Praat and at the same time can convert the annotations in textual form as a CSV (Comma Sparated Values) or save them as *.xls files for compatibility with Microsoft Office Excel and Apache OpenOffice Calc. We believe that this will allow more flexibility for researchers, without forcing them to use a specific tool.

3. *Conclusions*

At the age of 50, it is time for Voice Onset Time to enter the field of big corpora analyses. The access to larger linguistic datasets allows researchers to ground their understanding of this sub-segmental feature on more quantitatively realistic observations; to date, this approach has proven to benefit not only acoustic phonetics and sociophonetics, but also cognitive laboratory methodologies. It is therefore necessary to abandon the traditional time-consuming research routines based on manual annotations and automatize the preparation of the materials. Our contribution aims to increase the accessibility of already existing tools for phonetic analysis,

with the final goal of assisting the researcher through the processes of text-audio alignment, VOTs segmentation and durations extraction. VOTEUS is currently in development for Linux and Windows operating systems. Future work will focus on providing quantitative estimates about the time saved using our interface, as well as the results of usability tests. A preview version of VOTEUS will be available in early 2019 at the following link https://github.com/fedebecat/VOTEUS.

Bibliography

Abramson, A.S., Whalen, D.H. (2017). Voice Onset Time (VOT) at 50: Theoretical and Practical Issues in Measuring Voicing Distinctions. In *Journal of Phonetics*, 63, 75-86.

Bane, M., Graff, P. & Sonderegger, M. (2010). Longitudinal Phonetic Variation in a Closed System. In *Proceedings of the Annual Meeting of the Chicago Linguistics Society*, 46, 43-58.

Bang, H.-Y. (2017). The Structure of Multiple Cues to Stop Categorization and Its Implications for Sound Change. Ph.D. Dissertation, McGill University.

Bang, H.-Y., Sonderegger, M., Kang, Y., Clayards, M. & Yoon, T.-J. (2015). The Effect of Word Frequency on the Time Course of Tonogenesis in Seoul Korean. In *Proceedings of the 18th International Congress of Phonetic Sciences.*

Bang, H.-Y., Sonderegger, M., Kang, Y., Clayards, M. & Yoon, T.-J. (2018). The Emergence, Progress, and Impact of Sound Change in Progress in Seoul Korean: Implications for Mechanisms of Tonogenesis. In *Journal of Phonetics*, 66, 120-144.

Berez-Kroeker, A., Gawne, L., Kung, S., Kelly, B.F., Heston, T., Holton, G., Pulsifer, P., Beaver, D.I., Chelliah, S., Dubinsky, S., Meier, R.P., Thieberger, N., Rice, K. & Woodbury, A.C. (2018). Reproducible Research in Linguistics: A Position Statement on Data Citation and Attribution in our Field. In *Linguistics*, 56(1), 1-18.

Bigi, B. (2015). SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. In *the Phonetician*, 111-112, 54-69.

Bigi, B., Meunier, C. (2018). Automatic Speech Segmentation of Spontaneous Speech. In *Revista de Estudos da Linguagem. International Thematic Issue: Speech Segmentation*, 26(4), 1489-1530.

Boersma, P. (2001). Praat, a System for Doing Phonetics by Computer. In *Glot International*, 5, 341-345.

Chen D., Kozbur, D. & Yu, A. (2015). Pandering vs. Persuasion? Phonemic Accommodation in the U.S. Supreme Court. Working Paper.

Chen X., Xiong Z. & Hu J. (2018). The Trajectory of Voice Onset Time with Vocal Aging. In *Proceedings of INTERSPEECH 2018*, 1556-1560.

Cheng, A. (2017). VOT Merger and f0 Contrast in Heritage Korean in California. In *UC Berkeley PhonLab Annual Report*, 13(1), 281-311.

Cho, T., Docherty, G. & Whalen, D.H. (Eds.) (2018). Special Issue: Marking 50 Years of Research on Voice Onset Time. In *Journal of Phonetics*, 71.

Chodroff, E., Godfrey, J., Khudanpur, S. & Wilson, C. (2015). Structured Variability in Acoustic Realization: A Corpus Study of Voice Onset Time in American English Stops. In *Proceedings of the 18th International Congress of Phonetic Sciences.* Paper 632.

Chodroff, E., Wilson, C. (2017). Structure in Talker-specific Phonetic Realization: Covariation of Stop Consonant VOT in American English. In *Journal of Phonetics*, 61, 30-47.

Chodroff, E., Wilson, C. (2018). Predictability of Stop Consonant Phonetics Across Talkers: Between-category and Within-category Dependencies among Cues for Place and Voice. In *Linguistics Vanguard*, 4(2), 1-11.

Dijkstra, E.W. (1982). On the Role of Scientific Thought. In *Selected Writings on Computing: a Personal Perspective*. New York: Springer, 60-66.

Foulkes, P. (2015). Sociophonetics. In *Proceedings of the 18th International Congress of Phonetic Sciences*. Paper 1051.

Goldrick, M., Keshet, J., Gustafson, E., Heller, J. & Needle, J. (2016). Automatic Analysis of Slips of the Tongue: Insights into the Cognitive Architecture of Speech Production. In *Cognition*, 149, 31-39.

Heeringa, W., Van de Velde, H. (2017). Visible Vowels: A Tool for the Visualization of Vowel Variation. In *Proceedings of INTERSPEECH 2017*, 4034-4035.

Henry, K., Sonderegger, M. & Keshet, J. (2012). Automatic Measurement of Positive and Negative Voice Onset Time. In *Proceedings of INTERSPEECH 2012*, 871-874.

Keshet, J., Sonderegger, M. & Knowles, T. (2014). AutoVOT: A Tool for Automatic Measurement of Voice Onset Time Using Discriminative Structured Prediction. https://github.com/mlml/autoVOT/

Lisker, L., Abramson, A.S. (1964). A Cross-language Study of Voicing in Initial Stops: Acoustical Measurements. In *Word*, 20(3), 527-565.

McAuliffe, M., Stengel-Eskin, E., Socolof, M. & Sonderegger, M. (2017). Polyglot and Speech Corpus Tools: A System for Representing, Integrating, and Querying Speech Corpora. In *Proceedings of INTERSPEECH 2017*, 3887-3891.

Reddy, S., Stanford, J. (2015). A Web Application for Automated Dialect Analysis. In *Proceedings of NAACL-HLT 2015*, 71-75.

Schertz, J., Kang, Y. & Han, S. (2017). Cross-language Correspondences in the Face of Change: Phonetic Independence Versus Convergence in Two Korean-Mandarin Bilingual Communities. In *International Journal of Bilingualism*, 23(1), 157-199.

Sheena, Y., Hejná, M., Adi., Y. & Keshet., J. (2017). Automatic Measurement of Pre-aspiration. In *Proceedings of INTERSPEECH 2017,* 1049-1053.

Singh, R., Keshet, J. & Hovy, E. (2016). Profiling Hoax Callers. In *Proceedings of the 2016 IEEE Symposium on Technologies for Homeland Security (HST)*, 1-6.

Singh, R., Keshet, J., Gencaga, D. & Raj B. (2016). The Relationship of Voice Onset Time and Voice Offset Time to Physical Age. In *Proceedings of the 41 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 5390-5394.

Singh, R., Raj, B. & Gencaga, D. (2016). Forensic Anthropometry from Voice: An Articulatory-phonetic Approach. In *39th International Convention on Information and Communication Technology Electronics and Microelectronics (MIPRO)*, 1375-1380.

Solanki, V.J. (2017). Brains in Dialogue: Investigating Accommodation in Live Conversational Speech for Both Speech and EEG data. Ph.D. Dissertation. University of Glasgow.

Sonderegger, M. (2012). Phonetic and Phonological Dynamics on Reality Television. Ph.D. Dissertation, University of Chicago.

Sonderegger, M. (2015). Trajectories of Voice Onset Time in Spontaneous Speech on Reality TV. In *Proceedings of the 18th International Congress of Phonetic Sciences*. Paper 903.

Sonderegger, M., Bane, M. & Graff, P. (2017). The Medium-term Dynamics of Accents on Reality Television. In *Language*, 93(3), 598-640.

Sonderegger, M., Keshet, J. (2010). Automatic Discriminative Measurement of Voice Onset Time. In *Proceedings of INTERSPEECH 2010*, 2242-2245.

Sonderegger, M., Keshet, J. (2012). Automatic Measurement of Voice Onset Time Using Discriminative Structured Predictions. In *The Journal of the Acoustical Society of America*, 132(6), 3965-3979.

Stuart-Smith, J., Rathcke, T., Sonderegger, M. & Macdonald, R. (2015). A Real-time Study of Plosives in Glaswegian Using an Automatic Measurement Algorithm: Change or Age-grading?. In Torgersen, E., Harstad, B., Maehlum, B. & Royneland, U. (Eds.) *Language Variation: European Perspectives V: Selected Papers from the 7th International Conference on Language Variation in Europe (ICLaVE 7).* Amsterdam: John Benjamins, 225-237.

Stuart-Smith, J., Sonderegger, M., Rathcke, T. & Macdonald, R. (2015). The Private Life of Stops: VOT in a Real-time Corpus of Spontaneous Glaswegian. In *Laboratory Phonology*, 6, 505-549.

Turnbull, R. (2015). Assessing the Listener-oriented Account of Predictability-based Phonetic Reduction. Ph.D. Dissertation. Ohio State University.

DOI: 10.17469/O2104AISV000017

PAOLO BRAVI

Prosit: a Praat plug-in for the search and inspection of corpora of annotated audio files

The paper presents the Prosit software, a plug-in for Praat (Boersma, Weenink, 1992-2018), one of the most renowned programs for carrying out phonetic research. The Prosit plug-in is designed to help researchers in (i) making a flexible search on corpora of sound files with Praat TextGrid annotations; (ii) building their own corpus by editing and managing sound files and TextGrids; (iii) listening, visualizing, inspecting and analyzing sounds that match the search criteria. The rationale behind the project is the observed use of different schemes in sound annotation carried out via Praat, an instrument whose potential benefits allow searches to be carried out on sound corpora built up with different approaches. The Prosit plug-in also makes it possible to apply a number of either batch or single operations on the search outcomes.

Keywords: Praat plug-in, Praat search engine, TextGrid, sound annotations, sound visualization.

1. *Introduction and motivation*

The Praat TextGrid is a widely diffused format for storing information related to audio files. In many cases audio file analysis needs previous manual or automatic annotation of the sound files, so that the corpus on which the analysis is carried out eventually comes to consist of a number of pairs of sounds and TextGrids.

That said, a typical analytical batch procedure consists of iterating some kind of procedure on all the files comprising the corpus, which are usually stored in one directory and have a consistent annotation. But what happens if the "corpus" – or a pre-version of it – is not (yet) built up using a consistent annotation scheme or if we simply want to explore what could be large and non-homogeneous groups of annotated sound files as a possible preliminary step towards the creation of a well-organized corpus?[1]

Prosit[2] is a Praat plug-in designed to face this kind of need: on the one hand, it is a search-engine (with no external dependencies) that allows step-wise flexible research within possibly large and non-homogeneous groups of TextGrids,

[1] "Standards are like toothbrushes, everyone agrees that they're a good idea but nobody wants to use anyone else's" is a brilliant statement attributed to Murtha Baca that well synthesizes the difficulties relevant to metadata definition and management (see Pomerantz, 2015: 65).

[2] The name is an acronym for the *Pr*aat *O*bject *S*earch and *I*nspection *T*ool. At the same time, not without a hint of irony, it expresses the author's wish to allow users to carry out a fruitful search and successful work with TextGrids and audio files.

while, on the other, it allows the user to either listen to, edit, visualize or analyze the audio files that match the TextGrid interval(s) requested and save the relevant output.[3] The lack of a proper internal database management system and a corresponding search engine has been noted in numerous cases.[4] This is linked both to the fact that Praat allows its users the freedom to develop their own system of file management by means of its built-in scripting language (BIM: scripting [1][5]), and also because complex research on documents and relevant metadata and annotations can be managed properly using a software specifically designed for database management and information retrieval.[6]

In many cases, the workflow of researchers studying phonetics involves the use of multiple software. A quite common procedure among researchers is to carry out a first step of sound analysis and annotation using Praat, and then to apply statistics and drawing relevant graphics with R (RDCT, 2011). Some specifically designed software has been developed so as to automatize and improve interoperability between the two programs (Albin, 2014; Boril, Skarnitzl, 2016). Other programs designed for media file annotation which provide utilities for corpus management and analysis are presented and described in Durand, Gut & Kristoffersen, (2014, part III).[7] The decision to avoid any dependency in the Prosit plug-in clearly has its costs in terms of efficiency and entails a number of limitations with regard to database systems explicitly designed for

[3] To date, while pending comprehensive and exhaustive testing and the writing of documentation with a detailed description of the available functions, the Prosit plug-in can be requested directly from the author at the address paolobravi.gm@gmail.com.

[4] The Praat command "Create Corpus..." described in Boersma, (2014) is conceived in view of providing important functions for corpus management and to overcome a previously observed limitation of the program regarding the fact that "Praat does not include a proper database system as such, so searching a speech corpus with Praat must be implemented through Praat scripts (which can become painfully slow)" (Lennes, 2005: 15). The scope and functions of the Prosit plug-in here described only partially overlap with those of the Praat in-built command.

[5] The Praat built-in manual (BIM) will be constantly cited throughout the paper. For convenience, the following method will be adopted to point the reader to the appropriate manual page (note: the Praat version used is 6.0.39). In the Praat *Objects window*, under the *Help* section, use the command *Search Praat manual* and then write the word given (in this case, the word "scripting") in the form field. Then, from the list of outputted links, choose the entry indicated by the number within square brackets (in this case, the entry is [1]). Hence, from now on, these references to the Praat manual will have the following format: BIM: query [number].

[6] Generally speaking, it has to be observed that DMS software based on relational databases, wherein metadata are stored in a principled way, is far more flexible (though harder to design and maintain) than what are known as flat file databases, which contain less strictly organized and more redundant information with respect to the former (Srivastava, 2014; Elmasri, Navathe, 2016).

[7] A comparison between functions provided in Prosit and those implemented in other software designed for similar aims is beyond the scope of this paper. However, the interested reader should be aware that among the programs that can read/import annotations in the TextGrid format and that provide functions specifically designed for corpus analysis, a basic list of the most notable ones comprehends at least ELAN (Wittenburg, Brugman, Russel, Klassmann & Sloetjes, 2006), EMU-SDMS (Winkelmann, Harrington & Jänsch, 2017), EXMERaLDA – in particular the EXAKT tool – (Schmidt, 2002).

information storage and searches, but it does allow a Praat user to drive – so to say – 'his/her' own car, managing data that are structured according to the TextGrid format (BIM: TextGrid [2]), with which he/she is most likely to be well acquainted.

2. *User control and interfaces*

Prosit is based entirely on the tools (types of objects, commands, etc.) and interfaces (editors, forms, windows) provided in Praat. Excluding external dependencies has its pros and cons: the most notable advantages are ease of installation and use,[8] since any Praat user with basic practice can use the program by utilizing its standard apparatus. Moreover, a Praat user with some knowledge and experience with Praat scripting can change or add parts to the plug-in according to his/her own needs.[9] In terms of disadvantages, Prosit, being based on a linear search mechanism, has not the efficiency necessary for the inspection of large corpora. Moreover, there are some practical limitations related to the characteristics of the interfaces and editors which, as part of the Praat GUIs, cannot be overcome as of yet.

User control of the parameters involved in any operation provided by Prosit is carried out on the typical Praat forms. In many cases they appear in sequence, based on the user's choices, and are implemented by means of the "beginPause /endPause" mechanism for user control (BIM: user [1]). Instead, search report and relevant Prosit commands are managed via "ManPages" that are dynamically created whenever a new search starts and at every step of the search (BIM: manpages [1]).

3. *Search: the metaphor of a shopping trip*

Searching intervals in Prosit come about in three phases. For ease of conceptualization, these phases are identified using the metaphor (and the relevant names) of a shopping trip. In phase 1, the user sets what is called the *district*, viz a directory (as a "simple" or "parent" one), and a *store type*, viz a file type capable of storing information relevant to an audio file. At the moment, the only searchable file types are Praat TextGrids, but searches within other file types can be envisaged in a future version of the plug-in. In phase 2, the user sets the *shelf* (or the shelves) that Prosit has to explore. When the store type is "TextGrid",

[8] The plug-in can be installed following the simple instructions contained in the relevant page of the Praat built-in manual (BIM: plug-ins [1]).

[9] Praat has a built-in introduction to the program (Help > Praat intro) and a detailed explanation of its scripting language (Help > Scripting language). Furthermore, several tutorials on the use of the Praat program are available, written in various languages and with a variety of reader proficiency and topic focus in mind. Among those written in English, a certainly non-exhaustive list comprises van Lieshout, 2017; Styler, 2017; Weenink, 2018; Wood, 1994-2018; Cangemi, Auris, submitted.

this means identifying the tier(s) that will be searched through. A spectrum of possibility is provided to establish which tiers are to be searched in. In phase 3, the user eventually sets the *item*(s) he/she is interested in.

Search results, together with a description of the research carried out, is made available to the user by means of Praat "ManPages" that allow the user to perform a number of different operations. These range from listening to the specified part of the sound, visualizing it through specific animation devices (also playing it at a speed lower than the original) and editing and manipulating it. It is also possible to perform a number of batch operations on all items.

Searches can be carried out in different steps. Users can add new items via a new search, refine the preceding search, retain or exclude selected items through successive steps. Every search step and relevant parameters and results are easily available at every moment, so that the overall search can also take different paths according to the researcher's needs.

3.1 Search phases

The paragraphs in this section describe each of the three search phases.

3.1.1 Phase 1: "District"

After the plug-in installation, a button named "Prosit" will appear in the Praat Object window under the "Praat" menu. Clicking on this button gets the search underway. The first form that appears regards the *district* where a defined *store type* is searched for – at the moment, as mentioned above, the only manageable information "stores" are of the TextGrid type. Two main options are available in this search phase. The first one allows the user to specify one of the three ways of stating the directory to be searched for: [a] choosing by browsing; [b] choosing among the last directories searched[10]; or [c] stating the directory address explicitly, either writing it manually on a form or choosing from a pre-set list of directories that the user can easily create or modify in advance.[11]

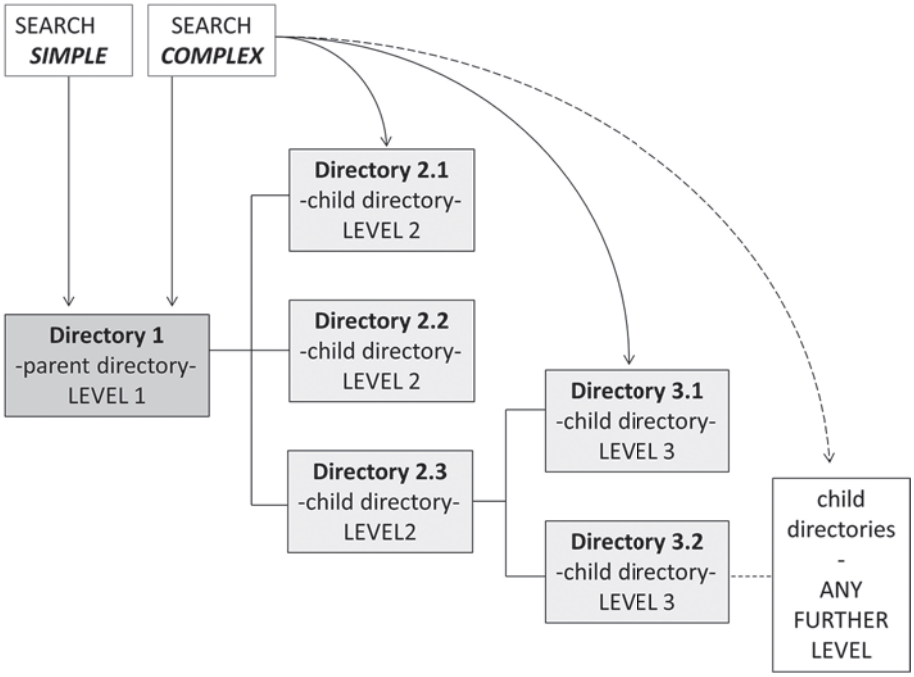
The second parameter allows the user to state the 'complexity' of the search, i.e. the number of levels of 'child directories' to be searched for beyond the main ('parent') directory. By default, this parameter is set to "complex" (i.e. the search will involve not only the directory chosen, but the 'child' directories contained in it) and "unlimited" (i.e. the search will proceed until no further 'child' directory is found). Users are allowed to either choose a "simple" search, i.e. a search limited to the directory stated, or to specify, when a "complex" search is chosen, a number corresponding to the levels of 'child' directories to be searched for (see Figure 1). This

[10] The number of directory addresses held in the memory is set to ten by default. This number may be modified according to the user's preferences.

[11] The list of pre-set directory addresses is stored in the file plugin_PROSIT/lists/Others/SearchDirOM_Input.txt. Each user will set his/her list of preferred addresses according to his/her own needs.

flexibility allows the search to be restricted or widened according to specific needs and to manage every possible different organization of file storage.

Figure 1 – *Sketch of the first phase of the search:two basic options ("Simple" and "Complex") are available and, in the latter case,a definite number of "child" directory levels can be set by the user*



3.1.2 Phase 2: "Shelf"

The second phase of the search is aimed at deciding which *shelves*, i.e. what type of annotations, are to be searched for. Dealing with 'stores' of annotations like TextGrids, 'shelves' are – outside the shopping metaphor – tier names. There are three possible ways of searching through them. The first method is to look up TextGrid tiers whose name matches a given string according to a specific chosen string-matching criterion.[12] The second method is to choose a tier name from a list comprising all tier names available in the TextGrids found in the first phase of the search. The third method envisages carrying out a search in all the tiers of the TextGrids. It goes without saying that the search may take longer with this third option, and that it could possibly yield a vast number of non-consistent results. In all cases the user can decide about the case-sensitiveness of the string search.

[12] The following series of options are available: "is equal to", "is not equal to", "contains", "does not contain", "starts with", "does not start with", "ends with", "does not end with", "matches (regex)". The last option, of course, permits much wider search flexibility, but requires a knowledge of how Praat manages regular expressions (see BIM: regular [1]).

If what is looked for is to carry out a search through two or more, but not all tiers, two options are available. The first way is to use the "match (regex)" option, which allows the search to be made in a single step. The second way is to carry the analysis through different steps, using the ADD method. For example, if one needs to find all non-empty intervals in tiers whose names contain either the text "Profilo" or the text "Strut", the user can either: (i) perform the search in one step, using the "match (regex)" option and writing "Profilo|Strut" in the search text field; or (ii) perform the search in two steps, using the "contains" option and writing "Profilo" in the search text field, and then repeating the option, after selecting the ADD method, searching for the text "Strut".

3.1.3 Phase 3: "Items"

The third phase of the search is aimed at deciding which *item*, i.e., beyond the metaphor, what label is to be searched for. Two out of the three methods described in paragraph 3.1.2 for searching through tier names are also available for labels. This means that it is possible (i) to set a string which matches the annotations according to one of the established matching patterns (see note 12) by writing it manually in the form, or (ii) to choose the string to be searched for from a predefined list. A third method is available for searching through labels, that is (iii) choosing a label from a list of all the ones actually present on the "shelves" – i.e. in the tiers – retrieved in phase 2.

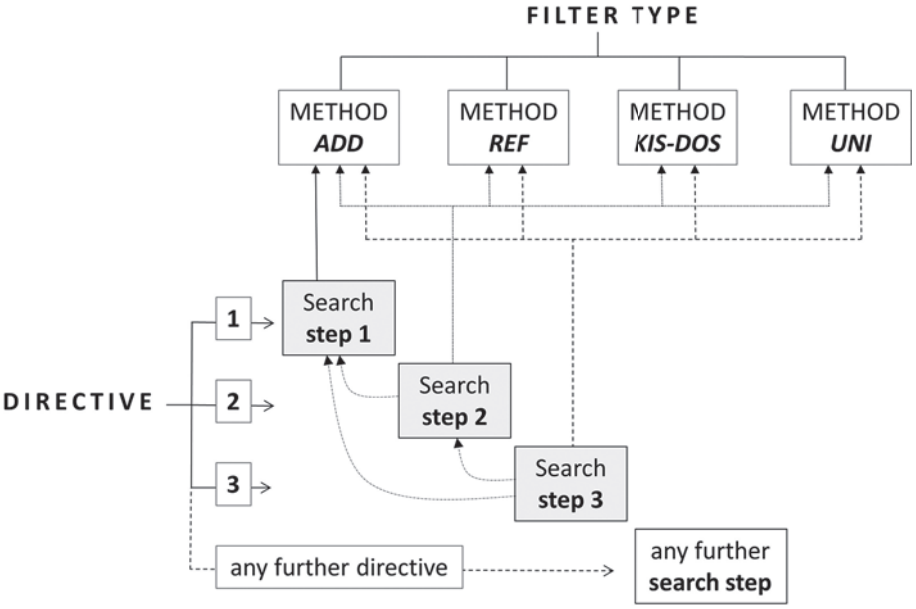
One of Prosit's crucial and potentially beneficial aspects is the possibility of choosing different ways to define the strings to be searched for in the TextGrid annotations, along with the flexibility offered by the chance to choose between various matching criteria, which also comprises the opportunity of using regular expressions. In particular, the method (iii), based on a preliminary examination of the annotation present in the selected tiers, may be useful in a preliminary phase of inspection of the corpus of TextGrids.

3.2 Search steps

Searches can be carried out in more steps, i.e. new directives may be given to modify the outcome of previous search steps. Each step (with the exception of the first one) can operate on the results of the previous steps in four modes. The first mode is addition (symbol: "ADD"). In this mode, new items are added which match the new directive criteria. This method is by rule adopted in the first search step. The second mode is refinement (symbol: "REF"). In this mode, items found in previous steps of the research are held if the outcomes also match the new search directive. The third mode regards selections (symbols: either "KIS" or "DOS"). In this mode, items selected by the user are either excluded (symbol: DOS – drop off selected items) or served, with exclusion of the non-selected one (symbol: KIS – keep in selected items). The fourth mode regards item reduplication (symbols: UNI). In this mode, two (or more) items

that refer to the same sound part outcoming from different directives are synthesized as a single item.

Figure 2 – *Prosit multiple search step method allows the user to conduct an inspection of the annotated sounds using different types of filter at each step*



The mechanism has been conceived with the aim of allowing the user great search flexibility, making it possible to rethink and carry out multiple tries during the search process. The Manpage interface makes it easy to pass from one search step to another and to account for the procedures involved in each step.

4. *Inspection, editing, actions*

Search results are reported at every step by means of Praat ManPages which are organized in three sections (see Figure 3a-b). The first section ("Search detail") gives a summary of the parameters used in each research step. In particular, it reports the filter type used (see Paragraph 3.2), the "district" where the search was carried out, the "store" type searched for (as of yet, the fixed type is "TextGrid"), the "shelf" (as of yet, string values referring to TextGrid tier names) and the query, as expressed in the three search phases outlined above. The second section ("User activity") contains a series of subsections that allow the user to improve or change his/her search and to 'navigate' in the various steps of his/her search and may also help him/her in building a corpus and using the search for specific aims (see below, par. 4.1). The third section ("Search results") lists the outcomes of the search. Each item is accompanied by three sub-

sections. The first ("Filters") reports a short summary of the search matching. The second ("Time") gives time details. The third one ("Operations") provides a list of words (or sentences) linked to the plug-in scripts that allow the user to listen, view and edit the sound through the Praat Sound & TextGrid editor, to select or deselect the item and to operate with various types of editor providing animation for sound visualization and manipulation (see par. 4.2). From here on, these links that appear in bold blue characters on the ManPages (see figs. 3a-b) will be referred to as "WL".

Figure 3a – *Example of a two-step search - step 1: the search directive 1 looks up into a parent directory (in this case, the built-in "test" one) for TextGrids comprising tier names that contain the string "R0-tierTc" and looks in these tiers searching for labels matching the regular expression "1|[3-5]" (85 items are found)*


PROSIT · step 001 · page 002

PRAAT OBJECTS SEARCH AND INSPECT TOOL | © 2017-2018 by Paolo Bravi | Version 1.0

§ 1 - *Search details:*

÷ DIRECTIVE n. **1** – FILTER: type **♣ADD** | DISTRICT: directory [complex] **test** | STORE: type **TextGrid** | SHELF: tier name contains [CI] **R0-tierTc** | QUERY: label matches (regex) [CI] **1|[3-5]**

§ 2 - *User activity:*

→ STEPS – improve your search: **♣Add to search** | **♦Refine search** | **♥Keep in selected** | **♠Drop off selected** | **≠ Exclude replicated items** |
→ REVISION – change your search: | **Restart from scratch** | **Select** |
→ UTILITIES – build your corpora: **Rename** | **Analyze** |
→ ACTIONS – use your search: **Examine** | **Save** | **Print** |
→ NAVIGATION – check the steps of your search: **Previous page** | **Next step** | Jump directly to whatever search step using "Go to" > "Search for page (list)"
→ INFO – get basic info about Prosit looking in the **About Prosit** page

§ 3 - *Search results:* **85** items found | **51 - 85** items displayed on page **2 - 2**

• ITEM n. **51** - §Filters: [**♣Directive 1**: *Found label*: **4** in tier: R0-tierTc(4) - *Store*: Lu010506-4 (id: 40819)] §Time: From=14.638 To=15.140 Duration=**0**.503 (s) §Operations: **Play** | **Stop** | **View & Edit** | **Add to selection** | **Remove from selection** | **Vis** |
• ITEM n. **52** - §Filters: [**♣Directive 1**: *Found label*: **5** in tier: R0-tierTc(4) - *Store*: Lu010506-4 (id: 40819)] §Time: From=15.140 To=15.940 Duration=**0**.800 (s) §Operations: **Play** | **Stop** | **View & Edit** | **Add to selection** | **Remove from selection** | **Vis** |
• ITEM n. **53** - §Filters: [**♣Directive 1**: *Found label*: **4** in tier: R0-tierTc(4) - *Store*: Lu010506-4 (id: 40819)] §Time: From=15.940 To=16.227 Duration=**0**.287 (s) §Operations: **Play** | **Stop** | **View & Edit** | **Add to selection** | **Remove from selection** | **Vis** |

Figure 3b – *Example of a two-step search - step 2: the directive 2 refines the retrieving adding a sec-ond condition: labels that do not contain the string "a" or "A" are searched for in tiers whose name is equal to "R2-LIN-VV-PT" (100 items are found)*


PROSIT · step 002 · page 001

PRAAT OBJECTS SEARCH AND INSPECT TOOL | © 2017-2018 by Paolo Bravi | Version 1.0

§ 1 - *Search details:*

÷ DIRECTIVE n. **1** – FILTER: type **♣ADD** | DISTRICT: directory [complex] **test** | STORE: type **TextGrid** | SHELF: tier name contains [CI] **R0-tierTc** | QUERY: label matches (regex) [CI] **1|[3-5]**
÷ DIRECTIVE n. **2** – FILTER: type **♦REF** | REFINE: search results in Step **1** | STORE: type **TextGrid** | SHELF: tier name is equal to [CI] **R2-LIN-VV-PT** | QUERY: label does not contain [CI] **a**

§ 2 - *User activity:*

→ STEPS – improve your search: **♣Add to search** | **♦Refine search** | **♥Keep in selected** | **♠Drop off selected** | **≠ Exclude replicated items** |
→ REVISION – change your search: | **Restart from scratch** | **Select** |
→ UTILITIES – build your corpora: **Rename** | **Analyze** |
→ ACTIONS – use your search: **Examine** | **Save** | **Print** |
→ NAVIGATION – check the steps of your search: **Previous page** | **Previous step** | **Next page** | **Next step** | Jump directly to whatever search step using "Go to" > "Search for page (list)"
→ INFO – get basic info about Prosit looking in the **About Prosit** page

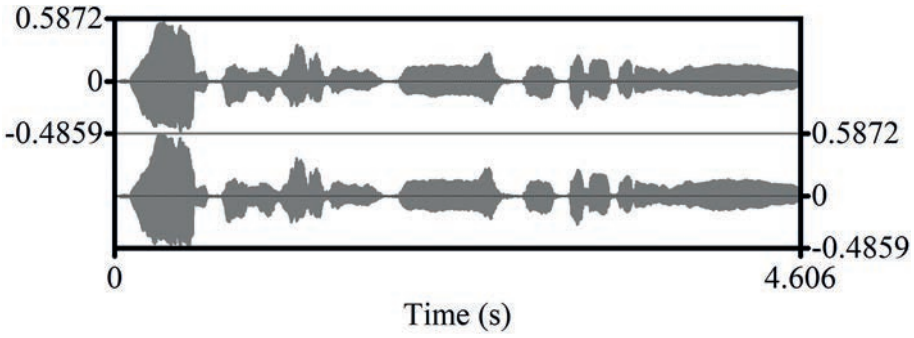
§ 3 - *Search results:* **100** items found | **1 - 50** items displayed on page **1 - 2**

• ITEM n. **1** - §Filters: [**♣Directive 1**: *Found label*: **5** in tier: R0-tierTc(4) - *Store*: Br170163-8 (id: 40816)][**♦Directive 2**: *Found label*: **es** in tier: R2-LIN-VV-PT(1) - *Store*: Br170163-8 (id: 40816)] §Time: From=0.766 To=1.319 Duration=**0**.553 (s) §Operations: **Play** | **Stop** | **View & Edit** | **Add to selection** | **Remove from selection** | **Vis** |

4.1 Batch operations

Prosit allows operations to be carried out in batch mode on the items resulting from retrieving. In the "user activity" paragraph, a number of operations are available that can be of use when dealing with (or building up) a corpus.[13] Among these, two sub-sections are available in the "Utilities" section. The first is accessible via the WL "Rename", which opens a pop-up form that allows the establishment of some parameters relevant for batch tier rename in all the files matching the search criteria.[14] The second may be reached through the WL "Analyse", which allows the resulting pitch in the part of the sound output from the search on the annotation to be extracted (and saved). In the "Actions" section, the user can "Examine" in a synthetic way the output of his/her search: by virtue of the Praat function "Concatenate recoverably" (see BIM: concatenate [2]), the user can listen to all the sounds in sequence, aligned in a single file with an accompanying TextGrid where each label indicates the origin of each portion of sound present in the chain. Other batch operations are accessible under the WL "Save", particularly as far as TextGrid and Sound file are concerned. Either TextGrid or Sound parts relevant to the retrieved intervals can be saved, with different options relevant to amplitude peaks and audio format. The WL "Print" offers three possibilities: waveform, pitch contour and pitch histogram of retrieved items can be printed, with a number of parameters relevant to each type of graph (see figs. 4a-c).

Figure 4a - *Batch operation on outcomes from search: printing of waveform excerpt*



[13] In this paper, only a brief summary of the operations that may be carried out on the retrieved items is provided. A detailed description of these functions will be given in a future contribution.

[14] This may be of help when TextGrid files are not stored in just one directory or when they have different tier names since they are part of different collections of files, whereby tiers are annotated according to non-homogeneous rules and/or in the light of different aims. Obviously, the user is asked to decide whether to overwrite the original TextGrid files or to save the new ones in a separate place.

Figure 4b - *Batch operation on outcomes from search: printing of a pitch contour excerpt*

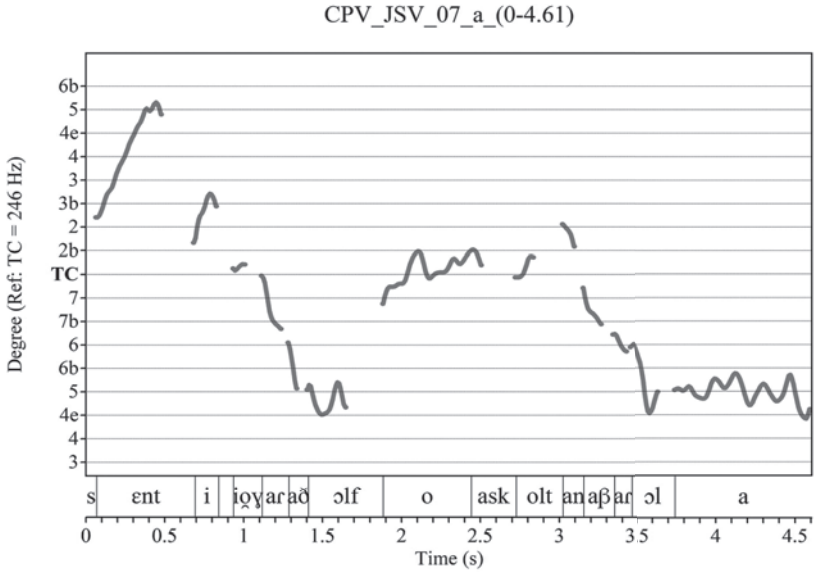
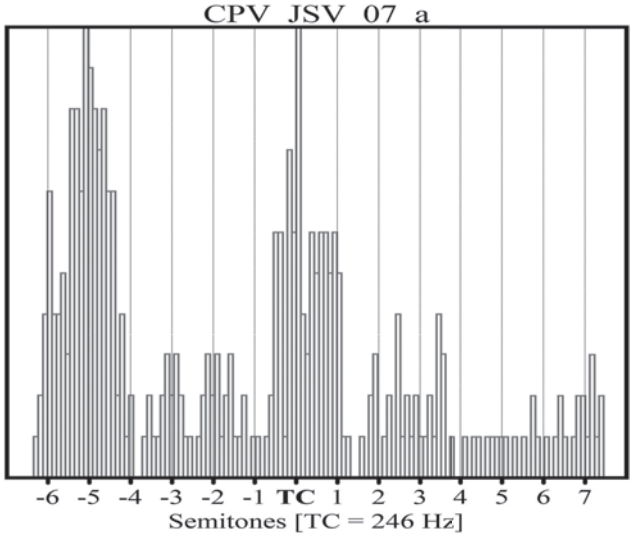


Figure 4c - *Batch operation on outcomes from search: printing of a histogram of pitch values (bins = 10 cents) found in the contour excerpt*

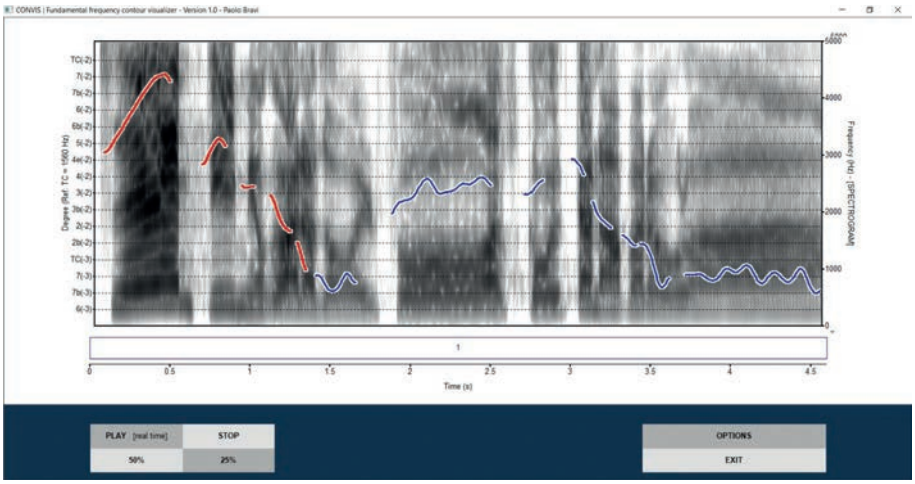


4.2 Single item operations

Items resulting from user retrieving are described in terms of their match to the search directive(s) and their time features (see above, Paragraph 4 and Figures 3a-b). A series of operation may be carried out on each item. Firstly, each sound segment can be quickly listened to via a pair of WLs named "Play" and "Stop". Secondly, the "View & Edit" WL allows access to the Praat TextGrid Editor with the time

boundaries relevant to the item, thereby accessing all commands available in this editor. This WL makes it possible to visualize just the tier searched for or all the tiers actually present in the TextGrid, and also to observe in which context the retrieved annotation is placed. Thirdly, another pair of WLs ("Add to selection" and "Remove from selection") allow the list of selected objects to be managed in the light of further search steps or batch operations. Fourthly, the WL "Vis" allows the user to visualize and/or manipulate the sound relevant to each item in either of the following two ways: either via (i) the "ConVis" panel or (ii) the Praat Manipulation Editor. The (i) ConVis panel is an application designed within Prosit on the Praat Demo Window with the objective of visualizing, through an animation, the evolution of the pitch contour (either at normal speed or delayed up to 4 times the original length) on top of the spectrogram and above a TextGrid tier relevant to the item indicated by the user (see Figures 5).

Figure 5 – *Single item operation: dynamic visualization of the pitch contour excerpt*



'Visualization' via (ii) the Praat Manipulation Editor is actually much more than a means to visualize sounds. In fact, this kind of editor permits each sound to be manipulated through pitch stylization and/or modification and duration changes (detailed information on the use and functionalities of this editor are in BIM, manipulation [1]).

5. *Conclusions*

The plug-in Prosit is at its first steps. This is obvious to the potential user who will find some "one-option menus" that, at first sight, do not seem to have any meaning. In most cases, this single-choice (i.e. pseudo-) menu is part of the lists of commands that are under elaboration or which have yet to be properly tested. However, even at this early stage, Prosit contains a basic infrastructure that allows future develop-

ments along lines that are the same (or analogous) to the ones that have already been set and are in use.

Future developments may regard different parts of the plug-in. In particular, three aspects could be improved or developed to better suit prospective user needs. The first one is to widen the range of possibilities of the search engine, which might look in "stores" other than TextGrids, in particular allowing searches on different metadata sets conceived according to standard procedures.[15] The second one is to enlarge the number of available batch operations, either aimed at helping in the constitution of a specific corpus or at managing and analyzing data and sounds resulting from the search. The third one is to provide other ways of fostering inspection capacity, of editing single items and of analyzing and visualizing the acoustical features of the relevant sound files.

Acknowledgements

I wish to thank Ignazio Macchiarella, Francesco Capuzzi and Francesco Cangemi for their support and advice.

Bibliography

Albin, A. (2014). PraatR: An Architecture for Controlling the Phonetics Software "Praat" with the R Programming Language. In *Journal of the Acoustical Society of America*, 135(4), 2198.

Boersma, P. (2014). The Use of Praat in Corpus Research. In Durand, J., Gut, U. & G. Kristoffersen (Eds.), *The Oxford Handbook of Corpus Phonology*. Oxford: Oxford University Press, 342-360.

Boersma, P., Weenink, D. (1992-2018). Praat: Doing Phonetics by Computer. http://www.fon.hum.uva.nl/praat / Accessed 15.04.18.

Boril, T., Skarnitzl, R. (2016). Tools rPraat and mPraat. In Sojka, P., Horák, A., Kopeček, I. & Pala, K. (Eds.), *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, Brno, Czech Republic, 12-16 September 2016, 367-374.

Cangemi, F., Auris, B. (submitted). *Praat Handbook*. Berlin: Language Science Press.

Durand, J., Gut, U. & Kristoffersen, G. (Eds.) (2014). *The Oxford Handbook of Corpus Phonology*. Oxford: Oxford University Press.

DCMI Usage Board (2012). DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms / Accessed 15.04.18.

[15] Neither a unique standard or framework, nor a general agreement exists on how to describe the contents and characteristics of digital audio files. Based on the current diffusion of metadata sets and on the main area of research activity of the Prosit developer, the possibility of searching for Dublin Core (DCMI Usage Board, 2012) and BDI (ICCD, 2007) metadata can be envisaged in a version- to- come of the plug-in. For a general survey on the conception and use of metadata in linguistic corpora, see Durand, Gut & Kristoffersen, 2014 (part I).

Elmasri, R., Navathe, S.B. (2016). *Fundamentals of Database Systems* (7th edition). Boston: Pearson.

ICCD (2007). Scheda BDI. Beni demoetnoantropologici immateriali versione 3.01 - Struttura dei dati e normativa di compilazione unificata e integrata - gennaio 2007. http://www.iccd.beniculturali.it/getFile.php?id=284 / Accessed 15.04.18.

Lennes, M. (2005). Hands-on Tutorial: Using Praat for Analysing a Speech Corpus. http://www.helsinki.fi/~lennes/vispp/lennes_palmse05.pdf / Accessed 15.04.18.

Pomerantz, J. (2015). *Metadata.* Cambridge MA – London: The MIT Press.

RDCT (2011). R: A Language and Environment for Statistical Computing. http://www.R-project.org / Accessed 15.04.18.

Schmidt, T. (2002). EXMARaLDA – ein System zur Diskurstranskription auf dem Computer. In *Arbeiten zur Mehrsprachigkeit*, B(34), 1-23.

Srivastava, R. (2014). *Relational Database Management System.* New Delhi: New Age International Private Limited.

Styler, W. (2017). Using Praat for Linguistic Research. http://savethevowels.org/praat / Accessed 15.04.18.

van Lieshout, P. (2017). PRAAT Short Tutorial – An introduction. https://www.researchgate.net/publication/270819326_PRAAT_--_Short_Tutorial_--_An_introduction / Accessed 15.04.18.

Weenink, D. (2018). Speech Signal Processing with Praat. http://www.fon.hum.uva.nl/david/sspbook/sspbook.pdf / Accessed 15.04.18.

Winkelmann, R., Harrington, J. & Jänsch, K. (2017). EMU-SDMS: Advanced Speech Database Management and Analysis in R. In *Computer Speech & Language*, 45, 392-410.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. & Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, Genoa, IT, 22-28 May 2006, 1556-1559

Wood, S. (1994-2018). *Introduction to Praat.* https://swphonetics.com/praat/introduction / Accessed 15.04.18.

Autori

Cinzia Avesani – National Research Council (ISTC - CNR), Padua
cinzia.avesani@pd.istc.cnr.it

Leonardo Badino – Multiscale Brain Communication, Istituto Italiano di Tecnologia, Ferrara, Italia
leonardo.badino@iit.it

Camilla Bernardasci – Phonogrammarchiv, Università di Zurigo
camilla.bernardasci@uzh.ch

Federico Becattini – Media Integration and Communication Center (MICC) – Università degli Studi di Firenze. Viale Morgagni 65 – Firenze
federico.becattini@unifi.it

Ilaria Brai – Institute of Cognitive Science and Technologies, CNR, Padua
ilaria.brai@gmail.com

Paolo Bravi – Martin Behaim Gymnasium, Nuremberg (DE)
paolobravi.gm@gmail.com

Francesco Cangemi – Institut für Phonetik, Universität zu Köln
fcangemi@uni-koeln.de

Christopher Carignan – Institut für Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität München, Germany
c.carignan@phonetik.uni-muenchen.de

Chiara Celata – Scuola Normale Superiore di Pisa, Italy
chiara.celata@sns.it

Benedetta Colavolpe – Dept. of Neuroscience, University of Padua
benedettacolavolpe@gmail.com

Piero Cosi – Institute of Cognitive Science and Technologies, CNR, Padua
piero.cosi@cnr.it

Luca Cristoforetti – Fondazione Bruno Kessler (FBK), 38123 Povo, Trento, Italia
cristofo@fbk.eu

Francesco Cutugno – Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, Università degli Studi di Napoli "Federico II", via Claudio 21, 80125 Napoli
cutugno@unina.it

Maria Di Maro – Dipartimento di Studi Umanistici, Università degli Studi di Napoli 'Federico II'
maria.dimaro2@unina.it

Dalila Dipino - Romanisches Seminar, University of Zurich, Switzerland
dalila.dipino@gmail.com

Cecilia Di Nardi – School of Postgraduates studies & Research, Royal College of Ireland, Dublino 2, Irlanda
ceciliadinardi@rcsi.com

Sara Falcone – CIMeC, Università degli Studi di Trento
sara.falcone@studenti.unitn.it

Vincenzo Galatà – National Research Council (ISTC - CNR), Padua
vincenzo.galata@pd.istc.cnr.it

Barbara Gili Fivela – Università del Salento, CRIL – DREAM, Lecce, Italia

Roberto Gretter – Fondazione Bruno Kessler (FBK), 38123 Povo, Trento, Italia
gretter@fbk.eu

Martine Grice – Institut für Phonetik, Universität zu Köln
martine.grice@uni-koeln.de

Alberto Inuggi – Robotics Brain and Cognitive Sciences Unit, Istituto Italiano di Tecnologia, Genova, Italia
alberto.inuggi@iit.it

Khalil Iskarous - University of Southern California
kiskarou@usc.edu

Martina Krüger – Institut für Phonetik, Universität zu Köln
m.krueger@uni-koeln.de

Cristian Leorin – Dept. of Neuroscience, University of Padua
cristian.leorin@gmail.com

Ilaria Mauri – Logopedia, IRCCS Ospedale San Raffaele, Milano, Italia
mauri.ilaria@hsr.it

Daniela Mereu - Libera Università di Bolzano | Alpine Laboratory of Phonetic Sciences
daniela.mereu@unibz.it

Stefano Negrinelli - Romanisches Seminar, Università di Zurigo
stefano.negrinelli@uzh.ch

Francesca Nicora - National University of Ireland, Galway (NUIG) - Ireland

Stefania Nigris – Institute of Cognitive Science and Technologies, CNR, Padua
stefy.nigris@gmail.com

Maurizio Omologo – Fondazione Bruno Kessler (FBK), 38123 Povo, Trento, Italia
omologo@fbk.eu

Antonio Origlia – Dept. of Information Engineering, University of Padua
antori@gmail.com

Antonio Origlia – Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, Università degli Studi di Napoli "Federico II", via Claudio 21, 80125 Napoli
antonio.origlia@unina.it

Duccio Piccardi – Dipartimento di Filologia, letteratura e linguistica. Sezione di linguistica – Università di Pisa. Palazzo Venera, Via S. Maria 36 – Pisa
duccio.piccardi@fileli.unipi.it

Maria Cristina Pinelli – Università degli Studi di Firenze
pinellimariacristina@gmail.com

Cecilia Poletto – Università degli Studi di Padova / Goethe-Universität Frankfurt
cecilia.poletto@unipd.it

Nilo Riva – Dipartimento di Neurologia, Neurofisiologia Clinica e Neuroriabilitazione, IRCCS Ospedale San Raffaele, Milano, Italia
riva.nilo@hsr.it

Antonio Rodà – Dept. of Information Engineering, University of Padua
roda@dei.unipd.it

Valentina Schettino – Dipartimento di Studi Letterari, Linguistici e Comparati, Università degli Studi di Napoli "L'Orientale", via Duomo 219, 80138 Napoli.
vschettino@unior.it

Piergiorgio Svaizer – Fondazione Bruno Kessler (FBK), 38123 Povo, Trento, Italia
svaizer@fbk.eu

Ottavia Tordini – Department of Philology, Literature and Linguistics, University of Pisa
ottavia.tordini@fileli.unipi.it

Rosanna Turrisi – Multiscale Brain Communication, Istituto Italiano di Tecnologia, Ferrara, Italia
rosanna.turrisi@iit.it

Mario Vayra – Department of Classical Philology and Italian Studies, University of Bologna
mario.vayra@unibo.it

Simon Wehrle – Institut für Phonetik, Universität zu Köln
simon.wehrle@uni-koeln.de

Claudio Zmarich – Institute of Cognitive Science and Technologies, CNR, Padua
claudio.zmarich@cnr.it

Revisori del volume

Cinzia Avesani (ISTC-CNR, Padova)
Leonardo Badino (IIT, Ferrara)
Chiara Bertini (Scuola Normale Superiore, Pisa)
Silvia Calamai (Università di Siena)
Francesco Cangemi (Universität zu Köln)
Chiara Celata (Scuola Normale Superiore, Pisa)
Piero Cosi (ISTC, Padova)
Francesco Cutugno (Università Federico II, Napoli)
Maria Paola D'Imperio (CNRS-Aix-Marseille Université)
Silvia Dal Negro (Libera Università di Bolzano)
Mauro Falcone (Fondazione Ugo Bordoni, Roma)
Vincenzo Galatà (ISTC-CNR, Padova)
Barbara Gili Fivela (Università del Salento, Lecce)
Pietro Maturi (Università Federico II, Napoli)
Daniela Mereu (Libera Università di Bolzano)
Maurizio Omologo (FBK, Trento)
Antonio Origlia (Università Federico II, Napoli)
Irene Ricci (Scuola Normale Superiore, Pisa)
Antonio Romano (Università di Torino)
Stephan Schmid (Universität Zürich)
Lorenzo Spreafico (OAW, Vienna)
Antonio Stella (Google)
Mario Vayra (Università di Bologna)
Alessandro Vietti (Libera Università di Bolzano)
Enrico Zovato (Nuance Communications)

Studi AISV è una collana di volumi collettanei e monografie dedicati alla dimensione sonora del linguaggio e alle diverse interfacce con le altre componenti della grammatica e col discorso. La collana, programmaticamente interdisciplinare, è aperta a molteplici punti di vista e argomenti sul linguaggio: dall'attenzione per la struttura sonora alla variazione sociofonetica e al mutamento storico, dai disturbi della parola alle basi cognitive e neurobiologiche delle rappresentazione fonologiche alle applicazioni tecnologiche. I testi sono selezionati attraverso un processo di revisione anonima fra pari e vengono pubblicati nel sito dell'Associazione Italiana di Scienze della Voce con accesso libero a tutti gli interessati.

Alessandro Vietti è ricercatore in linguistica e responsabile di ALPS (Alpine Laboratory of Phonetic Sciences) presso la Libera Università di Bolzano. Nel campo della fonetica sperimentale si occupa principalmente di variazione sociofonetica, parlato di bilingui, sviluppo di metodi statistici di analisi multivariata sia in prospettiva acustica che articolatoria.

Lorenzo Spreafico è ricercatore in linguistica presso l`Università degli Studi di Bergamo. I suoi attuali interessi di ricerca includono i processi di acquisizione di lingue in soggetti monolingui e bilingui, con particolare riguardo alla dimensione della fonetica articolatoria.

Daniela Mereu è assegnista di ricerca in linguistica presso la Libera Università di Bolzano. Le sue ricerche vertono principalmente su temi legati alla variazione linguistica dell'italiano e del sardo, secondo un approccio sociofonetico.

Vincenzo Galatà è assegnista di ricerca presso la sede di Padova dell'Istituto di Scienze e Tecnologie della Cognizione del Consiglio Nazionale delle Ricerche. Nel campo della fonetica sperimentale si occupa di parlato di soggetti bilingui e di variazione sociofonetica, di acquisizione del linguaggio e di sviluppo fonetico-fonologico in bambini prescolari italiani nativi e non nativi.

AISV - Associazione Italiana Scienze della Voce
sito: www.aisv.it
email: aisv@aisv.it| redazione@aisv.it
ISBN: 978-88-97657-28-6

Edizione realizzata da
Officinaventuno
info@officinaventuno.com | sito: www.officinaventuno.com
via F.lli Bazzaro, 18 - 20128 Milano - Italy