ISSN: 2612-226X

Studi AISV 12

LA VOCE NEI MEDIA E NELLE NUOVE TECNOLOGIE: PRODUZIONE E PERCEZIONE

THE VOICE IN THE MEDIA AND NEW TECHNOLOGIES: PRODUCTION AND PERCEPTION

a cura di

Valentina De Iacovo, Bianca Maria De Paolis e Daniela Mereu

AISV
Associazione Italiana Scienze della Voce

LA VOCE NEI MEDIA E NELLE NUOVE TECNOLOGIE: PRODUZIONE E PERCEZIONE

THE VOICE IN THE MEDIA AND NEW TECHNOLOGIES: PRODUCTION AND PERCEPTION

a cura di
Valentina De Iacovo, Bianca Maria De Paolis
e Daniela Mereu

Milano 2024

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: 978-88-97657-73-6
ISSN: 2612-226X

Sommario

DOI: 10.17469/O2112AISV000001

VALENTINA DE IACOVO, BIANCA MARIA DE PAOLIS, DANIELA MEREU

Prefazione

Il parlante comunica utilizzando (anche) la propria voce, che veicola, con diversi gradi di controllo, informazioni di carattere biologico, geografico, sociale, culturale, psicologico, emotivo e stilistico. Il contesto in cui l'evento comunicativo ha luogo ne influenza fortemente l'uso. Tra questi, i media rappresentano un ambito di studio ancora poco esplorato rispetto a tematiche che trattano aspetti di fonetica e che si presta quindi a ricerche di carattere interdisciplinare. Il presente volume raccoglie una serie di contributi su questo tema, che costituiscono una selezione dei lavori presentati durante il XX Convegno Nazionale AISV (Associazione Italiana Scienze della Voce) – *Voci, media e nuove tecnologie*, ospitato all'Università di Torino nel febbraio 2024. I saggi selezionati sono incentrati prevalentemente sul tema della voce nei media – tradizionali, come la radio e la televisione, e più recenti, come le piattaforme web, i social network e le applicazioni destinate alla messaggistica vocale – e nelle nuove tecnologie. Seguendo la tradizione della collana AISV, inoltre, anche in questo volume sono state accolte proposte a tema libero, dedicate a qualsiasi aspetto della ricerca sulla voce e sul parlato. La struttura di questo volume riflette pertanto questa varietà di temi.

Apre il volume il contributo di Josiane Riverin-Coutlée e Jonathan Harrington. I due autori sviluppano il tema dell'uso del parlato nei media come risorsa per lo studio dei cambiamenti fonetici nell'arco della vita di singoli individui. L'articolo illustra alcuni studi longitudinali condotti su tre personaggi pubblici: la Regina Elisabetta II, Alistair Cooke e Michaëlle Jean. La discussione dei risultati sulle vocali dimostra che il cambiamento tende a essere graduale, selettivo e condizionato da diversi fattori, quali i cambiamenti fonetici in atto nella comunità, la mobilità geografica e gli eventi biografici specifici del parlante. Il lavoro mette in luce il prezioso contributo del parlato mediatico negli studi di carattere fonetico, sottolineando la dovuta cautela che occorre adottare durante le fasi di selezione e analisi dei dati.

Il contributo di Crochiquia, Eriksson e Madureira si concentra invece su due esperimenti di percezione che indagano il ruolo delle voci dei personaggi dei film d'animazione nel fornire agli ascoltatori informazioni indicative sulla propria individualità. Il primo esperimento si basa sulla percezione auditiva delle caratteristiche fisiche, psicologiche e sociali. Il secondo mostra invece in che modo il contenuto lessicale influenzi la valutazione delle caratteristiche dell'oratore, facendo emergere il ruolo del parlato nella valutazione percettiva.

Nel capitolo successivo, Devauchelle, Rilliard, Doukhan e Ondel Yang illustrano un'analisi delle voci dei parlanti che compongono un *corpus* diacronico basato su programmi televisivi o radiofonici francesi riconducibili al periodo compreso dagli

in L1 o L2 possa influire sulla qualità dell'acquisizione linguistica. In particolare lo studio indaga l'influenza di tre condizioni di sottotitoli sull'apprendimento della pronuncia delle vocali /ʌ/ e /æ/ in inglese come seconda lingua: i risultati suggeriscono un miglioramento nella produzione fonetica per il gruppo che è stato sottoposto ai sottotitoli in lingua originale. La competenza ritmica di parlanti anglofoni che apprendono l'italiano come lingua straniera è invece il tema dello studio di Maffia, Vacchiano e De Meo. I risultati mostrano che la durata delle vocali, calcolata attraverso diverse metriche ritmiche, gioca un ruolo cruciale nella definizione della competenza ritmica; inoltre, alcune differenze emergono anche in relazione al livello di competenza degli apprendenti, suggerendo una complessità nella caratterizzazione del ritmo dell'interlingua.

Si prosegue quindi con una breve sezione che riguarda più da vicino alcuni aspetti tecnologici legati alla voce. L'articolo di Gay e Combei affronta il tema del divario esistente tra la percezione di infallibilità delle tecnologie vocali e l'effettiva realtà dei loro limiti, imprecisioni e problemi etici. Gli autori dimostrano come alcuni fattori (nel caso specifico, l'età) possano influenzare il grado di accuratezza nell'identificazione delle classi della variabile oggetto di indagine (il genere). L'obiettivo più generale del lavoro è incrementare la consapevolezza del potenziale di errore che questi sistemi presentano e di promuovere un approccio più prudente, inclusivo e informato al loro sviluppo e utilizzo. All'esplorazione del ruolo della sillaba nei sistemi di riconoscimento automatico del parlato (ASR) è invece dedicato il contributo di Picciau, De Cristofaro e Vietti. Attraverso l'addestramento di un modello ASR neurale e un'analisi linguistica, gli autori osservano come fattori quali la frequenza dei token di sillaba, la posizione dell'accento lessicale, il tipo di sillaba e le parti del discorso influenzano la rappresentazione neurale delle sillabe.

Infine, il contributo di Fontanella, Balsemin, Gili Fivela, Perrier, Savariaux, Tisato e Zmarich indaga lo statuto fonologico delle geminate in italiano rispetto alla loro organizzazione temporale con le vocali adiacenti. In particolare, gli autori si concentrano su quali siano i meccanismi articolatori coinvolti nell'organizzazione gestuale delle geminate italiane, basandosi sui dati degli intervalli temporali dei gesti consonantici rispetto a quelli vocalici attraverso lo sfruttamento di due modelli di sincronizzazione gestuale, il "Combined Vowel and Consonant" e il "Vowel-to-Vowel".

La raccolta dei lavori qui brevemente delineati rende, a nostro parere, la pubblicazione di questo volume un prezioso contributo per la riflessione scientifica della comunità dei fonetisti e dei linguisti, offrendo validi strumenti metodologici e aprendo nuove strade di ricerca. Il merito di questo sforzo va quindi a tutti gli autori e le autrici dei saggi, che desideriamo ringraziare vivamente.

Un sentito ringraziamento va infine ai membri del Comitato Scientifico per la loro generosa collaborazione alla revisione dei lavori:

Giovanni Abete (Università degli Studi di Napoli "Federico II")
Cecilia Maria Andorno (Università degli Studi di Torino)
Paolo Baggia (Nuance Communications)
Maria Grazia Busà (Università degli Studi di Padova)

DOI: 10.17469/O2112AISV000002

JOSIANE RIVERIN-COUTLÉE, JONATHAN HARRINGTON

Phonetic change over a lifetime in the media

Speech in the media is increasingly exploited in phonetic studies, as it constitutes unprecedented amounts of data easily accessible to researchers. In this contribution, we exemplify how speech in the media can be used to advance our understanding of phonetic change over the lifespan, an issue long hampered by data scarcity. We present longitudinal studies on three public figures – Queen Elizabeth II, Alistair Cooke, and Michaëlle Jean – in which recordings of single individuals over several decades were analyzed to explore the extent of phonetic flexibility during adulthood. Focusing on vowels, these studies showed that change tends to be gradual, non-linear, selective, and conditioned by social factors such as sound changes in progress in the community, geographic mobility, and speaker-specific life events. Speech in the media can contribute valuable data to a wide range of phonetic studies, provided that some caution is exercised when selecting and analyzing materials to avoid common pitfalls.

Keywords: speech in the media, public figures, phonetic change, longitudinal, social factors.

1. *Introduction*

This contribution is concerned with how speech in the media can be and has been utilized to examine issues in phonetic sciences for which the short time spans typically covered by traditional speech resources are a limitation. Our focus is on phonetic changes which may occur during adult life, but which are not due to the physiological effects of age. We first describe in § 2 a relatively recent paradigm shift in phonetics, wherein adulthood used to be considered a period of stability until longitudinal work questioned this assumption, thereby giving rise to a new generation of longitudinal studies. The concept of speech in the media as the object of phonetic investigation is then framed in § 3, with several examples of recent longitudinal studies which have exploited this type of speech materials. We cover three series of such studies in § 4 on Queen Elizabeth II (§ 4.1), Alistair Cooke (§ 4.2) and Michaëlle Jean (§ 4.3). We discuss the theoretical implications of these three cases in § 5, after which we broaden the discussion to the opportunities offered by speech in the media for phonetic studies, as well as basic precautions to take when using this type of resource.

2. *Speech over the lifespan*

Speech is known to change over the course of an individual's life. The study of child language acquisition, in particular, is concerned with how speech evolves from

cooing and babbling to full proficiency as children develop their cognitive and motor skills (Chung, Kong, Edwards, Weismer, Fourakis & Hwang, 2012; Clark, 2009; Lee, Potamianos & Narayanan, 1999). Physiological changes occurring during childhood and adolescence, but also much later in life, sometimes due to illness, can have an effect on some speech parameters such as the fundamental frequency (Linville, 2001; Markova, Richler, Pangelinan, Schwartz, Leonard, Perron, Pike, Veillette, Chakravarty, Pausova & Paus, 2016; Perry, Ohde & Ashmead, 2001). Social factors can contribute to some other changes within the individual. The life stage of adolescence is well known for the "vernacular peak" and related phenomenon of age grading (Eckert, 1997; Kirkham, Moore, 2013; Labov, 2001).

In contrast, healthy active adulthood has long been considered a period of linguistic stability (Chambers, 2009; Eckert, 1997; Pichler, Wagner & Hesson, 2018), partly due to the influential "critical period hypothesis" (Lenneberg, 1967) which suggested that major changes were unlikely after puberty. However, an emerging body of longitudinal studies have more recently questioned this assumption. In a benchmark study, Sankoff, Blondeau (2007) investigated longitudinally variants of /r/ in the "Montreal project", a large corpus of sociolinguistic interviews with French-speaking Montrealers which included 120 participants in 1971, 60 of whom were recorded a second time in 1984, and 12 a third time in 1995 (Sankoff, 2018; Sankoff, Cedergren, 1971; Thibault, Vincent, 1990; Vincent, Laforest & Martel, 1995). One of the main findings of Sankoff, Blondeau (2007) was that many adult participants produced a higher rate of dorsal variants [ʁ]/[ʀ] in 1984 than in 1971, in line with a change in progress in Montreal French, where the apical variant [r] was rapidly replaced by [ʁ]/[ʀ]. This type of work fueled an interest for the hitherto poorly understood phonetic flexibility of adult speakers, which seemed greater than previously assumed (e.g. Gerstenberg, Voeste, 2015; Buchstaller, Wagner, 2018; Beaman, Buchstaller, 2021).

However, an obstacle to this endeavor has been the scarcity of longitudinal corpora such as the Montreal project (Hazan, 2017; Sankoff, 2018). This is because of difficulties in keeping prolonged contact with participants, who may move away, fall ill, decide to withdraw; limited financial resources; research teams changing focus; insufficient quality of older recordings; etc. (Gerstenberg, Voeste, 2015; Young, Powers & Bell, 2006). We focus in this paper on a solution adopted by various researchers to overcome this issue: using speech in the media.

3. *Speech in the media*

In the context of this contribution, "speech in the media" as the object of phonetic studies refers to pre-recorded or live speech delivered by a public figure to their audience (Goffman, 1981). Public figures include journalists, actors, musicians, politicians, royalty, athletes, and people from various other professions who engage with an audience on a regular basis. Nowadays, the speech of such public figures is regularly recorded, at known time stamps, and is often made available on open

digital platforms (e.g. national broadcasters, YouTube and other social media). In addition, more or less detailed biographic information about these individuals is usually disclosed. All this makes speech in the media well suited for investigating the extent of phonetic flexibility during adulthood, first because of the availability of large quantities of recorded speech materials from unique individuals over potentially very long time periods. Second, the relative weight of various external factors on speech across the lifespan can also be addressed thanks to details of the speakers' professional and personal lives being known.

Indeed, recent years have seen a spurt in longitudinal studies of public figures' speech which have examined a variety of parameters and phenomena, e.g. vowels, consonants, suprasegmental features, style shifting, second dialect acquisition and bilingualism. To name but a few, these include studies on Muhammad Ali (Berisha, Liss, Huston, Wisler, Jiao & Eig, 2017), Piero Angela (Giannini, Pettorino, 2009), David Attenborough (MacKenzie, 2017), Queen Beatrix (Quené, 2013), Dagmar Berghoff (Reubold, Harrington, 2018), Noam Chomsky (Kwon, 2018), Alistair Cooke (Reubold, Harrington, 2015), Queen Elizabeth II (Harrington, Palethorpe & Watson, 2000a), Ruth Bader Ginsburg (Shapp, LaFave & Singler, 2014), Stefanie Graf (de Leeuw, 2019), Michaëlle Jean (Riverin-Coutlée, Harrington, 2022), Helmut Kohl (Braun, Friebis, 2009), Margaret Lockwood (Reubold, Harrington & Kleber, 2010), Jennifer Lopez (Bauernfeind, 2020), Ed Miliband (Kirkham, Moore, 2016), Barack and Michelle Obama (D'Onofrio, Stecker, 2022; Holliday, 2017), King Rama IX (Yang, Pittayaporn, Kirby & Jitwiriyanont, 2021), Oprah Winfrey (Hay, Jannedy & Mendoza-Denton, 1999) and Fareed Zakaria (Sharma, 2018).

The following section presents in a more detailed manner longitudinal studies on three of the aforementioned public figures: Queen Elizabeth II, Alistair Cooke and Michaëlle Jean. These studies were conducted by the authors and their collaborators, and have in common that they are all concerned with changes in vowel quality over several decades of the speakers' lives. By means of acoustic analyses of speech in the media, these studies address the issue of the extent of phonetic flexibility during adulthood. They also consider the influence of external factors such as sound change in progress in the community, geographic mobility, and important milestones in an individual's life.

4. *Longitudinal studies of three public figures*

4.1 Queen Elizabeth II

The first series of case studies we present focused on the late Queen Elizabeth II (1926-2022). Each year of her reign, the Queen delivered a 5-minute Christmas message to the UK and Commonwealth, totaling 70 broadcasts (1952-2021) by the same speaker, to the same audience, on relatively similar topics, even using the same words (e.g. "I wish you all [...] a very happy Christmas"). The first few broadcasts were on radio only, then televised starting in 1957, but always in good recording conditions for the time. While some of these are nowadays available on public

platforms such as YouTube, to carry out the studies described hereafter, Harrington and collaborators obtained recordings of the Queen's Christmas broadcasts from the BBC, with permission from Buckingham Palace.

Broadly speaking, the Queen's pronunciation could be described as an upper-class version of the Received Pronunciation (RP) of British English, RP being a variety of English typically associated with speakers of the BBC and people with higher levels of education more generally (Roach, 2004). Given that many RP features changed over the 20th century, Harrington et al. (2000a; 2000b) raised the question of whether the Queen's pronunciation changed over her lifetime as well. They first analyzed stressed monophthongs in broadcasts from three decades: the 1950s (with broadcasts from 1952, 1954 and 1957), the 1960s (1967, 1968 and 1972) and the 1980s (1983, 1985 and 1988) (Harrington et al., 2000b). To compare the Queen's vowels with those of mainstream RP (that typically characterized a more middle class accent), the same analysis was made of vowels produced by five female BBC broadcasters recorded in the 1980s.

The results showed changes in either F1, F2, or both, for most of the Queen's vowels, and usually in the direction of the BBC control group. Changes were generally larger between the 1950s and 1960s than between the 1960s and 1980s. Two vowels in particular changed substantially: /u/ was fronted and /æ/ was lowered. The results from other acoustic studies of vowels on the effects of aging (e.g. Reubold et al., 2010; Reubold, Harrington, 2018) suggested that these two specific changes found in the Queen's speech could not be attributed to the physiological maturation of the vocal tract. With increasing age, the fundamental frequency of female speakers tends to decrease, as well as non-low vowels' F1 frequency; whereas /u/-fronting involves a higher F2 and /æ/-lowering a higher F1. In addition, both /u/-fronting and /æ/-lowering were changes observed in RP over the 20th century (Bauer, 1985). The Queen's speech was thus likely influenced by phonetic changes taking place in the community. Given that the role of royalty in society is arguably more to safeguard traditions and institutions (including the so-called Queen's English; see Harrington et al., 2000a; 2000b) than to be avant-garde, the finding that the Queen's pronunciation changed along with that of her subjects made a strong case for seriously considering lifespan change in phonetic studies. Moreover, Harrington et al. (2000b) argued that the concept of apparent time (Cukor-Avila, Bailey, 2013), widely used to study language change and which is based on the assumption that adult speech is stable (Eckert, 1997), probably underestimated the rate of change within a community (see also Labov, 1994).

The results additionally showed that lifespan change was not necessarily linear: while some changes were observed over the two decades separating the 1960s and 1980s, these were smaller than the changes observed in the one decade separating the 1950s from the 1960s (see also Bowie, 2015). Moreover, while the Queen's /u/ fronted and /æ/ lowered, and generally *tended* towards those of the BBC control group, they only tended towards, without attaining the positions of the BBC control group. The Queen's participation in sound change appeared only partial, with vowel

qualities neither quite like those produced by the Queen in the 1950s, nor those of the 1980s' RP (Harrington et al., 2000b). Both trends suggest that lifespan change is more gradual than categorical, i.e. vowels could acquire intermediate qualities after decade-long drifts.

To summarize, studies of the Queen's Christmas broadcasts provided evidence that adults retained some phonetic flexibility across the lifespan. Even speakers expected to be relatively conservative may come to adopt innovative variants in the context of a change in progress in the community. In addition to their theoretical implications (e.g., relative to the apparent time concept), these studies opened up new questions about the potentially non-categorical nature of sound change and non-linearity of individual trajectories.

4.2 Alistair Cooke

The second public figure covered in this section is Alistair Cooke, a journalist and radio broadcaster born in 1908 near Manchester, UK. In 1937, Cooke moved to the USA, where he lived until his death in 2004, at the age of 95 years old. He is best known for *Letter from America*, a weekly 15-minute radio show broadcast by the BBC in which he presented to a (mostly) British audience topics related to American society, culture, politics and latest news. *Letter from America* is the longest-running radio progam with the same single presenter (1946-2004 without interruption), making it extremely rich longitudinal speech data, now available in full on the BBC website. An official biography of Alistair Cooke was also published (Clarke, 2000).

After a first analysis of the effects of age on Cooke's fundamental and formant frequencies (not summarized here; see Reubold et al., 2010), the influence of geographic relocation on his vowels was addressed in Reubold, Harrington (2015). For this purpose, the authors selected broadcasts of *Letter from America* from 1951, 1960, 1970, 1981, 1993 and 2004 (ages 42 to 95), in which they measured the first two formants of stressed monophthongs. They found phonetic changes over time in the quality of the BATH, THOUGHT, LOT and DRESS vowels, which differ between RP and American English (see Wells, 1982). For THOUGHT, LOT and DRESS, Cooke progressively changed from an American pronunciation in the 1950s-to-1970s to RP by 2004. For BATH, the change was more abrupt: from an American quality in 1970, it had shifted to RP by 1981. In other words, the more time Cooke spent in the USA, the less American his BATH, THOUGHT, LOT and DRESS vowels were. On the other hand, other vowels differing between RP and American English never qualified as American, i.e. the FORCE, NORTH, NURSE and START vowels remained non-rhotic as in RP over the entire dataset, resulting in a hybrid pronunciation during the 1950s to 1970s (Munro, Derwing & Flege, 1999; Nycz, 2015; Siegel, 2010).

In a follow-up study, Reubold, Harrington (2018) added to the radio corpus described above an interview conducted in 1934, when Alistair Cooke was 25 years old and still living in the UK. The analysis of this additional recording showed that

Cooke then spoke a variety of English that qualified as RP, which confirmed that his American BATH, THOUGHT, LOT and DRESS vowels of the 1950s-to-1970s were acquired after moving to the USA. Moreover, Reubold, Harrington (2018) suggested that RP was probably the speaker's second dialect, adopted as a young adult when he left the north of England to study at Cambridge. To further analyze the apparently abrupt shift in Cooke's BATH vowel, Reubold, Harrington (2018) examined 14 additional broadcasts of *Letter from America* from the years 1975 to 1977. Over this short time frame, when the speaker was aged 66 to 68 years old, the fronted American variant [æ] backed into the RP variant [ɑː]. The change was gradual, but took place over a much shorter period of time than those of the THOUGHT, LOT and DRESS vowels. As for why this accent reversal happened at this particular time, Reubold, Harrington (2015; 2018) pointed out that it coincided with a period of disillusion for Cooke. His initial enthusiasm about his host country (he rapidly became an American citizen, married an American and started a family, traveled the country for work) faded with the difficult political climate of the mid-1970s perhaps leading to a projection of himself as an outsider (Clarke, 2000). Cooke was also in contact with a greater diversity of Americans in earlier years, when he traveled the country, than in later years.

To summarize, these longitudinal studies of Alistair Cooke's vowels revealed the maintenance of great phonetic flexibility during his life. A reversal of some phonetic changes was also observed, with his pronunciation following the path: Northern British English > RP > American English > RP. Changes in vowel quality were mostly gradual, although they occurred over different time periods across vowels, and did not affect Cooke's entire vowel system. Finally, external factors contributed to explain these shifts: geographic mobility, but also speaker-specific events.

4.3 Michaëlle Jean

The third public speaker we are concerned with is Michaëlle Jean, a Canadian journalist and diplomat born in Haiti in 1957. Jean moved from Haiti to Quebec with her family at the age of 11 years old, and was brought up speaking French and Creole. Her career as a public speaker started in 1988 when she became a journalist for Canada's national French-speaking broadcaster Radio-Canada. As will be detailed below, Jean's professional occupations changed several times after 1988, but she always remained a public figure whose speech was regularly recorded. This allowed us to investigate longitudinally the influence of these career shifts on her pronunciation. For this purpose, a five-hour corpus spanning three decades of Jean's public career (1989-2022) was collated from various open sources such as YouTube and the Radio-Canada archives, totaling 65 recordings varying between 18 seconds and 29 minutes in length.

Riverin-Coutlée, Harrington (2022; in press) started from the hypothesis that since much of adult life revolves around work, turning points in an individual's career could influence the way they speak (Eckert, 1997; Pichler et al., 2018; Wirtz, Pickl, 2025). Jean was a particularly interesting case to explore this further, first

because she had many career changes at well-defined points in time: she was a journalist in Quebec from 1988 until 2005; she was governor general of Canada from 2005 until 2010; she was UNESCO's special envoy to Haiti from 2010 until 2014; she was secretary general of the international organization La Francophonie from 2014 until 2018; and since 2019, she has been promoting and sponsoring social causes through her foundation, and is regularly invited to speak at public forums. Second, these stages of Jean's career had linguistic relevance: as secretary general of La Francophonie, she acted as the global representative of 321 million French speakers from 93 states and countries (Organisation internationale de la francophonie, 2025), while as a journalist, her francophone identity was arguably less enhanced and the audience she typically engaged with was less international, i.e. mostly composed of Quebecers. This raised the question of whether Jean would produce more features of Quebec French as a journalist than as secretary general of La Francophonie.

In an acoustic analysis of the high vowels /i y u/, which in Quebec French split into tense [i y u] and lax [ɪ ʏ ʊ] variants depending on the consonantal context (Sigouin, Arnaud, 2014), Riverin-Coutlée, Harrington (2022) found that Jean produced this phonetic feature during her time as a journalist in Quebec, but progressively suppressed it by the time she became secretary general of La Francophonie. After 2019, a reversal of this trend was observed. In a follow-up study, Riverin-Coutlée, Harrington (in press) examined other features, among which the phonological difference in Quebec French between the FAITE and FÊTE vowels, where the former is a short monophthongal /ɛ/ and the latter is a long diphthongal /a͡ɛ/ (Riverin-Coutlée, Roy, 2022). Contrary to the tense-lax split of /i y u/, this feature exhibited stability over Jean's career. Part of these results confirmed that changes in a person's professional life could have an effect on the way they spoke; integrating analyses of career changes into the sociolinguistic toolkit (Wirtz, Pickl, 2025) could therefore benefit a field which already had a tradition of examining variation across social classes largely defined relative to people's income and occupation (Ash, 2002; Labov, 2001). On the other hand, the results revealed constraints on Jean's flexibility: while phonetic variants (or allophones) merged and unmerged, a phonemic contrast showed stability.

Jean's speech also presented some differences with Quebec French described in other studies. During her years as a journalist, she produced tense and lax variants of /i y u/ which were not as acoustically distinct as those found in the community (Sigouin, Arnaud, 2014). Moreover, the contrast between FAITE and FÊTE was realized through length, not the vowels' mono- or diphthongal quality. One possible explanation for this was that Jean only partially adopted Quebec French after having spent the first 11 years of her life in Haiti (Munro et al., 1999; Nycz, 2015; Siegel, 2010). Alternatively, this could reflect a normative style designed for mass media (Chalier, 2021; Reinke, Ostiguy, 2016). Whichever the explanation, these studies revealed the presence of intermediate variants in Jean's speech, e.g. her not-so-distinct and further progressively merging tense and lax vowels. In addition,

at all points in time, her pronunciation was hybrid: not entirely typical of Quebec French during her years as a journalist due to the monophthongal quality of FÊTE, but not completely devoid of Quebec French features as secretary general of La Francophonie given the long duration of FÊTE.

To summarize, these longitudinal studies of Michaëlle Jean's vowels suggested that career may influence speech across the lifespan. They also uncovered some possible limitations to Jean's flexibility, i.e. a phonological contrast did not change, contrary to a phonetic feature. As with the Queen, intermediate variants were produced; as with Cooke, a reversal of a change occurred and a hybrid pronunciation was observed.

5. *Discussion and conclusions*

The purpose of this paper was to exemplify how speech in the media can be exploited as research materials in (socio)phonetics. Speech in the media has been shown to be a valuable resource for taking a lifespan perspective on phonetic changes occurring during adulthood, a period of life previously assumed to be largely stable. The longitudinal studies of three public figures – Queen Elizabeth II, Alistair Cooke and Michaëlle Jean – covered in § 4 brought together acoustic analyses of vowels taken from corpora spanning multiple decades with social factors to better understand adults' phonetic flexibility. While at first sight generalizations that can be made from these exceptional speakers may appear limited, the events which seem to have influenced their phonetic trajectories are likely to be experienced by many people, e.g. sound change in progress in the community, geographic relocation and career changes. The theoretical and methodological implications of these studies are thus manifold, as discussed below.

A noteworthy finding from these studies, as well as from other longitudinal work (see Bowie, Yaeger-Dror, 2015 for a review), is that adults change their pronunciation over the lifespan. As mentioned in § 2, this was not a given in earlier literature, where speech was assumed to be stable after puberty (Eckert, 1997). This finding exposes a paradox surrounding apparent time, a methodological tool frequently used to investigate language change, where speakers of two generations are compared; if the older generation's speech characteristics differ from those of the younger generation, it is assumed that a change has occurred (Cukor-Avila, Bailey, 2013; Labov, 1994). However, it seems that older speakers' phonetic characteristics may be influenced by the very change examined. As Harrington et al. (2000b) suggested, apparent time comparisons could underestimate the rate at which change sweeps through a speech community, because the older generation may not have remained entirely stable since puberty and may, to a certain extent, participate in the change, as observed with the Queen. While apparent time remains a useful tool, its outcome should be interpreted with caution and ideally complemented with longitudinal data, because as argued by Sankoff (2018: 43), "[in] studying language change, nothing can replace real time". The finding that adult speech can change also has an impact for dialectology

(Chambers, Trudgill, 1998), where older, non-mobile speakers have typically been prioritized because their conservative speech characteristics offered a window on a previous state of the language (e.g. Poplack, St-Amand, 2007). However, in light of the research summarized here, which suggests that life events of a more modest magnitude than geographic relocation may influence adult speech (Reubold, Harrington, 2018; Riverin-Coutlée, Harrington, 2022), it cannot be assumed that the phonetic characteristics of an 80-year-old person in 2025 are exactly those of that same 20-year-old person in 1965. As succinctly noted by Pichler et al. (2018: 2), older speakers are not "homogeneous diachronic data repositories".

The studies summarized in § 4 also contribute to defining the extent of adults' phonetic flexibility. First, there is evidence that change is gradual. The Queen was found to participate in two phonetic changes in progress in RP, i.e. /u/-fronting and /æ/-lowering, but the variants she produced in the 1980s were still different from those of RP speakers. In an analysis of more recent Christmas broadcasts, Harrington, Reubold (2021) measured a small reversal in these changes in /u/ and /æ/ after the 1980s, but not to the extent that they were back to their 1950s' qualities. Acoustic changes measured in Cooke's speech also involved intermediate vowel qualities. For THOUGHT, LOT and DRESS, the change from American English to RP between the 1970s and 2000s was slow and gradual; for BATH, change occurred more rapidly between 1975 and 1977, but it was still not categorical. The /i y u/ vowels analyzed over Jean's career were found to be split between tense and lax variants when she was a journalist (1988-2005), merged during her time as secretary general of La Francophonie (2014-2018), but were at intermediate degrees of split during the other periods examined (2005-2010, 2010-2014 and 2019-2022). Perhaps some other speech features are more prone to categoricity than vowel quality; for example, in Sankoff, Blondeau's (2007) study on Montreal French, it is not the case that there is a continuum of variants between the innovative dorso-uvular fricative [ʁ] and conservative apico-alveolar trill [r]. Nonetheless, Sankoff, Blondeau (2007) calculated non-categorical *rates* of production of these two variants. We thus conclude from these various observations that phonetic change over the lifespan seems more gradual than categorical.

Furthermore, the studies summarized in § 4 showed that change during adulthood may lead to a hybrid pronunciation. During the period where the quality of Cooke's BATH, THOUGHT, LOT and DRESS vowels was closer to American English, his FORCE, NORTH, NURSE and START vowels were non-rhotic as in RP. Nycz (2015: 470) mentioned that adults moving to a new geographic area and adopting a second dialect often did so "à la carte", that is, picking up on some features but not others, as with Cooke (see also Evans, Iverson, 2007). Given that Michaëlle Jean spent the first years of her life in Haiti and then moved to Quebec at the age of 11 years old, this may be a possible explanation for why she only partially produced Quebec French features during her years as a journalist (i.e., why the tense-lax split of /i y u/ was less distinct than that found in the community, and why FAITE-FÊTE were distinguished via length but not quality). Taken together, all these studies suggest that although adults

retain some phonetic flexibility, the types of changes may be limited. In particular, phonological contrasts may be more resistant to change than phonetic differences: this could explain why Jean's phonetic tense-lax split merged and unmerged while the phonological length difference between FAITE-FÊTE remained stable over her whole career. Kwon (2018) also observed that Noam Chomsky tended to exhibit phonetic change but phonological stability across time.

Another important finding is that a reversal of previously observed changes happened for all of the Queen, Cooke and Jean. As mentioned above, Harrington, Reubold (2021) found less fronted /u/ and less lowered /æ/ in Christmas broadcasts from the 1990s to 2010s, compared to those of the 1980s. Some of the vowels of American English which Cooke adopted after moving to the USA reverted to RP after the 1970s (Reubold, Harrington, 2018). The merged tense and lax high vowels of Jean slightly unmerged in the last stage of her career (Riverin-Coutlée, Harrington, 2022). A similar trend was reported by Shapp et al. (2014) for Ruth Bader Ginsburg, who reverted to producing more non-standard variants of New York City English in the 2000s than she did in the 1970s. MacKenzie (2017) observed an increase in intervocalic /r/-tapping in David Attenborough's later narrations, whereas the tapped variant was produced less and less in the speech community (Hughes, Trudgill & Watt, 2012). While social factors may at least partly explain some of these reversals, as hypothesized for Cooke and Jean (see also Pichler et al., 2018), Harrington, Reubold (2021) connected this recurring tendency to exemplar theory and the cognitive processing of speech. According to exemplar theory, listeners memorize speech with extensive amounts of phonetic and non-linguistic detail (Drager, Kirtley, 2016; Foulkes, Docherty, 2006; Johnson, 1997; Pierrehumbert, 2016; *inter alia*). The new exemplars which listeners keep memorizing over the course of their lives have an impact on their phonological representations and speech productions (e.g. Harrington, Kleber, Reubold, Schiel & Stevens, 2018; Hay, Drager & Warren, 2010), which in certain cases may cause phonetic change, for example when moving to a new dialect area (Nycz, 2013). However, declining cognitive functions with aging perhaps mean that the memorization of new exemplars is less robust later than earlier in life, and therefore, that older exemplars influence to a greater extent older speakers' speech representations and productions (Harrington, Reubold, 2021; see also Pichler et al., 2018).

Harrington, Reubold's (2021) interpretation of the Queen's longitudinal data relative to cognitive processes reflects that speech in the media offers scope for more research than that covered in this paper, which was concerned with phonetic changes in vowels and their connection to a few social factors. Diversifying the pool of languages represented in this line of research is an important step towards understanding lifespan stability and change for other linguistic features, for example tones, as investigated by Yang et al. (2021) in Thai-speaking King Rama IX. Studies of speech in the media may also help uncover influential social factors that are not necessarily part of standard variationist studies, for instance political affiliation (Hall-Lew, Coppock & Starr, 2010). Moreover, because it potentially covers very

long time periods that are difficult to match in the laboratory, speech in the media is a precious resource for documenting the effects on speech of physiological changes due to aging. In an analysis of recordings of Piero Angela's broadcasts 40 years apart, Giannini, Pettorino (2009) observed numerous age-related changes, e.g. a slower speech rate, longer segments and pauses, a larger vowel space, a reduced distinction between stressed and unstressed syllables, a higher fundamental frequency, etc. Beyond healthy aging, speech in the media can occasionally be used to investigate the effects of certain pathologies. For instance, capitalizing on decades of recordings of Muhammad Ali, Berisha et al. (2017) found symptoms typical of hypokinetic dysarthria (Duffy, 2020) up to seven years before he was diagnosed with Parkinson's disease. This has important clinical implications in suggesting that practitioners could make use of speech to diagnose Parkinson's disease earlier in some patients, provided that changes belonging to healthy versus unhealthy aging are well defined, which longitudinal studies on speech in the media can contribute to.

Despite its many possibilities, longitudinal materials can also be confounded with so many different factors. For example, professional speakers with voice training may cope differently with physiological changes normally affecting speech in older age, as suggested for Margaret Lockwood (Reubold et al., 2010) and Dagmar Berghoff (Reubold, Harrington, 2018), whose non-low vowels' F1 did not follow the same downward trend as that observed in aging speakers without voice training. In addition, over the years, public figures are inevitably recorded with various equipment in different settings, which can affect acoustic measurements in different ways (e.g. Calder, Wheeler, 2022; Zhang, Jepson, Lohfink & Arvaniti, 2021). Such variation can sometimes be controlled for with statistical modeling, for instance by introducing random effects per recording (Lee, Nelder & Pawitan, 2006). Similarly, since some public figures such as Ed Miliband (Kirkham, Moore, 2016), Oprah Winfrey (Hay et al., 1999) or Fareed Zakaria (Sharma, 2018) have been found to alter their speech when speaking to different crowds, thus, audience and interlocutor effects should be controlled for, even if these are not the main concern of a given study. Provided that such basic precautions are taken, we suggest, in light of the research summarized in this paper, that speech in the media can be a valuable resource to advance our understanding of outstanding questions in phonetics (e.g. Kendall, Pharao, Stuart-Smith & Vaughn, 2023).

Acknowledgements

We would like to thank the organizers of AISV 2024 for their invitation and warm welcome to Torino, as well as the audience for thoughtful discussions. We acknowledge the financial support of the following organizations at various stages of research and writing: Fonds de recherche du Québec – Société et culture (https://doi.org/10.69777/270174), Deutsche Forschungsgemeinschaft (520195671).

References

Ash, S. (2002). Social class. In Chambers, J.K., Trudgill, P. & Schilling-Estes, N. (Eds.), *The Handbook of Language Variation and Change* (1st ed.). Oxford: Blackwell, 402-422.

Bauer, L. (1985). Tracing phonetic change in the received pronunciation of British English. In *Journal of Phonetics*, 13(1), 61-81. https://doi.org/10.1016/S0095-4470(19)30726-0

Bauernfeind, L. (2020). Exemplifying the language change of Jennifer Lopez: Is she still "Jenny from the block"? In *Lifespans and Styles*, 6(2), 11-21. https://doi.org/10.2218/ls.v6i2.2020.5217

Beaman, K.V., Buchstaller, I. (Eds.) (2021). *Language Variation and Language Change across the Lifespan: Theoretical and Empirical Perspectives from Panel Studies*. London: Routledge.

Berisha, V., Liss, J., Huston, T., Wisler, A., Jiao, Y. & Eig, J. (2017). Float like a butterfly sting like a bee: Changes in speech preceded Parkinsonism diagnosis for Muhammad Ali. In *Proceedings of Interspeech 2017*, Stockholm, Sweden, 20-24 August 2017, 1809-1813. https://doi.org/10.21437/Interspeech.2017-25

Bowie, D. (2015). Phonological variation in real time: Patterns of adult linguistic stability and change. In Gerstenberg, A., Voeste, A. (Eds.), *Language Development: The Lifespan Perspective*. Amsterdam: John Benjamins, 39-58.

Bowie, D., Yaeger-Dror, M. (2015). Phonological change in real time. In Honeybone, P., Salmons, J. (Eds.), *The Oxford Handbook of Historical Phonology*. Oxford: Oxford University Press, 603-618.

Braun, A., Friebis, S. (2009). Phonetic cues to speaker age: A longitudinal study. In Grewendorf, G., Rathert, M. (Eds.), *Formal Linguistics and Law*. Berlin: De Gruyter Mouton, 141-162.

Buchstaller, I., Wagner, S.E. (2018). Introduction: Using panel data in the sociolinguistic study of variation and change. In Wagner, S.E., Buchstaller, I. (Eds.), *Panel Studies of Variation and Change*. London: Routledge, 1-18.

Calder, J., Wheeler, R. (2022). Is Zoom viable for sociophonetic research? A comparison of in-person and online recordings for sibilant analysis. In *Linguistics Vanguard*, 20210014. https://doi.org/10.1515/lingvan-2021-0014

Chalier, M. (2021). *Les normes de prononciation du français. Une étude perceptive panfrancophone*. Berlin: De Gruyter.

Chambers, J.K. (2009). *Sociolinguistic Theory: Linguistic Variation and Its Social Significance* (3rd ed.). Chichester: Wiley-Blackwell.

Chambers, J.K., Trudgill, P. (1998). *Dialectology* (2nd ed.). Cambridge: Cambridge University Press.

Chung, H., Kong, E.J., Edwards, J., Weismer, G., Fourakis, M. & Hwang, Y. (2012). Cross-linguistic studies of children's and adults' vowel spaces. In *Journal of the Acoustical Society of America*, 131(1), 442-454. https://doi.org/10.1121/1.3651823

Clark, E.V. (2009). *First Language Acquisition* (2nd ed.). Cambridge: Cambridge University Press.

Clarke, N. (2000). *Alistair Cooke: The Biography*. New York: Arcade Publishing.

Cukor-Avila, P., Bailey, G. (2013). Real time and apparent time. In Chambers, J.K., Schilling, N. (Eds.), *The Handbook of Language Variation and Change* (2nd ed.). Chichester: Wiley-Blackwell, 237-262.

DE LEEUW, E. (2019). Native speech plasticity in the German-English late bilingual Stefanie Graf: A longitudinal study over four decades. In *Journal of Phonetics*, 73, 24-39. https://doi.org/10.1016/j.wocn.2018.12.002

D'ONOFRIO, A., STECKER, A. (2022). The social meaning of stylistic variability: Sociophonetic (in)variance in United States presidential candidates' campaign rallies. In *Language in Society*, 51(1), 1-28. https://doi.org/10.1017/S0047404520000718

DRAGER, K., KIRTLEY, M.J. (2016). Awareness, salience, and stereotypes in exemplar-based models of speech production and perception. In BABEL, A.M. (Ed.), *Awareness and Control in Sociolinguistic Research*. Cambridge: Cambridge University Press, 1-24.

DUFFY, J.R. (2020). *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management* (4th ed.). St. Louis: Mosby.

ECKERT, P. (1997). Age as a sociolinguistic variable. In COULMAS, F. (Ed.), *The Handbook of Sociolinguistics*. Oxford: Blackwell, 151-167.

EVANS, B.G., IVERSON, P. (2007). Plasticity in vowel perception and production: A study of accent change in young adults. In *Journal of the Acoustical Society of America*, 121(6), 3814-3826. https://doi.org/10.1121/1.2722209

FOULKES, P., DOCHERTY, G. (2006). The social life of phonetics and phonology. In *Journal of Phonetics*, 34(4), 409-438. https://doi.org/10.1016/j.wocn.2005.08.002

GERSTENBERG, A., VOESTE, A. (2015). Investigating the lifespan perspective. In GERSTENBERG, A., VOESTE, A. (Eds.), *Language Development: The Lifespan Perspective*. Amsterdam: John Benjamins, 1-8.

GIANNINI, A., PETTORINO, M. (2009). L'età della voce. In ROMITO, L., GALATÀ, V. & LIO, R. (Eds.), *Atti del IV Convegno Nazionale dell'Associazione Italiana di Scienze della Voce*, Arcavacata CS, Italy, 3-5 December 2007, 165-178.

GOFFMAN, E. (1981). *Forms of Talk*. Philadelphia: University of Pennsylvania Press.

HALL-LEW, L., COPPOCK, E. & STARR, R. (2010). Indexing political persuasion: Variation in the Iraq vowels. In *American Speech*, 85(1), 91-102. https://doi.org/10.1215/00031283-2010-004

HARRINGTON, J., KLEBER, F., REUBOLD, U., SCHIEL, F. & STEVENS, M. (2018). Linking cognitive and social aspects of sound change using agent-based modeling. In *Topics in Cognitive Science*, 10(4), 707-728. https://doi.org/10.1111/tops.12329

HARRINGTON, J., PALETHORPE, S. & WATSON, C.I. (2000a). Does the Queen speak the Queen's English? In *Nature*, 408, 927-928. https://doi.org/10.1038/35050160

HARRINGTON, J., PALETHORPE, S. & WATSON, C.I. (2000b). Monophthongal vowel changes in Received Pronunciation: An acoustic analysis of the Queen's Christmas broadcasts. In *Journal of the International Phonetic Association*, 30(1–2), 63-78. https://doi.org/10.1017/S0025100300006666

HARRINGTON, J., REUBOLD, U. (2021). Accent reversion in older adults: Evidence from the Queen's Christmas broadcasts. In BEAMAN, K.V., BUCHSTALLER, I. (Eds.), *Language Variation and Change Across the Lifespan: Theoretical and Empirical Perspectives from Panel Studies*. London: Routledge, 119-137.

HAY, J., DRAGER, K. & WARREN, P. (2010). Short-term exposure to one dialect affects processing of another. In *Language and Speech*, 53(4), 447-471. https://doi.org/10.1177/0023830910372489

Hay, J., Jannedy, S. & Mendoza-Denton, N. (1999). Oprah and /ay/: Lexical frequency, referee design and style. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, San Francisco, USA, 1-7 August 1999, 1389-1392.

Hazan, V. (2017). Speech communication across the life span. In *Acoustics Today*, 13(1), 36-43.

Holliday, N. (2017). "My Presiden(t) and Firs(t) Lady were black": Style, context, and coronal stop deletion in the speech of Barack and Michelle Obama. In *American Speech*, 92(4), 459-486. https://doi.org/10.1215/00031283-6903954

Hughes, A., Trudgill, P. & Watt, D. (2012). *English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles* (5th ed.). London: Routledge.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In Johnson, K., Mullennix, J.W. (Eds.), *Talker Variability in Speech Processing*. San Diego: Academic Press, 145-165.

Kendall, T., Pharao, N., Stuart-Smith, J. & Vaughn, C. (2023). Advancements of phonetics in the 21st century: Theoretical issues in sociophonetics. In *Journal of Phonetics*, 98, 101226. https://doi.org/10.1016/j.wocn.2023.101226

Kirkham, S., Moore, E. (2013). Adolescence. In Chambers, J.K., Schilling, N. (Eds.), *The Handbook of Language Variation and Change* (2nd ed.). Oxford: Wiley-Blackwell, 277-296.

Kirkham, S., Moore, E. (2016). Constructing social meaning in political discourse: Phonetic variation and verb processes in Ed Miliband's speeches. In *Language in Society*, 45(1), 87-111. https://doi.org/10.1017/S0047404515000755

Kwon, S. (2018). Phonetic and phonological changes of Noam Chomsky: A case study of dialect shift. In *American Speech*, 93(2), 270-297. https://doi.org/10.1215/00031283-6926146

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

Labov, W. (2001). *Principles of Linguistic Change: Social Factors* (Vol. 2). Chichester: Wiley-Blackwell.

Lee, S., Potamianos, A. & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. In *Journal of the Acoustical Society of America*, 105(3), 1455-1468. https://doi.org/10.1121/1.426686

Lee, Y., Nelder, J.A. & Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Boca Raton: Chapman and Hall/CRC.

Lenneberg, E.H. (1967). *Biological Foundations of Language*. New York: John Wiley and Sons.

Linville, S.E. (2001). *Vocal Aging*. San Diego: Singular Publishing Group.

MacKenzie, L. (2017). Frequency effects over the lifespan: A case study of Attenborough's r's. In *Linguistics Vanguard*, 3(1), 20170005. https://doi.org/10.1515/lingvan-2017-0005

Markova, D., Richer, L., Pangelinan, M., Schwartz, D.H., Leonard, G., Perron, M., Pike, G.B., Veillette, S., Chakravarty, M.M., Pausova, Z. & Paus, T. (2016). Age- and sex-related variations in vocal-tract morphology and voice acoustics during adolescence. In *Hormones and Behavior*, 81, 84-96. https://doi.org/10.1016/j.yhbeh.2016.03.001

Munro, M.J., Derwing, T.M. & Flege, J.E. (1999). Canadians in Alabama: A perceptual study of dialect acquisition in adults. In *Journal of Phonetics*, 27(4), 385-403. https://doi.org/10.1006/jpho.1999.0101

Nycz, J. (2013). Changing words or changing rules? Second dialect acquisition and phonological representation. In *Journal of Pragmatics*, 52, 49-62. https://doi.org/10.1016/j.pragma.2012.12.014

Nycz, J. (2015). Second dialect acquisition: A sociophonetic perspective. In *Language and Linguistics Compass*, 9(11), 469-482. https://doi.org/10.1111/lnc3.12163

Organisation internationale de la francophonie. (2025). *The Francophonie in brief*. https://www.francophonie.org Accessed 30 May 2025.

Perry, T.L., Ohde, R.N. & Ashmead, D.H. (2001). The acoustic bases for gender identification from children's voices. In *Journal of the Acoustical Society of America*, 109(6), 2988-2998. https://doi.org/10.1121/1.1370525

Pichler, H., Wagner, S.E. & Hesson, A. (2018). Old-age language variation and change: Confronting variationist ageism. In *Language & Linguistics Compass*, 12(6), 1-21. https://doi.org/10.1111/lnc3.12281

Pierrehumbert, J.B. (2016). Phonological representation: Beyond abstract versus episodic. In *Annual Review of Linguistics*, 2(1), 33-52. https://doi.org/10.1146/annurev-linguistics-030514-125050

Poplack, S., St-Amand, A. (2007). A real-time window on 19th-century vernacular French: The "Récits du français québécois d'autrefois". In *Language in Society*, 36(5), 707-734. https://doi.org/10.1017/S0047404507070662

Quené, H. (2013). Longitudinal trends in speech tempo: The case of Queen Beatrix. In *Journal of the Acoustical Society of America*, 133(6), EL452-EL457. https://doi.org/10.1121/1.4802892

Reinke, K., Ostiguy, L. (2016). *Le français québécois d'aujourd'hui*. Berlin: Walter de Gruyter.

Reubold, U., Harrington, J. (2015). Disassociating the effects of age from phonetic change: A longitudinal study of formant frequencies. In Gerstenberg, A., Voeste, A. (Eds.), *Language Development: The Lifespan Perspective*. Amsterdam: John Benjamins, 9-38.

Reubold, U., Harrington, J. (2018). The influence of age on estimating sound change acoustically from longitudinal data. In Wagner, S.E., Buchstaller, I. (Eds.), *Panel Studies of Variation and Change*. London: Routledge, 129-151.

Reubold, U., Harrington, J. & Kleber, F. (2010). Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers. In *Speech Communication*, 52(7), 638-651. https://doi.org/10.1016/j.specom.2010.02.012

Riverin-Coutlée, J., Harrington, J. (2022). Phonetic change over the career: A case study. In *Linguistics Vanguard*, 8(1), 41–52. https://doi.org/10.1515/lingvan-2021-0122

Riverin-Coutlée, J., Harrington, J. (in press). Stability, change and reversal in public speech: A longitudinal case study. In Buchstaller, I., Beaman, K.V. (Eds.), *Connecting the individual and the community: Contributions from sociolinguistic panel research*. London: Routledge.

Riverin-Coutlée, J., Roy, J.-P. (2022). A descriptive account of the Quebec French diphthong fête. In *Journal of the International Phonetic Association*, 52(2), 228-245. https://doi.org/10.1017/S0025100320000195

Roach, P. (2004). British English: Received Pronunciation. In *Journal of the International Phonetic Association*, 34(2), 239-245. https://doi.org/10.1017/S0025100304001768

Sankoff, G. (2018). Before there were corpora: The evolution of the Montreal French project as a longitudinal study. In Wagner, S.E., Buchstaller, I. (Eds.), *Panel Studies of Variation and Change*. London: Routledge, 21-51.

Sankoff, G., Blondeau, H. (2007). Language change across the lifespan: /R/ in Montreal French. In *Language*, 83(3), 560-588. https://doi.org/10.1353/lan.2007.0106

Sankoff, G., Cedergren, H. (1971). Some results of a sociolinguistic study of Montreal French. In Darnell, R. (Ed.), *Linguistic Diversity in Canadian Society*. Edmonton: Linguistic Research, 61-87.

Shapp, A., LaFave, N. & Singler, J.V. (2014). Ginsburg v. Ginsburg: A longitudinal study of regional features in a Supreme Court Justice's speech. In *University of Pennsylvania Working Papers in Linguistics*, 20(2), 149-158.

Sharma, D. (2018). Style dominance: Attention, audience, and the "real me". In *Language in Society*, 47(1), 1-31. https://doi.org/10.1017/S0047404517000835

Siegel, J. (2010). *Second Dialect Acquisition*. Cambridge: Cambridge University Press.

Sigouin, C., Arnaud, V. (2014). Les voyelles fermées tendues, relâchées et allongées du français québécois : la contribution d'indices statiques/dynamiques et absolus/normalisés à la détermination de leur identité acoustique. In *Actes des XXX^es Journées d'études sur la parole*, Le Mans, France, 23-27 June 2014, 567-575.

Thibault, P., Vincent, D. (1990). *Un corpus de français parlé. Montréal 84 : Historique, méthodes et perspectives de recherche*. Québec: Université Laval, Département de langues et de linguistique.

Vincent, D., Laforest, M. & Martel, G. (1995). Le corpus de Montréal 1995 : Adaptation de la méthode d'enquête sociolinguistique pour l'analyse conversationnelle. In *Dialangue*, 6, 29-45.

Wells, J.C. (1982). *Accents of English*. Cambridge: Cambridge University Press.

Wirtz, M.A., Pickl, S. (2025). Major life events as drivers of perceived linguistic change across adulthood. In *Language Variation and Change*, 1-24. https://doi.org/10.1017/S095439452400019X

Yang, C., Pittayaporn, P., Kirby, J. & Jitwiriyanont, S. (2021). Change and stability in the tonal contours of King Rama IX of Thailand, 1959-1997. In *Proceedings of the 1st International Conference on Tone and Intonation (TAI)*, University of Southern Denmark, Denmark, 6-9 December 2021, 66-70.

Young, A.F., Powers, J.R. & Bell, S.L. (2006). Attrition in longitudinal studies: Who do you lose? In *Australian and New Zealand Journal of Public Health*, 30(4), 353-361. https://doi.org/10.1111/j.1467-842X.2006.tb00849.x

Zhang, C., Jepson, K., Lohfink, G. & Arvaniti, A. (2021). Comparing acoustic analyses of speech data collected remotely. In *Journal of the Acoustical Society of America*, 149(6), 3910-3916. https://doi.org/10.1121/10.0005132

DOI: 10.17469/O2112AISV000003

ALICE CROCHIQUIA, ANDERS ERIKSSON, SANDRA MADUREIRA

Dubbing in animated films: challenges and the impact of voice design

Listeners rely on speech vocal cues to judge speakers' age, size, personality, and other paralinguistic and extralinguistic features. In this paper, we focus on two perception experiments investigating how characters' voices in animated films provide indexical information about their individuality to listeners. In the first experiment, we examined how listeners perceived physical, psychological, and social features from voice characteristics. The stimuli consisted of speech utterances by five animated film characters. The second experiment aimed to investigate the potential influence of lexical content on the evaluation of speaker characteristics. This experiment was identical to the first experiment, except the stimuli were transcriptions of the utterances used in the first experiment. Our hypothesis was that the results from the listening test and the reading test would not be congruent, but they turned out to be the exact opposite. This finding is intriguing and claims for further research.

Keywords: dubbing, animated films, vocal stereotypes, voice quality, voice analysis.

1. *Introduction*

As most of us have experienced and as backed by extensive accounts on the subject (Kreiman & Sidtis, 2011), listeners frequently rely on vocal cues to judge a speaker's characteristics, such as physical appearance and demeanour. At the same time, we also create expectations for how someone will sound based on these features.

Over the last century, research on those links has shed light on the complex interplay between physiological factors and cultural expectations in voice perception. While some expectations, such as women generally having higher-pitched voices, appear universal due to physiological differences, other correlations, such as physical attractiveness and voice quality, seem to vary between languages and cultures (Weirich, 2008; Xu & Lee, 2018).

A common finding in many studies dealing with voice identification has been the high degree of agreement among listeners regarding particular impressions of speakers from their voices, even if they are often inaccurate. These findings of shared expectations and assumptions point towards the phenomenon of vocal stereotypes and their consistent influence on impressions based on voice (Aronovitch, 1976; Imhof, 2010; Mahrholz, Belin & McAleer, 2018).

These insights have significant implications for fields such as voice design and animation, where understanding the nuances of voice perception can enhance the creation of more authentic and engaging characters.

For several reasons, animated films and TV shows are a particularly interesting kind of media for investigating those vocal stereotypes and their nature. Firstly, they allow creators to match physical, social, and psychological features with the expected correlating vocal characteristics, presumably also shared with the intended audience, which works as an effective and quick way to help establish a character to the viewers. These productions also welcome exaggerated vocal performances, which would be out of place in most live-action productions, as a way to enhance the intended impressions of the characters. Lastly, such productions are usually dubbed for distribution in international markets, thus allowing for comparing how the same character sounds in many different languages.

Our research takes a unique approach by exploring how vocal cues are perceived in the context of animated films (Crochiquia, 2020; Crochiquia, Andersson, Fontes & Madureira, 2021; Crochiquia, Andersson, Barbosa & Madureira, 2022; Crochiquia, Andersson, Madureira & Barbosa, 2023), a field that has not been extensively studied. In this paper, we detail the steps in developing the methodological procedures to investigate the role of voice in the construction of animated characters and their features, the use of vocal stereotypes in these types of performances, and how they may match or differ in multiple languages.

Zootopia, the film animated chosen for our investigation, heavily relies on preconceived notions of how its characters are meant to behave, both to confirm and to subvert expectations. The investigation of the vocal performances in it allows us to explore how vocal features are used to reflect biological and social meanings as outlined by hypotheses such as the Frequency Code (Ohala, 1984) and the Sirenic Code (Gussenhoven, 2016) in the different languages in which the film is dubbed. The Frequency Code associates low f_0 values with bigness, power, force, hostility, and aggressiveness. In contrast, high f_0 values are associated with smallness, submission, and fragility. The Sirenic Code (Gussenhoven, 2016) is related to the voice qualities characterised by air escaping between the vocal folds (breathy and whispery voices). Paralinguistically, breathy and whispery voices can be associated with low excitment and seduction.

Our work investigates those correlations by combining perceptual experiments to assess listeners' impressions of the characters from speech samples and the analysis of their acoustic parameters, perceptual assessment of voice quality, as well as text-only perceptual experiments with transcriptions of the audio samples to investigate the influence of lexical content in those judgments.

2. *Dubbing*

Dubbing, as a type of language transfer, most commonly refers to a kind of revoicing where the original voice track of an audiovisual product is entirely removed and

replaced by a new voice track in the target language while keeping the original soundtrack and sound effects in place. This is also called "synchronisation", as this type of audiovisual translation (AVT) often requires the spoken text in the target language to follow the lip movements on screen as closely as possible (Díaz-Cintas, 1999). Dubbing is one of the two most widespread forms of AVT, along with interlingual subtitling.

Generally, national markets are broadly categorised according to which AVT format is preferred for the distribution of foreign productions, usually dubbing or subtitling. The reasons for such preferences are many; among them, we can find population size, rates of literacy, cultural and ideological factors, the potential audience and profit of the production, and foreign language proficiency (Danan, 1991; Díaz-Cintas, 2019). However, the boundaries of this division have become increasingly blurred in the age of digitalisation (Chaume, 2018).

Traditionally, Brazil is one of the markets categorised as a dubbing country, with most foreign media productions arriving on TV, movie theatres, home video and streaming services dubbed in Brazilian Portuguese, including productions from other lusophone countries. On the other hand, Nordic countries are firmly on the subtitling side of that division, where the use of dubbing is almost exclusively limited to productions aimed at younger audiences and computer-animated films (Tveit, 2009; Pedersen, 2010).

Dubbing is a collaborative process involving translators, dialogue writers, dubbing directors, voice actors and sound engineers. Dubbing directors are perhaps the most crucial link in this chain, as they are responsible not only for the initial selection of voice actors and guiding them in their performances, providing them with information about the narrative, (which is often not available to actors beyond the scenes they perform in) but they can also make changes to dialogues to ensure a better fit (Bosseaux, 2018). In choosing a dubber for a character in a film, directors have to consider not only how they match the original voice and the character on the screen but also their ability to work within the highly constrained context of dubbing and to deliver a vocal performance that fits well with the onscreen gestures (Brisola, 2024).

Unsurprisingly, given the greater number of steps and professionals involved in the process of dubbing a film, this type of AVT is both more time-consuming and more expensive than subtitling. Keeping in mind that the quality of a dub can affect how the audiences engage with the production and impact its reception (Bosseaux, 2019), the dubbing of an audiovisual production is an investment that demands careful consideration from studios and distribution companies, and often the final choices for voice actors falls to them (Brisola, 2024).

The choice of voice actors and directions given to them during their performance is often made so with the intent to support the character's features and attitudes displayed within the narrative (Chaume, 2014) as to "enhance the audience's recognition of the character's image" (Liu, Zhang & Liang, 2022: 02) and preserve the so-called 'character synchrony' (Whitman-Linsen, 1992).

However, 'mismatched' voices can also be used in original voice acting and dubbing, usually for comedic purposes, like a physically attractive character having an unpleasant and shrill voice.

As such, dubbing directors and voice actors often rely on vocal stereotypes to portray these characters (Bosseaux, 2015), both to confirm and to subvert expectations (Liu et al., 2022) Such stereotypes are often based on particular voice qualities and prosodic elements, like a grave voice belonging to a large-sized character, but they can also rely on impressions associated with dialects and accents, like the use of the Bergen dialect for tough, brash characters in Norwegian dubbed versions of animated films (Nybakk, 2014).

Despite the complexity of its process, its widespread use and its commercial and cultural impact (especially in dubbing strongholds like Brazil, Germany, Italy and Spain), dubbing has lacked attention and recognition from both academic and professional perspectives for much of its history, especially in comparison with other types of AVT (Sánchez-Mompeán, 2020; Díaz-Cintas, 2015). This landscape has suffered an important, if still slow, shift in the past decade or so.

Recent research works on dubbing, especially within the so-called dubbing countries, have explored many facets of this mode of AVT (Sánchez-Mompéan, 2016; Fasoli, Mazzurega & Sulpizio, 2017), demonstrating that the field is fertile ground for exploration and showing its potential for intersection with many disciplines. Considering the growing interest in dubbing and how it can readily welcome different perspectives, it would be expected that non-lexical elements in speech, like prosodic features and voice quality, would emerge as one of the topics of concern, given how these features carry linguistic, paralinguistic and extralinguistic information and can affect the expression and comprehension of a speaker's identity and attitudes (Kreiman & Sidtis, 2011). However, the subject is still 'virtually unexplored' in AVT studies, with the modest number of academic works on the subject mostly coming from research in phonetics and speech sciences (Sánchez-Mompeán, 2020) and often focused on single-parameter analysis and correlations with character traits (Liu et al., 2022) and within a single language.

3. *Our experiments with dubbing data*

A listening and a reading experiment are reported in this paper. These two experiments are part of a project that focuses on the analysis and comparison of the vocal performances of the four main characters in the 2016 computer-animated film Zootopia (also titled Zootropolis in European markets), from Walt Disney Studios, in its original version in English and its dubbing in Brazilian Portuguese and Swedish. Zootopia tells the story of a metropolis of the same name, inhabited by anthropomorphic mammals, and it serves as an allegory about human prejudices and biases. This paper details and discusses the methodological procedures applied to the investigation of the Brazilian Portuguese dubbing of the film.

The film first presents the city as a utopia (as referenced by its name), as we see it through the eyes of Judy Hopps, an idealistic bunny who becomes the city's first prey police officer, a position exclusively occupied by predators until then. However, that utopian view quickly washes away as Judy (and the audience) uncovers the complex dynamics of Zootopia, where prey, despite making up the large majority of the city's population, are often underestimated. At the same time, predators, due to their nature, are often regarded with suspicion by prey.

The film focuses on anthropomorphic mammals whose characterisations and roles in the narrative are directly influenced by the behaviour displayed by those animals in nature or associated with them in popular culture. Among the large cast of characters in the film, we find a lion as the mayor of Zootopia and a sloth as a worker at an overly bureaucratic and slow-moving government agency.

We have chosen these characters for our corpus for two reasons: as the main characters, more material from their speech productions is available for investigation, especially samples that do not feature music or sound effects in the background, as the isolated track with the dialogue lines of the characters was not available, and removing the music and effects would possibly result in some acoustic information from the characters' voices being removed as well; the second reason for our choice of characters was the fact that their physical, psychological and social features were discussed by the film's writers, directors and voice actors in interviews and promotional material for the film, which gave us insight into the choices that guided the actors and dubbers in their performances and the intended impressions these characters are supposed to have on the audience.

Alongside the lead character of Judy Hopps (JH), we have Assistant Mayor Bellwether, who is first presented as a meek but friendly little sheep (B1) who then reveals herself as bitter and conniving (B2). The film's deuteragonist is a red fox named Nick Wilde (NW), who is an easygoing con artist. Chief Bogo (CB), Zootopia's police chief, is a gruff African buffalo. Figure 1 shows the four characters.

As Assistant Mayor Bellwether is revealed as the film's villain, we consider the samples from her productions after that to belong to a different character from the one in the first portion of the film. As such, five different characters are part of our experiments and analyses.

Figure 1 - *The main characters of Zootopia: Judy Hopps & Nick Wilde (top row), Chief Bogo & Assistant Mayor Bellwether (bottom row)*

3.1 Material, data collection, and stimulus preparation procedures of the experiments

As the first step for the experiments and analyses, the video file from the Blu-ray disc of the Brazilian version of Zootopia was extracted using *Format Factory* (FORMATZ, version 4.3.0.0, 2018) to a .mp4 file. The film's English and Brazilian Portuguese tracks were then extracted to PCM files using Adobe Premiere. The same procedures were later used to extract the Swedish audio from the Nordic Blu-ray disc.

The files were then imported into *ELAN* (Max Planck Institute for Psycholinguistics, version 5.4, 2018), where the entire film was transcribed into English and Brazilian Portuguese. Figure 2 shows the transcription process of the dialogue lines in Brazilian Portuguese in *ELAN*.

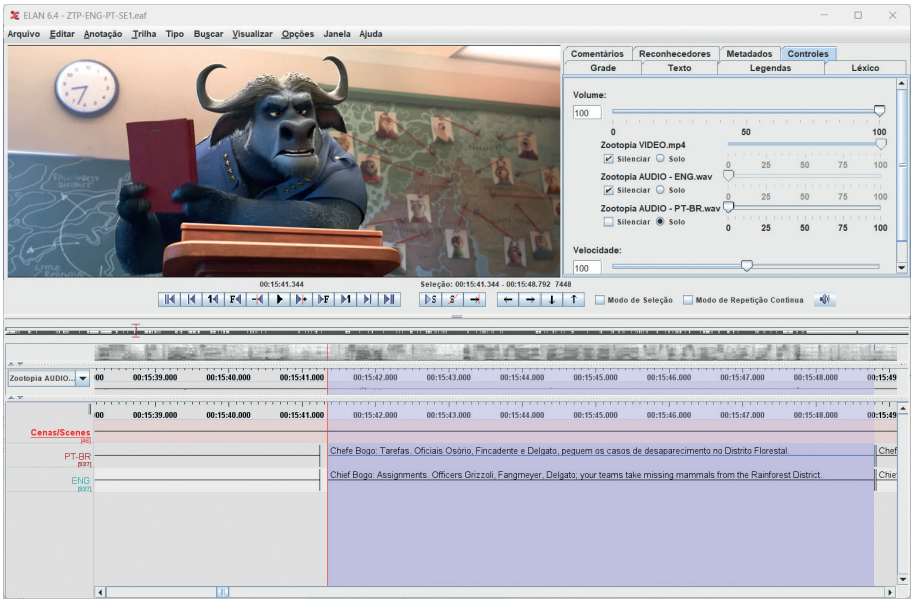
The audio files and annotation tiers were then exported to *PRAAT* (Boersma & Weenink, 2019) for further editing and the selection of samples. Later, the transcription of the Swedish audio was done on *PRAAT*.

For the experiments, five audio samples (and later the five corresponding text samples) were produced, one for each character, around 27 seconds in length. Each sample featured three dialogue lines from different scenes in the film.

No masking techniques were used to isolate voice from linguistic information in the samples as we believed the audio would feel odd to the listeners and affect their answers. In section 3.4, we will discuss our solution to investigate the influence of semantic and syntactic content in the audio samples.

The audio samples were then uploaded to the audio streaming platform SoundCloud, as the tool chosen for the experiments, SurveyMonkey.com, does not allow audio to be directly uploaded to the platform.

Figure 2 - *Transcription of dialogue lines in ELAN*



3.2 Listening Experiment

The Listening Experiment comprised a perceptual test, a phonetic description of voice quality settings, and acoustic data analysis.

3.2.1 The perceptual test

The test was conducted entirely online on SurveyMonkey.com. Listeners were initially recruited from the Graduate Program of Applied Linguistics and Language Studies at PUC-SP and were emailed a link to the test. Participants were also told they were welcome to share the link with family, friends, and whoever else they believed would be interested in participating in the test.

Participation in the test was voluntary. Participants were informed of their right to withdraw at any given time and that any data collected in the test would be anonymous past the collection stage. They received no compensation for their participation. Seventy-seven Brazilian Portuguese native speakers, 46 women and 31 men, aged between 20 and 50, participated in the test.

The test was divided into eight pages. The first page contained a brief description of the test, the study, its aims, and a questionnaire to collect background information on the respondents, including whether they had any hearing impairments or had seen the film before.

The instructions for the test and an example of the slide feature they would use for the evaluations were presented on the second page. Respondents were instructed to base their evaluations solely on the voices in the samples and their impressions of them. They were also told to use headphones to complete the test, and to indicate if that had not been the case.

The following five pages each presented a speech sample from one of the characters, using the embedded player from SoundCloud on the page, and 14 sliding scales for rating.

The respondents were asked to evaluate each of the five samples according to the bipolar scales regarding the speakers' size, age, temper, attitudes, character, and social and vocal features. They had to do that by using the sliding tool to indicate the point of the scale they felt best fit the character, which was translated into a score from 0 to 100; respondents were not made aware of the score during the test.

The descriptors chosen for the test were based on features displayed by the characters in the film and on descriptions given by the writers, directors, and voice actors in interviews, discussions, and promotional material for the film.

The test did not include pictures or videos of the characters. Table 1 lists the scales and pairs of descriptors in Brazilian Portuguese used in the test, and their translations to English, and Figure 3 shows a few of the sliding scales used in the Experiment, translated to English.

Respondents could only move to the next page once they had made a choice for all 14 scales; however, they were able to return to previous pages if they wished to review or change their answers. The final page featured a message warning the respondents that their answers were final and could not be changed once they had finished the test.

Table 1 - *Scales and Descriptors in Listening Experiment*

Scale	*Descriptors*
Size	*Pequeno - Grande* (Small - Big)
Age	*Jovem - Velho* (Young - Old)
Temper	*Agressivo - Dócil* (Aggressive - Docile)
Attitudes	*Dominador - Submisso* (Dominant - Submissive)
	Preguiçoso - Ativo (Lazy - Energetic)
	Durão - Gentil (Tough - Gentle)
	Charmoso - Desinteressante (Charming - Uninteresting)
	Bondoso - Maldoso (Kind - Mean)
	Cauteloso - Audacioso (Cautious - Bold)
	Alegre - Triste (Happy - Sad)
Character	*Honesto - Desonesto* (Honest - Dishonest)
Social features	*Sofisticado - Simples* (Sophisticated - Simple)
Voice	*Animada - Monótona* (Lively - Monotonous) *Agradável - Desagradável* (Pleasant - Unpleasant)

Figure 3 - *Sliding scales in the Listening Experiment*

3.2.2 Description of the voice quality settings

For the perceptual evaluation of voice quality, we have used the Voice Profile Analysis, or VPA, developed by Laver and Mackenzie Beck (2007) to outline the characters' vocal profiles.

The VPA is a comprehensive tool for voice assessment as it considers phonatory, supralaryngeal and prosodic features of a speaker's voice. The analytic unit of the protocol is the setting. The vocal features are evaluated based on their deviation from the neutral setting. This condition indicates no contraction or expansion, no shortening or lengthening of the vocal tract articulators, no extreme variation in muscular tension, no nasality beyond the necessary for linguistic purposes, periodical vibration of the vocal folds under moderate tension and compression and no audible friction (Laver, 1980; San Segundo & Mompean, 2017).

The protocol is divided into six parts, with the first three parts comprising features related to voice quality and the last three relating to vocal dynamics. The features are organised according to the location of indidence of the settings. Assessments are made primarily by listening to audio samples, which can be combined with corresponding video samples.

The assessment with the protocol is done in two stages. The first step is to identify the presence of neutral and non-neutral settings in a speech sample. In the second step, the marked non-neutral settings are rated in a scalar degree, from lesser (1) to more significant (6). Degrees 1 to 3 are moderate, and 4 to 6 are considered extreme; however, some settings are marked only for their absence or presence. Figure 4 shows the voice quality settings, and Figure 5 shows the prosodic elements.

Two phoneticians with extensive experience with VPA described the voice quality characteristics of the characters chosen for our analyses.

Figure 4 - *Vocal Tract, Muscular Tension and Phonation settings in VPA (Laver & Mackenzie Beck, 2007)*

	FIRST PASS		SECOND PASS						
	Neutral	Non-neutral	SETTING	moderate			extreme		
				1	2	3	4	5	6
A. VOCAL TRACT FEATURES									
1. Labial			Lip rounding						
			Lip spreading						
			Labiodentalization						
			Extensive range						
			Minimised range						
2. Mandibular			Close jaw						
			Open jaw						
			Protruded jaw						
			Extensive range						
			Minimised range						
3. Lingual tip/blade			Advanced tip/blade						
			Retracted tip/blade						
4. Lingual body			Fronted tongue body						
			Backed tongue body						
			Raised tongue body						
			Lowered tongue body						
			Extensive range						
			Minimised range						
5. Pharyngeal			Pharyngeal constriction						
			Pharyngeal expansion						
6. Velopharyngeal			Audible nasal escape						
			Nasal						
			Denasal						
7. Larynx height			Raised larynx						
			Lowered larynx						
B. OVERALL MUSCULAR TENSION									
8. Vocal tract tension			Tense vocal tract						
			Lax vocal tract						
9. Laryngeal tension			Tense larynx						
			Lax larynx						

C. PHONATION FEATURES

	SETTING	Present		Scalar degree					
		Neutral	Non-neutral	Moderate			Extreme		
				1	2	3	4	5	6
10.Voicing type	Voice								
	Falsetto								
	Creak								
	Creaky								
11. Laryngeal frication	Whisper								
	Whispery								
12. Laryngeal irregularity	Harsh								
	Tremor								

Figure 5 - *Prosodic, Temporal Organization and Other Features in VPA (Laver & Mackenzie Beck, 2007)*

			moderate			extreme		
			1	2	3	4	5	6
D. PROSODIC FEATURES								
13. Pitch	Mean	High						
		Low						
	Range	Extensive range						
		Minimised range						
	Variability	High						
		Low						
14. Loudness	Mean	High						
		Low						
	Range	Extensive range						
		Minimised range						
	Variability	High						
		Low						
E. TEMPORAL ORGANIZATION								
15. Continuity		Interrupted						
16. Rate		Fast						
		Slow						
F. OTHER FEATURES								
17. Respiratory support		Adequate						
		Inadequate						
18. Diplophonia		Absent						
		Present						

3.2.3 Acoustic analysis

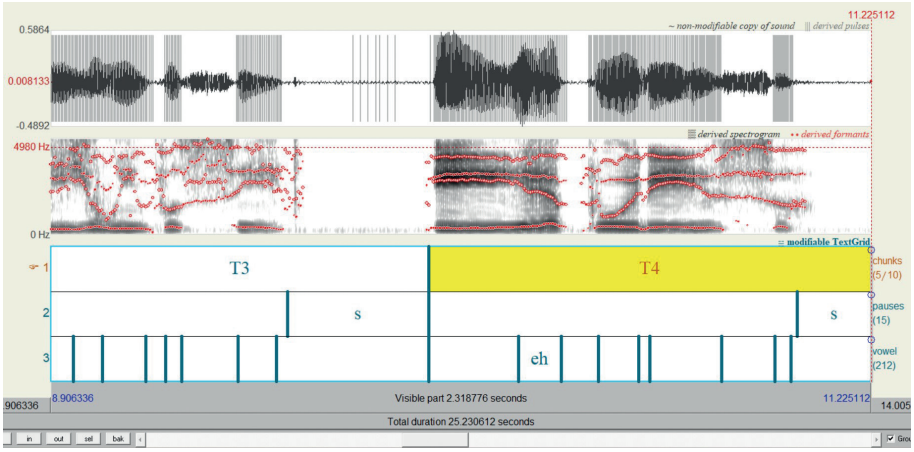
A modified version of the Prosody Descriptor Extractor script for PRAAT (Barbosa, 2021) was used for the acoustic analysis of the speech samples. The original script computes 30 prosodic parameters related to statistical descriptors f_0, intensity, long-term spectrum, voice quality, duration of vowels and silence. The modified version used in this study comprised 11 parameters: the difference between harmonics H1 and H2 in the vowel intervals (H1-H2 dB); Harmonic-to-Noise ratio in dB (HNR dB); Spectral emphasis in dB (emph dB); f_0 minimum in semitones re 1 Hz and in Hertz (f_0 min s.t.); f_0 maximum in semitones re 1 Hz and in Hertz (f_0 max s.t.); $f0$ median in semitones re 1 Hz and in Hertz (f_0 med s.t.); LTAS slope between bands 0-1000 Hz and 1000/4000 Hz (slLTASmed); f_0 baseline in semitones (f base s.t.); f_0 standard-deviation in semitones/Hertz (fSD s.t.); Local shimmer in per cent (shimmer %); Local jitter in percentual values (jitter %).

Many acoustic parameters selected here are among the most used for voice assessment and often appear in the literature investigating vocal expressivity. In addition, a few of these parameters have also been correlated in previous research to some of the traits we selected for our perceptual experiments; for example, impressions of size and dominance are often attributed to parameters related to f_0, as described in Ohala's Frequency Code (1984),

We chose to use the scale of semitones for most of the parameters related to f_0, as we expected the variation in frequency for female and male speakers to be roughly the same and to reduce skewness as they are more representative of a speaker's perceptual variability.

To run the script, a TextGrid file was created for each of the audio files, segmenting and annotating chunks of speech (Tier 1), their following silent pauses (Tier 2), and the stressed oral vowels (Tier 3), using an ASCII code in direct correspondence with the IPA symbols for Brazilian Portuguese. Figure 6 shows this process in PRAAT.

Figure 6 - *Annotation for extraction of acoustic parameters in PRAAT*



3.3 Reading Experiment

The main concern of our research on vocal stereotypes is to investigate the influence of vocal features on the impressions caused by the characters on the listeners. However, as no masking techniques were used in our experiments, we cannot ignore the semantic, syntactic, and vocal interacting factors potentially influencing listeners' judgments of the speech stimuli. As such, it became necessary to investigate how lexical content affects these judgments.

The design for the Reading Experiment was nearly identical to the Listening Experiment, with questions and instructions related to the aural aspect of the first test being removed. The experiment was also conducted online on SurveyMonkey.com.

The subjects were recruited in the same way as for the Listening Experiment. Like with the first experiment, participation was voluntary. Participants were informed of their right to withdraw at any point if they wished, and that any data collected in the experiment would be anonymised. They received no compensation for their participation. The respondents in the experiment were 73 Brazilian Portuguese-native speakers, 43 women and 30 men, aged between 19 and 58.

For this experiment, respondents were instructed to read the transcriptions of the five speech samples presented in the previous test and rate the characters in the 14 bipolar scales presented. The majority of the scales remained the same as in the first experiment, and the only changes were made to the last two scales, which were renamed from *Voz* (Voice) *to Maneira de expressão* (Manner of expression). The descriptors all remained the same.

Figure 7 shows a page of the Reading Experiment with the text sample and the same sliding scales shown in Figure 2.

Figure 7 - *Sliding scales in the Reading Experiment*

4. *Results*

4.1 Results of the interrater reliability tests in both the Listening and Reading Experiments

We used Cronbach's Alpha with both perceptual experiments to measure the reliability of the interrater agreement. With this method, values can range from 0 to 1, indicating complete disagreement and perfect agreement, respectively. Values above 0.7 are considered good, and scores above 0.9 are considered excellent. Interrater agreement scores were excellent for all characters except one in both the Listening Experiment (Table 2) and the Reading Experiment (Table 3), regardless of whether or not the respondents had seen the film. Nick Wilde's scores were slightly below excellent in some cases but still good.

Table 2 - *Interrater reliability scores for all judges in the Listening Experiment*

Voices	*Male*	*Female*	*All*
Bellwether 1	0.984	0.988	0.993
Bellwether 2	0.957	0.976	0.985
Chief Bogo	0.981	0.986	0.992
Judy Hopps	0.939	0.977	0.982
Nick Wilde	0.815	0.848	0.908

Table 3 - *Interrater reliability scores for all judges in the Reading Experiment*

Voices			*Male*	*Female*
	z-norm	*M/F*	*S/nS*	*S/nS*
Bellwether 1	0.987	0.990	0.980	0.988
Bellwether 2	0.934	0.984	0.973	0.986
Chief Bogo	0.951	0.965	0.916	0.968
Judy Hopps	0.955	0.967	0.873	0.911
Nick Wilde	0.940	0.931	0.890	0.927
All	0.961	0.976	0.953	0.965

4.2 Results of the perception tests of the Listening Experiment

Table 4 presents the mean values for the characters' perception scores. Higher values indicate an overall score closer to the descriptor listed, while lower values indicate a rating closer to the descriptor on the opposite end of the scale (See Table 1).

We have highlighted in bold the features with the strongest scores of each character, higher than 75 and lower than 25, as we consider them significant for the description of their profiles. For Age and Size, however, all scores were considered, as the values around the mid-points of those scales do not represent a neutral reading

of the character on those scales but rather an impression of the speaker as mid-sized or middle-aged.

Chief Bogo and the nice version of Bellwether usually appear on opposite ends of the scales, which is expected considering their physical appearance and demeanour in the film. The scores for the villainous Bellwether are similar to those of her façade in physical appearance but closer to Chief Bogo's in temperament and attitude.

Judy has one strong score, for Energetic, while Nick falls somewhere around the middle for most of the scales. Charming, Dominant and a Pleasant voice were the only features for which Nick received moderately high scores.

Table 4 - *Mean values for perception scores for the characters' features*

	B1	*B2*	*CB*	*JH*	*NW*
Big	30	**24**	**87**	34	53
Old	**23**	**22**	65	28	60
Aggressive	**13**	69	**86**	49	43
Dominant	44	**88**	**91**	69	66
Energetic	**88**	**84**	**78**	**83**	44
Tough	**13**	69	**91**	58	45
Charming	**78**	63	51	68	66
Kind	**82**	37	72	62	52
Bold	50	74	72	68	46
Happy	**89**	65	45	70	51
Honest	73	39	41	73	46
Sophisticated	45	64	50	59	55
Lively	**93**	74	48	78	42
Pleasant	74	56	51	71	67

4.3 Acoustic analysis results of the Listening Experiment

Table 5 shows the mean values of the acoustic parameters that presented the highest correlation with the perceptual descriptors. To help the reader get a feel for the f_0 levels for the different characters at a glance, we have added a translation of the original results extracted in semitones to Hz.

Among the highest-valued correlations were f_0 median and Size, f_0 baseline, spectral emphasis and Age, spectral emphasis and Toughness, and jitter and Honesty.

Table 5 - *Mean values for acoustic parameters*

	B1	*B2*	*CB*	*JH*	*NW*

emph dB	5.52	6.05	8.75	5.63	5.48
f0 med Hz	222	173	127	203	134
f0 base Hz	180	111	85	165	95
f med s.t.	83.50	89.25	83.83	92	84.75
f base s.t.	89.87	81.50	77	88.42	78.75
fSD s.t.	2.49	5.50	4.63	2.64	4.20
shimmer %	10.63	16.58	14.90	14.04	14.30
jitter %	2.40	3.93	4.08	2.67	3.76

For the first parameter, Chief Bogo stands out with a value about 2 dB higher than the other characters, whose values are typical for normal speech. We can look at the role Chief Bogo plays and the scenes he usually appears in for clues to explain that discrepancy. His lines usually involve him giving orders to the police officers, and in that situation, speaking louder is quite normal and expected.

Looking at the group in general, we find that spectral emphasis is positively correlated with the ratings for Aggressive (0.77), Dominant (0.67) and Tough (0.76), which were the features where Chief Bogo stood out, and negatively correlated with Charming (-0.86), Kind (-0.72) and Pleasant (-0.81), which were among Bellwether 1's strong scores.

Meanwhile, f_0 base value is negatively correlated with Size (-0.71), Aggressive (-0.75), Dominant (-0.73) and Tough (-0.70) and positively correlated with Charming (0.84), Kind (0.87), Happy (0.94), Honest 0.93), Lively (0.91) and Pleasant (0.81).

4.4 Results of the description of the VPA

Below are described the non-neutral voice quality settings and prosodic elements that characterise the characters' voice profiles in the audio samples included in the Listening Experiment.

Bellwether 1: Lip Spreading, Extensive Labial Range, Raised Larynx, Whispery Voice, High Mean Pitch, Extensive Pitch Range, High Pitch Variability, High Mean Loudness, Fast Speaking Rate.

Bellwether 2: Extensive Labial Range, Extensive Mandibular Range, Raised Larynx, Tense Vocal Tract, Whispery Voice, Harsh Voice, High Mean Pitch, Extensive Pitch Range.

Chief Bogo: Lip Rounding, Lowered Larynx, Tense Vocal Tract, Creaky Voice, Low Mean Pitch, Extensive Pitch Range, High Pitch Variability, High Mean Loudness.

Judy Hopps: Extensive Labial Range, Advanced Lingual Tip/Blade, Fronted Tongue Body, Raised Larynx, Whispery Voice, High Mean Pitch, Extensive Pitch Range, High Pitch Variability, High Loudness Variability, Fast Speaking Rate.

Nick Wilde: Lax Vocal Tract, Whispery Voice, Extensive Pitch Range, High Pitch Variability, Slow Speaking Rate.

4.5 Results of the Reading Experiment

Overall, the results from the Reading Experiment were quite similar to the results from the Listening Experiment. Table 6 shows a comparison between the scores for the Listening Experiment (**L**) and the scores from respondents of the Reading Experiment who had not seen the film (**RnS**), who, presumably, had no other contact with the characters from the film beyond the text samples presented to them in the Experiment.

Table 6 - *Comparison of perception scores between Listeners and Readers who did not see Zootopia*

Scales	*B1*		*B2*		*CB*		*JH*		*NW*	
	L	*RnS*	*L*	*RnS*	*L*	*RnS*	*L*	*RnS*	*L*	*RnS*
Big	30	23	24	27	87	56	34	38	53	47
Old	23	33	23	36	65	54	28	38	60	31
Aggressive	13	19	69	68	86	74	49	55	43	68
Dominant	44	54	88	78	91	83	69	68	66	80
Energetic	88	81	84	84	78	66	83	73	44	64
Tough	13	23	69	76	91	80	58	53	45	71
Charming	78	65	63	60	51	36	68	62	66	55
Kind	82	78	37	48	72	40	62	62	52	41
Bold	50	38	74	71	72	62	68	50	46	67
Happy	89	82	65	64	45	54	70	59	51	62
Honest	73	72	39	58	41	46	73	70	46	44
Sophisticated	45	47	64	54	50	45	59	60	55	45
Lively	93	83	74	70	48	56	78	58	58	65
Pleasant	74	79	56	55	51	33	71	60	67	42
Correlation	*0.966*		*0.921*		*0.767*		*0.893*		*-0.215*	

The correlations show that even in the absence of their voices and without previous knowledge of the characters and the narrative, the reading profiles are similar to the listening profiles, indicating that the textual information is enough to give the audience an idea of what these characters are like. However, the slightly stronger scores for most of the descriptors in the Listening Experiment indicate that vocal performance does help intensify the impressions given by the characters' lines, especially for Chief Bogo.

5. *Discussion*

Overall, the results of the Listening and the Reading Experiments showed congruence. However, the results have not been enough to explain in detail how the lexical content may have been a deciding factor in the perception of these characters, as only one of the samples made an explicit reference to a feature being investigated in the study. We hope to gain more insight into the influence of the lexical content by expanding the contexts of analysis.

The results of the Listening Experiment attest to the relevant role of voice characteristics in the attribution of physical and personality characteristics by the listeners.

Chief Bogo is a big character, Judy and Bellwether are small, and Nick is midsize. The scores obtained are according to their physical types. For the descriptor Big, Chief Bogo obtained a high score (87), and Nick an intermediate score (53). For the same scale, Bellwether 1, Bellwether 2 and Judy got, respectively, the following scores (30), (24) and (34). These results show the strong explanatory power of the Frequency Code. Looking at the acoustic parameters, we find a Low base f_0 value (85Hz) for Chief Bogo's voice, Mid f_0 values for Nick (95Hz) and high f_0 values for Bellwether 1 (180 Hz), Bellwether 2 (111 Hz) and Judy (165 Hz). Furthermore, Bogo's low f_0 voice was considered Aggressive and Tough, whereas Bellwether's High f_0 voice was considered Docile. Low Mean Pitch might also have influenced judgments of Charm. Chief Bogo, whose voice profile was characterised by Low Mean Pitch, Lowered Larynx and Lip Rounding voice quality settings, got the lowest score regarding Charm. Lowered Larynx and Lip Rounding expand the vocal tract and the acoustic result is lowered f_0.

Judgments of Dominance and Boldness were not related to high or low f_0 values, but the ones regarding Happiness, Pleasantness, and Liveliness were. The voices characterised by higher f_0 values were considered happier, more pleasant, and more lively. Bellwether 1 got the highest score for Happiness. This judgment might have been influenced by the Lip Spreading voice quality setting that the listeners frequently perceive as a smile (Emond, Rilliard & Trouvain, 2016).

Judgments regarding the characters' ages were also made according to their age profiles. Results show that the listeners of the perception tests could identify age differences. From older to younger, we find Bogo, Nick, Judy, and Bellwether 1 and 2.

Even in Nick's case, where none of his perception ratings met the threshold for strong scores in either the Listening Experiment or the Reading Experiment, voice features may have provided cues for personality judgements as he is portrayed in the film as a hard-to-read character. His vocal profile included Lax Vocal Tract, combined with Whispery Voice and Slow Speaking Rate. These characteristics might have influenced the low scores he obtained in judgments concerning Liveliness, Energy, Aggressiveness, Toughness and Boldness. Nick was the only character with a Slow Speaking Rate.

Judgments on Sophistication did not vary as much from character to character and from listening to reading. Associations between vocal profiles or lexical

characteristics and judgments of Honesty and Kindness are not clear-cut and may require investigation of contextual presuppositions.

6. *Conclusions*

The results of both experiments were compared and found to be congruent. The Listening Experiment performed better than the Reading Experiment in terms of revealing the characters' physical and psychological features. The vocal performances matched the characters' personality features displayed in the film and, in most cases, amplified them.

The f_0 baseline and spectral emphasis measures were helpful in considering the higher intensity of characteristics attributed to Size, Age, and attitudinal features such as Aggressiveness and Toughness.

Regarding the acoustic analysis of the material, some measurements aligned with the impressionistic patterns discussed and evidenced previously in the literature on sound symbolism (Hinton, Nichols & Ohala, 1995). At the same time, other correlations found in the study are not as efficiently explained by findings in previous studies and require further inspection, like whispery voices and higher scores for Charm (Gussenvhoven, 2004).

Some refinement concerning the analysis of the acoustic measurements may be necessary to eliminate measurements that do not offer any significant clues regarding the correlations between acoustic parameters and the physical and psychological features investigated here.

We believe that applying the experiments to data from different languages would also offer some insight into these issues and help formulate solutions to them. The Listening Experiment in Swedish is already underway. As correlations between vocal characteristics of a speaker and impressions of physical, psychological, and social features may vary depending on language and culture, we foresee an expansion of our research project by integrating researchers exploring other languages.

Acknowledgements

The authors wish to thank Plinio A. Barbosa for his contributions to the study. The first author thanks the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* – Brasil (CAPES) – Finance Code 001 [grant #88887.630904/2021-00, grant #88881.689915/2022-01]. The second author acknowledges a grant from The Swedish Foundation for International Cooperation in Research and Higher Education (STINT) (IB2015-6488). The third author thanks PIPEq-PUCSP [grant #31279, 2024]

References

Aronovitch, C.D. (1976). The voice of personality: stereotyped judgments and their relation to voice quality and sex of speaker. In *The Journal of Social Psychology*, 99, 207–220. https://doi.org/10.1080/00224545.1976.9924774

Bosseaux, C. (2015). Dubbing. In Bosseaux, C. *Dubbing, Film and Performance: Uncanny Encounters*. Oxford: Peter Lang, 55–84. https://doi.org/10.3726/978-3-0353-0737-5

Bosseaux, C. (2018). Investigating dubbing: Learning from the past, looking to the future. In, Pérez-González, L. (Ed.), *The Routledge Handbook of Audiovisual Translation.* New York: Routledge, 48–63. https://doi.org/10.4324/9781315717166

Bosseaux, C. (2019). Voice in French dubbing: the case of Julianne Moore. In *Perspectives,* 27(2)*,* 218–234. https://doi.org/10.1080/0907676X.2018.1452275

Brisola, E. (2024). *Um estudo experimental sobre o estilo de fala em dublagem.* Master's thesis. Pontifical Catholic University of São Paulo.

Campos Plaza, N. (Eds.), *Interdisciplinarity in translation studies. Theoretical models, creative approaches and applied methods*. Bern: Peter Lang, 259–276. https://10.3726/978-3-0351-0954-2

Chaume, F. (2014). *Audiovisual translation: dubbing*. New York: Routledge.

Chaume, F., Ranzato, I. & Zanotti, S. (2018). The challenges and opportunities of audiovisual translation: An interview with Frederic Chaume. In *CULTUS,* 2018(11)*,* 10-17.

Crochiquia, A. (2020). *A voz na construção de personagens em um desenho animado.* Master's thesis. Pontifical Catholic University of São Paulo.

Crochiquia, A., Eriksson, A., Fontes, M.A.S. & Madureira, S. (2020). A phonetic study of Zootopia characters' voices in Brazilian Portuguese dubbing: the role of stereotypes. In *DELTA: Documentação e Estudos em Linguística Teórica e Aplicada*, 36(3), 1-46. https://doi.org/10.1590/1678-460X2020360311

Crochiquia, A., Eriksson, A., Barbosa, P.A. & Madureira, S. (2022). A perceptual and acoustic study of dubbed voices in an animated film. In Frota, S., Cruz, M. & Vigário, M. (Eds.), *Proceedings of the 11th International Conference on Speech Prosody*, Lisbon, Portugal, 23-26 May 2022, 565-569. https://doi.org/10.21437/SpeechProsody.2022-115

Crochiquia, A., Eriksson, A., Madureira, S. & Barbosa, P.A. (2023). Animated film character profile: the roles of voice and lexical content. In Skarnitzl, R., Volín, J. (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic, 7-11 August 2023, 466-47. https://guarant.cz/icphs2023/862.pdf

Danan, M. (1991). Dubbing as an Expression of Nationalism. In *Meta,* 36(4)*,* 606–614. https://doi.org/10.7202/002446ar

Díaz-Cintas, J. (1990). Dubbing or subtitling: The eternal dilemma (film, translation techniques). In *Perspectives,* 7*,* 31–40. https://doi.org/10.1080/0907676X.1999.9961346

Díaz-Cintas, J. (2015). Preface. In Ranzato, I. (Ed.), *Translating Culture Specific References on Television: The Case of Dubbing*. New York: Routledge.

Díaz-Cintas, J. (2019). Film censorship in Franco's Spain: the transforming power of dubbing. In *Perspectives,* 27*,* 182-200. https://doi.org/10.1080/0907676X.2017.1420669

Emond, C., Rilliard, A. & Trouvain, J. (2016). Perception of smiling in different modalities by native vs. non-native speakers. In *Proceedings of the 8th International Conference on*

Speech Prosody, Boston, USA, 31 May / 1-3 June 2016, 639-643. https://doi.org/10.21437/SpeechProsody.2016-131

Fasoli, F., Mazzurega, M. & Sulpizio, S. (2017). When Characters Impact on Dubbing: The Role of Sexual Stereotypes on Voice Actor/Actress' Preferences. In *Media Psychology*, 20, 450-476. https://doi.org/10.1080/15213269.2016.1202840

Gussenhoven, C. (2016). Foundations of Intonational Meaning: Anatomical and Physiological Factors. In *Topics in Cognitive Science,* 8*,* 425-434. https://doi.org/10.1111/tops.12197

Hinton, L., Nichols, J. & Ohala, J. (1995). Introduction: Sound-symbolic processes. In Hinton, L., Nichols, J. & Ohala, J. (Eds.), *Sound Symbolism*. Cambridge: Cambridge University Press, 1–12. https://doi.org/10.1017/CBO9780511751806.001

Imhof, M. (2010). Listening to Voices and Judging People. In *International Journal of Listening*, 24, 19–33. https://doi.org/10.1080/10904010903466295

Kreiman, J., Sidtis, D. (2011). Voice, Emotion and Personality. In Kreiman, J., Sidtis, D., *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. Oxford: Wiley-Blackwell, 302–360.

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

Liu, W., Zhang, X. & Liang, C. (2022). An acoustic study on character voices of dominators and subordinates: A case study on male characters in Empresses in the Palace. In *Frontiers in Communication,* 7. https://doi.org/10.3389/fcomm.2022.1088170

Mahrholz, G, Belin, P. & McAleer, P. (2018). Judgements of a speaker's personality are correlated across differing content and stimulus type. In *PLoS One*, 13, e0204991. https://doi.org/10.1371/journal.pone.0204991

Nybakk, L.M. (2014). *Representations of linguistic variation in audiovisual translation: A study of American animated films and their Norwegian dubbed translations.* Master's thesis. Norwegian University of Technology and Science.

Ohala, J.J. (1984). An Ethological Perspective on Common Cross-Language Utilization of F0 of Voice. In *Phonetica,* 41(1)*,* 1–16. https://doi.org/10.1159/000261706

Pedersen, J. (2010). Audiovisual translation – in general and in Scandinavia. In *Perspectives,* 18(1)*,* 1–22. https://doi.org/10.1080/09076760903442423

Sánchez-Mompeán, S. (2016). "It's not what they said; it's how they said it": A corpus-based study on the translation of intonation for dubbing. In Rojo-López, A.M., Sánchez-Mompeán, S. (2020). Introduction: Unhiding the Art. In Sánchez-Mompeán, S. *The Prosody of Dubbed Speech: Beyond the Character's Words*. New York: Palgrave Macmillan, 1-10. https://doi.org/10.1007/978-3-030-35521-0_1

San Segundo, E., Mompean, J.A. (2017). A Simplified Vocal Profile Analysis Protocol for the Assessment of Voice Quality and Speaker Similarity. In *Journal of Voice,* 31(5)*,* 644. e11–644.e27. https://doi.org/10.1016/j.jvoice.2017.01.005

Tveit, J.E. (2009). Dubbing versus Subtitling: Old Battleground Revisited. In Díaz-Cintas, J., Anderman, G. (Eds.), *Audiovisual Translation: Language Transfer on Screen*. New York: Palgrave Macmillan, 85–96. https://doi.org/10.1057/9780230234581_7

Weirich, M. (2008). Vocal Stereotypes. In *Proceedings of 2nd Tutorial and Research Workshop on Experimental Linguistics*. Athens, Greece, 25-27 August 2008, 229–232. https://www.isca-archive.org/exling_2008/weirich08_exling.pdf

Xu, A., Lee, A. (2018). Perception of vocal attractiveness by Mandarin native listeners. In Klessa, K., Bachan, J., Wagner, A., Karpiński, M. & Śledziński, D. (Eds.), *Proceedings of the 9th International Conference on Speech Prosody*. Poznan, Poland, 13-16 June 2018, 344–348. https://doi.org/10.21437/SpeechProsody.2018-70

Whitman-Linsen, C. (1992). *Through the Dubbing Glass: The Synchronization of American Motion Pictures into German, French and Spanish*. Bern: Peter Lang.

DOI: 10.17469/O2112AISV000004

SIMON DEVAUCHELLE, ALBERT RILLIARD, DAVID DOUKHAN,
LUCAS ONDEL YANG

Variation of Perceived Voice Pitch Across Time Periods, Gender, and Age in French Media Archives

Have female voices lowered with time? A diachronic corpus based on French television or radio shows, constructed around four Periods (1955-56, 1975-76, 1995-96, 2015-16), grouping 1028 cisgender female and male speakers of four age categories. The speakers' voices were automatically segmented at the phone level, and acoustically analyzed. The fundamental frequency (F_0) and the formants were estimated; from the formants, the vocal tract length was calculated. Linear mixed regressions were fitted to these acoustic cues to estimate the role of the speaker's Age, Gender, and Period. The models showed few gender-specific changes in pitch-related parameters across Periods. Mean F_0 decreases with age for females, while it increases for males. The first formant decreases for open vowels over time Periods. These observations do not support a lowering of female voices over time. Results are discussed in the light of previous studies and changes in the structure of the French population.

Keywords: Voice pitch, Gender, Diachrony, Audiovisual archives, Fundamental frequency, Vocal Tract Length, Formants.

1. *Introduction*

Beyond the features relevant to the linguistic message *stricto sensu*, our voices carry various information relative to individual and social characteristics, thus participating in the indexicalization of one's social representation (Eckert, 2008). Our social roles and identities vary across contexts as we play different *characters* depending on interlocutors and social roles (Sadanobu, 2015); these complex identities are constructed during infancy and adolescence (Drager, 2015). Audiovisual media shape one's behavior through their prescriptive use of language and the display of social types (Stuart-Smith, 2017). The present work proposes a series of measurements related to the voice's perceived pitch in a diachronic perspective (here, we'll consistently use "pitch" to refer to the perceived tone – high or low – of a voice, not to fundamental frequency or other acoustic parameters). We view the voice pitch as an essential constituent of one's distinctive vocal characteristics, which typically participates in constructing gender display (Cartei et al., 2019, 2022).

Attributing a gender category to an interlocutor is important for spoken interactions in many cultures. It notably allows the use of adequate terms of address or honorifics (Tawilapakul, 2022), which are embedded in cultural systems and the localized language varieties (for terms of address in varieties of Spanish, see,

e.g., Rebollo Couto & Lopes, 2011). Gender categories may be attributed rapidly when cues are not ambiguous (Doukhan et al., 2023). This attribution is linked to a complex system of visual and acoustic cues (Richoz et al., 2017). Among the acoustic cues, various aspects were studied, and it has been shown that a combination of factors shall be used to decide gender (Merritt & Bent, 2022). However, the perceived pitch of a speaker's voice constitutes a fundamental aspect of gender indexicality in its complex realities (Leung et al., 2018; Pépiot & Arnold, 2021; Weirich & Simpson, 2018).

Individuals construct their voices during their lifespan, and the vocal characteristics are notably tuned to target the gender representation each wants to display in a given context (Holmes, 1998; Podesva, 2007, 2011). This vocal display is trained during childhood, even before puberty (Guzman et al., 2014), and of course, there is a significant vocal evolution during adolescence (Markova et al., 2016; Yamauchi et al., 2015). Later, during adulthood, we adapt our voices to the expectations of the sociocultural context (Ohara, 2001). The socially constructed characteristics of voices are thus part of the cues produced and perceived to interact smoothly within our society, and displaying more prototypical vocal characteristics impacts the speech perception process (Johnson, 2006). As cultural constructions, our voices are not purely intrinsic characteristics of our phonatory system but are tuned to how we use them.

As well as being an important indicator of gender, the voice pitch plays a varied role in vocal communication. A fundamental aspect of the voice's pitch during interactions was captured by Ohala's *Frequency Code* (Ohala, 1983, 1994), which describes the symbolic relation between the pitch of vocalizations and its behavioral interpretation. It is based on a systematic review of literature (Morton, 1977) that shows, in mammal and bird calls, a strong relation between low/descending pitch associated with hostile, aggressive calls vs. high/rising pitch associated with submissive, appeasement calls. The symbolic use of this vocal difference is based on the relation between an animal's size and the pitch of its voice; larger animals tend to have lower voices and to be more potent (Ohala, 1983, 1994). In human communication, this symbolic code, among others, is used for various behaviors (Gussenhoven, 2004) and may explain, e.g., cross-linguistic trends like the prevalence of raising pitch for questions. Such interpretations of voice pitch also resonate with the social judgments given to higher and lower voices regarding psychological powerlessness/power, intersecting with the speaker's gender – with a preference for higher-pitched voices for femininity that vary across cultures (van Bezooijen, 1995). Low voice pitch, a sexual dimorphism correlate, was described as an important feature for males' reproductive success and attractiveness (Apicella et al., 2007; Quené et al., 2021). The reverse holds for high-pitched female voices being judged as more sexy or attractive, with large variations linked to behavioral and sociocultural factors (Barkat-Defradas et al., 2021).

The social contexts within which a speaker is evolving thus create strong expectations of their vocal characteristics. In light of the preceding paragraphs,

these expectations may be divergent when female speakers behave in public arenas, typically in the political field, where a higher voice carries "feminine" attributes linked to powerlessness ("short, weak, dependent, modest" in the terms of van Bezooijen, 1995), and a lower voice is associated to "masculine" characteristics and power, credibility, etc. (Puts et al., 2006). As Coulomb-Gully (2022) puts it, female voices in the (French) political arena (a typical place for the expression of patriarchal power) are criticized for entering there and expressing themself: their voices are "*de trop*" – always too much of something (too high, too loud, etc.). Such mismatches in sociocultural expectations led some female politicians to follow voice coaching to deepen their voice; similar social impositions led others to modify their accents to fit the perceived expectations linked with a given position (Riverin-Coutlée & Harrington, 2022). These impositions also apply differently in different places, where prototypically feminine features may be relevant for being successful. Cinema and TV actresses also receive varying demands for their voices, depending on the producers' cultural background. French producers have different criteria than companies from the USA, looking after voice talents for dubbing that match the original voices (Le Fèvre-Berthelot, 2015).

Within this context that puts the voice's pitch as a central cue indexing gender, there are growing claims that the female voice's pitch has lowered over time, and most of them (e.g., Ellis et al., 2023; Krahé & Papakonstantinou, 2020) base their claim on the same reference (Pemberton et al., 1998). These claims are repeated over several media platforms and sometimes twisted in the process. A newspaper article (Nebelsztein, 2018), for example, cherry-picked two scientific papers (Berg et al., 2017; Pemberton et al., 1998) presenting studies supporting a lowering of the fundamental frequency of females to conclude female voices have lowered while male voices haven't – and it is not isolated (e.g., Robson, 2018; Taylor, 2018). There are several problems with this claim, firstly because it does not consider other studies with contrasting conclusions (e.g., Hollien et al., 1997, or the more recent work by Suire & Barkat-Defradas, 2020); secondly because the study by Berg et al. (2017) does not present diachronic evidence and uses a relatively peculiar production protocol that may influence their measure; third, because these works test the pitch of Australian (Pemberton et al., 1998) and German (Berg et al., 2017) females and the voice pitch is culturally encoded, so it is difficult to generalize to other populations; finally, the diachronic work of Pemberton et al. (1998) finds a difference amounting to a lowering in the mean pitch of 1.8st (which is arguably limited) and does not test male speakers, nor female of different age, so it is more complex to know if there are some other factors (like vocal effort – e.g., Berg et al., 2017; Titze & Sundberg, 1992) that may be involved in the described difference.

To provide referential values for the case of voices spoken and perceived in France, we present here a diachronic study that measured two correlates of voice pitch – the voice fundamental frequency (F_0) and an estimation of the vocal tract length (VTL) based on its resonances – within the speakers of a cross-sectional dataset based on 1028 cis-gender female and male speakers between 20 and 80 years old broadcasted

in French national television and radio channels along 60 years spanning between 1955 and 2015. The effects of the speakers' gender, age, and broadcasting periods are evaluated to observe potential variations. The paper first presents the corpus construction process, its evaluation, the acoustic measurements, and the statistical analysis; it then presents the changes observed for the three controlled factors and discusses them in the light of the literature.

2. *Methods*

2.1 Corpus

2.1.1 Speaker characteristics and identification process

This work is based on the diachronic corpus collected by Uro et al. (2022), composed of extracts from broadcasted television or radio shows from the French National Institute of Audiovisual (INA) archives. Programs broadcasted through regional channels were excluded from analyses to keep only voice representations at the national level. Each type of show has a specific stylistic feature. As speech style impacts voice characteristics, studio-based talk show interviews were preferred to maximize sound recording quality (to avoid environmental noises) and gather data from interactive discussions where the participants shall have reasonable speaking time and use question/answer interactions. Note that on a corpus of such a scale (more than 850 different archives entries were used), the topic in each show could not be controlled, but many domains are represented, from news and politics to arts, science, education, etc. The shows were selected for being broadcasted at four different time *Periods* spaced by 20 years (1955-1956, 1975-1976, 1995-1996, 2015-2016), and potential speakers were identified by INA's archivists to reach a target of thirty speakers of both genders in four age categories (20-35, 36-50, 51-65, over 65). Thirty speakers for each of the 32 categories of four time Periods * four Age categories * two Genders amounts to a theoretical target of 960 speakers. The documentalists identified more than 6,000 potential target speakers within INA archives.

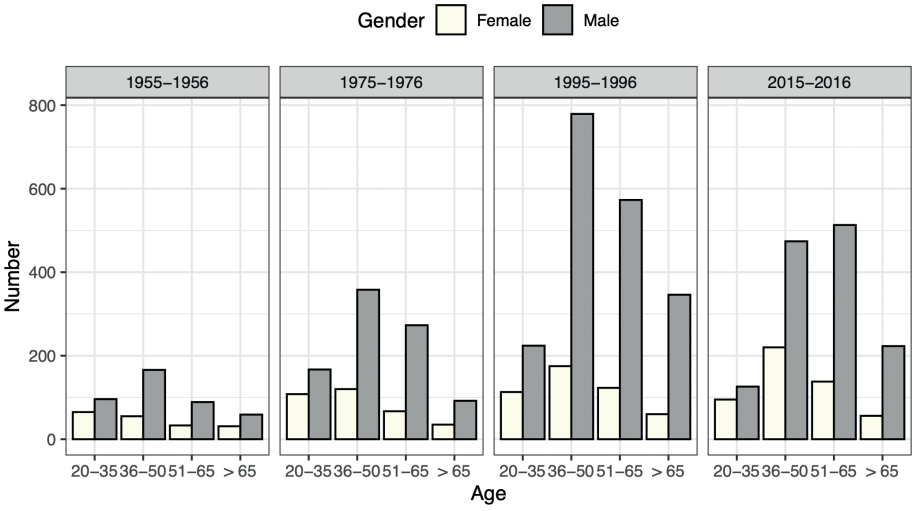
A semi-automatic process, through diarization and speaker identification, allowed the effective identification of the targets in actual recordings (Uro et al., 2022) and led to a dataset comprising 1028 speakers with a minimal amount of three minutes of recordings per speaker; these speakers are spread across time Periods, Age categories, and Gender according to Table 1. It is worth emphasizing the solid and consistent bias of female representation in the media (Coulomb-Gully, 2011; Doukhan et al., 2018): even trying to identify an even number of speakers of each gender, the dataset has about 50 % more males than females. This is explained by the difficult task of finding female speakers, especially in the youngest and oldest age categories, in the 1950s and the 1970s: the archivists managed to find a sufficient number of targets, but during the identification stage, many of the females were given minor roles, often only appearing on stage while males spoke on their behalf. The process led to identifying more males in some categories than the 30-speaker

initial goal because only the male target(s) speak in some archives featuring several target speakers.

Table 1 - *Number of speakers for each of the time Periods (by row), Age category (column), and Gender (Female or Male, in columns), plus marginal sums*

	20-35		*36-50*		*51-65*		*> 65*		*Row sum*		
	F	*M*	*F*	*M*	*F*	*M*	*F*	*M*	*F*	*M*	*Both*
1955-56	18	41	21	72	18	51	19	15	76	179	255
1975-76	18	18	23	42	28	37	18	25	87	122	209
1995-96	31	30	33	45	29	48	29	34	122	157	279
2015-16	31	31	30	53	29	49	31	31	121	164	285
Column sum	98	120	107	212	104	185	97	105	406	622	1028

Figure 1 - *Number of potential target speakers identified within INA databases across Periods (individual plots), age categories (x-axis), and gender (white or grey bars)*



The representational bias in the media also applies to age categories, with the youngest and oldest categories underrepresented. This explains the fewer targets in these age categories (Table 1), as illustrated in Figure 1, representing the number of potential targets identified by the archivists within INA databases. The overall aging of the French population may be observed with the growing proportion of the 51-65 y. o. category over the Periods, and the decrease of the 20-35 one. The difference in the number of targets between time Periods is linked to the smaller size of the oldest archives, with fewer channels and fewer shows recorded at that time.

For each target speaker identified in the corpus, a minimum of three minutes of speech was selected, removing passages where non-speech events exceed a given signal-to-noise ratio threshold to keep only relatively clean speech. A similar process also removed overlapping speech events. This cleaning procedure was essential

to keep only qualitative speech recordings; it removed about 2/3 of the spoken segments (Uro et al., 2022).

For each speaker, the selection process ended with passages of speech of more than two seconds (the different passages amounted to more than 3 minutes of speech), but not necessarily corresponding to complete speech turns, as the noise removal procedure may delete parts of the turns, e.g., where music was found in the background. These passages of speech are mainly rid off pauses, which were removed at the diarization stage. Table 2 presents the total duration of speech considered for each category of period, age, and gender.

Table 2 - *Total duration (in minutes) of speech passages selected for all speakers belonging to the same time Period (by row), Age category (column), and Gender (Female or Male, in columns), plus marginal sums*

	20-35		*36-50*		*51-65*		*> 65*		*Row sum*		
	F	*M*	*F*	*M*	*F*	*M*	*F*	*M*	*F*	*M*	*Both*
1955-56	21	98	52	200	39	168	102	161	214	527	741
1975-76	38	86	43	120	172	130	147	272	400	608	1008
1995-96	238	194	253	390	369	402	454	339	1314	1325	2639
2015-16	287	240	175	439	211	325	291	342	964	1346	2310
Column sum	584	618	523	1149	791	1025	994	1014	2892	3806	6698

2.1.2 Automatic transcription, phone segmentation, and evaluation

The entire speech corpus was automatically transcribed using Whisper (Radford et al., 2023). Then, the output transcripts were phonetized and force-aligned on the speech signals to extract vocalic excerpts, using pre-trained models for French from Montreal Forced Aligner (MFA; McAuliffe et al., 2017) and its French phonetic dictionary and language models (McAuliffe & Sonderegger, 2022c, 2022b, 2022a, 2022d). Timing indications proposed by the ASR predictions were also used to strengthen the alignments. Only oral vowels were selected by taking the middle frame of the detected vowel, resulting in the selection of over 1,300,000 frames. Twelve French oral vowels were selected: /a/, /ɑ/, /e/, /ɛ/, /œ/, /ə/, /ø/, /i/, /o/, /ɔ/, /u/ and /y/. Frames corresponding to vowels more than 200 milliseconds long were discarded from the study, preserving 96 % of the detected vowels. Automatic misalignment or poor automatic transcription can result in longer durations. We also removed frames corresponding to unvoiced phones using the voiced/unvoiced decision described in the next section.

An evaluation of the automatic transcription was conducted to evaluate the quality of the pipeline. While Whisper has been tested on French material (Radford et al., 2023), the evaluation corpora were based on read speech, potentially differing from the more spontaneous productions of our audiovisual corpus. A subset of 81 speech segments (about 20 minutes) was randomly selected across all periods, gender, and age categories. Four L1 French speakers manually transcribed the segments, which were then phonetized and forced-aligned by MFA. Our main concern was that the ASR system would make up words that had not been pronounced,

especially for older materials where sound artifacts are more common. Excerpts from old audio archives are less likely to have been used for Whisper training as they are less common. A further assumption was that Whisper tended to "smooth" transcriptions by removing hesitations and repetitions. To get a clearer idea of the system efficiency for this application, two scores were used corresponding to two analysis levels: the word error rate (WER) and the phone error rate (PER). Results are presented in section 3.1. In the following analysis, the central 25ms frame of each vowel was targeted for the acoustic analysis.

2.2 Acoustic analysis

The parameters targeted in this paper are known correlates of perceived pitch (Laver, 1980; Ohala, 1994) and are related to anatomical differences between females and males (Gisladottir et al., 2023; Sataloff, 2017; Simpson, 2009; Titze, 1989): from each vowel, segmented as described above, we estimated the voice fundamental frequency (F_0) and the first four formants. We also estimated the Vocal Tract Length (VTL) from the formants used as proxies of the vocal tract resonances (Titze et al., 2015).

For F_0 estimations, it has been shown that pitch detection algorithms perform differently on difficult signals (Vaysse et al., 2022). We used three algorithms to robustly estimate the fundamental frequency: Praat's AC algorithm (Boersma & Weenink, 2022) with an F_0 range of 65-650 Hz, REAPER (Talkin, 2015) with the default setting and with Hilbert transform (thus estimating two voicing estimates), and FCN-F0 (Ardaillon & Roebel, 2019) with its default pitch range (30-1000Hz) and Viterbi smoothing. We used F_0 values estimated by Praat and FCN-F0, plus the voiced/unvoiced decision produced by Praat and REAPER. Only the frames classified as voiced by Praat and REAPER were considered; among these frames, those who received F_0 values from Praat and FCN-F0 that do not differ by more than a 20 % gross-error-rate threshold (e.g., Jouvet & Laprie, 2017) were considered valid. This strategy considers only 52.2 % of valid frames in the original signal, less than if considering only the Praat algorithm output (61.1 %). On these frames, the value returned by Praat was considered.

The first four formants (F_1, F_2, F_3, F_4) were estimated by applying Praat's implementation of the Burg algorithm (Boersma & Weenink, 2022). As this work is focused on an estimate of acoustic differences linked with a possibly varying expression of gender through vocal cues, it seems dubious to use the speaker's gender to change the algorithm parameters – and typically, the frequency ceiling: Praat's recommendation is to set the ceiling to 5500Hz for female, and 5000Hz for males, but this proved problematic, e.g. for /u/ vowels produced by females (Gendrot & Adda-Decker, 2005). We thus followed here the strategy proposed by Escudero et al. (2009), which consists of extracting, for each vowel, several potential sets of formants with different ceilings and then deciding an optimal ceiling for each speaker and each vowel category. The best ceiling minimizes the sum of the variances of each of the first three formants (using the logarithm of their values in

Hertz) for this speaker and this vowel; three formants were considered here (only two in Escudero et al., 2009) as in French, the third formant is relevant for rounding (e.g., Ménard et al., 2009). The set of ceilings that was used is based on the default ceiling for each gender (5.5 and 5 kHz for females and males), plus twenty ceilings above and twenty below this default value, spaced by percentual of the default ceiling value; this was done using Praat's *FormantPath* function, that implements part of Escudero et al. (2009) algorithm, applying the different ceiling according to equation 1 (from Praat help), where the frequency ceiling (*FC*) at the i^{th} step above or below the middle-frequency ceiling (MFC) equals MFC multiplied by a factor determined by plus or minus (if above or below) *i* times the *step* size.

$$FC_i = MFC * e^{(\pm i * step)} \quad (1)$$

In our case, *i* varies from 1 to 20, and the *step* factor was set to 0.01, so to have steps of approximately 50Hz below or above the middle ceiling frequency – for ceilings ranging from above 4000Hz up to 6100 Hz for males, and from 4500 up to 6700Hz for females. Formants were estimated frame by frame; those on frames classified as unvoiced during the F_0 estimation process were discarded. The F_0 and the formant values estimated at the middle of each vowel are considered for the remainder of this work; F_0 was expressed in semitones relative to 1Hz and formants in Bark to fit their perception.

We used the first four formants to estimate the vocal tract length (VTL). This was done by using equations 2 and 3 from Lammert & Narayanan (2015), where F_1, F_2, F_3, and F_4 are the first four formant values (in Hz), and C=34000cm/s is the speed of sound. One estimation of VTL was done on each vowel of the corpus.

$$\phi = \frac{0.089\,F_1}{1} + \frac{0.102\,F_2}{3} + \frac{0.121\,F_3}{5} + \frac{0.669\,F_4}{7} \quad (2)$$

$$VTL = \frac{c}{4*\phi} \quad (3)$$

2.3 Statistical models

On each vowel, the measurements of the F_0, the first three formants, or the VTL were used as the dependent variable (DV) of linear mixed models fit with R's *lme4* library (Bates et al., 2015; R Core Team, 2023), with as fixed factors the speaker's Age (in years, averaged over the different shows from which their speech was extracted, if applicable), the speaker's Gender (Female, Male), and the time Period (1955-56, 1975-76, 1995-96, 2015-16). For models based on the F_0 and the VTL dependent variables, the Speaker and the Vowel (the 12 oral vowels considered here) factors were used as random factors in the model, with Vowel nested within Speaker, to account for the individual variability in vowel performances (see eq. 4 for the maximal model based on these measurements, where *DV* is either F_0 or VTL). For the models fitting formants, the Vowel factor was used as a fixed factor to allow describing formant variations linked to vowels, and the Speaker factor was

kept as a random factor (see eq. 5 for the maximal models based on formants, where F_i stands for formant formants F_1, F_2 or F_3).

$$DV \sim Age * Gender * Period + (Speaker/Vowel) \quad (4)$$

$$F_i \sim Age * Gender * Period * Vowel + (Speaker) \quad (5)$$

A model simplification procedure was then applied to each model described by equations 4 and 5, using R's *step()* function (Kuznetsova et al., 2017). For the model based on VTL, the simplification procedure deleted the triple and double interactions, leaving only the main fixed factors (see Table 3). For the model based on F_0, the triple interaction and the double interaction between Age and Period were deleted, leading to a model with three fixed factors and two double interactions: between Gender and Age, and between Gender and Period (see Table 4). For the models based on formants, no simplification was done.

Table 3 - *Output of the backward reduction process for the model fitted on VTL, for Random factors (first part: number of eliminated factors, number of parameters, log. Likelihood, Akaike Information Criterion, Likelihood Ratio Test, degrees of freedom, and p-value), and for Fixed factors (second part: number of eliminated factors, Sum of squares, mean square, numerator and denominator degrees of freedom, F value, and p-value)*

Random	*Eliminated*	*npar*	*logLik*	*AIC*	*LRT*	*Df*	*p*
<none>		19	-1670409	3340857			
(1 | Vow:Spk)	0	18	-2044955	4089946	749092	1	< 2.2e-16
(1 | Spk)	0	18	-1670770	3341575	721	1	< 2.2e-16
Fixed	*Eliminated*	*Sum Sq*	*Mean Sq*	*NDF*	*DenDF*	*F value*	*p*
Age:Gen.:Per.	1	6.20	2.07	3	999.05	1.9621	0.11801
Age:Period	2	0.94	0.31	3	1002.36	0.2970	0.82756
Gen.:Per	3	2.31	0.77	3	1002.88	0.7319	0.53308
Age:Gender	4	2.75	2.75	1	1002.03	2.6074	0.10668
Age	0	6.71	6.71	1	1003.95	6.3746	0.01173
Gender	0	2369.13	2369.13	1	1001.65	2250.1529	< 2e-16
Period	0	100.53	33.51	3	1006.18	31.8271	< 2e-16

Table 4 - *Output of the backward reduction process for the model fitted on F_0, for Random factors (first part: number of eliminated factors, number of parameters, log. Likelihood, Akaike Information Criterion, Likelihood Ratio Test, degrees of freedom, and p-value), and for Fixed factors (second part: number of eliminated factors, Sum of squares, mean square, numerator and denominator degrees of freedom, F value, and p-value)*

Random	*Eliminated*	*npar*	*logLik*	*AIC*	*LRT*	*Df*	*p*
<none>		19	-3085128	6170293			
(1 | Vow:Spk)	0	18	-3097069	6194174	23883	1	< 2.2e-16
(1 | Spk)	0	18	-3094811	6189658	19367	1	< 2.2e-16

Fixed	*Eliminated*	*Sum Sq*	*Mean Sq*	*NDF*	*DenDF*	*F value*	*p*
Age:Gen.:Per.	1	18.430	6.143	3	998.47	0.4760	0.69909
Age:Period	2	71.255	23.752	3	1001.64	1.8401	0.13817
Age:Gender	0	246.772	246.772	1	1004.04	19.1181	1.357e-05
Gender:Period	0	105.208	35.069	3	1004.40	2.7169	0.04356

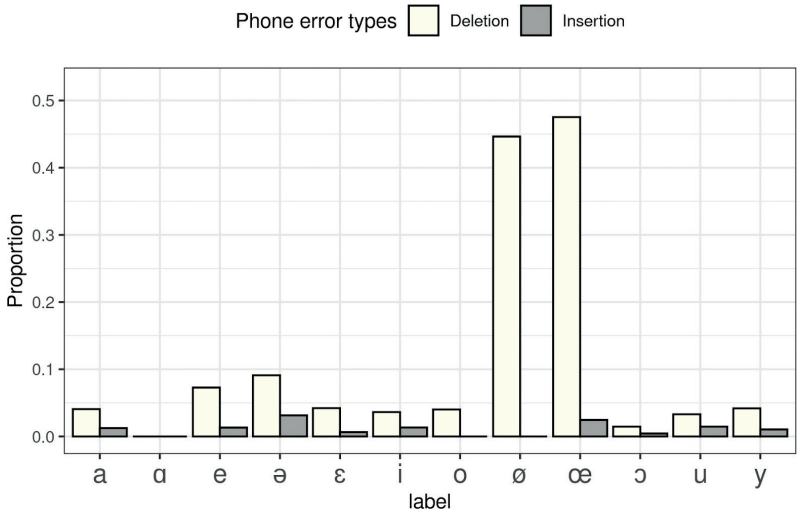
3. *Results*

3.1 Evaluation of the automatic transcription and phonetic alignment

On the subset of 81 speech segments, and compared to the human transcription, Whisper reached a 12.8 WER, using Whisper large v2 and after text normalization. The WER on our dataset is thus slightly higher than the WER on FLEURS and on Common Voice described in the OpenAI paper (Radford et al., 2023), but it is still a close performance. As we are here interested in the evolution of some phonetic characteristics of French spoken in the media, the PER appears to be more adapted to evaluate the efficiency of the ASR system. Some difficult-to-decide transcription differences may not lead to pronunciation differences (e.g., for French: plurals where spelling differs but not pronunciation) – and indeed, the PER was 7.9 % when all phones were considered and 9.8 % for oral vowels only.

Let's emphasize that the precision at the phone level is 97.4 %; in other words, Whisper produces few hallucinations in its predictions, at least once phonetized. 5.4 % of vowels in the evaluation subset were deleted during the automatic alignment process compared to the manual transcription – showing a reduced effect of Whisper in smoothing its outputs. Indeed, most of the errors at the phone level are deletions (68.2 %). As shown in Figure 2, the most frequently deleted vowels were /œ/ (46.7 % deletion), /ø/ (44.6 %), /ə/ (8.6 %), and /e/ (7.2 %). One-quarter of deleted oral vowels are derived from the French word for hesitation, "*euh*" (pronounced /œ/ or /ø/), followed by the function words "et" (6.5 %) and "de" (4.7 %). As shown in Figure 2, the vowel class the most "hallucinated" is the /ə/: 33 % of its insertions come from the negation word "*ne*" which is generally not produced in casual conversations but added by Whisper, producing more formal sentences in its outputs.

Figure 2 - *Proportion of vowels deleted (white) or inserted (grey) during the automatic transcription process*



Similar to Barras et al. (2002) findings for another ASR system evaluated for the transcription of French media archives, we do not notice differences in the phone or word error rates of Whisper transcriptions over time. Varying performances linked to gender may be related to the specific characteristics of the selected speakers (speaking style, recording conditions, etc.) and not to gender itself.

3.2 Variations of VTL with Age, Gender, and Period

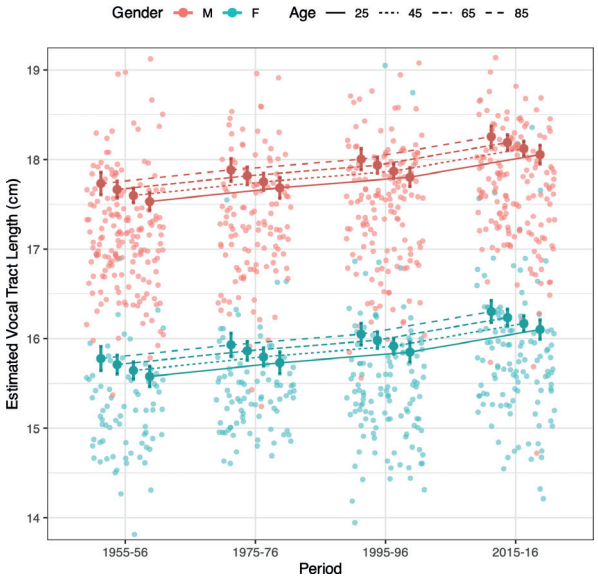
The mean tendencies of the linear mixed model fitted on values of VTL are displayed in Figure 3 for the three fixed factors together (let's recall there were no significant interactions between them). The main effect in the model is –of course– related to the speaker's Gender and reflects the mean difference of about two centimeters observed between adult females and males in vocal tract length, corresponding to several descriptions in the literature (De Pinto & Hollien, 1982; Hollien et al., 1994; Vorperian et al., 2009, 2011, 2019). The two other factors have much smaller effect sizes if they show significant and continuous trends. There is a tendency for vocal tract resonances to decrease with the speakers' Age, that amount, in terms of estimated VTL, to a lengthening of about 0.25cm between 25 and 85 years old. Comparatively, the effect linked to Period is twice as large, with an increase of about 0.5 cm of estimated VTL between the 1955s and 2015s.

3.3 Variations of F_0 with Age, Gender, and Period

The model fitted to the speakers' estimated F_0 values is based on two interactions: one between the Gender and Age factors, and the second between the Gender and Period factors. As for VTL, the Gender * Age interaction (see Figure 4) corresponds to the description in the literature of F_0 changes with these two factors, with Sataloff

et al. (2017) describing a decreasing trend of F_0 with age for females (descending pitch with age) and an increase for males (rising with age; see also Gisladottir et al., 2023). The trend of F0 change with age observed in our dataset corresponds to a decrease of 1.7st for 60 years or 0.3st for 10 years of age in female voices. For male speakers, the mean F_0 value tends to increase at a slower pace, with a rise of 1.3st for 60 years corresponding to a rate of 0.2st for 10 years.

Figure 3 - *Mean VTL (with confidence intervals) estimated by the model based on speaker Age (values for 25, 45, 65, and 85 y. o. are plotted), Gender, and time Period; the values estimated for each speaker are plotted as individual lightly colored points in the background, to give a better idea of the observed variation*

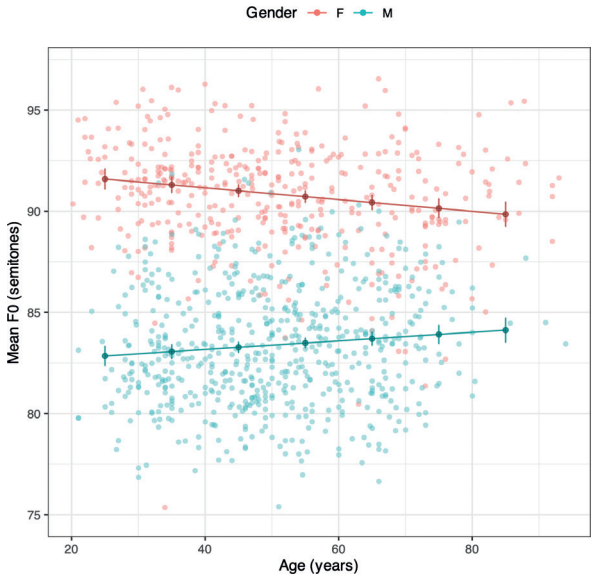


Note that the study by Leung et al. (2022) describes a downward trend of F_0, independent of gender, but the authors do not include the interaction between these two factors in their model and have a population sample with about twice as many females as males. The study by Berg et al. (2017), with a population between age 40 and 79 years old, observed a downward trend of F_0 for females but no significant change with age on this measurement for male speakers.

The interaction between Gender and Period is represented in Figure 5. A post-hoc comparison (with Bonferroni adjustment) between periods, fixing gender, showed that the only significant difference is an increase of F_0, observed between 1995-96 and 2015-16 for male speakers ($\chi^2 = 8.44$, $p = 0.044$). This limited importance of the interaction (the observed differences are smaller than one semitone), as well as the fact that the changes across periods do not show a clear trend but rather differences from one period to the other, not always in the same direction, may lead to conclude these differences may be linked to others, uncontrolled, factors (e.g., different characteristics of the shows, like the position of the microphones, the

ambient noise, etc.), than a result that can be interpreted as a gender-specific trend over time. We can nonetheless observe that the cloud of points corresponding in Figure 5 to females from the years 1955-1956 seems to be somehow higher than the other clouds, without this leading to a significant difference in their mean.

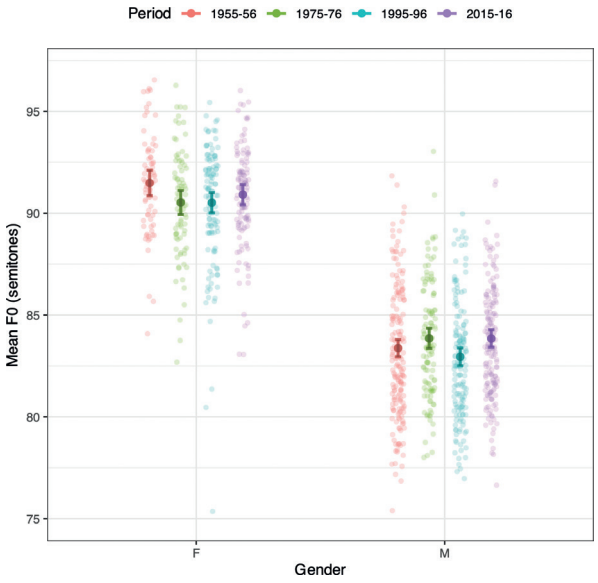
Figure 4 - *Mean F0 (with confidence intervals) estimated by the model based on speaker Age (x-axis), and Gender (colors and separate lines); the mean values estimated for each speaker are plotted as lightly colored points in the background to give a better idea of the observed variation*



3.4 Variations of Formants with Age, Gender, and Period

None of the three models based on formants could be simplified: it shows there are some complex interactions between the four fixed factors that will be difficult to fully understand, notably because they relate to age and phone. We are not aware of theoretical reasons for such differences (Hermes et al., 2023), apart from potential changes in the phonetic implementation of the vowel inventory (Cecelewski et al., 2023) – but such changes are not sufficiently described to provide stable hypotheses. In the remainder of the description, the Age factor will thus be overlooked to simplify the interpretation and understanding of the other factors (it was set to 40 years).

Figure 5 - *Mean F0 (with confidence intervals) estimated by the model based on speaker Gender (x-axis), and time Period (colored points for each gender); the mean values estimated for each speaker are plotted as lightly colored points in the background, to give a better idea of the observed variation*



The shapes of the vocalic spaces (not shown) show some differences across Gender and Periods. The larger effect (after the obvious effect of vowels, which fit the expected distribution of oral vowels in French across the (F_1, F_2) plane (e.g., Apostol et al., 2004) is related to Gender, with higher formant values for female speakers, as expected. For the interaction of the factor Period with the other factors, the vowels tend to be shifted toward lower F_1 values with more recent time Periods. This change is shown in Figure 6: it is possible to observe higher values for this formant during the 1955-56 Period compared to the latter ones, especially for open vowels /a/ and /ɑ/. This is less marked for closed vowels, especially for males. Higher formants for the same vowel are typically observed with a larger mandible opening (Erickson, 2002), which may be linked to higher vocal effort (Rilliard et al., 2018).

Variations of the F_2 across time Periods are plotted in Figure 7. The effect of Period is restricted to a few vowels, /ɑ/ and /ø/. For /ɑ/, we observe a rising of F_2 and thus an anteriorization tendency with time corresponding to the loss of the phonemic distinction between /ɑ/ and /a/ during the second half of the twentieth century (Cecelewski et al., 2023). The effect on /ø/ goes in the other direction, with a lowering of F_2. It is unclear which articulatory change may produce this difference if a centralization of the articulation is possible; we are unaware of a description of such a change in the literature. These two changes in the second formant do not seem gender specific, as they are observed for both genders, with comparable time pace and amplitude.

Figure 6 - *Mean F_1 (and confidence intervals) predicted by the model based on speaker Age (values of 40 y. o.) for both Genders (two plots) and the four time Periods (colors); the means estimated for each speaker and each vowel are plotted as lightly colored points, to give a better idea of the observed variation*

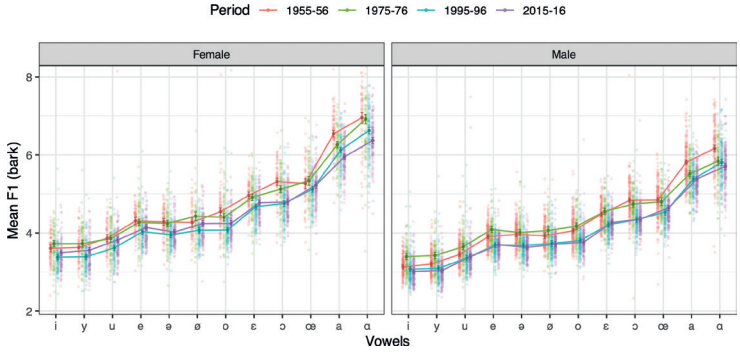


Figure 7 - *Mean F_2 (and confidence intervals) predicted by the model based on speaker Age (values of 40 y. o.) for both Genders (two plots) and the four time Periods (colors); the means estimated for each speaker and each vowel are plotted as lightly colored points, to give a better idea of the observed variation*

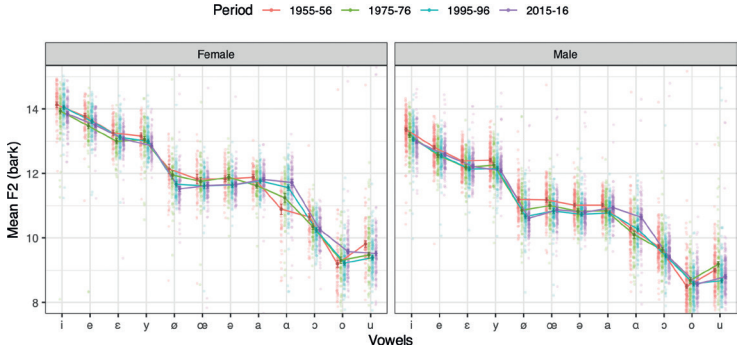
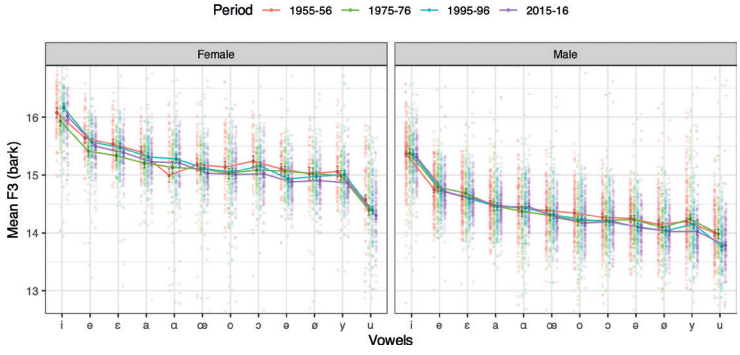


Figure 8 - *Mean F_3 (and confidence intervals) predicted by the model based on speaker Age (values of 40 y. o.) for both Genders (two plots) and the four time Periods (colors); the means estimated for each speaker and each vowel are plotted as lightly colored points, to give a better idea of the observed variation*



The changes in F_3 are presented in Figure 8. There are few interpretable tendencies for this formant, as the main change related to Period applies to the /ɑ/ produced by females in 1955-56. Still, it is difficult to interpret this change as something other than a bias potentially linked to formant estimation.

4. *Discussion & Conclusions*

This work presented the use of an extensive (111 hours) corpus of spontaneous speech (Uro et al., 2022) produced mainly in studio conditions in French national media archives, spanning 60 years in four periods. It is somehow balanced for gender and age – with limits related to strong gender and age bias in the representation of speakers featured in audiovisual media. An automatic transcription was made and evaluated. The evaluation showed a robust quality of the Whisper ASR system for French and media speech (targeting here interventions during talk show interviews), with a slightly higher word error rate than the one presented in their evaluation (Radford et al., 2023) on another speech genre. It also shows Whisper tends to remove the hesitations and repetitions typical of spontaneous speech productions: this shall not be a problem for our goal, but it should be carefully considered for studies interested in such phenomena.

We have presented an analysis of some measurements related to the perceived pitch of a voice: its fundamental frequency and formants, measured on oral vowels only. A preceding work (Rilliard et al., 2023) on the same corpus extracted similar measures but without phonemic segmentation but reached a comparable conclusion if limited by the absence of information on phones. The formants were used to estimate the vocal tract length that is related to its resonances. These source and resonance characteristics are cited in the literature as important correlates of perceived pitch (Ohala, 1983, 1994) and constitute major acoustic cues for the perceptual attribution of a gender to a voice (Guzman et al., 2014; Leung et al., 2018). In relation to the display of gender and a possible secular trend supporting an evolution tendency in France – and especially, in our case, through the national media, our data analysis does not support the view of a change (toward a lowering) in voice pitch specific to female voices, as in Hollien et al. (1997) for another linguistic and cultural context. let's recall these authors found an effect of style on pitch, but not a diachronic tendency. This is significant in our case, and we would like to restate that the corpus was based on talk shows featuring interviews – thus our results shall be restrained to this kind of material.

For the VTL, the model captured an increase in its estimated length across Periods, which corresponds to a lower pitch – but this lowering is observed for both genders; this impedes an interpretation as a gender-specific change. For F_0, the changes captured by the models in relation to Periods are minimal and do not follow a time trend but rather tend to vary from one period to the other in different directions, with amplitude below the threshold of statistical significance in most cases (but in one case for male speakers). Finally, for formants, the main

finding related to Period is linked to F_1, which has higher values for open vowels in 1955-56 compared to more recent Periods – and this observation is once again not gender specific if it may contribute to a lower pitch. Other findings are more related to changes in the vocalic system of French (loss of the /ɑ a/ opposition) than to changes in the voice pitch.

More robust variations in the acoustic parameters were nonetheless observed, which cues to a lowering of pitch. The most striking one is the change in F_0 with the speaker's age, with a descending trend for females and a rising trend for males. This leads to a remarkable reduction of about 3 st in the mean difference between male and female mean F_0 values from 20 to 80 years of age; about one-third of the difference between young adults is lost when comparing the older part of the population. These gender-specific changes correspond to several descriptions in the literature, either on acoustic measurements (Gisladottir et al., 2023) or based on physiological considerations (Sataloff et al., 2017). They are, however, not observed systematically (Berg et al., 2017). It is known that the voice production task influences F_0 changes (Hollien et al., 1997) thus, the controlled production task proposed by Berg et al. (2017) may explain some of the observed discrepancies. Let's recall our dataset proposes spontaneous productions during broadcasted talk shows that are clearly a different style of voice than the excerpts proposed in the literature just cited hereabove (which, with the exception of Hollien et al. 1997, do not include a diachronic dimension). A more comparable dataset was described by Suire & Barkat-Defradas (2020) and is based on vox pop interviews in French media archives. Unfortunately, this corpus was not controlled for age (for obvious reasons) thus, it cannot be used as a comparable reference here – if this study also does not observe trends of pitch lowering for female voices. In our study, a trend of VTL lengthening was also observed with age, which was gender-independent but supported the same interpretation as for F_0 of lowered pitch voices for older females. This trend does not depend on Periods.

The dataset presented here does not support the claim that the pitch of female voices has lowered since the 1950s, independently of other controlled factors – and typically their age (a conclusion similar to Hollien et al., 1997). In the same age range, females participating in talk shows on national French media tend to have comparable pitches in 1955 and in 2015. Let's also note that the mean pitch of voices within a given gender category varies by about one octave across individuals (see Figure 5). This must be highlighted, as the overall means (about 210Hz for females and 130 Hz for males in this dataset) are not necessarily representative of what female or male voices necessarily are: the mean observed for one speaker varies between about 130 and 260Hz for females, and 90 and 200Hz for males. There are thus no clear-cut F_0 differences that separate both gender voices. There are, nonetheless, notable differences in the distribution of individuals participating in television and radio shows in these two periods. As shown in Figure 1, based on a selection of 6,000 individuals from INA databases, there were more females between 20 and 35 years old and fewer females between 51 and 65 years old in the

1955-56 period than in 2015-16 (this also applies to males). This may be explained by the aging of the population in France during this period (INSEE, 2023). In the more recent Period (2015-16), there are more chances to observe a female speaker in the media (the proportion of females in the shows has increased since the 1950s; see Doukhan et al., 2018), but the chances of this individual being above 50 years old also increased. Not controlling for the population's age and sampling speakers randomly from TV and radio shows would thus lead to an age difference for the observed population. It is possible that auditors in the 2010s were exposed to female voices that were, on average, lower than the voices presented in a random sample in the 1955-56 period just because these 2010s females are, on average, older than in the 1950s. This may induce an erroneous (on the basis of the dataset presented here) conclusion about lower female voices in recent periods.

An important feature of this work is that it also looks at male speakers (unlike Pemberton et al., 1998), and this allowed us to describe trends across periods that are not gender specific but may be interpreted as features correlated with a lowering of voices across this time period. This is the case of VTL or of the first formant for open vowels, with lower resonances in more recent periods. To explain such a tendency, the changes in the studio settings across these periods may give an important clue, with the microphones' positions being more distant from the speaker's mouth in older archives, while they are closer in recent shows: this impacts the speaking style, and typically the vocal effort, with more projected voices in older archives (Boula de Mareüil et al., 2012). Producing higher vocal effort is known to increase the voice's F_0 (Alku et al., 2024; Berg et al., 2017; Titze & Sundberg, 1992), and it impacts the first formant, typically for the open vowels (Rilliard et al., 2018). Thus, changes in voice pitch observed in the media that are not gender- or age-dependent may be explained by situational variables and changes in sound recording practices.

Has the pitch of female voices changed with time? This dataset does not support this claim; meanwhile, this dataset certainly presents a series of possible biases that limit the reach of the conclusions one may draw from it. For example, it is a speech genre (speech presented in talk shows from the French audiovisual media) that features mostly individuals from high socioeconomic and prestige status (for example, almost all the speakers have their Wikipedia page). Thus, the dataset does not represent the general French population (a difference with studies like Berg et al., 2017; Gisladottir et al., 2023). The dataset is more adequate to describe the type of voices that are presented, with prescriptive characteristics, to the French population through the national audiovisual media. More specifically, the absence of a global trend does not mean some individuals cannot have lowered their voices: reports exist that some female personalities have decided or felt compelled to lower their voices (Coulomb-Gully, 2022). Typically, they are females with social positions of power (in the political sector or in the economy). It shall be possible to label the speakers of the current dataset in terms of professional occupation and to check if this allows observing a pitch lowering for a subpart of the (female) population;

this is an ongoing work that shall be presented soon. A problem with this approach is that females with political or economic responsibilities will constitute a very reduced or even absent population group in the 1950s.

The proposed dataset is also worth other types of analyses than trying to find a secular trend in voice pitch. An obvious body of research is linked with modeling potential articulatory changes in a diachronic perspective, as the convergence of the two /a/ of French (Cecelewski et al., 2023).

Acknowledgements

The authors wish to thank INA's documentalists, Laetitia Larcher and Anissa-Claire Adgharouamane, for their comprehensive research within the archives and their competent advice for selecting speakers within relevant shows. This work has been partially funded by the French National Research Agency (project Gender Equality Monitor – ANR-19-CE38-0012).

References

Alku, P., Kodali, M., Laaksonen, L. & Kadiri, S.R. (2024). AVID: A speech database for machine learning studies on vocal intensity. In *Speech Communication*, *157*, 103039. https://doi.org/10.1016/j.specom.2024.103039

Apicella, C.L., Feinberg, D.R. & Marlowe, F.W. (2007). Voice pitch predicts reproductive success in male hunter-gatherers. In *Biology Letters*, *3*(6), 682–684. https://doi.org/10.1098/rsbl.2007.0410

Apostol, L., Perrier, P. & Bailly, G. (2004). A model of acoustic interspeaker variability based on the concept of formant–cavity affiliation. In *The Journal of the Acoustical Society of America*, *115*(1), 337–351. https://doi.org/10.1121/1.1631946

Ardaillon, L., Roebel, A. (2019). Fully-Convolutional Network for Pitch Estimation of Speech Signals. In *Proceedings of Interspeech 2019*, Graz, Austria, 15-19 September 2018, 2005–2009.

Barkat-Defradas, M., Raymond, M. & Suire, A. (2021). Vocal Preferences in Humans: A Systematic Review. In Weiss, B., Trouvain, J., Barkat-Defradas, M. & Ohala, J.J. (Eds.), *Voice Attractiveness*. Singapore: Springer, 55–80.

Barras, C., Allauzen, A., Lamel, L. & Gauvain, J.-L. (2002). Transcribing audio-video archives. In *Proceedings of IEEE Int. Conference on Acoustics Speech and Signal Processing*, Orlando, Florida, USA, 13-17 may 2002, I-13-I–16.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. In *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Berg, M., Fuchs, M., Wirkner, K., Loeffler, M., Engel, C. & Berger, T. (2017). The Speaking Voice in the General Population: Normative Data and Associations to Sociodemographic and Lifestyle Factors. In *Journal of Voice*, *31*(2), 257.e13-257.e24.

Boersma, P., Weenink, D. (2022). PRAAT: Doing phonetics by computer [Computer program]. Version 6.2.08. http://www.praat.org/

Boula de Mareüil, P., Rilliard, A. & Allauzen, A. (2012). A Diachronic Study of Initial Stress and other Prosodic Features in the French News Announcer Style: Corpus-based Measurements and Perceptual Experiments. In *Language and Speech*, *55*(2), 263–293.

Cartei, V., Banerjee, R., Hardouin, L. & Reby, D. (2019). The role of sex- related voice variation in children's gender- role stereotype attributions. In *British Journal of Developmental Psychology*, *37*(3), 396–409.

Cartei, V., Reby, D., Garnham, A., Oakhill, J. & Banerjee, R. (2022). Peer audience effects on children's vocal masculinity and femininity. In *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1841), 20200397.

Cecelewski, J., Gendrot, C., Adda-Decker, M. & Boula de Mareüil, P. (2023). A Diachronic Study of Vowel Harmony in French Broadcast Speech since 1940. In Skarnitzl, R., Volín, J. (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic, 7-11 August 2023, 798-802.

Coulomb-Gully, M. (2011). Genre et médias: Vers un état des lieux. In *Sciences de la Société*, *83*, 3–13.

Coulomb-Gully, M. (2022). *Sexisme sur la voix publique: Femmes, éloquence et politique*. Avignon: Éditions de l'Aube.

De Pinto, O., Hollien, H. (1982). Speaking fundamental frequency characteristics of Australian women: Then and now. In *Journal of Phonetics*, *10*(4), 367–375.

Doukhan, D., Devauchelle, S., Girard-Monneron, L., Chávez Ruz, M., Chaddouk, V., Wagner, I. & Rilliard, A. (2023). Voice Passing: A Non-Binary Voice Gender Prediction System for evaluating Transgender voice transition. In *Proceedings of Interspeech 2023*, Dublin, Ireland, 20-24 August 2023, 5207–5211.

Doukhan, D., Poels, G., Rezgui, Z. & Carrive, J. (2018). Describing Gender Equality in French Audiovisual Streams with a Deep Learning Approach. In *VIEW Journal of European Television History and Culture*, *7*(14), 103.

Drager, K.K. (2015). *Linguistic variation, identity construction and cognition*. Berlin: Language Science Press.

Eckert, P. (2008). Variation and the indexical field. In *Journal of Sociolinguistics*, *12*(4), 453–476.

Ellis, S., Goetze, S. & Christensen, H. (2023). Moving Towards Non-Binary Gender Identification Via Analysis of System Errors in Binary Gender Classification. In *Proceedings of IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes, Greece, 4-10 June 2023, 1–5.

Erickson, D. (2002). Articulation of Extreme Formant Patterns for Emphasized Vowels. In *Phonetica*, *59*(2–3), 134–149.

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

Gendrot, C., Adda-Decker, M. (2005). Impact of duration on F1/F2 formant values of oral vowels: An automatic analysis of large broadcast news corpora in French and German. In *Proceedings of Interspeech 2005*, Lisboa, Portugal, 4-8 September 2005, 2453-2456.

Gisladottir, R.S., Helgason, A., Halldorsson, B.V., Helgason, H., Borsky, M., Chien, Y.R., Gudnason, J., Gudjonsson, S.A., Moisik, S., Dediu, D., Thorleifsson, G., Tragante, V., Bustamante, M., Jonsdottir, G.A., Stefansdottir, L., Rutsdottir, G., Magnusson, S.H., Hardarson, M., Ferkingstad, E., Halldorsson, G.H., Rognvaldsson, S., Skuladottir, A., Ivarsdottir, E.V., Norddahl, G., Thorgeirsson, G., Jonsdottir, I., Ulfarsson, M.O., Holm, H., Stefansson, H., Thorsteinsdottir, U., Gudbjartsson, D.F., Sulem, P., Stefansson, K. (2023). Sequence variants affecting voice pitch in humans. In *Science Advances*, *9*(23), eabq2969.

Gussenhoven, C. (2004). *The Phonology of Tone and Intonation* (1st ed.). Cambridge: Cambridge University Press.

Guzman, M., Muñoz, D., Vivero, M., Marín, N., Ramírez, M., Rivera, M.T., Vidal, C., Gerhard, J. & González, C. (2014). Acoustic markers to differentiate gender in prepubescent children's speaking and singing voice. In *International Journal of Pediatric Otorhinolaryngology*, *78*(10), 1592–1598.

Hermes, A., Audibert, N. & Bourbon, A. (2023). Age-related vowel variation in French. In Skarnitzl, R., Volín, J. (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic, 7-11 august 2023, 2045–2049.

Hollien, H., Green, R. & Massey, K. (1994). Longitudinal research on adolescent voice change in males. In *The Journal of the Acoustical Society of America*, *96*(5), 2646–2654.

Hollien, H., Hollien, P.A. & De Jong, G. (1997). Effects of three parameters on speaking fundamental frequency. In *The Journal of the Acoustical Society of America*, *102*(5), 2984–2992.

Holmes, J. (1998). Signalling gender identity through speech. In *Moderna Språk*, *92*(2), 122–128.

INSEE. (2023, July 3). *Pyramide des âges par état matrimonial 2020 – France métropolitaine Séries de 1901 à 2020*. https://www.insee.fr/fr/statistiques/2418118

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. In *Journal of Phonetics*, *34*(4), 485–499.

Jouvet, D., Laprie, Y. (2017). Performance analysis of several pitch detection algorithms on simulated and real noisy speech data. In *Proceeedings of the 25th European Signal Processing Conference (EUSIPCO)*, Kos, Greece, 28 August – 2 September 2017, 1614–1618.

Krahé, B. & Papakonstantinou, L. (2020). Speaking like a Man: Women's Pitch as a Cue for Gender Stereotyping. In *Sex Roles*, *82*(1–2), 94–101.

Kuznetsova, A., Brockhoff, P.B. & Christensen, R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. In *Journal of Statistical Software*, *82*(13), 1–26.

Lammert, A.C., Narayanan, S.S. (2015). On Short-Time Estimation of Vocal Tract Length from Formant Frequencies. In *PLOS ONE*, *10*(7), e0132193.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

Le Fèvre-Berthelot, A. (2015). Doublage au féminin: Transposer le genre. In *Genre En Séries*, *2*, 50–72.

Leung, Y., Oates, J. & Chan, S.P. (2018). Voice, Articulation, and Prosody Contribute to Listener Perceptions of Speaker Gender: A Systematic Review and Meta-Analysis. In *Journal of Speech, Language, and Hearing Research*, *61*(2), 266–297.

Leung, Y., Oates, J., Papp, V. & Chan, S.-P. (2022). Speaking Fundamental Frequencies of Adult Speakers of Australian English and Effects of Sex, Age, and Geographical Location. In *Journal of Voice*, *36*(3), 434.e1-434.e15.

Markova, D., Richer, L., Pangelinan, M., Schwartz, D.H., Leonard, G., Perron, M., Pike, G.B., Veillette, S., Chakravarty, M.M., Pausova, Z. & Paus, T. (2016). Age- and sex-related variations in vocal-tract morphology and voice acoustics during adolescence. In *Hormones and Behavior*, *81*, 84–96.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proceedings of Interspeech 2017*, Stockholm, Sweden, 20-24 August 2017, 498–502.

McAuliffe, M., Sonderegger, M. (2022a). French language model v2.0.0a. [Computer program] https://mfa-models.readthedocs.io/en/latest/

McAuliffe, M., Sonderegger, M. (2022b). French MFA acoustic model v2.0.0a. [Computer program] https://mfa-models.readthedocs.io/en/latest/

McAuliffe, M., Sonderegger, M. (2022c). French MFA dictionary v2.0.0a. [Computer program] https://mfa-models.readthedocs.io/en/latest/

McAuliffe, M., Sonderegger, M. (2022d). French MFA G2P model v2.0.0a. [Computer program] https://mfa-models.readthedocs.io/en/latest/

Ménard, L., Dupont, S., Baum, S.R. & Aubin, J. (2009). Production and perception of French vowels by congenitally blind adults and sighted adults. In *The Journal of the Acoustical Society of America*, *126*(3), 1406–1414.

Merritt, B., Bent, T. (2022). Revisiting the acoustics of speaker gender perception: A gender expansive perspective. In *The Journal of the Acoustical Society of America*, *151*(1), 484–499.

Morton, E.S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. In *The American Naturalist*, *111*(981), 855–869.

Nebelsztein, M. (2018, August 26). *Bonne nouvelle: La voix des femmes est plus grave qu'avant*. terra femina. https://www.terrafemina.com/article/pourquoi-la-voix-des-femmes-est-plus-grave-qu-avant_a343484/1

Ohala, J.J. (1983). Cross-Language Use of Pitch: An Ethological View. In *Phonetica*, *40*(1), 1–18.

Ohala, J.J. (1994). The frequency code underlies the sound-symbolic use of voice pitch. In Hinton, L., Nichols, J. & Ohala, J.J. (Eds.), *Sound Symbolism* (1st ed.). Cambridge: Cambridge University Press, 325–347.

Ohara, Y. (2001). Finding one's voice in Japanese: A study of the pitch levels of L2 users. In Pavlenko, A., Blackledge, A., Piller, I. & Teutsch-Dwyer, M. (Eds.), *Multilingualism, Second Language Learning, and Gender*. Berlin: De Gruyter Mouton, 231-256.

Pemberton, C., McCormack, P. & Russell, A. (1998). Have women's voices lowered across time? A cross sectional study of Australian women's voices. In *Journal of Voice*, *12*(2), 208–213.

Pépiot, E., Arnold, A. (2021). Cross-Gender Differences in English/French Bilingual Speakers: A Multiparametric Study. In *Perceptual and Motor Skills*, *128*(1), 153–177.

Podesva, R.J. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona. In *Journal of Sociolinguistics*, *11*(4), 478–504.

Podesva, R.J. (2011). The California Vowel Shift and Gay Identity. In *American Speech*, *86*(1), 32–51.

Puts, D.A., Gaulin, S.J.C. & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. In *Evolution and Human Behavior*, *27*(4), 283–296.

Quené, H., Boomsma, G. & Van Erning, R. (2021). Attractiveness of Male Speakers: Effects of Pitch and Tempo. In Weiss, B., Trouvain, J., Barkat-Defradas, M. & Ohala, J.J. (Eds.), *Voice Attractiveness*. Singapore: Springer, 153–164.

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Radford, A., Kim, J.W., Xu, T., Brockman, G., Mcleavey, C. & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202), Hawaii, USA, 23-29 July 2023. PMLR, 28492–28518.

Rebollo Couto, L., Lopes, C.R. (Eds.). (2011). *As formas de tratamento em português e em espanhol: Variação, mudança e funções conversacionais = Las formas de tratamiento en español y en portugués: variación, cambio y funciones conversacionales*. Niteroi: Editora da UFF.

Richoz, A.-R., Quinn, P.C., Hillairet De Boisferon, A., Berger, C., Loevenbruck, H., Lewkowicz, D.J., Lee, K., Dole, M., Caldara, R. & Pascalis, O. (2017). Audio-Visual Perception of Gender by Infants Emerges Earlier for Adult-Directed Speech. In *PLOS ONE*, *12*(1), e0169325.

Rilliard, A., d'Alessandro, C. & Evrard, M. (2018). Paradigmatic variation of vowels in expressive speech: Acoustic description and dimensional analysis. In *The Journal of the Acoustical Society of America*, *143*(1), 109–122.

Rilliard, A., Doukhan, D., Uro, R. & Devauchelle, S. (2023). Evolution of voices in French audiovisual media across genders and age in a diachronic perspective. In R. Skarnitzl, & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic, 7-11 August 2023, 753–757.

Riverin-Coutlée, J., Harrington, J. (2022). Phonetic change over the career: A case study. In *Linguistics Vanguard*, *8*(1), 41–52.

Robson, D. (2018). *The reasons why women's voices are deeper today*. BBC. https://www.bbc.com/worklife/article/20180612-the-reasons-why-womens-voices-are-deeper-today

Sadanobu, T. (2015). "Characters" in Japanese Communication and Language: An Overview. In *Acta Linguistica Asiatica*, *5*(2), 9–28.

Sataloff, R.T. (2017). *Clinical assessment of voice* (Second edition). San Diego: Plural Publishing, Inc.

Sataloff, R.T., Kost, K.M. & Linville, S.E. (2017). Chapter 13. The Effects of Age on the Voice. In Sataloff, R.T. (Ed.), *Clinical assessment of voice* (Second edition). San Diego: Plural Publishing, Inc, 221–240.

Simpson, A.P. (2009). Phonetic differences between male and female speech. In *Language and Linguistics Compass*, *3*(2), 621–640.

Stuart-Smith, J. (2017). Bridging the gap(s): The role of style in language change linked to the broadcast media. In Thogersen, J., Coupland, N. & Mortensen, J. (Eds.), *Language (De)standardisation in Late Modern Europe: Style and media studies*. Nashville: Novus Press, 51–84.

Suire, A., Barkat-Defradas, M. (2020). Evolution of human pitch: Preliminary analyses in the French population using INA audiovisual archives of Vox Pops. In *Proceedings of 2020 IASA-FIAT/IFTA Joint Conference*, online, Ireland, 26-29 October 2020.

Talkin, D. (2015). REAPER: Robust epoch and pitch estimator. [Computer Program] https://github.com/google/REAPER

Tawilapakul, U. (2022). Face threatening and speaker presuppositions: The case of feminine polite particles in Thai. In *Journal of Pragmatics*, *195*, 69–90.

Taylor, V. (2018). *Women's Voices Are Getting Lower Over Time, Thanks To Sexism*. Romper. https://www.romper.com/p/womens-voices-are-getting-lower-over-time-thanks-to-sexism-study-says-9406100

Titze, I.R. (1989). Physiologic and acoustic differences between male and female voices. In *The Journal of the Acoustical Society of America*, *85*(4), 1699–1707.

Titze, I.R., Baken, R.J., Bozeman, K.W., Granqvist, S., Henrich, N., Herbst, C.T., Howard, D.M., Hunter, E.J., Kaelin, D., Kent, R.D., Kreiman, J., Kob, M., Löfqvist, A., McCoy, S., Miller, D.G., Noé, H., Scherer, R.C., Smith, J.R., Story, B.H., Švec, J.G., Ternström, S., Wolfe, J. (2015). Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. In *Journal Acoust. Soc. America*, *137*(5), 3005–3007.

Titze, I.R., Sundberg, J. (1992). Vocal intensity in speakers and singers. In *The Journal of the Acoustical Society of America*, *91*(5), 2936–2946.

Uro, R., Doukhan, D., Rilliard, A., Larcher, L., Adgharouamane, A.-C., Tahon, M. & Laurent, A. (2022). A Semi-Automatic Approach to Create Large Gender- and Age-Balanced Speaker Corpora: Usefulness of Speaker Diarization & Identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France, 20-25 June 2022. European Language Resources Association, 3271–3280.

van Bezooijen, R. (1995). Sociocultural Aspects of Pitch Differences between Japanese and Dutch Women. In *Language and Speech*, *38*(3), 253–265.

Vaysse, R., Astésano, C. & Farinas, J. (2022). Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech. In *The Journal of the Acoustical Society of America*, *152*(5), 3091–3101.

Vorperian, H.K., Kent, R.D., Lee, Y. & Bolt, D.M. (2019). Corner vowels in males and females ages 4 to 20 years: Fundamental and F1–F4 formant frequencies. In *The Journal of the Acoustical Society of America*, *146*(5), 3255–3274.

Vorperian, H.K., Wang, S., Chung, M.K., Schimek, E.M., Durtschi, R.B., Kent, R.D., Ziegert, A.J. & Gentry, L.R. (2009). Anatomic development of the oral and phar-

yngeal portions of the vocal tract: An imaging study. In *The Journal of the Acoustical Society of America*, *125*(3), 1666–1678.

Vorperian, H.K., Wang, S., Schimek, E.M., Durtschi, R.B., Kent, R.D., Gentry, L.R. & Chung, M.K. (2011). Developmental Sexual Dimorphism of the Oral and Pharyngeal Portions of the Vocal Tract: An Imaging Study. In *Journal of Speech, Language, and Hearing Research*, *54*(4), 995–1010.

Weirich, M., Simpson, A.P. (2018). Gender identity is indexed and perceived in speech. In *PLOS ONE*, *13*(12), e0209226.

Yamauchi, A., Yokonishi, H., Imagawa, H., Sakakibara, K.-I., Nito, T., Tayama, N. & Yamasoba, T. (2015). Quantitative Analysis of Digital Videokymography: A Preliminary Study on Age- and Gender-Related Difference of Vocal Fold Vibration in Normal Speakers. In *Journal of Voice*, *29*(1), 109–119.

DOI: 10.17469/O2112AISV000005

DUCCIO PICCARDI, SILVIA CALAMAI

Fear of FAIR? Towards a new Italian incentive to oral data curation[1]

This paper deals with the climate of distrust permeating contemporary society and how it may affect the researchers' adherence to initiatives dealing with the promotion of data FAIRness. A case in point is the perplexing state of a recent Italian action, a new census of Tuscan oral archives (*Gra.fo Reloaded* project) in which the respondents were asked to disclose several details about the oral documents they gathered. As a potential alternative strategy to get the ball of data curation rolling, this paper explores the current utilitarian mindset adopted by major international linguistics communities and details its application to the conception of a novel Italian academic outlet: *Oral Archives Journal* (OAr), a journal on oral archive conservation, description, and use, as well as on speech transcription, technologies, legal issues, and research ethics.

Keywords: Oral archives, data journal, trust research, data FAIRness, Tuscan census.

1. *"To be fair, I don't trust them". Research data in a cynical age*

As the basis of cooperative endeavors, mutual trust is a pivotal component of any kind of society. Alarmingly, several reports underline a global trust shortage, painting the picture of a contemporary distrusting world (Rotenberg, 2018). Even though the multifaceted nature of trust and its several measurement strategies (Bauer, Freitag, 2018) might cast a shadow of doubt on the generalizability of this crisis scenario (Brehm, Savel, 2019), its persistence in discourse indexes a great deal of concern and a constant fostering of ideas regarding institutional corruption and untrustworthiness (Rotenberg, 2018). According to Hardin (2006), trust has gradually gained relevance as a consequence of major societal trends, such as the growing prevalence of loose-knit networks and political hurdles in efficiently dealing with economic and security crises. Technology does its part, as the chaotic nature of information on the web favors a post-truth milieu in which institutional and expert claims are often met with skepticism (e.g., Harsin, 2018). Be that as it may, a recent contribution by Citrin and Stoker (2018) dubbed our times "a cynical age", in which the (political) actions of others are assumed to always have self-serving motives. Similar phenomena are described in other domains: for example, Lambin (2008) defined consumer "fear of marketing" as an insidious mistrust stemming

[1] This contribution is the result of the joint work of the two authors. Duccio Piccardi is responsible for writing the text. Silvia Calamai is responsible for supervising the text and the project *Gra.fo Reloaded*.

from the assumption of constantly being the target of manipulative strategies, whose inescapable threads are woven by firm experts.

In this context, the concept of openness emerged in twenty-first century science as an attempt to make research practices transparent and accessible, ultimately revamping trust (Stracke, 2020). Research data management is a core concern of open science. Most notably, the principles enucleated in Wilkinson et al. (2016) gained traction through their widespread adoption by public initiatives: among those, the European Research Infrastructure Consortium CLARIN defines itself as an "avant la lettre" promoter of keeping data Findable, Accessible, Interoperable and Reusable (FAIR; De Jong, Maegaard, Fišer, Van Uytvanck & Witt, 2020: 3406). Of course, requiring drastic changes in researchers' mindsets, open science is not something that can be achieved instantly. Several factors explain noticeable asymmetries in the endorsement of its tenets. Individual disciplinary communities of practice faced the issue of openness in different times and ways: linguistics, for example, is considered a "late-comer to the scientific frontier of replicable research" (Janda, 2021: 447); thus, a recent survey underlined a very slow uptake of open science practices and an overall unsatisfactory current situation (Bochynska, Keeble, Halfacre, Casillas, Champagne, Chen, Röthlisberger, Buchanan & Roettger, 2023).

More intriguingly, research has shown that the propensity to adhere to specific scientific habits, such as data sharing, is related to contextual and individual factors (research group norms, researcher attitudes, perceived barriers, sex, age, academic rank, and geographic location; Bamberger, Reeves, 2021). Gomes, Pottier, Crystal-Ornelas, Hudgins, Foroughirad, Sánchez-Reyes, Turba, Martinez, Moreau, Bertram, Smout & Gaynor (2022) conducted a comprehensive survey of the reported reasons for open science non-compliance, which can be roughly divided in three types (rearranging the original taxonomy):

a. Knowledge barriers with respect to data archiving, rights, and sensitiveness (these can be lowered through the activities of local research communities, as it will be shown below, § 3.1).
b. Lack of incentives, since data archiving takes time and effort, but has no obvious reward (this aspect is being dealt with through recent strategies of academic acknowledgement, as it will be shown below, § 3.2.1).
c. Trust issues. Sharing data makes researchers feel exposed to comments about their potential errors or flaws, which, in turn, may reduce others' trust in them. Moreover, a common concern pertains to original data misinterpretation and inappropriate use. Lastly, researchers tend to avoid the risks of scooping, i.e., being anticipated by others in conducting planned analyses, thus losing publication opportunities. This habit is grounded on false assumptions: in reality, researchers tend to prioritize original data collection over data re-use. For this reason, Dijkers (2019: 175) was led to the blunt statement that "the fear of "research parasites" likely is pathologic".

All in all, open science seems to be an enterprise aiming to turn a vicious circle of distrust in a virtuous circle of trust, but making the first move could be not as straightforward as it may appear.

On a positive note, investigating researchers' behavior as data donors (i.e., survey participants) Fichtner, Horstmeier, Brühmann, Watter, Binder & Knaus (2023) found that the explicit mention of data sharing does not scare them away from completing the requested task. In the remainder of this paper (§ 2), a different picture will be drawn, as a recent CLARIN-IT-powered initiative, i.e., a new Tuscan oral archive census (one of the activities planned in the *Gra.fo Reloaded* project) struggled to take off due to gargantuan nonresponse rates. Of course, context is the key to interpreting these discrepancies. Belliard, Maineri, Plomp, Ramos Padilla, Sun & Jeddi (2023) recently synthetized ten procedural points to develop FAIR discussions in specific environments: noticeably, among those, the authors include identifying the features of the target community and establishing adequate rewards for the adoption of the principles. Here we argue that in Italy, where utilitarianism is a central component of the attitudes expressed towards European institutions (Conti, Memoli, 2015), we could try to overcome trust barriers in open science by developing appealing incentives. For this reason, we will try to advance a local instantiation of the international discussion on making data curation a research product per se through dedicated publication outlets, not without mentioning what the Italian community has already accomplished to lower the knowledge barriers surrounding these issues in its country (§ 3).

2. *The perplexing state of affairs of a new oral archive census*

Gra.fo Reloaded was a project on oral archive preservation and re-use run by the Italian Regione Toscana between June 2022 and the end of November 2023; it was coordinated by Siena University, with the participation of Fondazione Sistema Toscana, Soprintendenza Archivistica e Bibliografica della Toscana, Ecomuseo della Montagna Pistoiese, Istituto di Linguistica Computazionale "Antonio Zampolli" (Pisa National Research Council) and CLARIN-IT. As a first step towards devising effective preservation strategies, the project launched a new census of the oral archives gathered in Tuscany, aiming to both update previous seminal works (e.g., Andreini, Clemente, 2007) and create a research community revolving around novel research tools and initiatives, such as *Archivio Vi.Vo*., a platform under construction for oral archive restoration, description, and access (Calamai, Piccardi, Pretto, Candeo, Stamuli & Monachini, 2022), and the *Vademecum* described in § 3.1, respectively.

After a pre-test phase (Piccardi, Calamai, 2023), the census was published on Google Forms on December 10th, 2022. At the beginning of the census, two layers of text tried to consolidate the participant trust in the initiative, underlining its community-oriented motives (Benson-Greenwald, Trujillo, White & Diekman, 2023), its compliance with CLARIN-IT, and the adherence to the FAIR principles

of the promoted tools. As the census was purely preliminary to any kind of intervention, by no means were participants asked to agree to rights transfer or oral data deposit or disclosure.

In a first phase, through a circular cover letter asking for individual completion and dissemination of the initiative, the census form was sent to 369 email addresses, including individual researchers (selected through the critical analysis of their publication history), Italian scientific associations (linguistics, oral/public history, anthropology, sociology, musicology), Ph.D. coordinators of Tuscan universities, the offices of all Tuscan municipalities, and a small selection of cultural foundations. Moreover, all the involved partners advertised the census through their platforms, and a related article was published in the local spin-off of the daily National newspaper La Repubblica (2023, February 11)[2]. Lastly, the census was promoted during related in-person events.

In a second phase, which began towards the end of March 2023 and was formally halted after the end of the project, the census was divulged to other target individuals through personalized emails, phone calls, or social media accounts. In addition to the aforementioned categories, we also reached for representatives of U.S. Italian cultural foundations and ex-M.o.A./Ph.D. students who appeared (in institutional thesis repositories) to have gathered oral documents in Tuscany, for a total of more than fifty additional contacts.

After four months and shortly after the beginning of the second phase (mid-April 2023), we gathered 14 responses; after almost a year, at the end of the project, this number increased to a total of 23. As a rough approximation, leaving aside opaque point-to-mass diffusion mechanisms and potential technical issues in individual email sending or receiving, the whole census publication period led us to a whopping 95 % nonresponse rate. In oral archive census endeavors, difficulties in response collection are far from being unheard of. In the introductory remarks of the first Italian census of the public oral archives, Mulè (1993: 35-36) noted that only half of the previously pinpointed institutes replied to the questionnaire. A few years later, in a preliminary report on the state of oral archives in Ireland, Ferriter (1998: 91) observed that "many did not reply". Nonetheless, our quantities might signal, so to say, an eloquent silence which should be adequately interpreted (Ephratt, 2008). Thus, we looked at both the objective (i.e., who has actually responded) and subjective (i.e., the perplexities and doubts expressed by cautious respondents through personal communications) elements at our disposal, ending up with a picture resembling the above-discussed three barriers against open science (cf. Gomes et al., 2022):

[2] https://www.dfclam.unisi.it/sites/st08/files/allegatiparagrafo/13-02-2023/2023021153753080-1.pdf. Accessed 08 October 2024.

a. Silence as "I don't know". Respondents expressed sincere unawareness on crucial circumstantial information on the data they possess, such as ownership, copyright, and privacy issues, hindering their willingness to disclose details about their materials[3].
b. Silence as "it's not (academically) worth it". Admittedly, because of its extensiveness, completing our census did require a certain level of dedication and time. Of our 23 responses, only five were sent by full-time active academic personnel, the others being written by students, temporary and independent researchers, retired individuals. This disproportion highlights the unattractive nature of data curation for those individuals who are already busy with many other institutional tasks and responsibilities, calling for effective incentive development actions.
c. Silence as "I don't trust". More than half of the respondents did have previous contacts and/or in-person encounters with one of the two authors of this paper. Perceived social closeness is a pivotal component of cooperativeness and trust (e.g., Brudner, Karousatos, Fareri & Delgado, 2022). However, in order to avoid circularity, an initiative striving for community building should not majorly rely on pre-existing relationships. Moreover, even though, leaving aside an informal requested declaration of interest for the *Archivio Vi.Vo*. archival platform, the census did not imply in any way the drafting of a data deposit agreement, we noticed that our form was often erroneously interpreted as such, triggering defensive behaviors (see above, § 1).

Overall, this perplexing state of affairs points to the necessity of urgently furthering the general awareness on data curation: in order to do so, researchers should be provided with the right knowledge toolkits and incentives.

3. *"To be FAIR, I might study and publish with them". A Vademecum and a journal*

In the previous section, we highlighted the putative role played by three widely discussed barriers slowing down the adoption of an open-science mindset in the abysmal response rate of an Italian Regional census of oral archives. In this section, we will briefly address the crux of the matter: how can researchers bootstrap their knowledge about oral data curation? Why should they spend their time and efforts to organize their data? What are their actual benefits in joining the trusting circle of data FAIRness?

3.1 Knowledge barrier: The *Vademecum per il trattamento delle fonti orali*

The *Vademecum per il trattamento delle fonti orali* (Tavolo permanente per le fonti orali, 2023) is, at present, a five-year collaborative enterprise, aiming at

[3] Indeed, according to the responses given to specific sections of our census, only one participant had some form of legal training before gathering her oral documents. Similarly, only two had some prior archival knowledge, while the majority of the respondents (13) were well aware of the technical aspects of recording (tools, audio formats, etc.). It seems evident that technical training alone cannot make up for the lack of expertise in other pertinent domains.

offering both the Italian scientific communities and non-academic stakeholders a bootstrap guide on the conservation, the description, the use, and the re-use of oral archives (Calamai, Casellato & Stamuli, 2022). Very different subjects from academia and from the main agencies of the Ministry of Cultural Heritage worked together with private and public institutions and CLARIN-IT in order to solve long-standing structural problems related to oral sources. After a collaborative process of co-creation and public revision, a set of good practices were defined in a cross-disciplinary approach with respect to production, management, access, and preservation of audio materials produced not only within research projects but also public history initiatives and citizen science data collection. A first digital version was finalized in 2021 and made available through the websites of all the involved institutions and the official pages of the work group representing the *Vademecum*[4]. In October, 2023, a printed version was released with a substantial new section on speech transcription.

As it was originally encouraged through the round table held at the XV AISV annual conference (Piccardi, Ardolino & Calamai, eds., 2019), the *Vademecum* undertaking has always been deeply connected with open discussion. At the present time, a yearly meeting is organized in order to showcase its latest developments and related projects; moreover, the *Vademecum* has been presented to Italian University students and the general public through dedicated lessons and other types of events. However, having a grasp of its actual reach is quite problematic, since works referencing its use are just beginning to be published (e.g., Manali, 2024; Clemente, 2023 for an example of Master of Arts thesis). Here we argue that probing the public awareness of the *Vademecum* existence through multiple methodologies (e.g., periodic literature reviews, items inserted in pertinent questionnaires, etc.) will be a necessary task in order to trigger word of mouth in specific communities and organize focused dissemination events. The *Gra.fo Reloaded* census did contain a section about the *Vademecum*. Crucially, the vast majority of the respondents (18) had never heard of it before filling in the census form, but was eager to know more through the subscription to the *Vademecum* newsletter (16).

3.2 Incentive barrier: the *Oral Archives Journal* (OAr)

3.2.1 Historical background

Being a freely available enchiridion to oral data curation, the *Vademecum* could be able to cover for the knowledge needs of the Italian community. Consequently, we started to envision a way to give more value to the efforts of local researchers to engage in this kind of activities by looking at relevant discussions emerging in the linguistics milieu. At the beginning of the 2010s, both Australian and American linguistic societies published resolutions (see Thieberger, 2012 for a recollection of these documents) to ensure the recognition of language documentation archives as scholarly contributions. A few years later, Haspelmath, Michaelis

[4] https://sites.google.com/view/tavolopermanenteperlefontioral/chi-siamo. Accessed 08 October 2024.

(2014) envisioned a "corpus journal" hosting peer-reviewed annotated corpora in order to incentivize (quality) data curation through the standard mechanisms of academic career-building. Of course, at the time, the idea of similar outlets was already piloted in other research fields (see Thieberger, Margetts, Morey & Musgrave, 2016). In particular, Lawrence, Jones, Matthews, Pepler & Callaghan (2011) listed a taxonomy of data publication models including overlay data journals, i.e., academic outlets publishing data description documents through a peer-review process encompassing both the document and the actual data curation efforts. Ideally, the journal should have some degree of control over the archival platform linked to the proposals.

Thieberger et al. (2016) made dramatic progress in this line of discussion by systematically addressing the major "stumbling block" (so to quote Peter Austin's comment to Haspelmath, Michaelis, 2014) to the realization of similar academic acknowledgement ideas, i.e., the lack of structured guidelines to review language data. Among the several points of their extensive proposal, Thieberger and colleagues stressed the importance of archiving raw (i.e., recordings) alongside primary (transcriptions) and structural (annotations) data while linking these layers and describing the documents through adequate sets of metadata. Following these works, the Linguistic Society of America (Fitzgerald, 2021) recently inaugurated a new section of its flagship *Language* journal, *Language Revitalization and Documentation*, dedicated to research ethics and methodology and, crucially, to corpus overviews. This effort was followed by a paper by Garrett and Harris (2022) on the assessment of scholarship in language documentation, in which archival collections are considered de facto research outputs. In the same years, but from the viewpoint of another linguistics sub-discipline, i.e., phonetics, the position paper by Garellek, Gordon, Kirby, Lee, Michaud, Mooshammer, Niebuhr, Recasens, Roettger, Simpson & Yu (2020) raised some perplexities on this emerging "dual reward system". The authors argue that the separation of the scholarly acknowledgement of data curation from the value of research publications making use of the curated data risks engendering new weighting asymmetries in assessment systems, with archival publications being inevitably valued less and thus implicitly discouraged. Moreover, given the different research traditions of language documentation and phonetics, the authors fear that "giving additional credit to researchers for the archiving of phonetic data sets would feel like mixing apples with oranges" (ibid.: 11). On the contrary, in their view, data archiving should be recognized as a prerequisite for the submission of any phonetics research article to any journal.

While Garellek et al.'s (2020) point on valuing discrepancies does imply a constant monitoring of researcher (and institution) mindsets in assessing data contributions, their concern on diverging academic traditions does not match the current position held by the Italian research initiative behind the *Vademecum* (i.e., the *Tavolo permanente per le fonti orali*), which tries to raise awareness on crucial issues concerning human voice data through a work rooted in interdisciplinarity.

3.2.2 OAr and its structure

Against this background, the above-synthetized academic discussion on data journals was first brought to the attention of the *Vademecum* work group on data conservation in an in-person meeting held in Padua on July 10th, 2019. On that occasion, the idea of an Italian instantiation of this trend was deemed premature. Now the time seems ripe to work towards a new Italian journal open to oral archive data papers. Indeed, the state-of-the-art on data journal made further progress through the intervention of the Linguistic Society of America. Meanwhile, in Italy, the *Vademecum* has been made available to the community, providing valuable guidelines for oral data curation. Moreover, the *Archivio Vi.Vo.* platform at CLARIN-IT is entering a new phase of development which, through multiple tangential projects, will lead to its public release[5]. By offering a validated framework for oral document archiving, *Archivio Vi.Vo*. could help both contributors with their submissions and reviewers with their evaluations.

The University of Siena has recently sealed a deal with a National academic publishing house to host the annual issues of *Oral Archives Journal* (OAr). OAr will be structured in two main sections, for a total of eight subsections. Inspired by the original work groups of the *Vademecum*, the first main section is named *The life cycle of the archives* and welcomes contributions on

a. Oral data production, encompassing topics such as conducting fieldwork, constructing an oral archive, and applying existing metadata schemes or developing new ones;
b. Oral data curation, that is, data papers on oral archives deposited in reviewable repositories;
c. Oral data conservation, dealing with strategies for the restoration of physical carriers and the speech signal, pipelines for digitization and data migration, and storage sustainability;
d. Oral data use and re-use, hosting both original research papers and recounts of specific initiatives (e.g., exhibitions, touristic sound walks, etc.) making use of oral archives.

The second main section is focused on *Cross-cutting dimensions*, i.e., topics that are relevant to more than just one life phase of an oral archive:

a. Speech transcription, from both theoretical and operative perspectives;
b. Speech technology, including presentations of new tools for speech recording, processing, and archiving, as well as benchmarks and usability tests;

[5] Currently, *Archivio Vi.Vo*. is in closed beta. Up to now, it has been used to process the archive of the Tuscan folk singer Caterina Bueno (Calamai et al., 2022) and the Anna Buonomini archive, mostly consisting of ethnological interviews conducted in the mountains of the *Provincia di Pistoia* (this was done during the *Gra.fo Reloaded* project; see Pozzebon, Biliotti & Calamai, 2016 for a brief, albeit slightly outdated, description of this archive). See Luzietti (2023) for a first evaluation of *Archivio Vi.Vo*. in its current state, which was performed through a processing test of the historian Angela Spinelli's oral archive.

c. Legal issues, from consent procedures, privacy, data sensitivity, licensing, etc., up to copyright issues;
d. Research ethics, fostering good practices for forging virtuous relationships between the researchers and the communities, including strategies for results dissemination and restitutionary work to the communities.

Given its peculiar relevance to the academic discussion leading to OAr (§ 3.2.1) and, more in general, to the ecosystem of oral archives, Oral data curation deserves additional explanation. Specifically, proposals submitted to this section will be evaluated following an adaptation of the rubric of Fitzgerald (2021: 5, based on Thieberger et al., 2016), reproduced below:

a. Accessibility (deposited in a repository committed to providing long-term curation and access, including a persistent identifier and a citation form for items within the deposit; has a landing page or file with a basic description; includes access to metadata and a clear path to accessing the data in the corpus; files are in formats that are nonproprietary; levels of accessibility are properly justified from a legal standpoint).
b. Quality (the nature and amount of contextual and background information; the structure of the deposit; metadata quality; the nature of linguistic annotation of the data; structural linking between raw data and their annotations, such as time-aligned transcriptions).
c. Quantity (content; amount of data).

Compared to Firzgerald's (2021) version, we revised the last point of the Accessibility section in order to stress that OAr does not endorse open access solutions which are not backed by a thorough legal feasibility study and, at the same time, does not necessarily reject archives with restricted access if this decision is properly motivated in their specific contexts of production. Moreover, OAr does not provide clear-cut interpretative guidelines to the Quantity section: the contribution of content and amount of data to the significance of the presented archive should be evaluated on a case-by-case basis.

3.2.3 Some alternatives and the specificity of OAr

The attentive reader may argue that, in addition to the above-mentioned Australian and American traditions, other lines of research in linguistics came to the establishment of renowned journals accepting descriptions of oral (but not only) data collections (linked to the actual data) as paper proposals. A notable example of this is *Language Resources and Evaluation* (Ide, Calzolari, 2005), which, under the auspices of the European Language Resources Association, is concerned with "language data and descriptions in machine readable form used to assist and augment language processing applications, such as written or spoken corpora and lexica, multimodal resources, [...], multimedia databases, etc. [...]"[6].

Other outlets with broader scopes but interested in linguistics contributions are, for example, the *Journal of Open Humanities Data*, which welcomes short data

[6] https://link.springer.com/journal/10579. Accessed 08 October 2024.

papers on "research object[s] with high reuse potential", while stating that this type of work "does not replace a traditional research article, but rather complements it"[7]; and the *Research Data Journal for the Humanities and Social Sciences*, which is currently funded by the Consortium of European Social Science Data Archives and "aims to contribute to the transparency of research, accelerate dissemination and foster the reuse of scholarly data"[8].

Descriptions of oral (or audiovisual) data can be read in journals conceived to appeal also to researchers from disciplines well beyond the humanities. *Data in Brief* hosts "short, digestible data articles", "contributes to open science", "improves reproducibility"[9], and also stresses that "without a thorough description, any initial value to the data is lost" (Shaklee, 2014: 5). While phonetics was not originally mentioned among the fields exemplifying its "broad spectrum of disciplines" (that is, "including Biology, Chemistry, Economics, Psychology and Physics"; Wang, 2014: 85), articles describing oral data of related interest are evidently considered for publication (e.g., Di Benedetto, De Nardis, 2022). The similarly named journal *Data* have a "Data descriptors" section which serves analogous purposes; it should be noted that its "Data in application" subject area includes research and experimental data along with data in healthcare, finance, etc[10]. Accounts of speech recorded for disparate ends and pertaining to heterogeneous disciplines can be found, e.g., in Ezzes, Schneck, Casilio, Fromm, Mefferd, de Riesthal & Wilson (2022) and Doyle, Şerban (2024).

This non-exhaustive list highlights a scientific trend kin to the intents of OAr while, on the other hand, reinforcing its specificity. In particular, OAr is centered around the orality of the discussed research data. This implies that the journal is not specifically tied to any individual discipline, and strives to be of interest to all the researchers involved with human voice (e.g., anthropologists, archivists, ethnomusicologists, jurists, linguists, oral historians, psychologists, sociologists, etc.). At the same time, this general theme excludes any description of data which is not oral (e.g., written or visual). OAr follows a line of archival discussion trying to go beyond the idea of oral data as cumbersome substitutes of transcriptions or limited to means of interpretation of written data (Valentini, Piccardi, Calamai & Stamuli, 2023). OAr not only puts oral data in the forefront, but also values them as documents before resources, i.e., the journal's take on the documentary value of oral data does not necessarily end with any anticipation of their possible utilization in, e.g., technological or scientific endeavors[11]. In accordance with the *Vademecum*, OAr primarily seeks to endorse an enhancement of the informative value of oral data.

[7] https://openhumanitiesdata.metajnl.com/about. Accessed 08 October 2024.
[8] https://brill.com/view/journals/rdj/rdj-overview.xml. Accessed 08 October 2024.
[9] https://www.sciencedirect.com/journal/data-in-brief/about/aims-and-scope. Accessed 08 October 2024.
[10] https://www.mdpi.com/journal/data/about. Accessed 08 October 2024.
[11] Note that this is not an understatement of Reusability as a core component of the FAIR principles: from OAr's perspective, oral data have to be set up for their reuse (through, e.g., adequate metadata) as a prerequisite of a deserving archival process.

This implies that the archives submitted to OAr should present a well-thought-out conceptualization of the communicative events, exhaustive descriptions, and well-constructed networks of relationships between the oral documents and contextual information (such as related documents of other kinds – for example, a concert flyer or a picture of an interviewee – if any). In the case of documents from the pre-digital era, OAr favors effective linkages between the physical carriers of the data, their digital versions, and the individual communicative events. Lastly, OAr considers archive data papers as scholarly contributions on par with other types of articles (see § 3.2.1).

4. *Closing remarks. Now it really comes down to trust*

A dedicated journal represents a missing piece of the emerging picture of Italian tools promoting a paradigm shift towards data curation. Hopefully, knowledge and incentives will lower the perceived minimum requirements to set in motion the trusting virtuous circle of open science. Nonetheless, countering a major societal trend as the one described in § 1 may require other widespread actions. For this reason, we do hope that an impactful discussion on oral data management will germinate in Italian universities. Through the development of, e.g., pertinent course modules and handbooks, the students of today should be able to grow in a (research) world without any fear of FAIRness.

Acknowledgements

The *Gra.fo Reloaded* project was funded by Regione Toscana (Fondo per lo Sviluppo e la Coesione – FCS 2014-2020, Giovanisì), Siena University, Fondazione Sistema Toscana, and Ecomuseo della Montagna Pistoiese. Soprintendenza Archivistica e Bibliografica della Toscana, Istituto di Linguistica Computazionale "Antonio Zampolli" (Pisa National Research Council) and CLARIN-IT also contributed to the project. The authors would like to thank all who have taken the Tuscan census into consideration and Maria Francesca Stamuli (Soprintendenza Archivistica e Bibliografica della Toscana) for having brought Mulè's (1993) passage on nonresponse to their attention.

References

Andreini, A., Clemente, P. (Eds.) (2007). *I custodi delle voci. Archivi orali in Toscana: primo censimento*. Florence: Regione Toscana.

Bamberger, M.R., Reeves, T.D. (2021). Individual and contextual factors associated with data sharing in the social sciences. In *The Social Science Journal*. Advance online publication. https://doi.org/10.1080/03623319.2021.1986663.

Bauer, P.C., Freitag, M. (2018). Measuring trust. In Uslaner, E.M. (Ed.), *The Oxford Handbook of Social and Political Trust*. New York: Oxford University Press, 15-36. https://doi.org/10.1093/oxfordhb/9780190274801.013.1.

Belliard, F., Maineri, A., Plomp, E., Ramos Padilla, A.F., Sun, J. & Jeddi, M.Z. (2023). A 10 step checklist for starting FAIR discussions in your community: Call for contributions. In *Fair Connect*, 1(1), 45-48. https://doi.org/10.3233/FC-230505.

Benson-Greenwald, T.M., Trujillo, A., White, A.D. & Diekman, A.B. (2023). Science for others or the self? Presumed motives for science shape public trust in science. In *Personality and Social Psychology Bulletin*, 49(3), 344-360. https://doi.org/10.1177/01461672211064456.

Bochynska, A., Keeble, L., Halfacre, C., Casillas, J.V., Champagne, I-A., Chen, K., Röthlisberger, M., Buchanan, E.M. & Roettger, T.B. (2023). Reproducible research practices and transparency across linguistics. In *Glossa Psycholinguistics*, 2(1), 1-36. https://doi.org/10.5070/G6011239.

Brehm, J., Savel, M. (2019). What do survey measures of trust actually measure?. In Sasaki, M. (Ed.), *Trust in Contemporary Society*. Leiden/Boston: Brill, 233-260. https://doi.org/10.1163/9789004390430_013.

Brudner, E.G., Karousatos, A.J., Fareri, D.S. & Delgado, M.R. (2022). Trust and reputation. How knowledge about others shapes our decisions. In Krueger, F. (Ed.), *The Neurobiology of Trust*. Cambridge: Cambridge University Press, 155-184. https://doi.org/10.1017/9781108770880.010.

Calamai, S., Casellato, A. & Stamuli, M.F. (2022). Collaborative best practices. An Italian Vademecum on the conservation, the description, the use and the re-use of oral sources. In *Bulletin de l'AFAS. Sonorités*, 48, 182-195. https://doi.org/10.4000/afas.7499.

Calamai, S., Piccardi, D., Pretto, N., Candeo, G., Stamuli, M.F. & Monachini, M. (2022). Not just paper: Enhancement of archive cultural heritage. In Fišer, D., Witt, A. (Eds.), *CLARIN. The Infrastructure for Language Resources*. Berlin/Boston: De Gruyter, 647-665. https://doi.org/10.1515/9783110767377-025.

Citrin, J., Stoker, L. (2018). Political trust in a cynical age. In *Annual Review of Political Science*, 21, 49-70. https://doi.org/10.1146/annurev-polisci-050316-092550.

Clemente, M.A. (2023). Genova ferita, una città all'interno di un evento internazionale. Master of Arts dissertation, Università di Pisa.

Conti, N., Memoli, V. (2015). Show the money first! Recent public attitudes towards the EU in Italy. In *Italian Political Science Review*, 45(2), 203-222. https://doi.org/10.1017/ipo.2015.11.

De Jong, F., Maegaard, B., Fišer, D., Van Uytvanck, D. & Witt, A. (2020). Interoperability in an infrastructure enabling multidisciplinary research: the case of CLARIN. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. & Piperidis, S. (Eds.), *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. Paris: ELRA, 3406-3413.

Di Benedetto, M-G., De Nardis, L. (2022). The GEMMA speech database: VCV and VCCV words for the acoustic analysis of consonants and lexical gemination in Italian. In *Data in Brief*, 43. https://doi.org/10.1016/j.dib.2022.108373.

Dijkers, M.P. (2019). A beginner's guide to data stewardship and data sharing. In *Spinal Cord*, 57, 169-182. https://doi.org/10.1038/s41393-018-0232-6.

Doyle, D., Şerban, O. (2024). Interruption audio & transcript: derived from group affect and performance dataset. In *Data*, 9(9). https://doi.org/10.3390/data9090104.

Ephratt, M. (2008). The functions of silence. In *Journal of Pragmatics*, 40(11), 1909-1938. https://doi.org/10.1016/j.pragma.2008.03.009.

Ezzes, Z., Schneck, S.M., Casilio, M., Fromm, D., Mefferd, A.S., de Riesthal, M. & Wilson, S.M. (2022). An open dataset of connected speech in aphasia with consensus ratings of auditory-perceptual features. In *Data*, 7(11). https://doi.org/10.3390/data7110148.

Ferriter, D. (1998). Oral archives in Ireland: a preliminary report. In *Irish Economic and Social History*, 25(1), 91-95. https://doi.org/10.1177/033248939802500108.

Fichtner, U.A., Horstmeier, L.M., Brühmann, B.A., Watter, M., Binder, H. & Knaus, J. (2023). The role of data sharing in survey dropout: a study among scientists as respondents. In *Journal of Documentation*, 79(4), 864-879. https://doi.org/10.1108/JD-06-2022-0135.

Fitzgerald, C.M. (2021). A framework for Language Revitalization and Documentation. In *Language*, 97(1), 1-11. https://doi.org/10.1353/lan.2021.0006.

Garellek, M., Gordon, M., Kirby, J., Lee, W-S., Michaud, A., Mooshammer, C., Niebuhr, O., Recasens, D., Roettger, T.B., Simpson, A. & Yu, K.M. (2020). Letter to the editor: Towards open data policies in phonetics: What we can gain and how we can avoid pitfalls. In *Journal of Speech Sciences*, 9, 3-16. https://doi.org/10.20396/joss.v9i00.14955.

Garrett, A., Harris, A.C. (2022). Assessing scholarship in documentary linguistics. In *Language*, 98(3), 156-172. https://doi.org/10.1353/lan.0.0266.

Gomes, D.G.E., Pottier, P., Crystal-Ornelas, R., Hudgins, E.J., Foroughirad, V., Sánchez-Reyes, L.L., Turba, R., Martinez, P.A., Moreau, D., Bertram, M.G., Smout, C.A. & Gaynor, K.M. (2022). Why don't we share data and code? Perceived barriers and benefits to public archiving practices. In *Proceedings of the Royal Society B*, 289(1987). https://doi.org/10.1098/rspb.2022.1113.

Hardin, R. (2006). *Trust*. Cambridge/Malden: Polity Press.

Harsin, J. (2018). Post-truth and critical communication studies. In *Oxford Research Encyclopedia of Communication*. Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780190228613.013.757.

Haspelmath, M., Michaelis, S.M. (2014). Annotated corpora of small languages as refereed publications: a vision. Diversity Linguistics Comment. https://dlc.hypotheses.org/691. Accessed 08 October 2024.

Ide, N., Calzolari, N. (2005). Introduction to the special inaugural issue. In *Language Resources and Evaluation*, 39(1), 1-7. https://doi.org/10.1007/s10579-005-2689-0.

Janda, L.A. (2021). Managing data and statistical code according to the FAIR principles. In Berez-Kroeker, A.L., McDonnell, B., Koller, E. & Collister, L.B. (Eds.), *The Open Handbook of Linguistic Data Management*. Cambridge/London: The MIT Press, 447-452. https://doi.org/10.7551/mitpress/12200.003.0042.

Lambin, J-J. (2008). *Changing market relationships in the internet age*. Louvain-la-Neuve: Presses universitaires de Louvain.

Lawrence, B., Jones, C., Matthews, B., Pepler, S. & Callaghan, S. (2011). Citation and peer review of data: moving towards formal data publication. In *International Journal of Digital Curation*, 6(2), 4-37. https://doi.org/10.2218/ijdc.v6i2.205.

Luzietti, R.B. (2023). Evaluation of the Archivio Vi.Vo architecture: A case study on the reuse of legacy data for linguistic purposes. In Erjavec, T., Eskevich, M. (Eds.), *Selected Papers from the CLARIN Annual Conference 2022*. Linköping: Linköping University Electronic Press, 90-98. https://doi.org/10.3384/ecp198009.

Manali, S. (2024). Il ritorno della Pantera. Un laboratorio universitario per la costruzione di un archivio orale. In *JLIS*, 15(1), 159-176. https://doi.org/10.36253/jlis.it-571.

Mulè, A. (1993). Censimento e presentazione dei dati. In Barrera, G., Martini, A. & Mulè, A. (Eds.), *Fonti orali. Censimento degli istituti di conservazione*. Roma: Ministero per i beni culturali e ambientali, Ufficio centrale per i beni archivistici, 33-49.

Piccardi, D., Ardolino, F. & Calamai, S. (Eds.) (2019). *Audio archives at the crossroads of speech sciences, digital humanities and digital heritage*. Studi AISV 6. Milano: Officinaventuno.

Piccardi, D., Calamai, S. (2023). How many oral archives are in your home? Piloting a new Tuscan census in the Gra.fo Reloaded project. In Carbé, E., Lo Piccolo, G., Valenti, A. & Stella, F. (Eds.), *La memoria digitale. Forme del testo e organizzazione della conoscenza. Atti del XII convegno annuale AIUCD*. University of Siena: AIUCD, 389-394.

Pozzebon, A., Biliotti, F. & Calamai, S. (2016). Places speaking with their own voices. A case study from the Gra.fo archives. In Ioannides, M., Fink, E., Moropoulou, A., Hagedorn-Saupe, M., Fresa, A., Liestøl, G., Rajcic, V. & Grussenmeyer, P. (Eds.), *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. EuroMed 2016*. Cham: Springer, 232-239.

Rotenberg, K.J. (2018). *The psychology of trust*. London and New York: Routledge. https://doi.org/10.4324/9781315558912.

Shaklee, P.M. (2014). Data in brief – Making your data count. In *Data in Brief*, 1, 5-6. https://doi.org/10.1016/j.dib.2014.09.001.

Stracke, C.M. (2020). Open science and radical solutions for diversity, equity and quality in research: a literature review of different research schools, philosophies and frameworks and their potential impact on science and education. In Burgos, D. (Ed.), *Radical Solutions and Open Science*. Singapore: Springer, 17-37. https://doi.org/10.1007/978-981-15-4276-3_2.

Tavolo permanente per le fonti orali (2023). *Vademecum per il trattamento delle fonti orali*. Roma: Ministero della Cultura, Direzione generale Archivi.

Thieberger, N. (2012). Counting collections. Endangered Languages and Cultures. https://www.paradisec.org.au/blog/2012/11/counting-collections/. Accessed 08 October 2024.

Thieberger, N., Margetts, A., Morey, S. & Musgrave, S. (2016). Assessing annotated corpora as research output. In *Australian Journal of Linguistics*, 36(1), 1-21. https://doi.org/10.1080/07268602.2016.1109428.

Valentini, C., Piccardi, D., Calamai, S. & Stamuli, M.F. (2023). Da cornice a soggetto. Il documento sonoro nell'infrastruttura Archivio Vi.Vo. In *Archivi*, 18(1), 88-125.

Wang, H-R. (2014). "Publish or perish": Should this still be true for your data? In *Data in Brief*, 1, 85-86. https://doi.org/10.1016/j.dib.2014.11.005.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J- W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In *Scientific Data*, 3. https://doi.org/10.1038/sdata.2016.18.

DOI: 10.17469/O2112AISV000006

EMANUELA CRESTI, MASSIMO MONEGLIA

Prosody and the Pragmatic Functions of Vocative in Spoken Italian Corpora

Vocatives are bare nouns that address the interlocutor. They are out of the argument structure but are recently considered syntactically integrated into the utterance. Vocatives function to *call on* or *strengthen social relationships.* The prosody of recalls (*Vocative chants*) has been studied in the Autosegmental model. They can appear isolated by prosody when starting the utterance or more integrated when at its end.
Assuming the Language into Act Theory and a corpus-based approach, the paper argues that Vocatives are Units of information and that their prosodic performance is predictive of their functions: a) isolated in a prosodic unit of a root type, they are speech acts (*Distal/ Proximal calls*, *Greetings*, *Regret, Protest*) with differential prosody; b) focused in the first position of a Pattern of illocutionary units, serve at *sharing the attention* with the addressee; c) when out of focus, both in first and final position, are dialogical units of Allocutive *strengthening the empathic relation* and are not referential. Semantic correlations support their functional distinction.

Keywords: vocative, prosody, information functions, Allocutive, spoken Italian.

1. *Introduction*

What is commonly understood with the term vocative is not as simple as it might appear. However, there is general agreement that expressions that address the interlocutor – reader or addressee – used in isolation, without an apparent syntactic link with the rest of the sentence, are identified as vocatives.

However, as soon as the limits of this superficial description are overcome, if we try to identify the semantic domain of these expressions, their morphological and syntactic features, the levels and means of their isolation, and their interpretations and explanations diverge.

The literature on the subject has gradually expanded over the years, starting with Zwicky (1974). It has allowed the inclusion, although marginal, of the institute in the most recent major grammars, such as Quirk, Greenbaum, Leech & Svartvik (1985), Renzi, Salvi & Cardinaletti (1995), Biber, Johansson, Leech, Conrad & Finegan (1999), Abeillé, Goddard (2021) or in targeted research. However, the vocative is still considered in the literature as a puzzle (Coene, D'Hulst & Tasmowski, 2019). Sonnenhauser, Noel (2013), which can still be considered the most complete treatment of the subject, highlights the "intricate nature" of the vocative and the problems it poses for current linguistic analyses. The volume collects studies of many specialists in language typology, morphology, and syntax, and the cross-

linguistic variation of forms of address is observed. Empirically oriented papers focus on data from different languages, although most papers bring the vocative noun phrases back to formal syntactic frames. The universal prosodic marking of the 'vocative chant' is also considered (Göksel, Pöchtrager, 2013). Still, we want to report the editors' reflections.

> The main problem in discussing the 'vocative' is the separation of levels – form vs. function and the subsystems of prosody, morphology, syntax, semantics, and pragmatics – as well as the differing assumptions about the structures constituting the object of observation – morphology only, or syntactic structures such as NPs, clauses, sentences or utterances and the contribution of prosodic signaling. There is no consensus on their structure and no consensus on how to encompass vocative structure within a cross-linguistic approach (Sonnenhauser, Noel, 2013:17)

The conclusion is that the complexity of the vocative depends on its nature, which lies *between language system* and *performance*.

We can summarize some crosslinguistic features of the vocative. The lexical selection is limited to 2nd person pronouns (both direct and reverential), bare nouns comprehending proper names, appellatives, titles, and role nouns, possibly characterized by a morphologic case as in Latin, but also nominal phrases integrated by restrictive relative clauses developing forms of address.

The syntactic isolation of the vocative can be explained in different ways, but it is commonly assumed out of the argument structure of the predicate (Moro, 2003; D'Alessandro, van Oostendorp, 2016). It is marked by specific distribution, mainly at the beginning or end of the sentence in written texts or at the beginning or end of the utterance in speech. Moreover, the isolation is encapsulated by space and commas in writing or by silence and prosodic boundaries in speech.

As for the function of the vocative, there are significant interpretative differences; since Zwicky (1974), it has been observed that although the functions can be many, two main ones can be identified: the *call*, like a real appeal for getting the attention of the addressee, and the *address* to reactivate the addressee's attention (Zwicky, 1974; Quirk et al., 1985; Mazzoleni, 1995; Corr, 2022). It's worth noticing that there is a shared pre-theoretical assumption that vocatives "encode" a second person since every alleged vocative expression can be retraced to a "you," which might underly its meaning (Bernstein, forthcoming; Micali, 2022). Accordingly, all vocatives are dealt with as deictic devices, as second-person pronouns are.

This set of features can be enriched by findings emerging from corpus-driven data on the American-English corpus reported by Biber et al. (1999), highlighting the importance of vocatives in defining and maintaining social cohesion that appears to be most significant in spoken conversations. Two distributions of vocatives result from the data:

– isolated or initial at the beginning of the utterance, primarily associated with a function of *Call*;
– final at the end of the utterance, primarily associated with *stressing the social cohesion* (with the highest frequency).

However, the association between distribution and functions is not predictive.

The prosodic performance of vocatives started to be faced only in the 2000s. Prosodic contours that are attributed to the recall of the addressee have been studied mainly within the framework of Auto-segmental theory and are reported under the name of *vocative chants* and *chanted calls* (Ladd, 2008; Frota, Prieto, 2015; Borràs-Comes, Sichel-Bazin & Prieto, 2015). It is worth noticing that the contribution of Göksel, Pöchtrager (2013), still a point of reference, proposed a difference between *calls* and what they consider *true vocatives* prosodically described.

Recently, in the chapter on Vocative in Corr (2022), remarks were also made on the different prosodic realizations of vocatives. At the beginning of the utterance, they correspond to a chunk marked by a prosodic boundary. In contrast, at the end of the utterance, they are more integrated into the prosodic contour, typically characterized by the falling movement of post-focal units.

D'Alessandro, van Oostendorp (2016), from a phonological point of view, and Corr (2022), at the syntactic structure level, face the problem of integrating the pragmatic functions of vocative within the grammar, trying to overcome the fracture "between system and performance" complained by Sonnenhauser, Noel. However, a general problem arises when formal grammar, which in principle does not treat the pragmatic aspect of language, must consider speech acts and their embodied dimensions expressed by prosody, as is the case with vocatives[1].

The literature does not clearly state a systematic relation between the various functions of vocatives and their prosodic properties. For instance, the sharp distinction between the *call* and the *address* – highest frequency, as reported in the Longman Grammar – relies on distributional properties and is not accompanied by any prosodic description.

This paper aims to make explicit the correlation between the vocative's functions and their prosodic performance through corpus-based analysis.

In § 2 we will sketch the main concepts of the Language into Act Theory (L-AcT), the framework adopted in this paper, which emphasizes the relation between prosody and pragmatics in speech corpora. In § 3 and 4, we will consider the language contexts of our Italian data set in which vocatives bear illocutionary functions. In § 3 we will look at vocatives filling one isolated prosodic unit (Comment unit) performing one speech act. In § 4 we will introduce the notion of Illocutionary pattern, a structure in which the vocative appears as a *call* within the *functional calling* pattern as a focused prosodic unit in the first position. In § 5 we will consider vocatives that cannot constitute independent speech acts but develop an information function of support (Allocutive). Regardless of their distribution within the utterance, they strengthen the empathic relation with the addressee and are consistently de-focused. Finally, in § 6 we will propose that the differential pragmatic values expressed by prosody correlate with semantic restrictions shown by vocatives, banning the referential and deictic reading in the Allocutive function.

[1] Rizzi (1997) considers the performative at the highest level of the sentence structure and is at the origin of this approach.

2. *The L-AcT framework and the relevance of prosody*

The research investigates the identification and definition of the vocative within L-AcT, a framework allowing the corpus-based study of prosody (Cresti, 2000; Moneglia, Raso, 2014; Cresti, Moneglia, 2018). Vocatives are analyzed in the LABLITA corpus of spoken Italian (Tuscan variety) not as lexical or propositional items but as information units developing a pragmatic function, which is considered in correlation with their necessary prosodic performance.

L-AcT proposes an organic frame, according to which the affect and the inherent pragmatic goal of the speaker are considered at the origin of speech and performed through a prosodic pattern. In the Austinian tradition (Austin, 1962), L-AcT takes a speaker-oriented perspective. The speaker's affective and pragmatic behavior toward the addressee determines the illocutionary act (Cresti, 2020). Prosody manifests the illocutionary information and behaves as an embodied interface with the locutionary act that consists of semantics and morpho-syntactic constituents, which find their place within the prosodic pattern. Therefore, the analysis of prosodic cues is at the core of the model.

Prosodic boundaries segment the speech flow into information units (Chafe, 1994) shaped by perceptively relevant prosodic contours ('t Hart, Collier & Cohen, 1990). Information units are framed in a prosodic pattern that performs the utterance.

The Comment is the information unit that expresses the illocutionary force and is necessary and sufficient to constitute the utterance. The other information unit types (Topic, Appendix, Parenthesis, and various units dedicated to managing the interaction with the addressee) depend on the Comment and cannot be pragmatically interpreted in isolation. The locutionary act fills the information units, whose syntactic and semantic selection depends on their functional goal.

In this frame, vocatives are supposed to develop two information functions with differential prosodic performance:

– Comment specifies the illocutionary activity, typically performing a *call* and other speech act types. In this case, from a syntactic point of view, vocatives can be considered independent nominal utterances. Comment units are realized through dedicated prosodic *root* units ('t Hart et al., 1990) presenting formal variants correlating with different illocutions they convey (Cresti, 2020; Cresti, Moneglia, 2023).
– Allocutive, developing dialogical support of the utterance, works as an *address* to reactivate the addressee's attention and specifically express the social relationship with the speaker. In this case, the utterance must have already established its independent illocution in the Comment. From a syntactic point of view, vocatives are nominal phrases linked to the Comment at the information structure level. Allocutives are always realized through a post-focal flat or falling prosodic unit without significant variations (Raso, 2014; Raso, Leite, 2010; Raso, Ferrari, 2020).

In the following paragraphs, corpus-based evidence will demonstrate these assumptions in detail.

3. *Illocutionary types enacted by Vocative expressions*

Deriving from corpus observation, Biber et al. (1999) already claimed that vocatives could correlate with many pragmatic uses: *call*, *alert*, *invocation*, *challenge*, *offense*, *appreciation*, *congratulation*, *welcome*, and *greetings*[2].

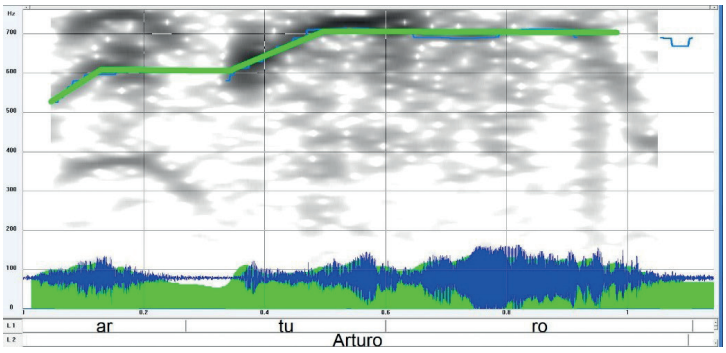
Our research on the LABLITA corpus confirms Biber's assumptions. It has allowed us to find many cases of utterances that are nominal expressions, clearly set off from any compositional structure, used in isolation, and fulfilled by typical vocative items. They can enact illocutions of different classes, specifically the directive, expressive, and ritual classes, according to the L-AcT taxonomy (Cresti, 2020). As foreseen, each illocutionary type is performed by a dedicated root unit.

The most common examples of illocutionary uses filled by vocatives concern different types of recall, but they also extend to other directive and expressive types. Let's see some examples and their f0 tracks[3] regarding *Call* illocutionary types, which, according to our classification, are a subtype of the directive class. They are pragmatically oriented at *opening the communication channel to get the addressee's appearance or involvement* and can be distinguished between *distal* and *proximal calls*.

The *Distal call* regards an addressee not visible in the context that is requested to show up.

A child calls her dog, who ran away in a public garden.
*NIP: Arturo //COM
%ill: distal call [prvcvl62-panc]

Figure 1



[2] See also Borràs-Comes et al. (2015).

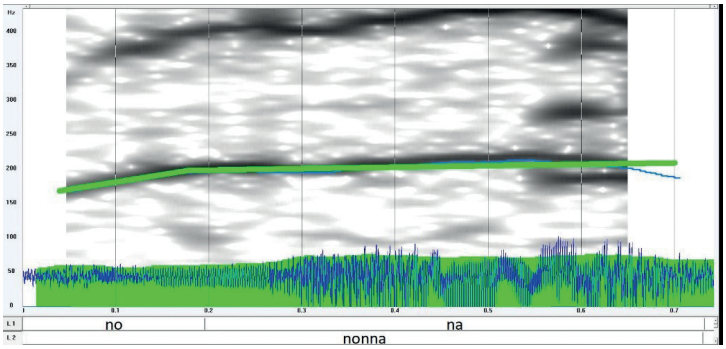
[3] Being derived from speech recorded in a natural environment, the f0 track may show errors (as in Fig. 3) or be partially not calculated. Here and below, the figures show the f0 track face to the first or second harmonic to verify the correctness of the contour, which is highlighted through its stylization. Layer 1 shows the alignment with syllables, and Layer 2 the segmentation into the prosodic units of the utterance, gathering the syllabic sequence. To the end of readability, the segmentation into prosodic units (if any) is traced onto the f0 track. F0 calculation and synthesis are achieved through Winpitch (Martin, 2011). The audio files of the full set of examples presented in the figures are available to the reader in .wav files from https://drive.google.com/drive/folders/1zTM4UmS4mZSnZAAy6Ze8pt-1FB1cZyz3M?usp=sharing.

A rising contour performs the unit to a high f0 level, followed by a prolonged hold on the last post-tonic vowel, characterized by high-intensity values.

The *Proximal call* regards an addressee present in the near context but not involved in the exchange.

The child calls her grandmother, who was talking with another person.
*NIP: nonna //COM
%ill: proximal call (with an attitude of irritation) [prvcvl62-panc]

Figure 2



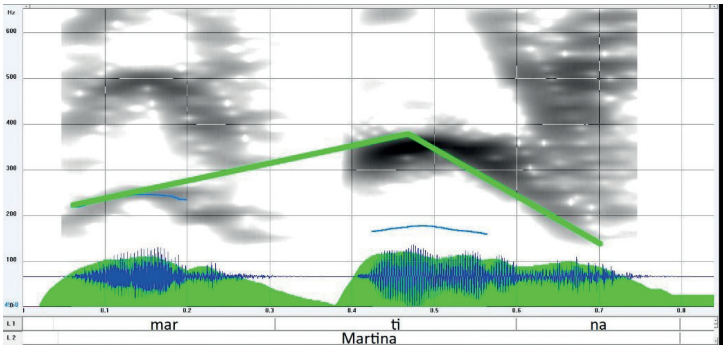
The *Proximal call* is performed by a rising contour similar to the *Distal call* but at a much lower f0 range, still marked by a lengthening on the last post-tonic vowel.

The alleged *Calls* are not limited to the previous ones but include the expressive and ritual acts presented below.

The *Protest* belongs to the expressive class, here characterized by the expression of a feeling of fear.

During a driving lesson, the novice risked a collision.
*MAX: Martina //COM ma che cazzo fai?COM
'Martina! What the fuck are you doing?'
%ill: 1 protest (emotionally marked by fear); 2 refusal (emotionally marked by anger) [ifamdl10-guid]

Figure 3

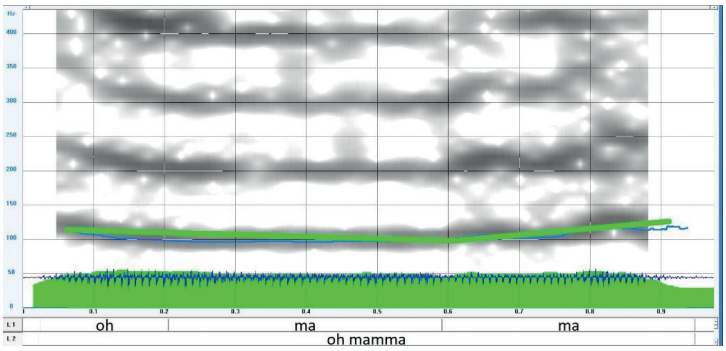


The *Protest* is characterized by a rising-falling contour at a high f0 range and strong intensity. The necessary falling movement occurs on the tonic vowel.[4]

Regret also belongs to the expressive class, characterized by expressing the corresponding feeling.

After a wrong Poker hand.
*MAX: oh mamma //COM
'oh mother'
%ill: regret [ifamcv14]

Figure 4



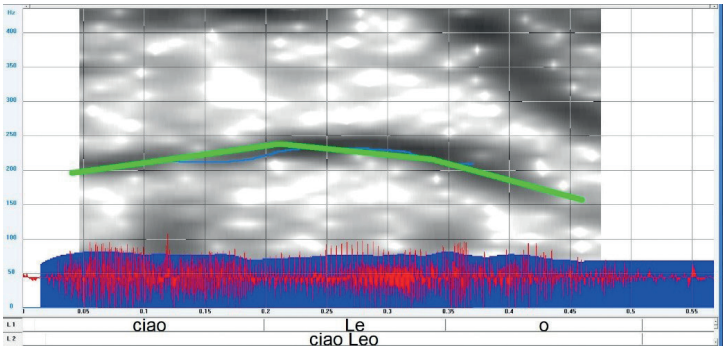
The *Regret* is performed by a platform contour at a low f0 range with a lengthening of the tonic syllable.

Greeting belongs to the ritual class, characterized by a friendly modality. When enacted as a parting greeting, it is preferably performed through a greeting formula and the addressee's name.

When going home after joint work.
*OLV: cia' Leo //COM
'so long Leo'
%ill: greeting [prvcvl63-plum]

[4] According to Ladd (1978), *Protest* or *Warning* is described as a "stepping down sequence of two-level tones." Even if realized with a proper name, it is not necessarily associated with *calling* but "conveys the implication that an utterance is stereotyped or stylized" expressing a nuance of meaning. In L-AcT, *Protest* is considered a specific illocutionary act of the expressive class performed by the dedicated contour just described, which differs from those of *Calling*.

Figure 5



The *Greeting* is performed by a rising-falling contour at a mid-f0 range with the falling movement on the vocative expression.

4. *Illocutionary patterns of* Functional call

Illocutionary patterns are utterances composed of a chain of at least two Comments, called Multiple Comments (CMM). In this case, the Comment unit does not stand alone. Illocutionary patterns are frequent in spontaneous speech and are conceived according to a *natural rhetoric model* of pragmatic units (*Reinforcement, List, Comparison, Alternative structures,* etc.). They are performed according to a prosodic and rhythmic pattern (Panunzi, Saccone, 2018).

Functional calling is a specific Illocutionary pattern composed of two CMMs: the first develops a directive *call* to share the attentional focus with the addressee, and the other is a kind of request whose force can vary.

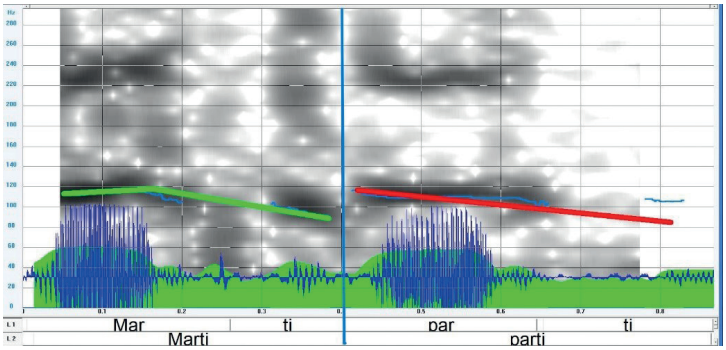
The *Call* within this pattern is *proximal*. Still, the communicative channel is foreseen already open, and it is prosodically distinguished from *Proximal calls* in isolation. Let's look at examples of *Functional calling* in which the *Call* is followed by *Order* or *Reproach* illocutions. The f0 track of the last example is in Fig. 6.

While dining.
*CLA: Giovanni /CMM l'acqua //CMM
'Giovanni ! Water'
%ill: call + order [ifamcv17]

During a car driving lesson.
*MAR: babbo /CMM falla finita di recitare //CMM
'Dad! Stop acting'
%ill: call + reproach [ifamdl19]

during a car driving lesson
* MAX: Marti /CMM parti //CMM
' Marti! go'
%ill: call + instruction [ifamdl19-234]

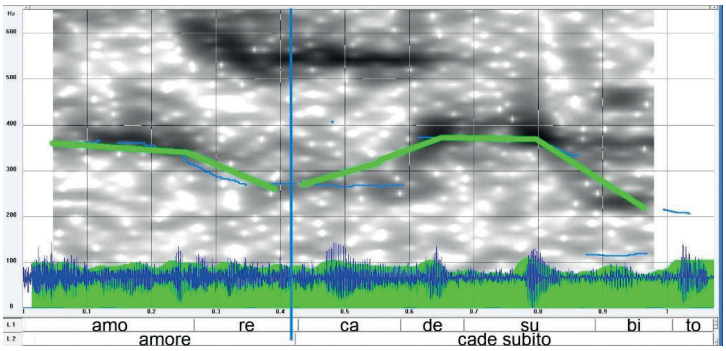
Figure 6



Both contours are characterized by a falling directive movement, where the *call* is shaped differently from the *Proximal call* in Fig. 2. The latter is performed in isolation and is rising. Below, the *Call* introduces an *Alert* illocution.

To a child playing with some breakable object.
*DAN: eh /^{PHA} amore /^{CMM} cade subito // ^{CMM}
'hey, darling! it falls right away'
%ill: call + alert [ifamcv15]

Figure 7



Also, in this pattern, the first CMM of *Call* is shaped by a directive falling contour, while the second CMM, bearing a different directive illocutionary type, is performed by a rising-falling contour.[5]

It must be highlighted that vocative expressions within a *Functional call* can include second-person deictic pronouns. For instance, the deictic *Call* in the

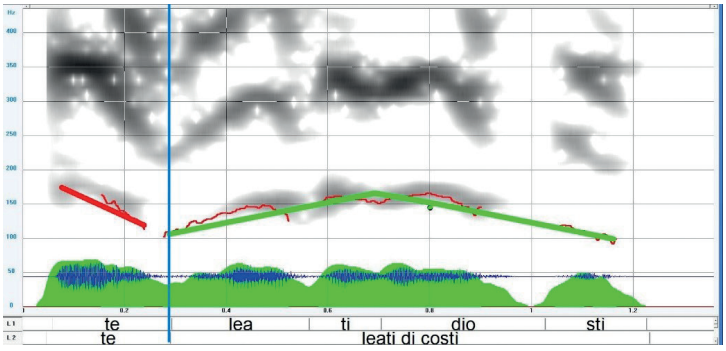
[5] All the vocatives developing a *Call* within illocutionary patterns are always at the beginning of the composed pattern. This distribution has been faced in generative frameworks, considering the vocative's connection with the major syntactic entity at the illocutionary level (D'Alessandro, van Oostendorp, 2016; Corr, 2022) According to L-AcT, the vocative accomplishing a *Call* in a *Functional calling* pattern is structured within a pragmatic model composed of two illocutions. Therefore, the pattern structure cannot be headed by the illocutionary force conveyed through the vocative.

following examples introduces a directive illocution, respectively, a *Partial question* and an *Order*. It shows its typical falling movement in both cases, as in Fig. 8.

Among many friends leafing through an album of old photos
*ELA: e te /CMM quanto tu c' a(vevi) ?CMM
'and you! how old were you?'
%ill: call + partial question [ifamcv01]

During a lively discussion at dinner
*MAX: te /CMM le(v)ati di costì //CMM
'you! get out of here!'
%ill: call + order [lab 05]

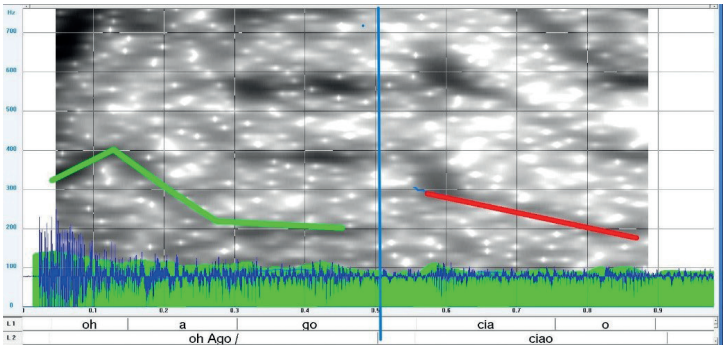
Figure 8



Also, greetings, preferably at the addressee's coming, are realized through an illocutionary pattern: the first CMM is still a *Call*, and the second is a proper greeting formula, both performed with a falling movement.

*SAB: oh Ago /CMM ciao //CMM
'oh Ago! Hi!'
%ill: call + greeting [ipubdl03]

Figure 9



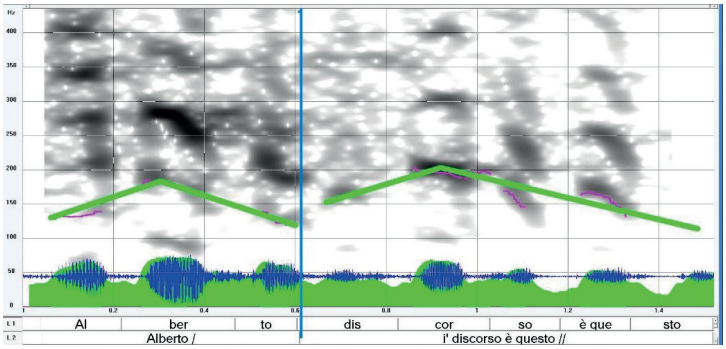
We also found *Functional callings* that don't match the previous model since they are composed of a *Call* followed by a representative act. The latter is characterized by a disagreement attitude with the addressee instead of a directive illocution. Let's consider some examples. The f0 track of the last one is in Fig. 10.

*OTT: Assan /CMM senti /CNT no //CMM
'Assan ! Look, no way'
%ill: call + assertion (disagreement attitude) [ipubcv01]

*OTT: < Mega > /CMM non ci siamo capiti //CMM
'Mega ! We did not understand each other'
%ill: call + assertion (disagreement attitude) [ipubcv01-287]

*ABC: Alberto /CMM i' discorso è questo //CMM
'Alberto ! This is the point'
%ill: call + assertion (disagreement attitude) [ipubcv04]

Figure 10



When vocatives develop a Comment unit – both in isolation and within an illocutionary pattern – they are pragmatically interpretable. It may have theoretical relevance to note that, as assumed in the literature, the addressee is semantically included, and vocatives may encode a second-person pronoun. This assumption is supported by evidence that, specifically in *Functional calls*, the vocative expression can be replaced by a second-person pronoun, as seen in the example in Fig. 8.

However, it is worth noticing that the substitution will lead to odd utterances when applied to vocatives working as illocutionary acts in isolation. As we will see in the next paragraph, vocatives in the Allocutive function cannot be substituted by a deictic 'you'.

5. *Allocutive*

Within the L-AcT framework, various Dialogical information functions of prompting for attention (*Incipit*, *Expressive*, *Conative*, *Allocutive*) have been identified through corpus-based observation (Cresti, 2000; Frosali, 2008; Raso,

2014; Raso, Ferrari, 2020). However, only vocatives are specialized to implement the Allocutive function. The Allocutive identifies the addressee, seeking his attention, but its specific function is socially cohesive, and its end is establishing an empathic connection with him (Moneglia, Raso, 2014).

It is necessary to recall the functional distinction between an Allocutive and a Comment since the Allocutive cannot be pragmatically interpreted in isolation and functionally depends on the latter. Consequently, the distribution of the Allocutive with respect to the Comment is a fundamental feature for its identification together with its prosodic realization.

Italian corpus data show that most Allocutives are distributed after the Comment unit and are performed through a dedicated prosodic unit, which, however, is integrated into the declination line of the utterance. The prosodic contour can be described as a gently falling or flat unit with a weak intensity value. Its phonetic execution remains accurate, unlike, for example, that of the Comment Appendixes (Cresti, 2021). This feature is necessary for the recognizability of the proper name or appellation of the addressee, which is required to carry out the function.

Let's consider some examples illustrating the final distribution, which appears to match the usage found to be the most frequent by Biber et al. (1999). In the following examples, the Allocutive is respectively filled by appellatives, family names, and proper names:

*LUC: lui disse /INT me la piglio io /COM-r nini // ALL-r
'he said, I'm the one taking it, boy'
%ill: reported speech [ifamcv22-286]

*VER: e' 'un lo so /COM mamma // ALL
'eh, I don't know it, mom'
%ill: conclusion [ifamdl14-2]

*GAL: di solito si fa così / COM Mar(co) // ALL
'it's usually done like this, Marco'
%ill: reproach [ifamcv14]

*WAL: ammazzai due vipere / COM ieri /APC / Leo // ALL
'I killed two vipers yesterday, Leo'
%ill: narration [prvcvl63-plum]

*LEO: la fascetta /TOP indo' la va / COM Marco ? ALL
'the band, where should it go, Marco?'
%ill: partial question [prvcvl63-plum]

Figures 11 and 12 show the f0 tracks of the last two examples. The vocative is performed through a flat-falling contour. Fig. 11 shows an Allocutive presenting a prosodic contour at a lower f0 range than the Appendix, which usually concludes the utterance.

Figure 11

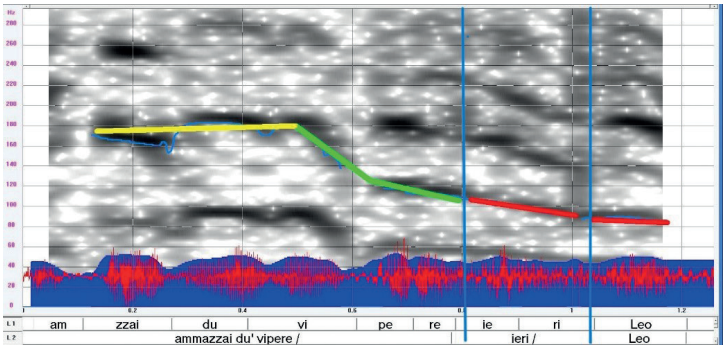
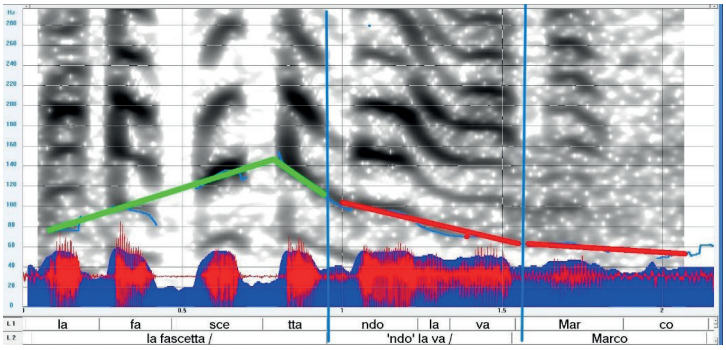


Figure 12



In the previous examples, the illocutions performed belong to the *assertive*, *expressive*, and *question* illocutionary types. Regardless of the illocutionary variation of the Comment, all Allocutives are realized by the flat-falling prosodic unit.

Concerning the rare instances of Allocutives with a distribution before the Comment, they, by preference, do not occur at the very start of the utterance. The following are some of the information patterns we found in corpora:

– after Topic, Incipit, and Conative

*IDA: guarda /^{CNT} Alessandro /^{ALL} credimi //^{COM}
'look, Alessandro, believe me'
%ill: directive [ifamdl18]

– interspacing i-COM and COM and CMM and CMM

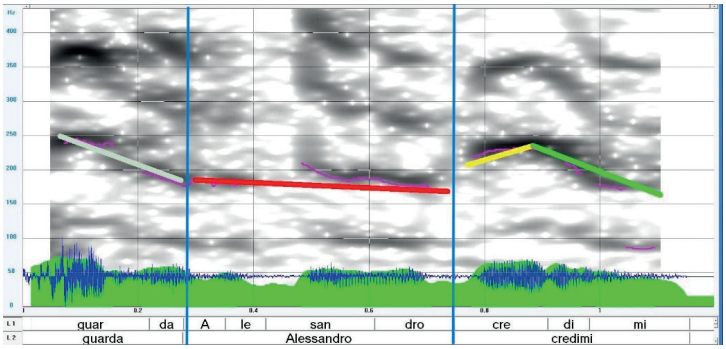
*ALE: lo sai /$^{i\text{-}COM}$ Daddo/^{ALL} s'è rotto il video-registratotore //^{COM}
'you know, Daddo, the video recorder broke'
%ill: complaining [ifamcv15]

– at the beginning of an utterance within a Reported episode

*IDA: allora dico /INT [Antonio /ALL_r guarda /CNT_r Pretty woman /TOP_r esiste in televisione //COM_r]
'then I say, Antonio, look, Pretty woman exists on television'
%ill: reported speech [ifamdl20-54]

Fig. 13 shows the f0 track of the first example. The prosodic contour of the Allocutive unit is still coherent with the features found in the final position (flat with low-intensity values) even though it cannot follow the declination line in this position.

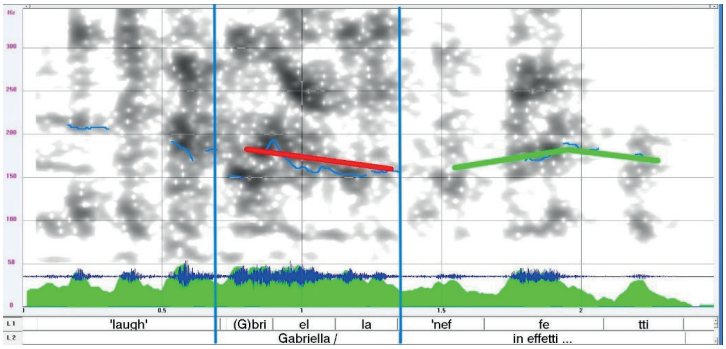
Figure 13



In our corpus, only a few examples of the Allocutive occur at the absolute beginning of the utterance, as in the example in Fig. 14.

During a chat
*ELA: <laugh> Gabriella /ALL in effetti...COM
' Gabriella, actually...'
% ill: expression of obviousness [ifammn03]

Figure 14



In this distribution, the Allocutive is characterized by a prosodic contour featuring a flat-falling movement.

Crucially, contrary to the *Functional calling* pattern, in which the vocative is also found in the first position, when the vocative develops the Allocutive function, it cannot be pragmatically interpreted in isolation as an act of calling because of its prosodic performance.

6. *Syntax and Semantics of Allocutives*

6.1 The Lexicon

Concerning the limited lexicon of Allocutives, only a restricted group of nouns are employed, and they are always characterized for making explicit the relationship between the speaker and the addressee.[6]

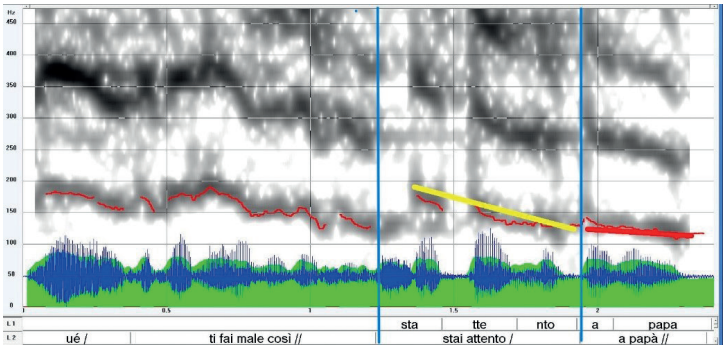
Frequent and cross-linguistically recurrent nouns are:

- Proper name (*addressee's label*)
- Titles (*addressee's social position*)
- Appellative (*affective speaker's labeling*)
- Role noun (*family position with respect to the speaker*)

Here, we can only quickly refer to the "inverse allocution", which is widespread in the southern dialects and varieties of Italian (Corr, 2022b). In the inverse Allocutive relation, the alleged addressee is the speaker himself. Inverse Allocutive equally aims to stress the existing relationship with the addressee. The speaker's self-citation is preceded by the morpheme *a*, which can be interpreted as a benefactive dative. In this construction, lexical selection is even more restricted, limited to proper names or family relationships. See Fig. 15:

Father's calling his little son back.
*ABC: uè /CNT ti fai male cussì //COM stai attent /COM a papà //ALL
'hey, you get hurt like this. Be careful, (to) dad'
%ill: 1 admonishing; 2 alert

Figure 15



[6] The lexical restriction is particularly evident if we compare the spoken usage to vocative in *dedications* or *invocations* in high register written texts.

Let's notice that the last prosodic unit, *a papà* (to dad), is performed coherently through a flat-falling contour like the other types of Allocutives.

6.2 Syntax and Semantics

For syntax, corpus data confirms what is foreseen in the literature: bare nouns fill vocatives and resist articles (Rohlfs, 1966; Moro, 2003).[7] This research does not confirm the assumption that vocatives "encode" a second person since every alleged vocative expression can be retraced to a "you".

Let's recall that Italian corpus data show that vocatives filled by second-person pronouns are not found to develop Illocutionary types in isolation. If second-person pronouns substitute actual examples, the resulting utterances sound odd to competence judgment. The deictic reference to the addressed interlocutor is found only in vocatives within *Functional calling* patterns following their directive illocutionary force.

Considering specifically vocatives filling Allocutives, we stress that no deictics second person pronoun can develop this function. So, the Allocutive determines two restrictions: referential phrases determined by articles and deictic expressions are banned in this function.

Therefore, it is necessary to abandon a generic consideration of vocative as a grammar category that implies a deictic nature. Both distribution and prosodic performance determine different functions with specific semantic restrictions.

An apparent logical contradiction emerges when vocatives work as Allocutives: the addressee's attention is drawn to the speaker by the latter's proper name, title, or role noun. These expressions, which are commonly used to refer to the addressee, consistently convey the relationship with the latter, whereas surprisingly, deictic and definite descriptions, which are inherently referential, are excluded.

Regarding the deictic second-person pronouns, they are not adequate because they are semantically empty and can be filled only by referring to the context. Therefore, the relationship between the speaker and the addressee is missed. As regards nominal phrases with determinate articles, they are not semantically empty but are mandatorily referential. We hypothesize that referentiality is the specific obstacle to performing the Allocutive function. The very semantic nature of the Allocutive emerges. Proper names, titles, and role nouns are designators that may not have a referential reading. Allocutives only cite the addressee's designator (Kripke, 1971) to strengthen the empathic relation, and this information function has nothing to do with an act of referring to someone in the world.[8] In other words, the semantic restriction on the Allocutive is because deixis and determined noun phrases necessarily have a referential interpretation.

---

[7] There are well-known exceptions. French typically accepts vocatives filled by determined nominal constituents (*allez les enfants !*) (Bernstaint et.al 2019; Bernstaint, forthcoming).

[8] Crucially unlike Topic. Lambrecht (1996), on the contrary, claims that Allocutive can be equated to Topic since both are nominal detached units. See Cresti (forthcoming) for a discussion.

A further distributional property of the Allocutive function concerns the possibility of its iteration within an utterance. Indeed, the iteration of vocatives is not attested in corpora and results in odd competence judgment (unlike Topic or Appendix units). For instance, if a second Allocutive unit is added to actual examples, the resulting utterance is rejected by competence.

*GAL: di solito si fa così /^{COM} Marco /^{ALL} * papà //^{ALL}
'it's usually done like this, Marco / * dad'
%ill: reproach [ifamcv14]

*LUC: lui disse /^{INT} me la piglio io /$^{COM\text{-}r}$ nini /$^{ALL\text{-}r}$ * Giovanni //$^{ALL\text{-}r}$
'he said, I'm the one taking it, boy / * John'
%ill: reported speech [ifamcv22]

*MAX: queste son belle /^{COM} mamma /^{ALL} * cara //^{ALL}
'these are nice, mom / * darling'
%ill: appreciation [ifamcv01]

The rejection of vocative iteration has been observed also by Corr (2022). However, according to competence examples, the Author assumes there can be iterations, provided they are not contiguous. We observe that the restriction does not depend on the contiguity of the two vocative expressions but on their functional diversity: an illocutionary function of *Call* in the first CMM within a functional calling pattern and a second at the end of the second CMM with an Allocutive function. For instance, the real example below can be easily integrated with an Allocutive at the end of the pattern.

*ABC: Alberto /^{CMM} i' discorso è questo / ^{CMM} caro mio //^{ALL}
'Alberto ! This is the point, my dear!'
%ill: call + assertion (disagreement attitude) [ipubcv04]

7. *Conclusions*

Assuming the L-AcT frame, vocatives are identified as units of information whose prosodic performance predicts their different pragmatic functions. Vocatives can develop Comment units with illocutionary force and Allocutive units supporting the utterance.

Comment units are performed by *root* prosodic units shaped by various contours correlating with specific illocutionary forces. On the contrary, Allocutive units are always performed by a prosodic unit characterized by a flat-falling contour.

Comment units can occur in isolation, giving rise to simple speech acts (*Distal/ Proximal calls, Greetings, Regret, Protest*) or be patterned within a *Functional calling* structure. In the second case, vocatives serves as a *call* to share *the attentional horizon with the addressee*. When patterned, the vocative unit is focused in the first position.

Allocutive units are aimed at *strengthening the empathic relationship with the addressee.*

They cannot occur in isolation. By preference, they are placed at the end of the utterance after the Comment or rarely in the first position before the Comment. In both cases, they are performed by a flat or gently falling unit.

Accordingly, the vocative is not a single grammatical notion "in between system and performance" (Sonnenhauser, Noel, 2013), but it plays different pragmatic functions foreseen by competence and embodied by prosody.

The functional distinction among the usages of vocative sheds light on the semantic selection that applies in each case. When Vocatives develop an illocutionary type within the *Functional calling* pattern, they can be filled by a deictic pronoun and interpreted as referring to the second person, as is largely assumed to be a general property of vocatives in the literature.

Conversely, when vocatives develop the Allocutive function, their lexical selection is restricted to bare proper names, titles, appellatives, and role nouns. Deictics and referential phrases are banned in this function, and iteration is excluded. Our interpretation can explain this, considering that *strengthening the empathic relationship* is implemented by citing the addressee's designator rather than through a referential activity.

References

Abeillé, A., Godard, D. (Eds.) (2021). *La Grande Grammaire du français*. Paris: Acte Sud Imprimerie nationale.

Austin, J.L. (1962). *How to do things with words*. Oxford: Oxford University Press.

Bernstain, J., Ordóñez, F & Roca, F. (2019), On the emergence of personal articles in history of Catalan. In Bouzoita M. (Ed.), *Cycles in language changes*. Oxford: Oxford University Press, 88-108.

Bernstein, J. (forthcoming). On the DP status of vocative expressions in Romance. In *Proceedings of Incontro di Grammatica Generativa* 48, University of Florence, February 16-18, 2023.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *The Longman Grammar of Spoken and Written English*. London: Longman.

Borràs-Comes, J., Sichel-Bazin, R. & Prieto, P. (2015). Vocative intonation preferences are sensitive to politeness factors. In *Language and Speech* 58(1), 68-83.

Chafe, W. (1994). *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: The University of Chicago Press.

Coene, M., D'Hulst, Y. & Tasmowski, L. (2019). "Allez, (mon) chou, on y va!". Twenty Years Later: Revisiting the Puzzle of French Vocatives. In *Bucharest Working Papers in Linguistics* XXI, 2, 101-120.

Corr, A. (2022). *The Grammar of the Utterance*. Oxford: Oxford University Press.

Corr, A. (2022a). Address Inversion in Southern Italian dialects. In *Isogloss* 8(4), 1-37.

Cresti, E. (2000). *Corpus di italiano parlato*. Firenze: Accademia della Crusca.

Cresti, E. (2020). The Pragmatic Analysis of Speech and its Illocutionary Classification according to the Language into Act Theory. In Izre'el, S., Mello, H., Panunzi, A. & Raso, T. (Eds.).*In Search of Basic Units of Spoken Language: A Corpus-Driven Approach*. Amsterdam, New York: John Benjamins. 177-216.

Cresti, E. (2021). The Appendix of Comment according to Language into Act Theory: corpus-based research. In *CHIMERA*, vol. VIII, 46-69.

Cresti, E. (forthcoming). Topic vs. Allocutive in the Language into Act Theory. Corpus-based research on spoken Italian. In Cimmino, D., Ozerov, P. (Eds.), *Disentangling Topicality Effects*, Special issue, *Linguistic Typology at the Crossroad Journal*.

Cresti, E., Moneglia, M. (2023). The role of prosody for the expression of illocutionary types. The prosodic system of questions in spoken Italian and French according to Language into Act Theory. In *Frontiers in Communication, Sec. Psychology of Language*, 8, 1-28.

D'Alessandro, R., van Oostendorp, M. (2016). When imperfections are perfect. Prosody, phi-features and deixis in Central and Southern Italian vocatives. In Carrilho, E., Fiéis, A., Lobo, M. & Pereira, S. *Romance Languages and Linguistic Theory 10: Selected papers from 'Going Romance' 28*. Amsterdam: Benjamins, 61-82.

Frota, S., Prieto, P. (2015). *Intonation in Romance*. Oxford: Oxford University Press

Frosali, F. (2008). Il lessico degli ausili dialogici, In Cresti, E. (Ed.) *Prospettive nello studio del lessico italiano*. Firenze: FUP, 417-424.

Göksel, A., Pöchtrager, M.A. (2013) The vocative and its kin: marking function through prosody. In Sonnenhauser, B., Noel, A.H.P. (Eds.) *Vocative! Addressing between System and Performance*. Berlin: De Gruyter, 87-108.

't Hart, J., Collier, R., Cohen, A. (1990). *A Perceptual Study on Intonation. An Experimental Approach to Speech Melody*. Cambridge Cambridge: University Press.

Kipke, S. (1971). Identity and necessity. In Munitz M.K. (Ed.), *Identity and individuation*. New York: New York University Press, 135.164.

Ladd, R. (1978). Stylized Intonation. In *Language. Linguistic Society of America*, 54(3), 517-540.

Ladd, R. (2008). *Intonational Phonology*. Cambridge: Cambridge University Press.

Lambrecht, K. (1996). On the formal and functional relationship between topics and vocatives. Evidence from French. In Golberg, E. (Ed.) *Conceptual structures, discourse and language*, Chicago: Chicago University Press, 267-288.

Martin, P. (2011). Winpitch: a multimodal tool for speech analysis of endangered languages. *Proceedings of Interspeech* 2011, 3273-3276.

Mazzoleni, M. (1995). Il vocativo. In Renzi, L., Salvi, G. & Cardinaletti, A. (Eds.), *Grande Grammatica italiana di Consultazione* III. Bologna: Il Mulino, 377-402.

Micali, I. (2022). L'uso dei pronomi allocutivi tra pragmatica e sociolinguistica. In *Rivista DILEF,* 2, 258-276.

Moneglia, M., Raso, T. (2014). Notes on the Language into Act Theory. In Raso, T., Mello, E. (Eds.), *Spoken corpora and linguistics studies*. Amsterdam: Benjamins. 468-494.

Moro, A. (2003). Notes on Vocative Case: a case study in clause structure. In Quer, J., Schroten, J., Scorretti, M., Sleeman, P. & Verheugd, E. (Eds.), *Romance Languages and Linguistic Theory* 2001. Amsterdam: John Benjamins, 251-265.

Panunzi, A., Saccone, V. (2018). Complex illocutive units in L-AcT: an analysis of non-terminal prosodic breaks of bound and multiple comments. *Revista de Estudos da Linguagem,* 26, 1647-1674.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A grammar of contemporary English*. London: Longman.

Raso, T., Leite, F. (2010). Estudo contrastivo do uso de alocutivos em italiano português e espanhol europeus e português brasileiro. In *Dominios de lingu@gem*, 4(1), 151-174.

Raso, T. (2014). Prosodic constraints for discourse markers. In Raso, T., Mello, H. (Eds.), *Spoken corpora and linguistics studies*. Amsterdam: Benjamins, 411-467.

Raso, T., Ferrari L.A. (2020). Uso dei Segnali Discorsivi in corpora di parlato spontaneo italiano e brasiliano. In Ferrari R., Birello M. (Eds.), *La competenza discorsiva e interazionale*. Napoli: Aracne, 61-107.

Renzi, L., Salvi, G. & Cardinaletti. A. (Eds.) (1995), *Grande Grammatica italiana di Consultanzione III*. Bologna: Il Mulino.

Rizzi, L. (1997). The Fine Structure of the Left Periphery. In: Haegeman, L. (Eds.) *Elements of Grammar*. Kluwer International Handbooks of Linguistics. Dordrecht: Springer, 281-337.

Rohlfs, G. (1966). *Grammatica storica dell'italiano e dei suoi dialetti.* Torino: Einaudi.

Sonnenhauser, B., Noel, A.H.P. (Eds.) (2013). *Vocative! Addressing between System and Performance*. Berlin: De Gruyter.

Zwicky, A. (1974). Hey, what's your name! In La Galy, M., Fox, R.A. & Bruck, A. (Eds.), *Papers from the Tenth Regional Meeting of the Chicago Linguistics Society*, Chicago: Chicago Linguistics Society, 787–801.

DOI: 10.17469/O2112AISV000007

GIOVANNI VINCIGUERRA, GLENDA GURRADO, PATRIZIA SORIANELLO[1]

"Sei un idiota!". Observations on prosodic mock impoliteness in the Italian language

This research aims to explore some prosodic aspects of mock impoliteness, a mismatch between the semantic meaning of an utterance and its actual communicative intention. Mock impoliteness has always been presented as the opposite of genuine impoliteness: while the former is considered a kind of impoliteness that does not intend to cause any offence to the recipient, the latter is a communicative strategy that aims to threaten the face of the recipient.

This study intends to examine the phonetic and phonological cues of mock and genuine impoliteness, with a particular focus on how pitch range and mean intensity help distinguish between these two attitudes. The data showed that mock impoliteness is characterised by a less marked prosody than genuine impoliteness, with lower values for pitch range and mean intensity. Moreover, the two parameters are positively correlated. Finally, the phonological analysis showed that mock and genuine impoliteness have different distributions of their nuclear pitch accents.

Keywords: mock-impoliteness, prosody, pitch range, intensity.

1. *Defining mock impoliteness*

Mock impoliteness is a pragmatic strategy involving a mismatch between the semantic meaning of the utterance and its actual communicative intention. Studies on humour, irony and impoliteness were the first to show an interest in pragmatic mismatch. In fact, these three topics often overlap considerably, with Taylor (2016: 13) stating that "they may be seen as sharing an interest in similar interactional phenomena but viewing them from different perspectives and with different goals" (Taylor, 2016: 13).

The studies on mock impoliteness (it. *scortesia canzonatoria*) are part of the broader research on linguistic impoliteness. This field of research appeared for the first time in 1996, within the archetypal framework outlined by Jonathan Culpeper in "Towards an anatomy of impoliteness". Since the first theories, mock impoliteness has been presented as the opposite of genuine impoliteness. If the genuine attitude is associated with a series of communicative strategies aimed at attacking the face of the listener and thus causing social conflicts, the mock attitude is defined as a type

[1] The paper is the result of a continuing collaboration between the authors. However, for academic purposes: *Study concept and design*: GV, GG, PS; *Data collection*: GV, GG; *Data analysis*: GV, GG; *Writing original draft* and *Writing-review*: GV, GG, PS.

of impoliteness which does not intend to offend the recipient. It is "a particular interpersonal relation in which the interlocutor's face is not compromised and tends to reinforce participants' trust and affective feelings" (Bernal, 2008: 272-273). In this regard, Culpeper clarifies that:

> [...] we can think of mock impoliteness in theoretical terms as involving the cancelling of impoliteness perlocutionary effects flowing from a conventionalised impoliteness formula when an obvious mismatch emerges with the context it is used in. (Culpeper, 2011: 208)

In fact, Geoffrey Leech (1983) already tried to capture this kind of phenomenon with the Banter Principle, which is widely considered to be the ancestor of mock impoliteness. Leech was the first to adopt the meta-pragmatic label of "mock impoliteness"[2]. Illustrating the differences between the concepts of "irony" and "banter", the author summarises that "while irony is an apparently friendly way of being offensive (mock-politeness), the type of verbal behaviour known as 'banter' is an offensive way of being friendly (mock-impoliteness)" (Leech, 1983: 144). According to the Banter Principle, in certain communicative circumstances, the speaker can produce a message with a false and impolite locutionary habit in order to show solidarity with the interlocutor. The adequate interpretation of the speech act can only take place if the recipient, after understanding the real communicative intentions, associates them with a meaning of opposite polarity, i.e. polite. As the author points out, the lack of politeness can acquire an advanced pragmatic function, becoming a mark of intimacy, since "the ability to be impolite to someone in jest helps to establish and maintain such a familiar relationship" (*ibid*.).

Similarly, mock impoliteness also finds a place in the model proposed by Manfred Kienpointner (1997): according to Kienpointner, mock impoliteness fits into a continuous space that goes from politeness to rudeness, called "cooperative rudeness". The author states that it "is a technique for creating a relaxed atmosphere [...] especially if there is a little social distance between the participants"[3].

2. *The functions of mock impoliteness*

Over time, scholars have observed how the speaker's use of mock impoliteness serves three focused illocutionary purposes[4]. The first function associated with mock impoliteness is the promotion of camaraderie among interactants, which Culpeper (2011: 215) suggests generally "takes place between equals, typically friends, and

[2] Mock impoliteness has also been the subject of a debate over terminology and definitions. Culpeper (1996; 2005; 2011), Bousfield (2008) and Haugh & Bousfield (2012) – following Leech (1983) – adopt the label of "mock impoliteness"; Bernal (2008) proposes the label of "non-authentic impoliteness" or "non-genuine impoliteness" (*Sp. descortesía no auténtica*); Zimmerman (2003) employs the Spanish term *anticortesìa*.

[3] Kienpointner (1997: 262).

[4] Cf. Culpeper (2011), Haugh & Bousfield (2012), Taylor (2016).

is reciprocal". For example, the *insultes de solidarité* investigated by Lagorgette & Larrivée (2004) could be considered a typical example of an impolite speech act used by the speakers to reinforce solidarity with the listener. In the paper, the authors list the possible means by which the listener can recognise all these types of insults: i.e. the use of adjectives modifying the epithet[5], the use of diminutive suffixes, etc.[6]. Other researchers studying mock impoliteness have also identified a number of cues that help guiding the recipients towards the correct interpretation of the messages. For example, according to Haugh (2010: 2108), these include: lexical exaggeration, formulaicity, topic shift markers, contrastiveness, prosodic cues, inviting laughter and facial/gestural cues. Taylor (2016) infers that many of these cues are the same as those identified for irony and humour, and – given this overlap – she hypothesises that these cues signal a shift to a pretence mode.

Secondly, mock impoliteness is also used by speakers to conceal coercion: humour is used "to minimally disguise the oppressive intent, i.e. as a repressive discourse strategy" (Holmes, 2000: 176).

The final function of mock impoliteness is that it may be deployed to entertain the audience, as a form of exploitative humour that "involves pain for the target but pleasure for other participants" (Culpeper, 2011: 215). In these frameworks, mock impoliteness exhibits a typical instance of "disaffiliative humour": especially when it is performed in public contexts, "the speaker [...] and non-targeted receivers affiliate against the target and may derive humorous pleasure from the target being genuinely disparaged in a humorous (i.e. witty/creative) manner" (Dynel, 2021: 27).

The relationship between impoliteness and amusement is a recurrent topic in the pragmatic and communication literature. Culpeper (2005: 45), who explored this relationship using data from the television quiz show "The Weakest Link", suggests that there are four general factors behind this intimate relationship between impolite interactions and public entertainment: 1. intrinsic pleasure; 2. voyeuristic pleasure; 3. audience superiority; 4. audience safety. Moreover, he specifies that, in addition to these general factors, there may be some more specific causes: for example, literary genres thrive on conflict as a means of advancing the plot and characterisation, creating dramatic entertainment.

Specifically, this study aims to analyse the first type of strategy in order to observe the prosodic behaviour of mock impoliteness when it is used by the speakers to promote friendship and solidarity with their interlocutors.

[5] Lagorgette & Larrivée note speakers' frequent use of the adjective *petit* (En. *small*) as a mark of solidarity attitude (e.g. *Petit menteur!*, "Little liar!").

[6] The paper's final section also mentions two further contexts in which insults are not a product of human aggressivity: verbal duels and pornographic contexts.

3. *(Im)politeness and prosody*

The connection between (im)politeness and prosody has not received enough attention from researchers. This lack of research seems to be mainly due to the difficulty of finding spontaneous audio material.

Culpeper et al. (2003) is the first and foundational study on the contribution of prosody to the development of the impoliteness strategy. The authors assumed that "no utterance can be spoken without prosody and it is, therefore, desirable at some point to include this dimension of speech in pragmatic analysis" (Culpeper et al., 2003: 1568).

The main focus of the study is "the contribution not of *what* was said but of *how* it was said" (*ibid*.), since sometimes prosody alone is responsible for the impolite realisation of the utterances. Along with the experimental sections of the article, Culpeper and his co-authors show how negative and positive impoliteness strategies can only be implemented through prosody mediation. For example, the *negative impoliteness strategy* "invade auditory space" prosodically involves the role of pitch and intensity. In ordinary conversation, *etiquette* requires the speaker to match the pitch and the intensity of their voice to that of the interlocutor. In impolite interactions, the speaker – moved by the arousal of emotions – can increase the two prosodic indices, simulating an invasion of the listener's auditory space.

Prosody is not a marginal component: it contributes to differentiating polite behaviours from political and impolite ones, also from a perceptual point of view (e.g. Culpeper et al. 2003, Álvarez et al., 2011)[7]. In addition, prosody makes it possible to differentiate genuine impoliteness from mock impoliteness[8], thus allowing the speaker to mitigate or intensify the degree of (im)politeness of their utterances (cf. Culpeper, 2011a; Gili Fivela & Bazzanella, 2014).

3.1. The prosodic features of mock impoliteness

Due to their multifunctional nature, genuine and mock impoliteness could be difficult to differentiate in some conversational contexts. Interpreting the mock attitude correctly depends on an interactive process, in which both interlocutors identify the locutionary speech act, such as a false impolite literal formula.

However, the contextual and paralinguistic components play a pivotal role in the process of interpretation, because mock impoliteness – like other "mixed messages"[9] (Culpeper et al., 2017) – is the result of a multimodal clash between words, structures and non-linguistic resources.

---

[7] "[...] politic behavior can be defined as socio-culturally determined behavior directed towards the goal of establishing and/or maintaining in a state of equilibrium the personal relationships between the individuals of a social group, whether open or closed, during the ongoing process of interaction" (Watts, 1989: 135).

[8] Cf. McKinnon & Prieto (2014); Andreeva et al. (2016); Xu & Gu (2020); Vinciguerra & Gurrado (*forthcoming*).

[9] Culpeper et al. (2017: 323) refer to *mixed messages* as "messages that contain features indicative of a polite interpretation mixed with features indicative of an impolite interpretation". Other *mixed*

At present, the largest portion of studies focusing on the prosodic patterns of (im)polite interactions has promoted the investigation of the perceptual role played by prosody. Only recently have scholars begun to investigate the role played by specific prosodic indices aimed at creating mock or genuine nuances (Caballero et al., 2018).

Currently, there are very few studies aimed at comparing genuine and mock impoliteness. In this regard, McKinnon & Prieto (2014) examined the role played by prosody and gestures in the perception of mock impoliteness in Catalan. The authors focused on the intonation patterns and the acoustic properties of the utterances, elicited by the Oral Discourse-Completion Task, and integrated the results with a perceptual test. They observed that the prosodic features associated with genuine impoliteness were typically those associated with anger. This is not surprising, given that highly arousing emotions have previously been associated with genuine impoliteness[10]. Anger has been the main focus of research into the vocal expression of emotions. Most researchers have found that this emotion typically produces an increase in speech rate and in mean f_0 and a widening of the pitch range. Statistical tests revealed that intonation contours did not occur in significantly different proportions (p. >.05), although they found that there were no relevant differences in intonation contours between the two experimental conditions, which is plausible given the small size of the corpus, as the authors postulated. For both genuine and mock impoliteness, the most recurring nuclear configurations were L* L % and H+L* L %, which in Catalan can be used for the command (Prieto, 2014). The acoustic analysis showed that the pitch range and peak intensity were significantly higher in genuine situations. However, the difference between mock and genuine was not significant for syllable duration and the average intensity for the whole utterance, even if they were quantitatively higher for genuine utterances.

Further data in this regard comes from the study carried out by Andreeva et al. (2016) in two different languages. They collected 32 utterances each in German and in Polish (4 speakers x 4 utterances x 2 impoliteness conditions) to conduct a production and perception analysis based on McKinnon & Prieto's previous work. Six acoustic parameters were extracted using Praat, including mean f_0, pitch range, mean intensity and speech rate. Voice pitch and pitch range did not differ significantly between the two languages or the attitudes, although there was a slight tendency for the pitch range to be wider in the mock condition in both languages[11]. The f_0 standard deviation significantly differed in both languages, with higher values in the mock utterances than in genuine utterances. Regarding the mean intensity values, the Polish group spoke louder than the German speakers, and the utterances produced in the derogatory condition were significantly louder than

messages typically fall under such label are *sarcasm*, *banter*, *teasing*, *jocular mockery*, *jocular abuse*, *ritual insults*, *mock politeness*, *insincere* or *manipulative politeness*, *pushy politeness*, *under politeness*.

[10] Culpeper (2011: 148-149) stated that anger and disgust are "two emotions particularly relevant to impoliteness events".

[11] German: 10.84 ST (Genuine) vs 12.32 ST (Mock); Polish: 12.76 ST (Genuine) vs 13.52 ST (Mock).

those produced in the supportive group. The Polish group showed higher intensity values (+ 4.56 dB) in the genuine condition than in the mock one, whereas the German group was equally loud in both experimental conditions. In addition, both groups of speakers realised genuine utterances with a wider range of intensity than mock utterances. Statistically, the impolite attitude had a significant effect on speech rate; in fact, the German group spoke faster in the mock condition, while the Polish group had comparable rates in both conditions. As in McKinnon & Prieto, the authors analysed and labelled the nuclear pitch accents of each utterance according to ToBI conventions. In the German data 5 pitch accents were identified: H*, L*, H+L* L+H*, and H+!H*. Both high and low pitch accents were more or less equally distributed across the mock and genuine attitude[12]. H* was the most frequent pitch accent in mock and genuine utterances. The nuclear pitch accents used by Polish speakers were H*, ^H*, !H*, L+H*, L+!H*, H+!H*, and L*. In contrast to the German H* pitch accent, the Polish H* was characterised by a pitch peak which was typically reached early in the syllable, producing a falling pitch over the syllable. In contrast with the German group, Polish informers modified their intonation systematically with intended impoliteness conditions. Andreeva et al. (2016: 1002) conclude that speakers tend to use a combination of phonetic-prosodic and categorical intonation properties which they exploit to varying degrees.

In the same vein, Xu & Gu (2020) explored the prosodic configuration of mock and genuine (im)politeness in the Mandarin language, and identified a more defined trend. In order to perform a quantitative analysis of prosodic characteristics, they conducted a context-elicited discourse completion task so that they would collect genuine and mock (im)polite Mandarin utterances in both imperative and interrogative modes. For each target utterance, the researchers measured f_0, intensity, time and voice quality. The data revealed that mock impoliteness was characterised by lower f_0 and intensity values; furthermore, pitch range was wider in genuine impoliteness. Finally, while jitter and shimmer were lower in mock impoliteness, HNR values were higher.

The only study available on this subject in the Italian language confirms the presence of some differences between the two attitudes (Vinciguerra & Gurrado, *forthcoming*). The authors compared two samples of utterances performed by a group of male speakers of Italian (Bari). For each target sentence, the researchers measured the following parameters: $f_0$mean, $f_0$min, $f_0$max and pitch range, intensity mean, intensity min, intensity max and intensity range, the speech rate (calculated by dividing the utterance duration, including pauses, by the number of syllables actually produced by the speaker in the utterance), and the duration of the nuclear stressed vowel. The quantitative and statistical analysis revealed that f_0 mean, f_0 maximum value and pitch range were significantly lower in mock utterances. In a comparison of mock and genuine impoliteness, the mean, maximum and range of intensity were significantly higher in genuine attitude utterances. These trends

[12] However, H+L* L % was more frequent in genuine impoliteness than in mock impoliteness but this tendency is not statistically significant.

confirm the results of previous studies. Diversely, speech rate and the duration of nuclear vowels showed no significant behaviour.

4. *The research*

The present study is a part of this line of research, and aims to shed light on the prosodic cues that characterise mock impoliteness especially in comparison with genuine impoliteness. This reason for this focus is the lack of studies on mock attitude in Italian language.

On this basis, the present research aims to answer the following main questions:

1. To what extent does the mock/genuine attitude affect the prosody of an utterance? Do these two attitudes differ in terms of prosody?
2. What role do pitch range and intensity play in discriminating between genuine and mock impoliteness?
3. What is the nuclear distribution of pitch accents in mock and genuine impoliteness?

To date, research has focused on genuine impoliteness from different perspectives, leaving mock impoliteness in a marginal position. In particular, the prosodic dimension has not been particularly investigated. For this reason, we aimed to determine the role played by some acoustic parameters in the communication of mock impoliteness, and to identify the extent to which each of them characterises the attitude. In particular, we focused on pitch range and intensity because, in a previous research these two parameters seemed to best differentiate mock impoliteness from genuine impoliteness (Vinciguerra & Gurrado, *forthcoming*).

We also investigated whether the distribution of the pitch accents could be influenced by the different types of impoliteness conveyed by the utterance: in particular, we wanted to establish whether mock and genuine impoliteness differ significantly in terms of nuclear contours.

4.1 Participants

Six male speakers of Bari Italian, aged between 27 and 38, were invited to participate in the experiment. At the time of the recording session, they had no phonetic or phonological skills; they gave their written informed consent, although they were not informed of the aim of the study. The choice of involving only male speakers is motivated by the fact that some studies have found that men tend to use mock impoliteness more than women[13], in order to show camaraderie with each other. This type of miscommunication is preferred by men since adolescence, as a strategy to contrast with the adult world and to reinforce their belonging to a group

[13] Some sociolinguistic studies have shown that women tend to adopt a verbal interaction model based on politeness and the expression of feelings and emotions, while men prefer a more practical and referential communicative model (Lakoff, 1973).

(Zimmerman, 2003). As a consequence, it was hypothesised that boys would be able to perform an impolite sentence in a more natural way than girls, because they are more comfortable with irony and mock impoliteness[14].

4.2 Materials and recording session

12 texts were created for this research. Each of them described a contextual situation ending with an impolite declarative sentence, such as "You're an asshole!" (Pron+Verb+Adj). According to the contextual situation described, six of the target sentences were labelled "genuine impolite" while the other six were labelled "mock impolite".

In order to select the most common epithets used by Bari Italian speakers in everyday conversation, a preliminary test was conducted on a Google Form. We asked a group of 36 young men from Bari to read a series of scenarios, imagine themselves in the situation and complete the sentence with an epithet related to the scenario described. By means of this procedure, we could identify the epithets that best matched the spontaneous speech in order to construct the target sentences. The pre-test produced homogeneous results. Almost all the participants selected the same six epithets, which were used to design the role plays. The six epithets were: asshole (it. *stronzo*), jerk (it. *cretino*), moron (it. *deficiente*), dick (it. *coglione*), idiot (it. *idiota*) and *trimone*, a peculiar Apulian epithet of uncertain etymology, meaning stupid or similar.

The following are two examples of role plays:

Genuine attitude

You are coming back from work. You are in your car, you are very stressed, tired and hungry, waiting for the light to change. You take your eye off it for one second, and as soon as the light turns green someone behind you starts honking the horn madly. You get angry and exclaim:

"You're an asshole"

Mock attitude

You are coming back from work. You are in your car, waiting for the light to change. Suddenly, someone behind you starts honking the horn intoning a jingle. After a couple of seconds you realize that your friend is driving the car and he wants to mess with you. It's fun, so as soon as he approaches your car, you exclaim:

"You're an asshole"

The recording session took place in a quiet room, with the distance between the speaker and the microphone set at 20 cm to ensure a correct level of intensity. The speakers were asked to read the role plays as naturally as possible, identifying with

[14] However, this issue could feed a gender stereotype. For this reason, we aim to perform the same test with female speakers in the future.

the situation described. First, they had to read the scenarios in silence, then they could read out the texts and each related target sentence.

The digital recording was run by means of a TASCAM DR-40, frequency sampling of 44kHz/24-bit resolution. The recording session provided 72 target sentences (36 genuine impolite sentences and 36 mock impolite sentences).

4.3 Acoustic, phonological and statistical analysis

The stimuli were analysed using PRAAT (6.3.10), taking into account the variables employed in this field of study:

– pitch range (ST)_PR
– mean intensity (dB)_INT

In order to explore the trends that emerged from the analysis, we extracted the mean (M) and the standard deviation (SD) of the following parameters: $f_0$mean (f_0x), $f_0$maximun ($f_0$max), Intensity max (Imax), Intensity Range (IR) and Total Duration (TD).

The phonological analysis was based on the autosegmental and metrical theory, using the ToBI transcription system. It aims to verify the nature of the nuclear pitch accents of the two pragmatic attitudes.

Statistical significance was assessed by means of the Paired-Samples T-Test (p. <.05). The correlation between pitch range and intensity was also tested using Pearson's correlation coefficient.

4.4 Results

The data revealed that the prosody of mock impoliteness is less marked than that of genuine impoliteness. As shown in Table 1, with reference to the pitch range, the difference between the two attitudes is about 1.7 ST, while the mean intensity difference is 3 dB.

Table 1 - *Pitch range and Intensity mean differences between Mock (M) and Genuine (G) Impoliteness*

	M/G mean differences	*t*	*df*	*p*
PR	- 1.67 ST	- 2.797	35	<0.01
INT	- 3 dB	- 3.047	35	<0.01

With reference to the Pearson test, our data showed a positive correlation between pitch range and mean intensity values in both mock (r = .465; p. <.01) and genuine impolite sentences (r=.478; p. <.01).

In order to investigate in more detail the comparison between these two utterances realisations, other prosodic parameters were measured. The results demonstrated that mock impoliteness is characterised by lower values of f_0x, $f_0$max, Imax, IR and TD, as shown in Table 2. The data suggest that, in this case, all prosodic parameters contribute to the differentiation between mock and genuine impoliteness. For this reason, some phonetic parameters were examined. First of all, the Onset was higher in genuine than in mock impoliteness (respectively: M=144 Hz, SD=24; M=127 Hz, SD=13; p. <.01). In addition, the initial verb also seemed to be influenced by the attitude, specifically the duration of the first consonant was longer in the 73 % of the mock impolite assertives considered (mock: M=167 ms, SD=61; genuine: M=118 ms, SD=37; p <.01).

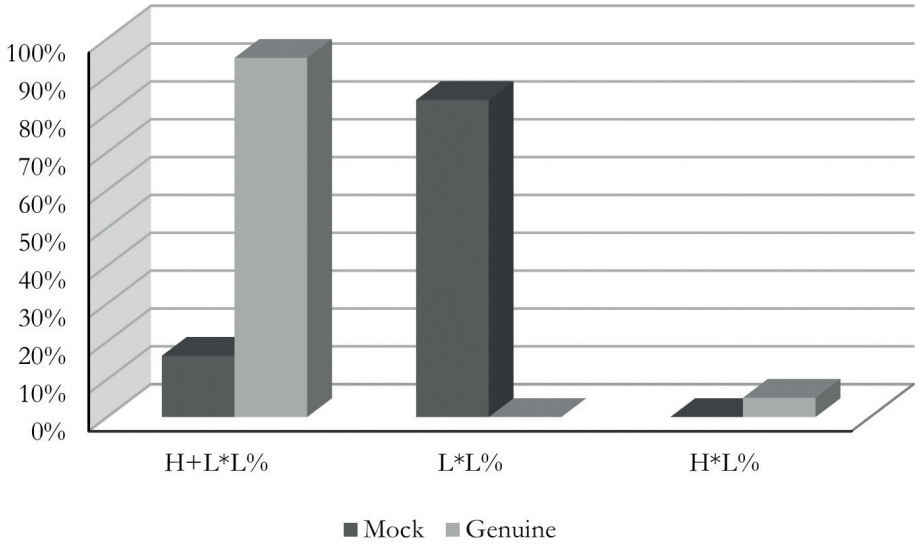
Table 2 - *f_0, Intensity and Duration mean differences between Mock and Genuine Impoliteness*

f0x	*f0max*	*Imax*	*IR*	*TD*
- 3 ST	- 1.8 ST	- 9.1 dB	- 8.4 dB	+ 130 ms

The intonation contours of the attitudes also showed clear patterns. The analysis was carried out to verify the distribution of nuclear pitch accents in the two attitudes under consideration. As can be seen in Fig. 1, the mock impoliteness mainly presents a L* L % (83.78 %) contour and a very small percentage of H+L* L % nuclear pitch configuration (16.2 %). This aligns with the emphatic nature of genuine impolite sentences and with the intonation patterns of Bari Italian[15]. By contrast, the mock impolite sentences show a flat nuclear contour in which the lack of prominence indicates that the attitude is not intended to have a literal meaning. This intonation contour, combined with a longer duration of nuclear vowels and a soft voice quality, often accompanied by a hidden smile, is enough to give the mock sentences an oblique and indirect meaning that differentiates them from neutral declarative sentences.

[15] There are numerous studies on the intonation of Bari Italian. Among them we refer to Gili Fivela et al. (2015) for its comprehensive and comparative nature.

Figure 1 - *Distribution of nuclear pitch accents in mock and genuine impoliteness*



The distribution of nuclear pitch accents indicates a more dynamic f_0 profile in genuine impoliteness assertives than in mock impoliteness ones. As an example, Figs. 2 and 3 show the spectrogram, the waveform and the orthographic transcription of the same impolite declarative sentence (*Sei un coglione!*/*You're an asshole!*) pronounced by the GM speaker with a mock (Fig. 2) and a genuine (Fig. 3) attitude, respectively. It can be noticed that the genuine impolite sentence is more dynamic and has an H+L* L % pitch contour, while the mock impolite sentence shows a L* L % configuration. In terms of pitch range, the difference between the two sentences is 4.2 ST. As for the intensity, the speaker used this parameter as a cue to distinguish between the two utterances, with a difference of 9.25 dB.

Figure 2 - *Mock Impoliteness GM_19 M "You're an asshole"*

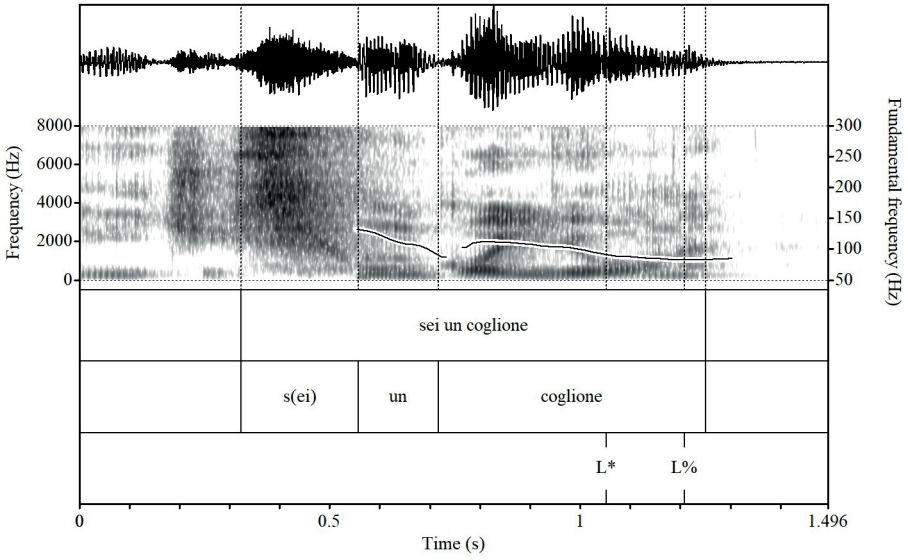
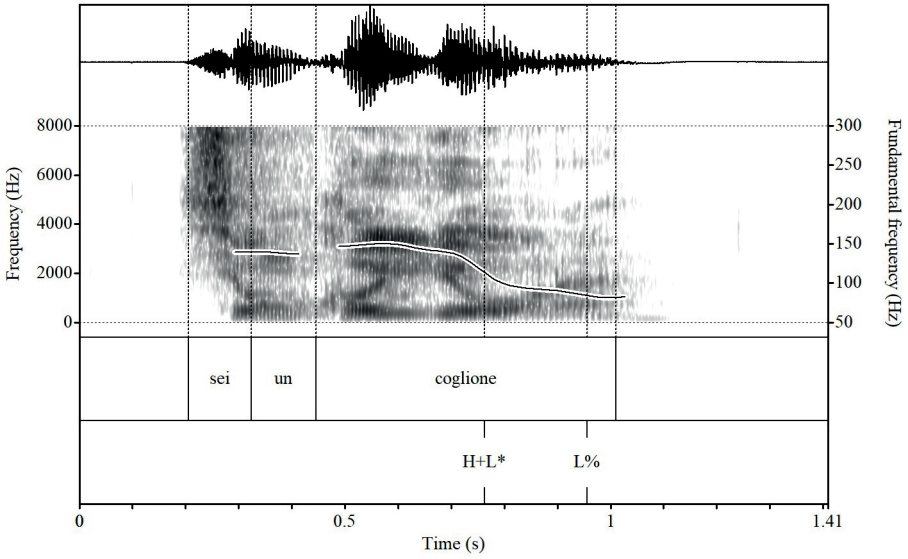


Figure 3 - *Genuine Impoliteness GM_14 G "You're an asshole"*



5. *Discussion and conclusions*

The present study focused on mock impoliteness, a form of indirect communication that has received less attention than genuine impoliteness. Specifically, mock impoliteness has not been considered in any study of the Italian language, particularly regarding the influence of prosody, which is believed to play a key role in the production and perception of impoliteness. Given this, we intended to detect the prosodic cues that characterise mock impoliteness, with the aim of identifying the strategies used by speakers to communicate this particular kind of attitude.

This goal required a comparison between mock and genuine impoliteness. While the former consists in the communication of an attitude that allows the speaker to attack the face of the interlocutor in a more or less voluntary way, the latter can be considered as a particular type of impoliteness that is not intended to offend the recipient, but is used to tease the interlocutor in a gentle way.

The experiment was designed around two main research questions. The first one was to determine whether the impolite attitude affected the prosody of utterances. The results were positive. In fact, speakers seemed to use different prosodic strategies for each attitude, which means that the two types of impoliteness do affect the suprasegmental level of the utterance. Pitch range and mean intensity showed a significant increase in genuine attitude more than in mock impoliteness. This tendency, supported by the results of the correlation test, answered the second research question. The data revealed that pitch range and intensity are positively correlated. As a consequence, it is not possible to affirm which one of the two parameters is more involved in the communication of impoliteness. This result contributes to the definition of the prosody of mock impoliteness. However, with regard to genuine impoliteness, these data seem to align with Culpeper et al. (2003), who found that the marked prosody usually noticed in genuine impoliteness is quite similar to that of highly arousing emotions. For instance, in some contexts, genuine impoliteness is accompanied by anger, an emotion that is generally expressed by an increase in f_0, intensity and speech rate (cf. Ekman, 2003; Scherer, 2003). Culpeper et al. (2003) connect the parallel increase in the pitch range and mean intensity parameters with the aforementioned "invade auditory space" strategy used by speakers to communicate the impolite attitude. Through this strategy, the speakers distance themselves from the prosodic level that is usually shared with the interlocutor, with the aim of prosodically trespassing on a higher and more appropriate level underlining the unfriendly intention. We hypothesise that even in the case of mock impoliteness, the speakers' need to mark the communication of the impolite intention is manifest in the positive correlation between the two parameters. A comparison with neutral speech, purged of any attitude or emotions, could help prove this hypothesis in the future.

In the light of the results collected so far, it can be affirmed that, in our study, genuine impoliteness is characterised by a significantly more marked prosody than mock impoliteness, in terms of pitch range and intensity.

The present research also investigated the intonation tendencies that characterise the sentences. Our data revealed two distinct and prevalent intonation contours related to the two attitudes analysed. Mock impoliteness tended to have a L* L % nuclear pitch profile, while genuine impoliteness is mostly characterised by the dynamic H+L* L % contour. Data are only partially consistent with McKinnon & Prieto (2014), who found that, with respect to Catalan, L* L % and H+L* L % are the most recurrent nuclear configurations in both genuine and mock impoliteness. Specifically, the H+L* L % configuration is twice as frequent in the genuine impoliteness utterances, whereas the L* L % configuration is equally distributed. The phonological tendencies of our study suggest that the speakers do use different pitch nuclear contours to differentiate between the two attitudes conveyed.

Furthermore, mock and genuine impoliteness initially to show some differences in the first part of the sentence corresponding to the verb ('be') and in the epithet. In mock impoliteness, the epithet is usually characterised by a low pitch and no dynamic pattern. On the contrary, genuine impolite assertive utterances seem to have a high-fall pattern, accompanied by a higher scaling between the two tonal targets. This cue associates the genuine epithet with a prominence and with an increase in intensity from a perceptual point of view. The analysis of other phonetic features also showed a higher onset of the f_0 curve and a longer duration of the verb (especially of its initial consonant) in mock impoliteness than in the genuine impoliteness. This tendency, which needs to be deeply investigated in the future, seems to suggest that the speaker intends to signal his mocking intention to the hearer from the first syllables of the utterance. This hypothesis is confirmed by some speakers who tend to associate the lengthening with a kind of "smiled speech" (Émond & Laforest, 2013; Barthel & Quené, 2015). In particular, speaker AI used this kind of strategy: in mock impoliteness, he deliberately produced a temporal lengthening of the verbal phrase, which is usually accompanied by a smile/laugh, and by a following lowering of f_0 and intensity values in the epithet. This can be considered as a means of highlighting the non-coincidence between the real intention of the utterance and its literal form. In this way, by means of the smile/laugh the hearer is immediately aware of the mocking attitude of the speaker and the epithet is completely freed from its conventional meaning. However, this strategy needs to be investigated in a further analysis.

In the light of the results collected so far, we would like to broaden the sample, by also including female speakers, in order to avoid the influence of gender stereotypes. Furthermore, it would be advisable to perform a comparison with the neutral speech (free of any emotion or attitude), with the aim to identify the strategies used by speakers to differentiate the mock impoliteness from a speech that is far from any impolite intention.

In the future, a perception test could be useful for analysing how prosodic parameters contribute to encoding the speaker's mocking or genuine insulting attitude.

References

Abelin, Å. (2017). Impolite prosody in Swedish and the importance of context. In Bączkowska, A. (Ed.), *Impoliteness in Media Discourse*. Wien: Peter Lang, 13-26.

Álvarez, A., Blondet, M.A. & Darcy, R. (2011). Sobre (des)cortesía y prosodia: una relación necesaria. In *Oralia*, 14, 437-450.

Andreeva, B., Bonacchi, S. & Barry, W. (2016). Prosodic cues of genuine and mock impoliteness in German and Polish. In *Proceedings of Speech Prosody 2016*, Boston, USA, 31 May-3 June 2016.

Barthel, H., Quené, H. (2015). Acoustic-phonetic properties of smiling revised – Measurements on a natural video corpus. In *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, Scotland, 2015.

Bernal, M. (2008). Do insults always insult? Genuine politeness versus non-genuine politeness in colloquial Spanish/¿Insultan los insultos? Descortesıa autentica vs. descortesía no auténtica en el español coloquial. In *Pragmatics*, 18(4), 775-802.

Bousfield, D. (2008). *Impoliteness in Interaction*, Philadelphia and Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.167

Brown, P., Levinson, S. (1987). *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Caballero, J.A., Vergis, N., Jiang, X. & Pell, M.D. (2018). The sound of im/politeness. In *Speech Communication*, 102, 39-53. https://doi.org/10.1016/j.specom.2018.06.004

Culpeper, J. (1996). Towards an anatomy of impoliteness. In *Journal of Pragmatics*, 25 (3), 349-367. https://doi.org/10.1016/0378-2166(95)00014-3

Culpeper, J. (2005). Impoliteness and entertainment in the television quiz show: 'The Weakest Link'. In *Journal of Politeness Research*, 1(1), 35-72. https://doi.org/10.1515/jplr.2005.1.1.35

Culpeper, J. (2011). *Impoliteness: using language to cause offence*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511975752

Culpeper, J., Haugh, M. & Sinkeviciute, V. (2017). (Im)politeness and Mixed Messages. In Culpeper, J., Haugh, M. & Kádár, D.Z. (Eds.), *The Palgrave handbook of linguistic (im)politeness*, London: Palgrave Macmillan, 323-356. https://doi.org/10.1057/978-1-137-37508-7_13

Dynel, M. (2021). Desperately seeking intentions: Genuine and jocular insults on social media. In *Journal of Pragmatics*, 179, 26-36. https://doi.org/10.1016/j.pragma.2021.04.017

Ekman, P. (2003). *Emotions revealed. Understanding faces and feelings*. London: Weidenfeld and Nicolson.

Émond, C., Laforest, M. (2013). Prosodic correlates of smiled speech. *Proceedings of Meetings on Acoustics, ICA 2013*, Montreal, Canada, 2-7 June 2013. https://doi.org/10.1121/1.4799490

Gili Fivela, B., Avesani, C., Barone, M., Bocci, G., Crocco, C., D'Imperio, M.P, 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. https://dx.doi.org/10.1093/acprof:oso/9780199685332.003.0005

Gili Fivela, B., Bazzanella, C. (2014). The relevance of prosody and context to the interplay between intensity and politeness. An exploratory study on Italian. In *Journal of Politeness Research*, 10/1, 97-126. https://doi.org/10.1515/pr-2014-0005

Hardaker, C. (2017). Flaming and trolling. In Hoffmann, C., Bublitz, W. (Eds.), *Pragmatics of Social Media*. Boston: De Gruyter, 493-522. https://doi.org/10.1515/9783110431070

Haugh, M. (2010). Jocular mockery (dis)affiliation and face. In *Journal of Pragmatics*, 42(8), 2106-2119. https://doi.org/10.1016/j.pragma.2009.12.018

Haugh, M., Bousfield, D. (2012). Mock impoliteness, jocular mockery and jocular abuse in Australian and British English. In *Journal of Pragmatics*, 44, 1099-1114.

Holmes, J. (2000). Politeness, power and provocation: how humour functions in the workplace. In *Discourse Studies*, 2, 159-185. https://doi.org/10.1016/j.pragma.2012.02.003

Kienpointner, M. (1997). Varieties of rudeness. Types and functions of impolite utterances. In *Functions of Languages*, 4(2), 251-287. https://doi.org/10.1075/fol.4.2.05kie

Lagorgette, D., Larrivée, P. (2004). Interprétation des insultes et relations de solidarité. In *Langue française*, 144, 83-103.

Lakoff, R. (1973). Language and woman's place. In *Language in Society*, 2/1, 45-80.

Leech, G. (1983). *Principles of Pragmatics*. London: Longman.

McKinnon, S., Prieto, P. (2014). The role of prosody and gesture in the perception of mock impoliteness. In *Journal of Politeness Research*, 10, 185-219. https://doi.org/10.1515/pr-2014-0009

Mills, S. (2011). Discursive approaches to politeness and impoliteness. In Linguistic Politeness Research Group (Ed.), *Discursive Approaches to Politeness*. Berlin and Boston: De Gruyter Mouton, 19-56. https://doi.org/10.1515/9783110238679.19

Murray, I.R., Arnott, J.L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotions. In *The Journal of the Acoustical Society of America*, 93(2), 1097-1108.

Prieto, P. (2014). The intonational phonology of Catalan. In Jun, S. (Ed.). *Prosodic typology 2. The phonology of intonation and phrasing*. Oxford: Oxford University Press, 43-80. https://doi.org/10.1093/acprof:oso/9780199567300.003.0003

Scherer, K.R. (2003). Vocal communication of emotion: a review of research paradigms. In *Speech communication*, 40, (1-2), 227-256. https://doi.org/10.1016/S0167-6393(02)00084-5

Taylor, C. (2016). *Mock Politeness in English and Italian*. Philadelphia and Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.267

Vinciguerra, G., Gurrado, G. (*Forthcoming*), I tratti prosodici della scortesia: un'indagine pilota sulla mock impoliteness. In Roccaforte, M. (Ed.), *La Comunicazione Parlata/Spoken Communication*. Rome: Aracne.

Watts, R.J. (1989). Relevance and relational work: Linguistic politeness as a political behaviour. In *Multilingua*, 8, 131-166.

Xu, C., Gu, W. (2020). Prosodic characteristics of genuine and mock (im)polite Mandarin utterances. *Prooceedings of Interspeech 2020*, Shangai, China, 25-29 October 2020.

Zimmerman, K. (2003). Constitución de la identidad y anticortesía verbal entre jóvenes masculinos hablantes de español. In Bravo, D. (Ed.), *La perspectiva no etnocentrista de la cortesía: Identidad sociocultural de las comunidades hispanohablantes. Actas del Primer Coloquio del Programa EDICE*, Estocolmo, Suecia, Septiembre, 2002, CD-ROM, 47-59.

DOI: 10.17469/O2112AISV000008

RICCARDO ORRICO, STELLA GRYLLIA, NA HU, AMALIA ARVANITI

The role of metrical structure in prominence perception

Despite the increased interest on prominence, there is still lack of consensus of what prominence is. A recent review (Ladd, Arvaniti, 2023) has pointed out that the work conducted so far has painted a contradictory picture: on the one hand, some studies report that prominence assessment changes as a function of acoustic properties of words, pointing to the interpretation of prominence as a universal phenomenon; other evidence, however, suggests that the role of prominence is determined by phonology, therefore language-specific. We report a prominence study in Greek with the aim of testing the hypothesis that phonological factors linked to metrical strength trump acoustic salience in the assessment of prominence. The results met our predictions and revealed that accented words are more likely to be selected as prominence than unaccented words, regardless of their acoustic properties and, crucially, even when accented words are less salient than unaccented words.

Keywords: Prominence, Metrical structure, Greek, Rapid Prosody Transcription.

1. *Introduction*

The last decades have witnessed an increased interest in the study of prosodic prominence. Such interest may have been encouraged by the development of experimental paradigms that have facilitated the investigation of prominence and other prosodic phenomena. A popular paradigm used in prominence studies is Rapid Prosody Transcription (henceforth RPT, Cole, Mo & Baek, 2010a; Cole, Mo & Hasegawa-Johnson, 2010b; see Cole, Shattuck-Hufnagel, 2016, for an overview). In RPT, linguistically untrained participants are asked to listen to utterances and mark prominent words and the location of boundaries on a transcript of those utterances. The researchers then examine the factors contributing to word prominence based on the characteristics of the words most commonly chosen by participants. Specifically, the results are analyzed by considering inter-rater agreement (typically, Fleiss' Kappa) and calculating p-scores, the percentage of participants who marked a word as prominent (a similar measure is also used to analyze the properties boundary placement, b-scores, see Cole et al., 2010a). The linguistic correlates of prominence are statistically validated by correlating participants' responses with the properties of individual words. The ultimate aim of many RPT studies is to test how likely a word is to be selected as a function of the properties of that word. Some studies used p-scores as the outcome variable of interest for their statistical testing, such as Cole et al. (2010b) who performed correlation analysis between p-scores and properties of the signal in American English, or Baumann and Winter (2018) who used p-scores to perform a random forest analysis to investigate the relative importance

of phonetic, phonological, and lexical cues on p-score variation in German. Other studies have used the raw binary response prominent/not prominent for individual words as the dependent variable of Generalized Linear Mixed-Effect Models (among others, Baumann, Winter, 2018; Cole, Hualde, Smith, Eager, Mahrt & De Souza, 2019; Bishop, Kuo & Kim, 2020; Arvaniti, Gryllia, Zhang & Marcoux, 2022; Im, Cole & Baumann 2023; Orrico, Gryllia, Kim & Arvaniti, 2023; 2025; Orrico, Gryllia, Hu, Kim & Arvaniti, 2024).

These studies have provided useful insight into the perception of prominence and have shown that prominence perception is affected by several parameters, including but not limited to acoustic and phonological factors (see also Wagner, Origlia, Avesani, Christodoulides, Cutugno, D'Imperio, Escudero, Gili Fivela, Lacheret, Ludusan, Moniz, Chasaide, Niebuhr, Rousier-Vercruyssen, Simon, Šimkz, Tesser & Vainio, 2015 on this point). Amplitude and duration of words have been strongly correlated to the perception of a word as prominent; this has been shown for American English (Cole et al., 2010b; 2019; Bishop et al., 2020; Im et al., 2023), German (Baumann, Winter, 2018; Riesberg, Kalbertodt, Baumann & Himmelmann, 2020), French and Spanish (Cole et al., 2019), and Italian (D'Imperio, 1997). The role of F0 in prominence perception has also been shown for different languages (for German, Baumann, Winter, 2018; for English, Bishop et al., 2020; Im et al., 2023; for English, Spanish, and French, Cole et al., 2019), though some studies have reported that its contribution is weaker than other acoustic variables (Cole et al., 2010b and Kochanski, Grabe, Coleman & Rosner, 2005 for American and British English respectively). Other prosodic factors include the presence of a pitch accent, which has been considered a strong prominence predictor in English (Bishop et al., 2020; Im et al., 2023) and German (Baumann, Winter, 2018) and the type of pitch accent, with low accents being considered less prominent than high ones, and high accents being considered less prominent than rising ones (see Baumann, Winter, 2018 for evidence in German, and Arvaniti et al., 2022 and Orrico et al., 2023, 2025 for evidence in British English). Finally, prominence perception is also affected by non-prosodic cues. Recent studies show the effect of pragmatics: Orrico et al. (2025) reported the effect of contrast on word selection in British English; Im et al., 2023 reported the effect of givenness and correction in American English. Other factors include word frequency (Cole et al., 2010b; Baumann, Winter, 2018) and syntactic properties (Calhoun, Wollum & Kruse Va'ai, 2019). Some studies show that the relative use of prosodic and non-prosodic cues for prominence selection during RPT depends greatly on individual variability. Orrico et al. (2025) reported that, in British English, the selection of a word as prominent varies as a function of both accents' shape and function, though individual variability has a role in defining which one of the two criteria is more important. Similar findings have also been reported for German: while both prosodic and non-prosodic factors have a role in predicting prominence, individuals may have different preferences towards the criteria used, with some preferring prosodic and others non-prosodic cues (Baumann, Winter, 2018).

1.1 Prominence beyond acoustic salience

The studies reported in the previous section have unveiled crucial aspects of the processes behind prominence perception; however, a clearer understanding of what constitutes prominence is still lacking (cf. Ladd, Arvaniti, 2023). Prominence is often defined very vaguely, as any property of a word that makes it stand out compared to others (see the definitions in Cole et al., 2010b and Bauman, Winter, 2018). The lack of a restrictive definition is partly the reason why researchers have followed exploratory approaches (correlation between responses and properties of words) rather than testing specific hypotheses. Ladd and Arvaniti's (2023) pointed out that recent results provide a contradictory picture of prominence rather than clarifying its properties. As reported above, some findings correlate prominence to gradient dimensions of the signal, like F0, duration, and amplitude, and suggest a straightforward relationship between the two. In other words, prominent words are expected to have specific features: they are longer, louder, or have higher pitch than non-prominent words. Additionally, when considering the role of pitch accent, factors linked to F0 shape (e.g. the phonetic difference between a L* and a L+H*) have been argued to play a crucial role in determining relative prominence, thereby indicating the importance of acoustic salience over phonological structure. This points towards a universal concept of prominence.

In contrast, other studies suggest that the role of prominence is language-specific. We can indeed find evidence that such a view of prominence – as pertaining to words that are acoustically more salient than others – does not apply across languages. Riesberg et al. (2020) report an RPT study in which German and Papuan Malay listeners marked prominent words in both German and Papuan Malay. Inter-rater agreement on prominence selection revealed that German listeners showed good levels of agreement in both languages, while Papuan Malay listeners showed i) higher agreement when rating prominence in German than in their native language and ii) lower agreement than the German participants when rating Papuan Malay. In addition, the cues used by Papuan Malay speakers were consistent with those used by the German ones when rating prominence in German. These results indicate that while prominence rating in German might be based on universally recognizable cues that make words acoustically salient (both German and Papuan Malayan listeners had similar intuitions about German), the same set of cues does not define prominence in Papua Malay. In fact, while German listeners recognized those cues as prominence lending also in Papuan Malay, Papuan Malayan listeners could not agree on what prominence is in their own language, showing that they do not have the same notion of prominence as German speakers. Related to this point, Cole et al. (2019) report an RPT investigation on English, Spanish, and French under different instruction conditions, with one group being given a definition of prominence based on acoustic factors and another being given a definition based on pragmatic criteria. The results show that while all three languages were affected by the instructions, the effect was larger in French when compared to English. This outcome is most likely related to typological differences between French and

English: French is a language that lacks stress, and in which pitch accents are not prominent-lending and are not used to signal focus (Welby, 2006; Cole et al., 2019) which may have led the French speakers to adhere more closely to the instructions.

Additionally, the investigation of prominence using other paradigms has suggested that the link between prominence and acoustic cues is mediated by phonological structure. Rump and Collier (1996) report two experiments in Dutch on the relationship between relative pitch height and focus. In the first experiment, participants were asked to adjust the F0 height of the pitch accents in the utterance *Amanda gaat naar Malta* 'Amanda goes to Malta', bearing one pitch accent on *Amanda* and another on *Malta*; the aim was to have participants adjust the F0 contour so that it corresponded to different focus conditions (broad focus, corrective focus on *Amanda*, *Malta*, or both). In the second experiment, listeners were asked to map pitch contours with the focus condition. Overall, the results showed that while the presence of a corrective focus on *Amanda* required a high peak on that word and the absence of f0 movement on *Malta*, a corrective focus on Malta did not require that *Amanda* had flat f0; in fact, the adjustments created for that condition still had a visible movement on *Amanda*. Such a relationship between focus and the relative F0 height on the two words cannot be explained within a framework in which prominence is directly linked to F0. On the contrary, this shows the role of metrical structure on the interpretation of prominence: the first utterance (focus on *Amanda*) is interpreted as having a nuclear accent (*Amanda*), which is followed by a deaccented *Malta*; the second utterance (focus on *Malta*) has a prenuclear accent on *Amanda*, which is followed by the nuclear accent on *Malta*. Thus, the interpretation of prominence does not depend on which of the two words has the highest pitch but rather on which word bears the metrically strongest element, i.e. the nuclear pitch accent (cf. Ladd, Arvaniti, 2023).

These findings, together with methodological investigations of the RPT paradigm, cast doubts on the purported connection between acoustic salience and prominence perception. Indeed, while it is true that RPT has been proven to be a robust paradigm (see, for example, Orrico et al. 2025 which showed that RPT results can be replicated, both in relation to the role of phonetics and pragmatics in prominence assessment and in relation to the individual variability in the responses), some studies have reported the relationship between RPT results and the type of task being used. As mentioned above, the instructions can have an effect on the way listeners approach the task (Cole, Mahrt & Hualde, 2014; Cole et al.; 2019). Additionally, Orrico et al. (2024) showed that RPT responses are also affected by the number of words that listeners are allowed to select. They reported two RPT studies in Greek, one in which they asked participants to mark as many prominent words as they saw fit (following the traditional paradigm, Cole et al., 2010a; 2010b), and one in which they asked participants to mark only the most prominent word (a task that had been tested by Calhoun et al., 2019 in English and Samoan). While the general picture of prominence did not change across tasks, the effect of acoustic cues on prominence selection was much smaller when listeners

were allowed to select only the most prominent word; e.g., F0 height did not affect responses. Such task-dependent results require that findings on prominence should be interpreted with caution.

In sum, recent studies have painted a picture of prominence that is connected to the acoustic properties of the signal. While this is a result that is reported in several studies employing RPT, therefore robust, the investigation of non-Germanic languages (e.g. Cole et al., 2019; Riesberg et al., 2020), as well as evidence collected using other paradigms (e.g. Rump, Collier, 1996), challenges such a view of prominence. One possible factor that can explain the results that do not fit the standard understanding of prominence is offered by Ladd and Arvaniti (2023). These authors highlight the need of taking into account language-specific metrical structure in the definition of prominence: prominent words are those that are parsed in metrically strong positions and not those that are the most salient from an acoustic perspective.

1.2 Overview of Greek intonation

In this section we provide a brief overview of Greek intonation couched within the Autosegmental-Metrical (AM) framework (Pierrehumbert, 1980; Ladd, 2008). We describe the tunes frequently used for questions and statements, showcasing that metrically strong words are not necessarily acoustically more salient than weaker words (e.g. unaccented words), as in polar questions with non-final focus. Most of these tunes were used in the current study.

In statements with broad focus, the default tune is (L*+H) H* L-L %, with the nuclear pitch accent H* being realized as a high and declining plateau on the last content word (Baltazani, 2003; Arvaniti, Baltazani 2005; Lohfink, Katsika & Arvaniti, 2019). All content words before the nucleus carry a L*+H prenuclear accent (Arvaniti, Ladd & Mennen, 1998). In statements with narrow focus, the nuclear pitch accent is L+H* followed by L-L % edge tones (Arvaniti, Baltazani, 2005; Arvaniti, Ladd & Mennen, 2006; Georgakopoulos, Skopeteas, 2010; Lohfink et al., 2019). As in broad focus statements, all content words before the nucleus have a L*+H prenuclear accent, while words following the nucleus are typically deaccented.

Greek polar questions are string identical to declaratives and are distinguished from them only by intonation. The default tune for polar questions is (L*+H) L* H-L %; all content words preceding the nuclear accent bear a L*+H accent, and words following the nucleus are typically deaccented. The nuclear L* accent is realized as a low F0 stretch on the stressed syllable of the word in focus. The edge tone configuration is realized as a rise-fall movement, whose alignment depends on the location of the focus. When the word in focus is the last word of the utterance, the L* accent is realized on the stressed syllable of that word, and the H- phrase accent is realized on its last syllable followed by the L % boundary tone. When the word in focus is not the final word of the question, the H- phrase accent is realized on the *stressed syllable* of the final word (Grice, Ladd & Arvaniti, 2000;

Arvaniti 2002). Note that polar question tunes with non-final focus represent a mismatch between metrical strength and acoustic salience, as opposed to narrow focus statements in which the two dimensions cooccur on the same word. The focused element in questions is indeed low in F0, while the final word, despite being unaccented, shows a rise-fall movement (H-L %) on the stressed syllable (where typically pitch accents are realized) that is acoustically salient. Despite its resemblance to a pitch accent, native speakers distinguish between the H- phrase accent and phonetically comparable accentual rises, like the L+H* of narrow focus statements (Arvaniti et al., 2006).

Finally, wh-questions are realized with the nuclear pitch accent associated with the stressed syllable of the wh-word, which is fronted, followed by a dip or low F0 stretch (depending on the length of the question), followed by the boundary tone, realized on the last vowel of the question (among others Baltazani, 2002; Baltazani, 2003; Arvaniti, Baltazani, 2005; Arvaniti, Ladd, 2009; Grice et al., 2000; Baltazani, Gryllia & Arvaniti, 2020). Different tonal configurations have been attested, depending on the pragmatic function of the question. The most frequent tonal patterns for wh-questions include either a L*+H pitch accent in nuclear position, typically followed by a L-H % combination of edge tones, or a L+H* pitch accent typically combined with flat L-L % edge tones. Fig. 2 in § 2.3 provides examples of some of the contours that have been discussed above and were used as stimuli.

1.3 Research question

The present study aims to investigate the role of metrical structure on prominence assessment. We focus on prominence at the sentence level (prominent words in an utterance) and do not consider lexical prominence (prominent syllable in a word); though see Ladd (2008: Ch. 2 and Ch. 7) and Ladd and Arvaniti (2023) for a discussion about the relationship between lexical and sentence prominence. Our study investigates the relationship between sentence prominence assessment in RPT and phonological cues linked to the presence and type of accents. The study does not follow the traditional use of RPT (i.e. investigate all possible properties of a word that make it prominent) but uses the paradigm to test the hypothesis that prominence perception reflects the metrical organization of the utterance regardless of the differences in the phonetic realization of prosodic events. Our study is couched within the AM theory of intonation, in which the metrical structure (presence or absence of an accent, whether that accent is nuclear or prenuclear) and the phonetic realization of tonal events are orthogonal dimensions. This allowed the design of stimuli in which metrical structure and acoustic cues do not always match (see differences in questions and statements in § 1.2) and tested the prediction that the selection of a word as prominent would be determined by whether that word is metrically strong (accented or not), regardless of what accent it bears (e.g. L* or L+H*) or its acoustic salience.

2. *Methods*

2.1 Participants

Twenty-three Greek speakers participated in the study, and responses of four participants were excluded because three of them self-reported a history of language- or hearing-related disorders and one selected more than 85 % of the words in more than 10 % of the stimuli, suggesting carelessness during the task.

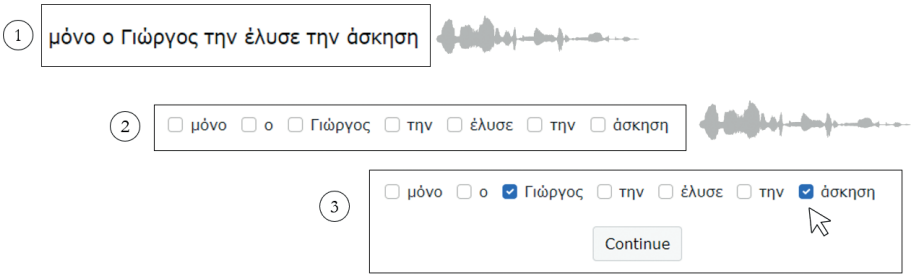
Therefore, the analysis involved the responses from 19 participants (12 females, 5 males, 1 non-binary, and 1 preferred not to say; ages ranged from 20 to 61 years, Median = 24.0, Mean = 31.8). They were all native speakers of Greek, brought up in Greece. They were recruited via personal networks and participated voluntarily.

2.2 Procedure

The study was conducted online using ROLEG (Radboud Online Linguistic Experiment Generator, https://www.roleg.nl/), a web platform developed at the Centre for Language Studies (CLS) at Radboud University. The participants were provided with a web link that they used to access the study page; they were asked to carry out the task in a quiet environment, listen carefully to the stimuli via headphones, and regulate the volume to a comfortable level; they were also encouraged to contact us in case of questions about the study or of technical problems. They were asked to select the words that in their opinion "stand out from the rest, that is, sound more important or stressed or as if the speaker has emphasized them" by checking a box next to that word. The wording of the instructions was chosen to avoid biasing participants towards focusing on either acoustic or semantic aspects of the words (cf. Cole et al., 2014; 2019). Participants were instructed to select as many words as they saw fit.

The study started with three practice trials followed by 74 experimental trials and included four prompts for self-paced breaks. Trial randomization was done by creating three lists containing all the stimuli in different orders; the lists were randomly assigned to participants. In each trial, participants listened to an utterance twice while seeing its transcript on screen; the transcript lacked punctuation (other than apostrophes in contractions) and capitalization (except for proper names). During the second repetition, a checkbox appeared next to each orthographic word, as shown in the schematic representation of a sample trial in Fig. 1.

Figure 1 - *Schematic representation of experimental trials. Participants first heard the utterance while seeing a transcription (1), then they heard a repetition of the same utterance and a selection box appeared next to each word to select prominent words (2)-(3)*

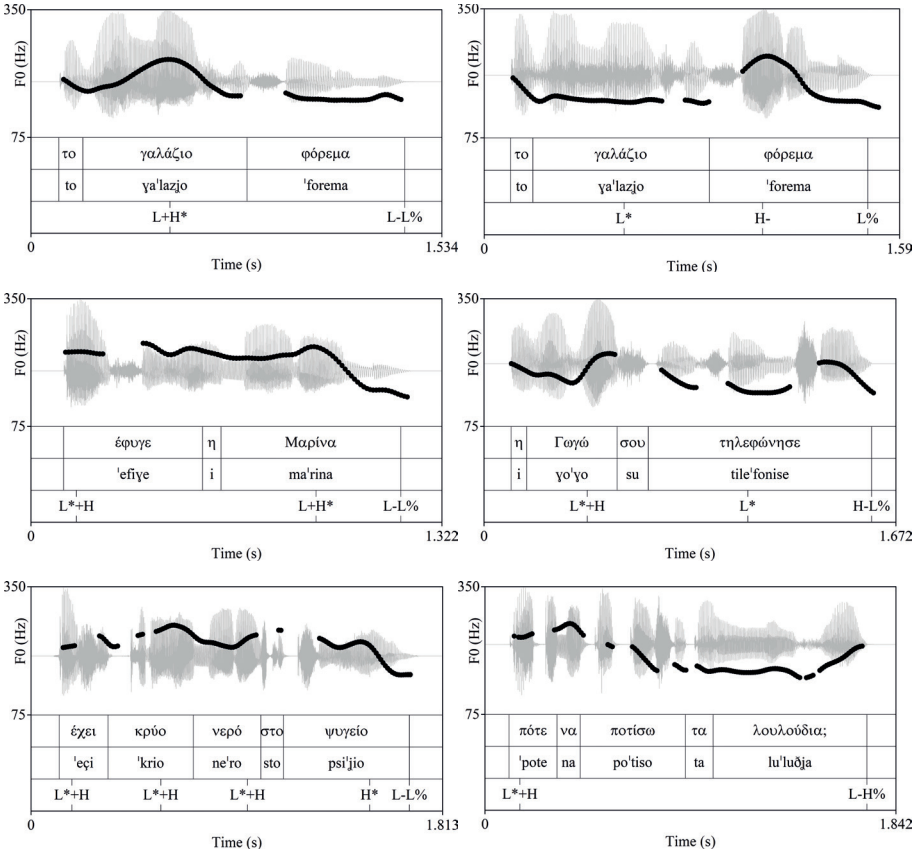


2.3 Stimuli

The stimuli were 74 utterances recorded from one female native speaker of Greek. They varied in length (2-8 orthographic words, 277 in total), sentence type, and focus position, all consisting of one intonational phrase with only one intermediate phrase; their tunes (pitch accents and edge tones) reflected those that are most typically used for the sentence types according to previous descriptions of Greek intonation (see § 1.2). After recording the stimuli, two authors listened to and inspected the utterances to ensure they were produced with the intended tune, focus location, and phrasing. The stimuli included the following utterance types, which were chosen to have stimuli with different types of nuclear accents. See Fig. 2 for examples.

- Narrow focus statements with final (10 items) and non-final focus (16 items); they were realized with a L+H* nuclear accent on the focused word, a prenuclear L*+H on all preceding content words, and L-L % edge tones. In items with final focus, the nuclear pitch accent was realized on the last content word in the utterance; in items with non-final focus, the nuclear pitch accent was realized on a non-final word, with the following content words being deaccented.
- Polar questions with final (10 items) and non-final focus (16 items), realized with a L* nuclear accent on the focused word, a L*+H on all preceding content words, and H-L % edge tones. In items with final focus, the nuclear accent was realized on the stressed syllable of the last content word, while the H-L % edge tones were realized on the last syllables of that word; in non-final focus items, the nuclear accent was realized on non-final words, while the H- rise was aligned with the stressed syllable of the last word.
- Broad focus statements (16 items), realized with a H* nuclear accent on the final word, L*+H prenuclear accents on all preceding content words, and L-L % edge tones.
- Information-seeking wh-questions (6 items), realized with a nuclear L*+H accent on the sentence-initial wh-word and L-H % edge tones.

Figure 2 - *Sample stimuli for the following sentences (from top-left to bottom-right): i) non-final narrow focus statement [to ɣaˈlaʒjo ˈforema]* 'The blue dress'*; ii) polar question with non-final focus [to ɣaˈlaʒjo ˈforema]* 'The blue dress'*; iii) narrow-focus statement with final focus [ˈefiɣe i maˈrina]* 'Marina left'*; iv) polar question with final focus [i ɣoˈɣo su tileˈfonise]* 'Did Gogo call you?'*; v) broad focus statement [ˈeçi ˈkrio neˈro sto psiˈʝio]* 'There's cold water in the fridge'*; and vi) wh-question [ˈpote na poˈtiso ta luˈluðja]* 'When should I water the flowers'



Here we report an analysis of these utterances that shows how the acoustic properties of the words relate to their phonological status; the analysis will confirm that metrical strength and acoustic salience do not always correlate (see § 1.3). We used Praat (Boersma, Weenink, 2023) to extract $F0_{max}$, duration, and Root-Mean-Square (RMS) amplitude from the stressed syllables of all words in all utterances and fitted three models in R, each having the acoustic value as the dependent variable, Accent (unaccented, prenuclear L*+H – henceforth pL*+H, – H*, L+H*, L*, L*+H) as the fixed effect and Item (the utterance) as random intercept. Results from multiple comparisons are shown in Tables 1 to 3 and plotted in Fig. 3. The results for duration showed that unaccented words were shorter than accented ones, though not shorter than L*+Hs; among accent types, differences in duration were

found between H*s and prenuclear accents and L*+Hs, with H*s being longer than both; additionally, L+H*s were longer than L*+Hs. In the case of $F0_{max}$, as expected, L* accents were the ones with the lowest values and they were significantly lower than all other categories, including unaccented words. Unaccented words also had low F0 max values, though not lower than H*s. The model for RMS showed that L+H* accents had the highest amplitude values; unaccented words did not have significantly lower amplitude than all accent types, but only L+H*s and pL*+H.

Overall, the results confirmed the arguments reported above for the analysis of target stimuli, showing that accented words are not necessarily more acoustically salient than unaccented words; additionally, it shows that different accents are not equally salient from an acoustic perspective.

Table 1 - *Pairwise comparison for the effect of Accent on duration*

contrast	*estimate*	*SE*	*t*	*p*
unaccented – pL+H*	-0.54	0.14	-3.82	<.01
*unaccented – H**	-1.30	0.24	-5.50	<.0001
*unaccented – L+H**	-1.14	0.19	-5.96	<.0001
*unaccented – L**	-0.93	0.19	-4.87	<.0001
unaccented – L+H*	0.05	0.37	0.13	n.s.
pL+H – H**	-0.76	0.25	-3.00	<.05
pL+H – L+H**	-0.60	0.21	-2.82	n.s.
pL+H – L**	-0.39	0.21	-1.83	n.s.
pL+H – L*+H*	0.59	0.38	1.55	n.s.
H – L+H**	0.16	0.28	0.57	n.s.
H – L**	0.37	0.28	1.30	n.s.
H – L*+H*	1.35	0.43	3.16	<.05
L+H – L**	0.21	0.25	0.84	n.s.
L+H – L*+H*	1.19	0.40	2.94	<.05
L – L*+H*	0.98	0.40	2.43	n.s.

Table 2 - *Pairwise comparison for the effect of Accent on F0 max*

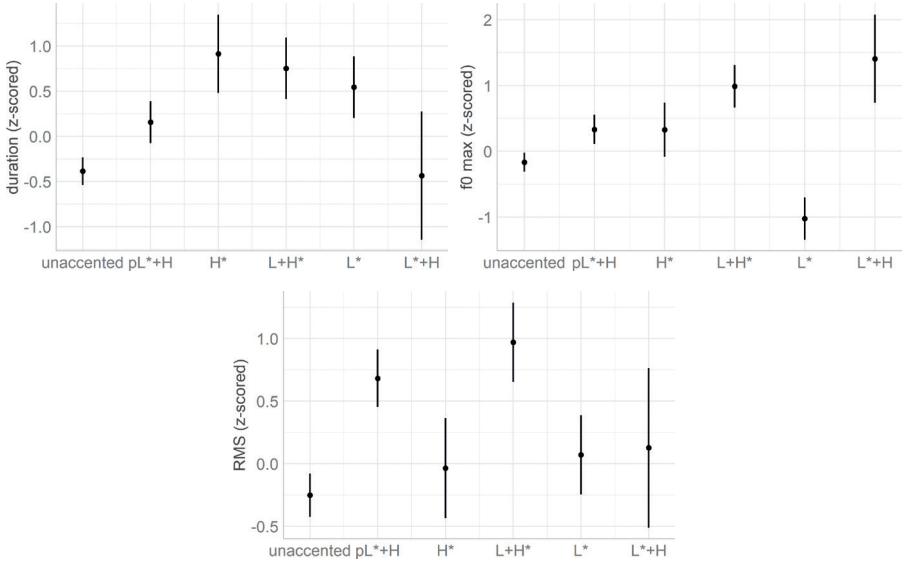
contrast	*estimate*	*SE*	*t*	*p*
unaccented – pL+H*	-0.50	0.14	-3.70	<.01
*unaccented – H**	-0.49	0.22	-2.22	n.s.
*unaccented – L+H**	-1.15	0.18	-6.37	<.0001
*unaccented – L**	0.85	0.18	4.72	<.001
unaccented – L+H*	-1.57	0.35	-4.47	<.001
pL+H – H**	0.00	0.24	0.02	n.s.
pL+H – L+H**	-0.65	0.20	-3.27	<.05
pL+H – L**	1.35	0.20	6.78	<.0001
pL+H – L*+H*	-1.07	0.36	-2.96	<.05
H – L+H**	-0.66	0.27	-2.46	n.s.
H – L**	1.35	0.27	5.04	<.0001

contrast	*estimate*	*SE*	*t*	*p*
H – L*+H*	-1.08	0.40	-2.67	n.s.
L+H – L**	2.01	0.23	8.59	<.0001
L+H – L*+H*	-0.42	0.38	-1.10	n.s.
L – L*+H*	-2.43	0.38	-6.35	<.0001

Table 3 - *Pairwise comparison for the effect of Accent on RMS amplitude*

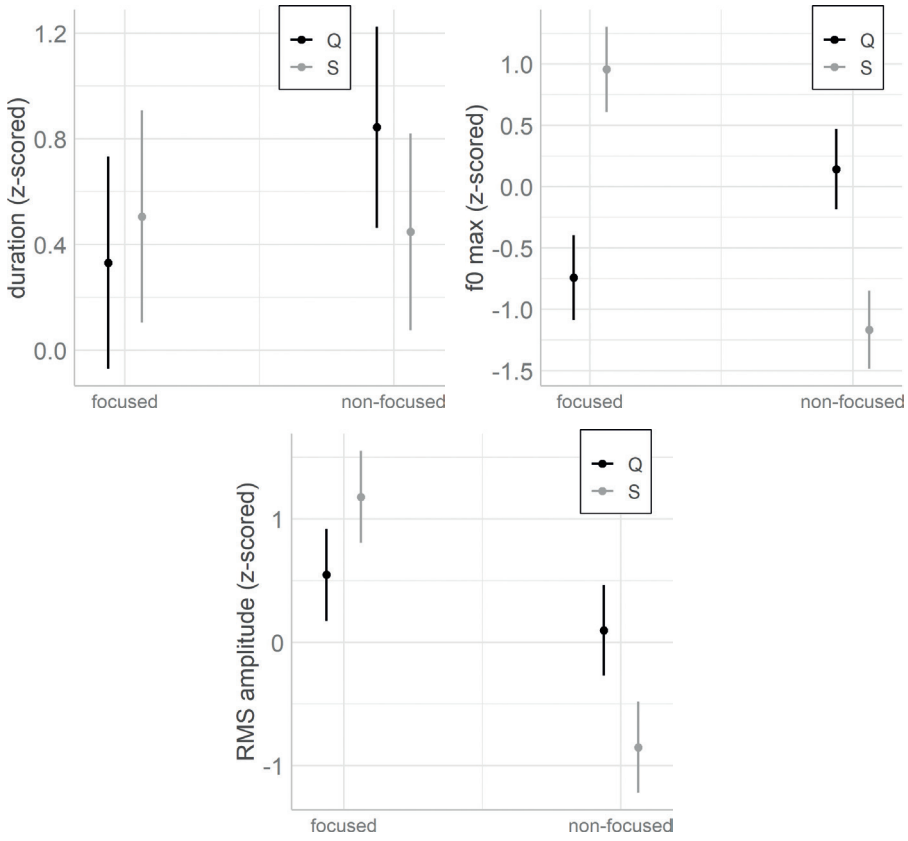
contrast	*estimate*	*SE*	*t*	*p*
unaccented – pL+H*	-0.94	0.13	-7.46	<.0001
*unaccented – H**	-0.22	0.21	-1.03	n.s.
*unaccented – L+H**	-1.22	0.17	-7.31	<.0001
*unaccented – L**	-0.32	0.17	-1.92	n.s.
unaccented – L+H*	-0.38	0.33	-1.17	n.s.
pL+H – H**	0.72	0.22	3.35	<.05
pL+H – L+H**	-0.29	0.19	-1.55	n.s.
pL+H – L**	0.61	0.19	3.31	<.05
pL+H – L*+H*	0.56	0.34	1.63	n.s.
H – L+H**	-1.01	0.26	-3.93	<.01
H – L**	-0.11	0.26	-0.41	n.s.
H – L*+H*	-0.16	0.38	-0.43	n.s.
L+H – L**	0.90	0.22	4.03	<.01
L+H – L*+H*	0.84	0.36	2.35	n.s.
L – L*+H*	-0.06	0.36	-0.16	n.s.

Figure 3 - *Duration, F0$_{max}$, and RMS amplitude values as a function of presence and type of accent, See Tables 1, 2, and 3 for information about statistical significance*



An additional set of analyses was performed considering only narrow focus statements and polar questions with non-final focus. As mentioned in § 1.3, these two utterance types represent a clear case of match (statements) and mismatch (questions) between metrical strength and acoustic salience, therefore they can provide crucial insights for the understanding of prominence. Their acoustic analysis was done by extracting duration, $F0_{max}$, and Root Mean Square (RMS) amplitude from the stressed syllable of the word in focus (the word bearing the nuclear accent) and the stressed syllable of the word in final position. The values of these measures were z-scored and used as the dependent variables in individual Linear Mixed-Effect Models (LMMs) in R (R Core Team, 2021); all three models had Focus (Focused, Non-focused), Modality (Question, Statement), and their interaction as fixed effects and item as random intercept; we report the model output for the interaction effect, which is the most informative in relation to our design. The models showed that duration was not affected by Focus or Modality (Focus × Modality: Est. = .57 (.39), t = 1.46, p > .1; see Fig. 4). For $F0_{max}$, the interaction between Focus and Modality was significant (Est. = 3.01 (.34), t = 8.81, p < .0001); specifically, a pairwise comparison showed that, in statements, the focused word had higher F0 max than the non-focused word (Est. = 2.12 (.24), t = 8.76, p < .0001), while in questions the focused words had lower $F0_{max}$ than their non-focused counterparts (Est. = .88 (.24), t =3.62, p < .001), see the middle panel in Fig. 3. Finally, there was a significant interaction effect for amplitude (Est. = 1.58 (.31), t = 5.15, p < .0001): amplitude was higher in focused words, but the effect was larger in statements than questions (Est. = 2.03 (.22), t = 9.27, p < .0001 and Est. = .45 (.22), t = 2.08, p < .05, respectively); see Fig. 4. In sum, these analyses confirm the prediction that in polar questions with non-final focus, acoustic salience and metrical strength do not match.

Figure 4 - *Duration, F0$_{max}$, and RMS amplitude as a function of focus × modality. All comparisons were significant except for duration*



2.4 Statistical analysis of the RPT responses

The RPT responses were analyzed using Generalized Linear Mixed-Effect Models (GLMMs) in R with the packages *lme4* (for the model fitting, Bates, Maechler, Bolker & Walker, 2015) and *emmeans* (for pairwise comparisons, Lenth, 2023). The analysis was conducted by running two models. The first model tested the binary RPT responses (whether a word was selected or not as prominent) as a function of Phonological status of words: unaccented, prenuclear accented, nuclear accented. The second model tested the binary responses as a function of Nuclear accent type: H* (broad focus statements), L+H* (narrow focus statements), L* (polar questions), L*+H (wh-questions). The random structure in both models included random intercepts for Speaker and Item.

We additionally ran a third model including only statements and polar questions with non-final focus, as they represented clear cases of matches and mismatches between metrical strength and acoustic salience. For this analysis we only included words bearing the nuclear pitch accent and those bearing the edge tones. Specifically,

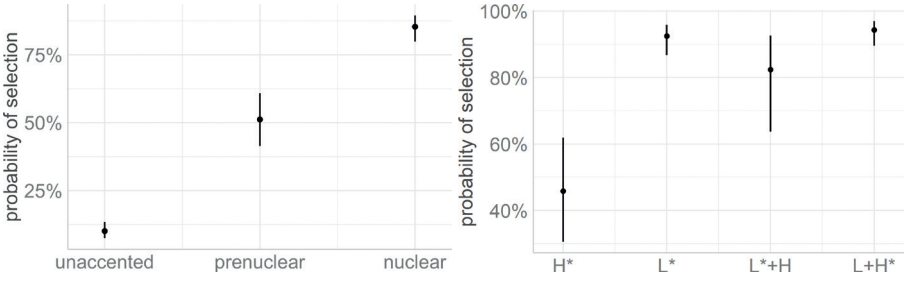
we analyzed responses for nuclear accented (L+H* and L*) and final words (L-L % and H-L %) in these two utterance types. The model was built with the binary response as dependent variable, and Modality (Statement, Question), Focus status (Focused, Non-focused), and their interaction as fixed effects; the random structure included random intercepts for Speaker and Item.

3. *Results*

The results of two GLMMs fitted for the aggregate RPT responses largely met our predictions. The first model tested the selection of a word as a function of the phonological status of the word (i.e. unaccented, prenuclear accented, nuclear accented) and showed that unaccented words are less likely to be marked as prominent than both prenuclear accented words (Est. = 2.18 (.20), z = 11.16, p < .0001) and nuclear-accented words (Est. = 3.83 (.19), z = 20.01, p < .0001). Additionally, nuclear-accented words were more likely to be selected than words bearing prenuclear accents (Est. = 1.64 (.22), z = 7.58, p < .0001).

The second model tested the effect of accent type in nuclear position. The only accent that was rated significantly different than the others was H*, the nuclear accent in broad focus statements. Specifically, H*-bearing words were less likely to be selected as prominent than words accented with L* (Est. = 2.66 (.37), z = 7.20, p < .0001), L*+H (Est. = 1.71 (.54), z = 3.20, p < .01), and L+H* (Est. = 2.98 (.39), z = 7.69, p < .0001). There were no differences among the other accents L* – L*+H: Est. = .96 (.52), z = 1.84, p > .1; L* – L+H*: Est. = .30 (.35), z = .85, p > .1; L*+H – L+H*: Est. = 1.26 (.53), z = 2.37, p > .05). The outputs of the two models are plotted in Fig. 5.

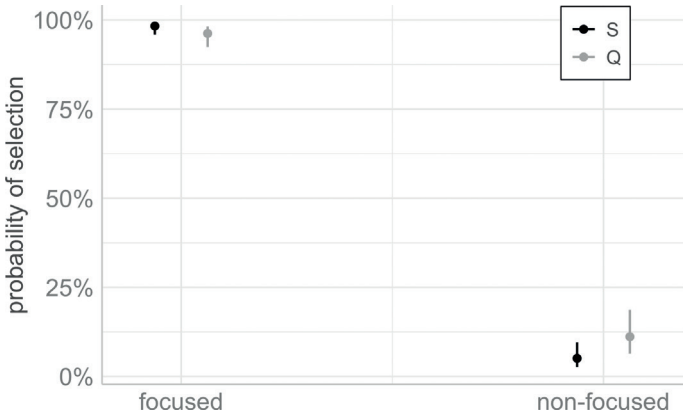
Figure 5 - *The left panel shows the probability of selection as predicted as a function of phonological status; all comparisons were significant; the right panel shows the same probability as a function of nuclear accent type; only H* in comparison with all other accents was significant, the other comparisons were not*



3.1 Statements and polar questions with non-final focus

The GLMM fitted for the statements and questions with non-final focus showed that, in line with our predictions, focused words were more likely to be selected than non-focused words, regardless of modality (for Statements: Est. = 6.98 (.49), z = 14.18, p < .0001; for Questions: Est. = 5.23 (.37), z = 14.21, p < .0001). The interaction between focus status and modality was significant (Est. = 1.75 (.54), z = 3.27, p < .01), and the pairwise comparison showed that the responses only differed for unaccented words in questions (bearing H-L %) vs statements (bearing L-L %), with higher probability of selection in questions than statements (Est. = .94 (.28), z = 3.31, p < .001). Crucially, no difference was found between L* and L+H* (Est. = .81 (.45), z = 1.80, p > .05). These effects are shown in Fig. 6.

Figure 6 - *Probability of selection as predicted by the GLMM for focused and non-focused words in questions (black) and statements (grey). All comparisons were statistically significant, except for the difference between Q and S in the focused condition*



4. *Discussion*

We reported an RPT study conducted on Greek with the aim of investigating the relationship between prominence assessment and phonological information linked to metrical structure. We hypothesized that the selection of a word as prominent would be determined by its metrical status and not by the phonetic shape of the accent it bore or how salient it sounded. To test this hypothesis, we designed a set of stimuli comprising both matching and mismatching cases between metrical strength and acoustic salience.

In general, our predictions were largely met. RPT responses varied as a function of the presence of pitch accent rather than specific acoustic cues. The general analysis showed that the likelihood of selecting a word aligned with the metrical organization of the utterances: unaccented words were the least likely to be selected, while nuclear-accented words were the most likely, with prenuclear-accented words falling in between. This outcome, accompanied by the fact that unaccented words

were not necessarily shorter, less loud, or lower in pitch than accented ones (see § 2.3), supports our predictions on prominence selection.

When looking at statements and questions with non-final focus, this result becomes even clearer. We showed that in polar questions with non-final focus, the word in focus is acoustically less salient than the final word, as opposed to statements, in which salience and metrical strength overlap. Specifically, the L+H*s in statements were always followed by a deaccented and less salient (much lower in pitch and amplitude) word, while the L* accents in questions were always followed by an unaccented word that was of comparable duration and amplitude but all had much higher pitch. The RPT results for questions, however, showed that the nuclear-accented words were the most likely to be selected in the utterance, despite not being the most salient; in contrast, the final word bearing the H-L % edge was much less likely to be selected than the accented words, despite being acoustically more salient. By and large, the size of the effect of focus condition on the probability of word selection was the same for the two modalities (see § 3.1 and Fig. 6), indicating that neither the presence nor the magnitude of the effect changed as a function of differences in acoustic characteristics between accented and unaccented words.

This shows that when Greek listeners assessed the prominences of these utterances, they did not select the words that stood out by virtue of being acoustically more salient, but interpreted the role of the words in the metrical organization of the utterance. As reported in § 1.1, similar results are reported in Rump and Collier (1996), in which Dutch listeners interpreted the focus of an utterance by considering the metrical strength of the words rather than their relative F0 height. This is in line with our view of prominence as determined by metrical strength rather than acoustic salience. This is also compatible with previous studies showing that the interpretation of prominence is mediated by the phonological representation of that utterance (cf. Ladd, Arvaniti, 2023).

Another important result is that L*s had the same probability of selection as L+H*s and that, more generally, the phonetic shape of a pitch accent was not an important drive for prominence selection (contrary to results reported in, among others, Baumann, Winter, 2018, and Cole et al., 2019). Indeed, despite differences in amplitude, F0 height, and duration, the effect of pitch accent type was very weak: the pairwise comparison showed that all nuclear accents were highly likely to be selected, with the only exception being H*. As a matter of fact, the likelihood for H* to be selected as prominent was the lowest among all accents, including the prenuclear L*+H. Though unexpected, this result is not entirely incompatible with our view of prominence. One possible explanation for this outcome could be related to the function that these accents have in the discourse. As reported in § 1.2, H* is the nuclear accent in broad focus statements, where the whole utterance represents new information conveyed to the addressee. In some frameworks, such utterances are considered to lack focus altogether (Katz, Selkirk, 2011). This represents a crucial difference between H*s and other accents. While accents like

L* or L+H* indicate narrow focus, narrowing down the focus to the accented word, H* accents have the function of presenting every word in the utterance as equally important. Furthermore, the fact that H* accents were not less salient than other accents also shows that their low likelihood of being selected does not simply depend on information related to the signal and is therefore linked to the way those accents are processed at a higher-order level. Such mechanisms involving the meaning of the accents have also been found in previous studies investigating other languages. Orrico et al. (2025) found that H* accents are less prominent than L+H* accents in British English, with a clear difference between the two observed only when the former had a non-contrastive function and the latter a contrastive function, confirming the importance of cues beyond the acoustic details. Similarly, D'Imperio's (1997) investigation of prominence perception in Neapolitan Italian showed that in narrow focus statements, the word bearing the nuclear pitch accent was consistently interpreted as the most prominent word in the utterance, but the interpretation of prominence in broad focus statements was around chance level, which is consistent with our results. Crucially, the fact that these results about prominence perception in Greek relate to results collected for other languages suggests that these mechanisms are not specific to Greek, but are part of prominence perception across languages.

We also found that acoustic salience did matter to some extent: a slightly higher probability of selection was observed for unaccented words with a H- phrase accent. Both the size of this effect and the fact that it was observed only for unaccented words show that the role of acoustic cues is subordinate to that of metrical strength. Additionally, this effect might also be explained as epiphenomenal, linked to methodological facts, like the type of stimuli used or the type of task we adopted. As mentioned in the introduction, RPT responses can be influenced by task details. Orrico et al. (2024) showed that when participants were asked to select *all* prominent words (as in the current study), the effect of acoustic variables (such as F0 height) was larger than when the participants were asked to select only the most prominent word. The fact that the effect of these acoustic variables is small and susceptible to changes in the task suggests that they are not the main drive for prominence selection.

In conclusion, the investigation reported here, following the lead of Ladd and Arvaniti (2023), provides empirical evidence in support of a view of prominence that considers acoustic elements of utterances as cues to prominence only by virtue of the fact that they are part of well-formed intonation contours, associated to the text according to language-specific metrical rules that ultimately define which words are prominent.

Acknowledgements

This work was supported by the European Research Council through grant ERC-ADG-835263 (SPRINT) to Amalia Arvaniti. We thank our participants for

voluntarily participating in the study, Vasiliki Kyritsi and Vasiliki Barouta for their assistance with data collection, and Bob Ladd and Jennifer Cole for many helpful discussions on theoretical and methodological aspects linked to prominence.

References

Arvaniti, A., Baltazani, M. (2005). Intonational analysis and prosodic annotation of Greek spoken corpora In Jun, S-A. (Eds.), *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford University Press. 84-117.

Arvaniti, A., Gryllia. S, Zhang, C. & Marcoux, K. (2022). Disentangling emphasis from pragmatic contrastivity in the English H* ~ L+H* contrast. *Proc. Speech Prosody* 2022, 837-841, doi: 10.21437/SpeechProsody.2022-170

Arvaniti, A., Ladd, D.R. & Mennen, I. (1998). Stability of tonal alignment: the case of Greek prenuclear accents. *Journal of Phonetics*, 26: 3-25.

Arvaniti, A., Ladd, D.R. & Mennen, I. (2006). Tonal association and tonal alignment: evidence from Greek polar questions and contrastive statements. *Language and Speech*, 49: 421-450.

Arvaniti, A., Ladd, D.R. (2009). Greek wh-questions and the phonology of intonation. *Phonology*, 26, 43–74.

Baltazani, M. (2002). Quantifier scope and the role of intonation in Greek. Unpublished Ph.D. dissertation, University of California, Los Angeles.

Baltazani, M. (2003) Broad Focus across sentence types in Greek. *Proceedings of Eurospeech-2003*, Geneva, Switzerland.

Baltazani, M., Gryllia, S. & Arvaniti, A. (2020). The Intonation and Pragmatics of Greek wh-Questions. *Language and Speech*, 63 (1), 56-94. doi: 10.1177/0023830918823236

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01

Baumann, S., Winter, B. (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics*, 70, 20-38.

Bishop, J., Kuo, G. & Kim, B. (2020). Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from Rapid Prosody Transcription. *Journal of Phonetics*, 82, 100977.

Boersma, P., Weenink, D. (2023). Praat: doing phonetics by computer [Computer program]. Version 6.3.13, retrieved 31 July 2023 from http://www.praat.org/

Cole, J., Hualde, J.I., Smith, C.L., Eager, C., Mahrt, T. & de Souza, R.N. (2019). Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish. *Journal of Phonetics*, 75, 113-147. DOI: https://doi.org/10.1016/j.wocn.2019.05.002

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Cole, J., Mahrt, T. & Hualde, J.I. (2014). Listening for sound, listening for meaning: Task effects on prosodic transcription. In *Proceedings of Speech Prosody*, Dublin, Ireland 20-23 May 2014, 859-863.

Cole, J., Mo, Y. & Baek, S. (2010a). The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and Cognitive Processes* 25 (7):1141. DOI: http://dx.doi.org/10.1080/01690960903525507

Cole, J., Mo, Y. & Hasegawa-Johnson, M. (2010b). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, 1(2), 425-452.

Cole, J., Shattuck-Hufnagel, S. (2016). New methods for prosodic transcription: Capturing variability as a source of information. *Laboratory Phonology*, 7(1).

D'Imperio, M. (1997). Breadth of focus, modality, and prominence perception in Neapolitan Italian. In Ainsworth-Darnell, K., D'Imperio, M. (Eds.) *Working papers in linguistics:* Volume 50, 19-39.

D'Imperio, M. (2000). Acoustic-perceptual Correlates of Sentence Prominence in Italian. In Muller, J.S., Huang, T. & Roberts, C. (Eds.) Working papers in linguistics: Volume 54, 59-77.

Georgakopoulos T., Skopeteas S. (2010). Projective versus interpretational properties of nuclear accents and the phonology of contrastive focus in *Greek. Linguistic Review*, 27(3), 319–346.

Grice, M., Ladd, D.R. & Arvaniti, A. (2000). On the place of phrase accents in intonational phonology. *Phonology*, 17(2), 143-185.

Im, S., Cole, J. & Baumann, S. (2023). Standing out in context: Prominence in the production and perception of public speech, *Laboratory Phonology*, 14(1). doi: https://doi.org/10.16995/labphon.6417

Katz, J., Selkirk, E. (2011). Contrastive focus vs. discourse-new: Evidence from phonetic prominence in English. *Language*, 771-816.

Ladd, D.R. (2008). *Intonational phonology*. Cambridge University Press.

Ladd, D.R., Arvaniti, A. (2023). Prosodic prominence across languages. *Annual Review of Linguistics*, 9, 171-193.

Lenth, R. (2023). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.8.8, <https://CRAN.R-project.org/package=emmeans>

Lohfink, G., Katsika, A. & Arvaniti, A. (2019). Variability and category overlap in the realization of intonation. In Calhoun, S., Escudero, P., Tabain, M. & Warren, P. (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia 2019, 2946-2950

Orrico, R., Gryllia, S., Hu, N., Kim, J. & Arvaniti, A. (2024). Prosodic prominence in Greek: methodological and theoretical considerations. *Proc. Speech Prosody 2024*.

Orrico, R., Gryllia, S., Kim, J. & Arvaniti, A. (2023). The influence of empathy and autistic-like traits in prominence perception. In Skarnitzl, R., Volín J. (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 1280-1284). GUARANT International.

Orrico, R., Gryllia, S., Kim, J. & Arvaniti, A. (2025). Individual variability and the H* ~ L+H* contrast in English. *Language and Cognition*, 17, e9. doi: 10.1017/langcog.2024.62

Pierrehumbert, J.B. (1980). The phonology and phonetics of English intonation (Doctoral dissertation, Massachusetts Institute of Technology).

Riesberg, S., Kalbertodt, J., Baumann, S. & Himmelmann, N. (2020). Using Rapid Prosody Transcription to probe little-known prosodic systems: The case of Papuan Malay, *Laboratory Phonology*, 11(1): 8. doi: https://doi.org/10.5334/labphon.192

Wagner, P., Origlia, A., Avesani, C., Christodoulides, G., Cutugno, F., D'Imperio, M., Escudero, D., Gili Fivela, B., Lacheret, A., Ludusan, B., Moniz, H., Chasaide, A., Niebuhr, O., Rousier-Vercruyssen, L., Simon, A., Šimkz, J., Tesser, F. & Vainio, M. (2015). Different Parts of the Same Elephant: A Roadmap to Disentangle and Connect Different Perspectives on Prosodic Prominence. In The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK: the University of Glasgow (10th-14th of August 2015). ISBN 978-0-85261-941-4. Paper number 0202.1-5

Welby, P. (2006). French intonational structure: Evidence from tonal alignment. *Journal of Phonetics*, 34(3), 343-371.

DOI: 10.17469/O2112AISV000009

EMANUELA GALLO, LOREDANA SCHETTINO

Acoustic characteristics of prolongations in spontaneous Italian speech[1]

This study concerns the analysis of segmental prolongations from an acoustic and functional perspective. Since the lengthening of sounds is a common feature of speech, recognizing disfluent-related occurrences is a complex operation and the main method commonly used is a "pragmatic approach" based on subjective perceptive judgment. However, there is a lack of investigations concerning acoustic characteristics that, in addition to duration, contribute to the perception of markedness and distinguish the portion of the segment identifiable as a prolongation from that identifiable in context as an unmarked part of the lexical element. From the results, it emerges that prolongation instances exhibit longer duration, reduced intensity, lower jitter and shimmer values, and higher Harmonics-to-Noise Ratio (HNR) measurements. Furthermore, their functional differences have been observed to correlate with different acoustic realizations.

Keywords: Prolongations, Disfluencies, Acoustic features, Functions.

1. *Introduction*

Human spontaneous speech is a complex activity that involves online planning, production, and reception processes. During these processes, speakers may produce phonetic cues to temporally suspend speech and gain extra time for planning or they may abandon utterances due to changes in plans and edit already uttered sequences considered somehow imprecise for the communicative intentions. These heterogeneous phenomena that interrupt the flow of speech making it discontinuous, are generally referred to with the term *disfluencies*. Early studies on these phenomena have been driven by the observation of their correlation with pathological conditions and aimed at distinguishing typical disfluencies from atypical disfluencies (a.o.; Johnson, 1961; Wingate, 1984, Zmarich, 2017). In fact, further studies have shown that disfluencies not only normally occur in typical speech but are not just occasional and individual production errors, but regularly occur in speech and are naturally involved in the economy of speech (Crocco, Savy, 2003), used as linguistic tools by speakers to control the quality of their production and to manage the online processes of speech planning, articulation, and reception (Allwood, Nivre & Ahlsén, 1990; Ginzburg, Fernández & Schlangen,

[1] This article is the result of the collaboration among the authors. However, for academic purposes only, Emanuela Gallo is responsible for sections 2, 3 and 5. Loredana Schettino is responsible for sections 1 and 4. Both are responsible for sections 6 and 7.

2014; Lickley, 2015; Voghera, 2017). Although different terminologies and classifications of disfluency phenomena have been developed according to specific research domains, interests and approaches, there is a certain agreement on Levelt's description of the general structure of disfluency events (Levelt, 1983) based on his proposed model of speech production (Levelt, 1989)[2]. According to this framework, speakers can monitor their own produced (overt) speech as well as their (inner) speech plan and thus detect and take action on issues either before or after articulation. Hence, two major subtypes of disfluencies have been distinguished, namely *repairs* and *hesitations*. Repairs or Backward-looking disfluencies, such as substitutions, insertions, and deletions, occur when speakers interrupt the ongoing speech and trace back to previously spoken material correcting and modifying it, or re-establishing continuity with the preceding utterance. Hesitations or Forward-looking disfluencies refer to a temporary suspension of flowing speech (Lickley 2015) and consist of a range of phonetic cues that allow speakers to gain extra time to deal with information selection, lexical retrieval, or other planning/production issues (Allwood et al., 1990; Ginzburg et al., 2014). Given that speech unfolds in real-time within diverse communicative contexts and conditions, hesitations are regarded as conversational tools that may be useful for speakers to gain extra time to manage their discourse and, at the same time, for listeners to process information. Their occurrence was observed to be mainly connected to moments that require a higher cognitive load in the speech planning process, such as the conceptualization and formulation of a new utterance, the search for a specific lexical item, and the selection of new information in discourse. These suspension devices may manifest through prospective repetitions, silent pauses, filled pauses (non-lexical and lexical fillers), and by prolonging speech sounds beyond their normal duration. As observed by Betz, Wagner (2016), lengthening is the first signal of hesitation, the primary measure a speaker can apply to solve problems in speech planning. The term *prolongations* has been used in the literature to refer to disfluent lengthenings which have been described as "phones that are longer than would be expected in normal, fluent speech" (Eklund, 2004: 163) or "a marked prolongation of one or more phones, resulting in a longer than average duration of syllables and words [...] This coincides with a local reduction in speaking speed which is not expected by the listener, causing an impression of disfluency and hesitation" (Betz, 2020: 14). Considering that the perception of the length of segments as being *marked* depends on the context of occurrence rather than actual duration, the identification of prolongations is a non-trivial operation[3]. This study aims to investigate to what

[2] Various models and theories of speech production have been proposed to account for the well-coordinated speech movements underlying the production of spoken language. However, it has been acknowledged that theories still struggle to properly account for all the properties of these movements. A collection of studies exploring the dynamics of speech motor control can be found in the volume "Models and theories of speech production" edited by Gafos, van Lieshout (2020).

[3] Grounded on the assumption that in spoken communication, speakers tend to make use of different available modes (vocal as well as gestural) in a cooperative and integrated manner, it has been argued that speech phenomena should be analyzed with a multimodal perspective to get a clearer picture of

extent acoustic characteristics, beyond duration, contribute to the perception of markedness of disfluent lengthenings in speech as well as to examine if different functions of prolonged speech segments influence acoustic parameters changes.

The article is structured as follows: § 2 presents the literature concerning prolongations and filled pauses; § 3 presents the general and specific aims of the study; § 4 describes the linguistic data, the annotation process, and the parameters of the linguistic analysis; in § 5 and § 6 the results of the analysis are reported and discussed.

2. *Related work*

2.1 Prolongations and filled pauses

Segmental prolongations have been largely described as a kind of stuttering behavior which, opposite to repeated movements, do not consist of atypical movements but of absence of typical movements (Onslow, 2023). However, this type of disfluency phenomena is also one of the most common form of non-pathological speech disfluency (Eklund, 2001, 2004; Williams, 2022) and frequently co-occur with other types of disfluent phenomena in everyday speech, reflecting the transitional speakers' lack of motor control during speech planning, i.e., when speakers need time to overcome some execution difficulty, or serving as a strategy to gain additional time to solve the problem in planning processes (Gósy, Eklund, 2018). Since the lengthening of sounds is a common feature of speech serving various functions, i.e., phrase-finality, accentuation, back-channeling and hesitation (Betz, 2020), a key feature to distinguish prolongations from other types of lengthening lies in their pitch contour and occurrence within syllables: disfluent lengthenings are generally realized with lower pitch range and flat pitch contour and occur in long vowel nuclei as well as in sonorant codas (Betz, 2020). Nevertheless, prolongations are hard to classify in a rule-based way as syllables vary in duration in fluent speech according to context making unclear how long they must be in order to be perceived as a hesitant disfluency element. According to Lickley (2015), a possible solution consists of adopting a "pragmatic approach", based on subjective perceptive judgment, supported by reliability measures of inter-annotator agreement. Even if duration appears to be the main acoustic cue to be used in prolongation recognition, it is not always the decisive factor. Firstly, it is important to consider that perceived prolongation is dependent on speech rate and thus needs to be calibrated against a durational framework, rather than being defined in absolute durational terms. Secondly, there is significant variation among listeners in their temporal thresholds for identifying prolongations: listeners may perceive elongated

their role and production. Different approaches have been proposed to tackle this, e.g., some studies rely on multimodal transcription systems (Kosmala, 2024), whereas other studies carry out the analyses at the speech and gestural levels independently and then merge these levels to observe their correlation in the speech flow, as in Schettino, Campisi & Origlia (2024).

sounds along a continuum wherein prolonged segments of a word are intentionally produced and hence not regarded as instances of disfluency (Gósy, Eklund, 2018). Therefore, several other factors may contribute to the perception of markedness and influence prolongation recognition, such as articulation rate, fundamental frequency, intensity, phonetic context, segment quality, and position in the phrase (Gósy, Eklund, 2018). Most of these markedness-contributing factors have been systematically analyzed for filled pauses, which similarly to prolongations, also create a suspension and delay production through vocalization and duration, distinguishing themselves from other hesitation phenomena (Eklund, 2004). Regarding their manifestations, filled pauses contain non-lexical material and are typically comprised of a central vowel, optionally followed by a nasal, e.g., *eeh, ehm, mhh* (Di Napoli, 2020). Although there is evidence indicating that filled pauses share standard acoustical features across different languages, such as a flat pitch and stable formant frequencies (Hamzah, Jamil & Seman, 2012), studies have shown that there is variability in the way these phenomena are produced, influenced by both language-specific factors and individual speaker characteristics (Giannini, 2003; Jabeen, Wagner, 2023). Studies on production mechanisms of filled pauses were conducted within the context of Japanese spontaneous speech, both in monologue and conversation settings, concentrating on the acoustic differences between vowel sounds in vocalizations and those belonging to lexical items. Maekawa, Mori (2017) highlighted that the voice quality of filled pauses seems to be used by speakers to transmit various paralinguistic information and that there is a systematic difference in voice quality between vowels in filled pauses and those in lexical items which correlates to different functions in speech. They showed that vowel sounds in filled pauses had relatively "softer" phonation, as indicated by higher spectral tilt and H1-H2 ratio values, i.e., the difference between the first and second harmonic, and were also unstable, as indicated by larger jitter values, i.e., the cyclic variation of the fundamental frequency of a vocal sound. Additionally, they were also marked by lower intensity, lower F0, and higher F1.

Similarly, Li, Ishi, Fu & Hayashi (2022) investigated the differences in filled pauses between Japanese native speakers and Chinese learners of Japanese as L2 employing both prosodic and voice quality measurements. Their findings indicate that vowels of filled pauses produced by Japanese speakers exhibit longer duration, lower intensity, and more aperiodicity, whereas Chinese learners of L2 Japanese show longer duration, lower F0 mean, and a tenser phonation type to distinguish filled pauses from ordinary lexical items. Moreover, their filled pause production patterns are influenced by their native language as indicated by features such as intensity, H1-A3 ratio, i.e., the difference in frequency between the first harmonic and the third formant, jitter and shimmer, i.e., the cycle-to-cycle variation of the vocal fold in relation to the amplitude of the sound wave.

These investigations have demonstrated that vowel differences in filled pauses from other lexical items involve both prosodic and vocal quality parameters, with duration and intensity contributing significantly to the distinction while parameters

linked to vocal quality such as jitter, shimmer, HNR (which measures the relative amount of additive noise in the vocal signal and the degree of acoustic periodicity), and H1-H2 ratio played a secondary role. Given these observations, it is reasonable to hypothesize that voice quality features could contribute to the perception of markedness of prolongation phenomena.

2.2 Cross-linguistic studies on prolongations

The literature on the characterization of segmental prolongations has mainly concerned their position within the word being prolonged, the relevant part of speech, segmental content, duration and pitch (see Tab. 1).

Studies concerning prolongations in different languages show both common patterns and language-specific traits. Regarding the distribution within the initial/medial/final segments of words, prolongations appear to be strongly dependent on syllable structure and language-specific phonotactic rules. Indeed, in the series of studies conducted by Eklund on prolongations in different languages, he observed differences and similarities in the word positioning of prolongations in relation to morphological complexity as well as phonotactic rules and syllable structure complexity. First, he proposed the "morphology matters" hypothesis (Eklund, 2004), which postulates that the distribution of prolongations is related to the morphological complexity of the language under consideration. This hypothesis was supported in Swedish and American English (Eklund, Shriberg, 1998), two languages with similar morphological complexity and which show an identical distribution of prolongation in initial, medial, and final positions, contrasting with Tok Pisin pidgin, which is considered much less complex and show a prevalence of initial and final occurrences. On the one hand, subsequent works on Japanese, Mandarin, and Hungarian support this hypothesis (Den, 2003; Lee, 2004; Gósy, 2017). On the other hand, investigations on German (Betz, 2017) presented an exception, as the distribution of prolongations resembled that of Tok Pisin, Japanese, and Mandarin despite showing morphology traits closer to Swedish, American English and Hungarian. Hence, Eklund proposed the "phonotactics matters hypothesis" (Gósy, Eklund, 2018). This new framework was further supported by a study on Hebrew (Silber-Varod, 2019), as well as Italian speech data, where prolongations occur almost exclusively in the final position of the word (Schettino, 2023). The conclusions that were drawn from the above-mentioned studies are that, although morphology and phonotactics may play some role in what position of the word prolongations might appear, other factors must also play a role.

For what concerns the part of speech of words, prolongations tend to occur more frequently on function words rather than content words in Swedish, Tok Pisin, German and Hungarian, while no clear distinction was found between Mandarin, Japanese, and Italian (Di Napoli, 2020; Schettino, 2023).

Regarding segmental content, all types of segments might be subject to prolongation, but vowels and sonorants are generally more prone, within language-specific phonological constraints. In German (Betz, 2017), sonorants are the target

of prolongations, followed by the combination of diphthongs and long vowels. In Hungarian, Hebrew, and Turkish (Altiparmak, Kuruoğlu, 2018), prolongations involve more vowels than consonants. The same happens in Mandarin (Lee, 2004), where vowels and vowel-coda pairs are more frequently prolonged than consonants. Likewise, in Italian (Schettino, Eklund, 2023), most cases of prolongations are made by the lengthening of vowel sounds, occurrences of lengthening of word-final consonants followed by a schwa sound, diphthongized and lengthened final vowels with schwa sound, as well as lengthening of continuous consonants or nasals, and rare cases of word-final vowels plus lengthening of the nasal sound. Maybe these cases of schwa insertion in word-final prolongations led to the perception and classification of prolongations as a particular type of filled pause in Italian studies (see Schettino, Eklund, 2023).

In terms of duration, prolongations are generally longer on vowels than on consonants in both Hungarian and Italian whereas no significant durational difference between vowels and consonants was found in Hebrew. For what concerns the other languages under investigation, prolongations have been analyzed in comparison to filled pauses, which were found to have greater average duration than lengthenings in different languages, including European Portuguese (Moniz, Mata & Viana, 2007), Swedish, Tok Pisin, German, and Italian.

Nevertheless, systematic investigations concerning acoustic characteristics that, in addition to duration, contribute to distinguishing the portion of the segment identifiable as a marked prolongation is still lacking.

Table 1 - *Summary of studies on prolongations and filled pauses with analyzed parameters*

Hesitation Type	Language	Studies	Position within the word	Part of speech word	Segmental Content	Duration	F0, Pitch	Intensity	Spectral Tilt	Jitter and Shimmer	HNR	F1,F2
PRL	Swedish, Tok Pisin	Eklund, R. (2001).	✓	✓		✓						
PRL	English	Shriberg, E. E. (1994).	✓									
PRL	Japanese	Den, Y. (2003).	✓	✓								
PRL	Mandarin	Lee, T.-L., He, Y.-F., Huang, Y.-J., Tseng, S.-C., & Eklund, R. (2004).	✓	✓								
PRL	Hungarian	Gosy, M., & Eklund, R. (2017).	✓	✓	✓	✓						
PRL	Hebrew	Silber-Varod, V., & Gosy, M., & Eklund, R. (2019).	✓		✓	✓						
PRL	German	Betz, S., Eklund, R., & Wagner, P. (2017).	✓	✓	✓	✓	✓					
PRL	Italian	Schettino, L., & Eklund, R. (2023).	✓	✓	✓	✓						
PRL, FP	Italian	Di Napoli, J. (2020).				✓						
FP	Japanese, Chinese	Li, X., Ishi C.T., Fu, C. & Hayashi, R. (2022).					✓	✓	✓	✓	✓	
FP	Japanese	Maekawa, K., & Mori, H. (2017).								✓		✓
FP	Punjabi	Jabeen, F., & Wagner, P. (2023).					✓					✓

3. *Research aim*

Besides the main acoustic cue of duration, other factors may contribute to perceiving prolonged sounds as disfluent-related phenomena. Given this framework, the present study aims to characterize the distinction between the disfluent portion of a prolonged segment (e.g. th**e<ee>**) from that identifiable in context as an unmarked part of the lexical item (e.g. th**e**<ee>) by deepening the analysis of phonetic-prosodic parameters as well as parameters related to voice quality, further supporting the recognition of the linguistic value of prolongations serving as a tool for managing spoken productions and the reliability of the functional approach commonly adopted for their identification.

Hence, the proposed study aims to address the following questions:

1. To what extent do acoustic characteristics, beyond duration, contribute to the perception of markedness of disfluent lengthenings?
2. How do different functions of prolonged speech segments influence acoustic parameter changes?

4. *Methodology*

4.1 Corpus and dataset

The analysis was conducted on the "Modokit-FROG" corpus (Sarro, 2023) consisting of audio-visual recordings of 10 Italian speakers (5 females, 5 males; Mean age = 25.1 SD = 1.91). As the corpus was specifically collected to identify and analyze Manner expressions through speech and gestures, speakers are not reported (nor observed) to be affected by speech pathologies. Participants come from different regions, namely Abruzzo (5), Lazio (2), Campania (2), and Apulia (1), and are all students at the University of L'Aquila. Following the Modokit protocol (Voghera, Mayrhofer, Ricci, Rosi, & Sammarco, 2020), the data collection involved eliciting both written and oral productions using the "frog stories" method, which consists of input taken from the picture book "Frog, ¿where are you?" by Mercer Mayer (1969). Each recording session comprised a speaker who actively participated by telling the story to a silent interlocutor, who was unaware of the story's content. The initial group, consisting of four males and one female, described the twenty-nine images first orally and then provided written descriptions (PS task). The second group, comprising four females and one male, did the opposite, initially writing the frog story and then narrating it orally to the addressee (SP task).

4.2 Annotation

Firstly, disfluency phenomena have been annotated on the ELAN software (Sloetjes, Wittenburg, 2008) employing a multilevel annotation scheme that provides information at formal and functional levels. The first level concerns the macro-structures of the disfluencies in which several regions are delineated following Shriberg's (1994) disfluency model, namely, the region to be repaired (RP), the

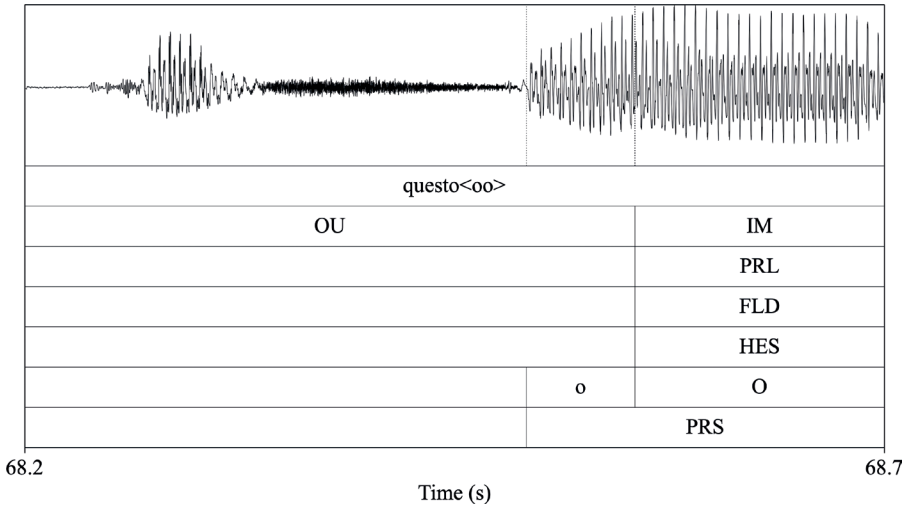
repaired one (RS), and the one in which the delay occurs (IM). The second level focuses on the micro-structure of disfluency, providing detailed information about the specific type of disfluency phenomenon. Among other phenomena, disfluency-related prolongation occurrences (PRL) were identified as marked lengthening of segmental material (Eklund, 2004; Betz, 2020) without establishing a minimum duration threshold and relying only on auditory judgment, i.e., once identified the unmarked-sounding portion of the lengthened word, the remaining portion was labelled as prolonged segment. At the third level, each item is assigned its macro-function, categorized as either Backward-Looking (BLD) or Forward-Looking disfluency (FLD) (Ginzburg et al., 2014). On a fourth level, Forward-Looking items are associated with more specific possible functions based on their context of occurrence:

- *Word Searching* (WS) for items that show difficulties in retrieving the target word.
- *Structuring* (STR) for items involved in structuring the discourse on a syntactic and informative level.
- *Focusing* (FOC) for items preceding semantically relevant elements.
- *Hesitative* (HES) is the label assigned to items that primarily serve the basic planning function of speech and it is applied to instances where no other overlapping function can be identified[4].

The annotations were performed by two expert judges and were evaluated by measuring the values of Cohen's k, i.e., 0.82 for the type of disfluency phenomenon and 0.77 for specific functions, standing respectively for "high agreement" and "substantial agreement" (Landis, Koch, 1977). Subsequently, all instances of prolongation were further labeled on Praat (Boersma, Weenink, 2021) at two separate levels (see Fig. 1): on one tier, the entire segment subject to prolongation was labeled as Prolonged Segments (PRS). Then, each PRS was further divided into a prolongation segment, which refers to the marked extension of segmental material corresponding to PRL, and a segment identified as unmarked (word segment, w).

[4] The scheme is described in detail in Schettino (2022).

Figure 1 - *Praat annotation system*



4.3 Linguistic and statistical analysis

At this stage, the analysis focused on prolongations of vocalic segments. For these, each PRL and w segments were analyzed concerning the following acoustic parameters:

- Duration (ms)
- Intensity (midpoint, z-scored)
- Fundamental frequency (midpoint, Normalized 100Hz)
- Formant frequencies (midpoint, Lobanov normalized, "PhonR")
- Jitter (local)
- Shimmer (local)
- Harmonics-to-Noise Ratio (mean, dB)

All the measurements selected as the object of investigation and the information encoded following the annotation scheme were extracted using a Praat script. The acoustic parameters and prolongation functions were processed and subjected to statistical analysis using the R software (R Core Team, 2021). To control for individual variability, the values obtained for each parameter were standardized by computing the z-score. For the formant values, the z-score method was employed, also referred to as the Lobanov normalization (Lobanov, 1971).

For statistical analysis, Linear Mixed Models ('lme4' package, Bates, Maechler, Bolker & Walker, 2015) were fitted with Speaker as random effect. Different models were built for each acoustic parameter selected as the dependent variable. Independent variables were: Prolonged Segment (levels: PRL, WS) and sociolinguistic factors, i.e., sex, age, and place of origin. For the PRL subset, the Function was introduced as a fixed factor. A post-hoc analysis was also conducted (emmeans package, Lenth,

Ingmann, Love, Buerkner & Hierve, 2018). The results yielded by the statistical analysis were interpreted taking into account descriptive measurements extracted with the "summarySE" function (Rmisc package, Ryan M. Hope, 2012).

5. *Results*

5.1 General results

In the selected dataset, consisting of 91 minutes of recordings, 641 phenomena of prolongation of word-final segments were identified. As previously mentioned, the study focused on prolongations of vocalic segments, which were analyzed with reference to standard Italian vocalic sounds, resulting in a total of 549 occurrences. Table 2 outlines the number of prolongation instances per speaker. On average, there were about 5 phenomena of prolongation per minute across all speakers, with a mean of 5,91 occurrences in the PS task and a mean of 5,87 in the SP task. Given the high variability among speakers in terms of the occurrence of prolongations, mixed models with the speaker as a random effect were deemed appropriate to support the analysis.

Table 2 - *Duration, number of words, average words per minute, prolongation occurrences, and average prolongation occurrences per minute*

n	*task*	*Dur(m)*	*word*	*word/m*	*PRL(n)*	*PRL/m*
1	PS	10,42	800	76,7	39	3,74
2	PS	10,67	1034	96,9	40	3,75
3	PS	12,78	1717	134,3	103	8,06
4	PS	15,15	1315	86,7	113	7,46
5	PS	10,07	816	81	66	6,56
6	SP	8,92	1270	142,3	9	1,01
7	SP	5,83	656	112,5	49	8,40
8	SP	3,78	385	101,8	5	1,32
9	SP	6,98	790	113,1	66	9,45
10	SP	6,45	685	106,2	59	9,15

5.2 Phonetic-prosodic parameters

In Fig. 2, the results for duration are depicted. As expected, prolongation segments are significantly longer than the unmarked portion of the prolonged segments as they consist of the lengthening of segmental material (see Tab. 3 and Tab. 4).

Figure 2 - *Duration values in word (w) and prolongation segments (PRL)*

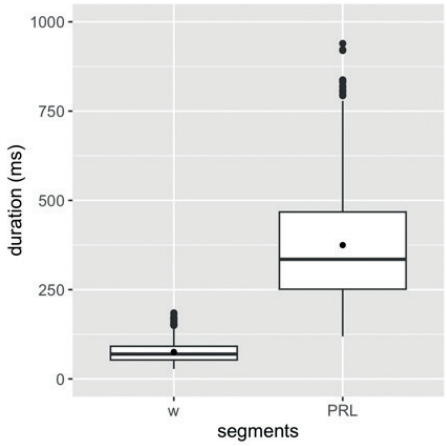


Table 3 - *Mean, standard deviation, standard error of the mean, and confidence interval of duration values in w and PRL*

Segm.	*Duration (ms)*	*sd*	*se*	*ci*
w	75	29	1.26	2.48
PRL	376	167	7.1	14.07

Table 4 - *Pairwise comparison among the levels of the Prolonged Segment for the duration model*

Contrast	*Estimate*	*SE*	*t-ratio*	*p. value*
w-PRL	-301	7.13	-42.200	<.00

Regarding intensity, Fig. 3 shows that there is a slight decrease in prolonged segments (see Tab. 5 and Tab. 6).

Figure 3 - *Z-scored intensity*

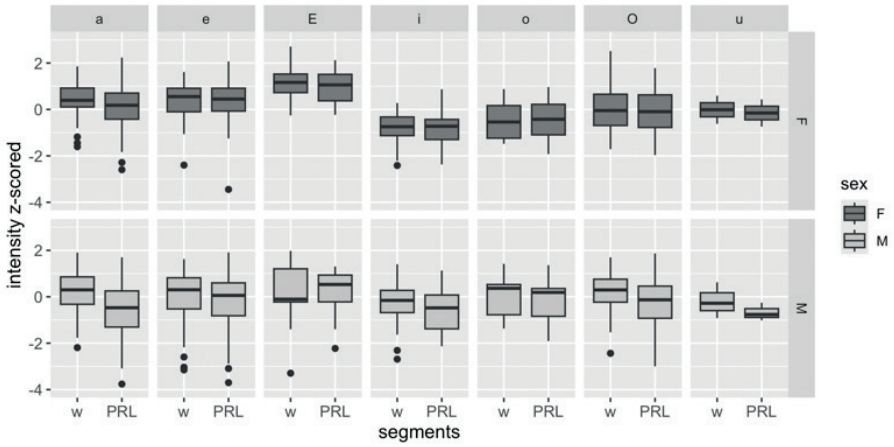


Table 5 - *Mean, standard deviation, standard error of the mean, and confidence interval of intensity values in w and PRL*

sex	*Segm.*	*Intensity midpoint (dB)*	*sd*	*se*	*ci*
F	w	61.24	5.25	0.33	0.66
F	PRL	60.73	5.23	0.33	0.65
M	w	60.74	5.04	0.28	0.56
M	PRL	58.49	5.43	0.31	0.61

Table 6 - *Pairwise comparison among the levels of the Prolonged Segment for the intensity model*

Fixed effects	*Estimate*	*SE*	*t-ratio*	*p. value*
w-PRL	-0.0236	0.008	-2.639	0.02

As shown in Fig. 4 and Tab. 7, a slight, though not statistically significant, decrease is also observed in fundamental frequency values of prolonged segments.

Figure 4 - *Fundamental frequency in word and prolongation segments*

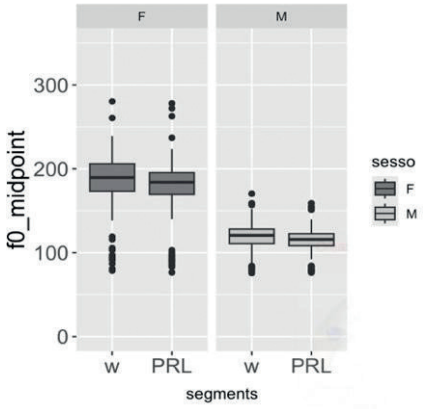


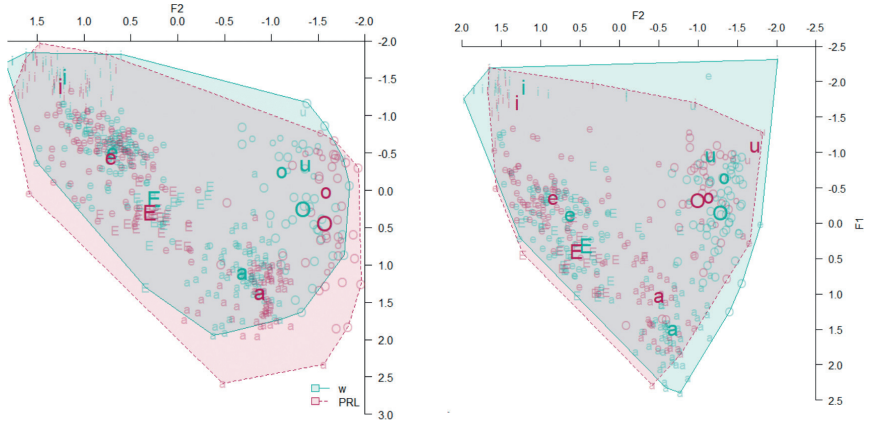
Table 7 - *Mean, standard deviation, standard error of the mean, and confidence interval of F0 midpoint values in w and PRL*

sex	*Segm.*	*F0 midpoint (Hz)*	*sd*	*se*	*ci*
F	w	189.20	41.89	3.02	5.96
F	PRL	187.9	32.95	2.18	4.30
M	w	122.82	32.24	2.45	4.83
M	PRL	115.51	13.22	0.95	1.87

5.3 Voice quality parameters

Concerning formant frequencies, no significant differences are found in the data between portions of prolongations and lexical items. As shown in Fig. 5, there is an overlap between the realizations of vowel elements belonging to the unmarked lexical item preceding the segment identified as prolongation in both men and women.

Figure 5 - *Formant frequencies of all vowels for men (left) and women (right)*



To examine the periodic disturbance of vocal fold vibration, jitter and shimmer measurements were utilized. As illustrated in Fig. 6 (see also Tab. 8 and 9), results indicate a decrease in jitter during prolongations when compared to previous lexical items in both male and female speakers for all vowel segments. The same decrease is shown in Fig. 7 (see also Tab. 10 and 11) for shimmer values.

Figure 6 - *Z-scored Jitter Local*

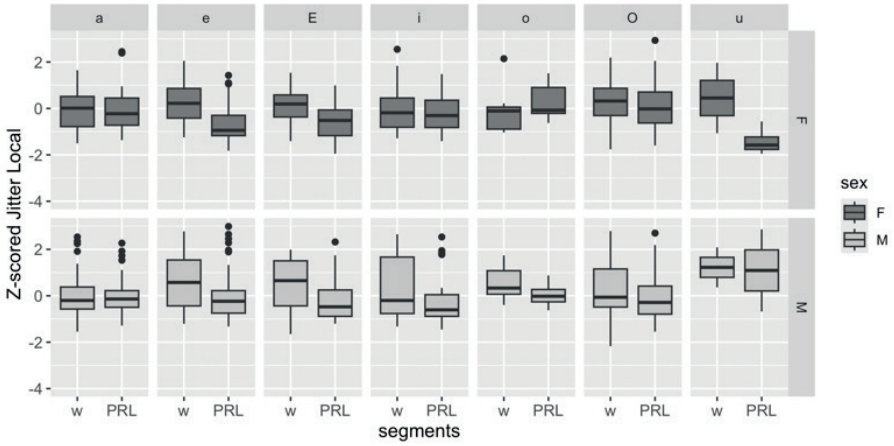


Table 8 - *Mean, standard deviation, standard error of the mean, and confidence interval of jitter local values in WS and PRL*

sex	*Segm.*	*Jitter local*	*sd*	*se*	*ci*
F	w	0.013	0.011	0.0007	0.0015
F	PRL	0.009	0.012	0.0007	0.0015
M	w	0.020	0.020	0.0016	0.0032
M	PRL	0.013	0.013	0.0011	0.0023

Table 9 - *Pairwise comparison among the levels of the Prolonged Segment for the jitter local model*

Fixed effects	*Estimate*	*SE*	*t-ratio*	*p. value*
w-PRL	0.269	0.08	3.104	<.00

Figure 7 - *Z-scored Shimmer Local*

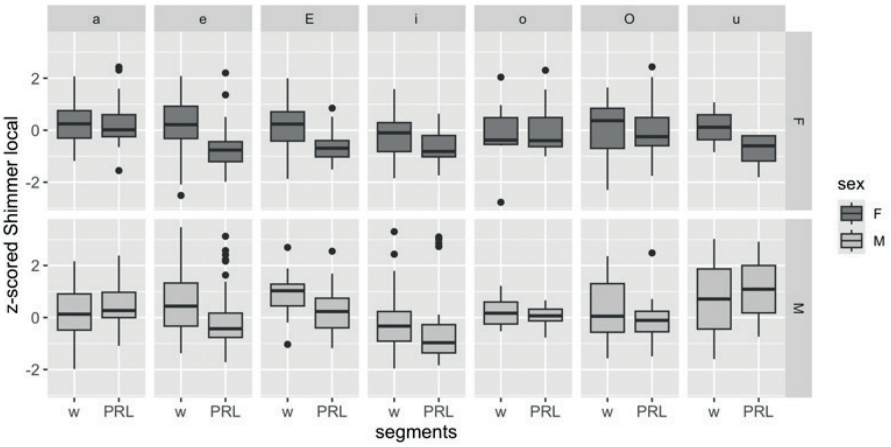


Table 10 - *Mean, standard deviation, standard error of the mean, and confidence interval of shimmer local values in w and PRL*

sex	*Segm.*	*Shimmer local*	*sd*	*se*	*ci*
F	w	0.068	0.04	0.003	0.006
F	PRL	0.045	0.03	0.002	0.004
M	w	0.081	0.06	0.004	0.009
M	PRL	0.063	0.05	0.003	0.007

Table 11 - *Pairwise comparison among the levels of the Prolonged Segment for the shimmer local model*

Fixed effects	*Estimate*	*SE*	*t-ratio*	*p. value*
w-PRL	0.327	0.108	3.26	<.01

As reported in Fig. 8, results indicate a statistically significant difference in HNR measurements between prolongations and word segments. The mean HNR is higher in prolongations than in word segments (see Tab. 12 and Tab. 13).

Figure 8 - *Mean HNR values*

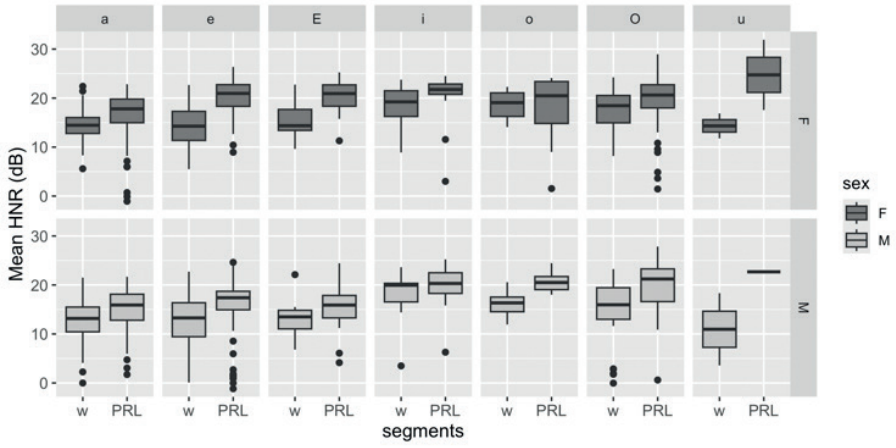


Table 12 - *Mean, standard deviation, standard error of the mean, and confidence interval of HNR values in w and PRL*

sex	*Segm.*	*Mean HNR*	*sd*	*se*	*ci*
F	w	15.48	4.24	0.29	0.58
F	PRL	19.15	5.22	0.33	0.66
M	w	13.19	5.55	0.39	0.78
M	PRL	16.34	5.81	0.40	0.79

Table 13 - *Pairwise comparison among the levels of the Prolonged Segment for the HNR model*

Fixed effects	*Estimate*	*SE*	*t-ratio*	*p. value*
w-PRL	-0.179	0.027	-6.444	<.00

5.4 Prolongation functions

All instances of prolongation of word-final vowels were analyzed in terms of functions based on the context of their occurrence. As shown in Fig. 9, these phenomena are primarily associated with the basic planning function (HES: 63 %), to a lesser extent connected with the discourse structuring function (STR: 22 %), lexical search (WS: 9 %), and the marking of semantically relevant elements (FOC: 6 %).

Figure 9 - *Prolongation functions frequency*

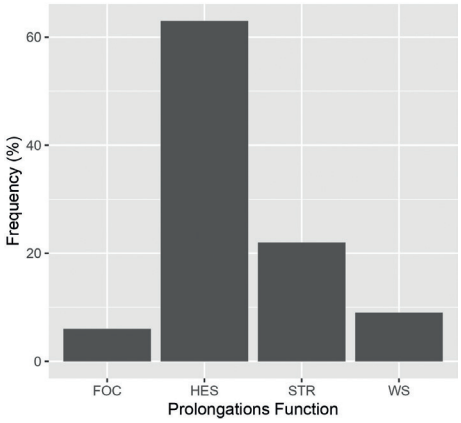


Figure 10 - *Phonetic-prosodic parameters of prolongations per function*

Figure a - *Duration*

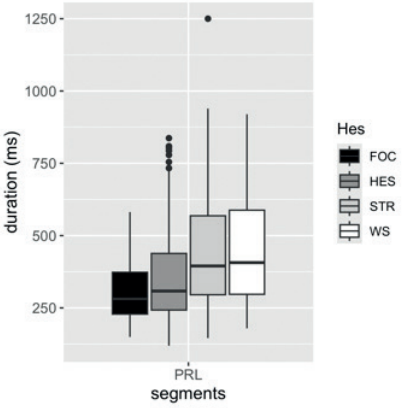


Figure b - *Intensity*

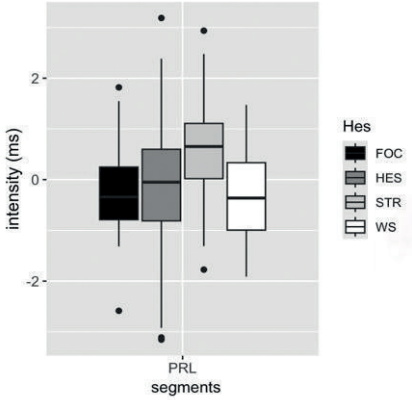
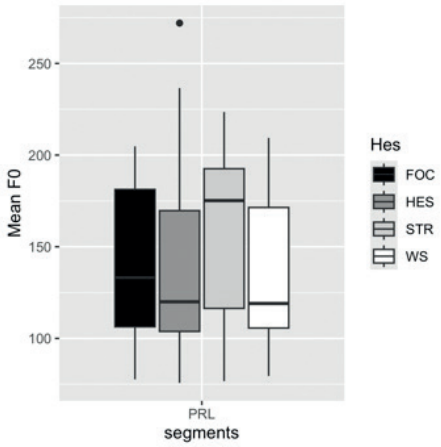


Figure c - *Mean Fundamental Frequency*



The analysis showed that the duration of prolongations varied significantly depending on the function they served within the speech. Figure 10a shows that prolongations exhibited a significantly longer duration when associated with the Word Searching (WS) function (Est = 136.23, SE = 36.02, t value = 3.782, p = 0.0002) and prolongations linked with the Structuring (STR) function (Est = 129.45, SE = 31.38, t value = 4.126, p <.00) as compared to those associated with Hesitative. Interestingly, prolongations associated with Focusing (FOC) function were even shorter (see Tab. 14).

Table 14 - *Mean, standard deviation, standard error of the mean, and confidence interval (default 95 %) of duration values per prolongations function*

Function	*Duration*	*sd*	*se*	*ci*
FOC	308	112	19.5	39.7
HES	350	147	7.94	15.6
STR	438	205	18.6	36.9
WS	447	169	23.7	47.6

As illustrated in Fig. 10b, prolongations linked with the structuring function showed significantly higher intensity values (Est =0.961, SE = 0.15, t value = 6.375, p <.00). Prolongations associated with hesitative function had slightly lower intensity values, closely followed by those associated with focusing function. Finally, prolongations associated with the word searching function had the lowest intensity values (see Tab. 15).

Table 15 - *Mean, standard deviation, standard error of the mean, and confidence interval (default 95 %) of intensity midpoint values per prolongations function*

Function	*Intensity*	*sd*	*se*	*ci*
FOC	58.2	4.9	0.8	0.8
HES	58.5	5.3	0.2	0.5
STR	63	4.4	0.4	0.8
WS	57	4.8	0.6	1.3

For what concerns fundamental frequency, as illustrated in Fig. 10c, prolongations linked with the Structuring function exhibited the highest mean F0 values (Est = 3.198, SE 0.55, t value = 5.761, p <.00), whereas prolongations associated with generic planning showed significantly lower mean F0 value (Est = 1.8541, SE = 0.5265, t value = 3.521, p = 0.00). Prolongations involved in lexical retrieval had the lowest mean F0 value (Est = 1.6292, SE = 0.6471, t value = 2.518, p = 0.01) (see Tab. 16).

Table 16 - *Mean, standard deviation, standard error of the mean, and confidence interval (default 95 %) of mean F0 values per prolongations function*

Function	*Mean F0*	*sd*	*se*	*ci*
FOC	161.1	33	4.7	9.6
HES	150.3	48	2.6	5.2
STR	177.5	51	4.8	9.6
WS	146.1	34	4.2	8.4

6. *Discussion*

Based on the findings presented in this study, statistically significant differences concerning phonetic-prosodic and voice quality parameters were observed between the vocalic segments identified as unmarked and the markedly prolonged segments. Firstly, the increase in duration observed in prolonged segments aligns with expectations, as duration serves as a primary acoustic cue for identifying prolongation instances.

As speakers elongate sounds, there is typically a decrease in the muscular effort required for articulation, leading to an energy reduction and lower intensity values during prolongation production. Moreover, lower jitter and shimmer values as well as a higher HNR ratio indicate a lower presence of noise and a higher level of periodicity in the speech signal during prolongation compared to vowels in regular lexical items.

These findings suggest that, on average, speakers tend to maintain a more periodic and stable voice during prolongation production possibly due to articulatory economy, which refers to the tendency of speakers to produce speech sounds with minimal effort during moments of uncertainty (Stepanova, 2007).

Also, the lack of significant statistical difference between word-final unmarked segments and prolongation segments in fundamental frequency and formant frequencies could suggest that speakers maintain similar vocal settings while lengthening. Furthermore, results on fundamental frequency align with prior research on hesitation markers in Italian discourse (Schettino, Betz, Cutugno, & Wagner, 2021a), in which prolongations typically displayed a flat pitch contour and had a narrower pitch range compared to other hesitation phenomena. However, the absence of variation in F0 may also be attributed to the relatively small portion identified as unmarked lexical items compared to prolongation segments.

Regarding formant frequencies, the data do not reveal any tendency to centralization, i.e., a phonetic phenomenon where the articulation of a vowel sound shifts towards a more centralized position in the oral cavity. Generally, articulatory economy and efforts reduction would be consistent with centralization and thus, changes in formant values. However, the data are consistent with previous studies on Italian speech, which noted only occasional instances of prolongations realized by the elongation of word-final vowels with a schwa sound and that timbric realization of prolongations can be traced back to articulatory models of the phonological inventory rather than to articulatory economy (Schettino, Eklund, 2023; Giannini, 2003; Schettino, Betz, Cutugno, Wagner, 2021b).

Moreover, to further understand the role of prolongation in the management of speech, the analysis concerned the variation of phonetic-prosodic parameters according to prolongations specific functions, namely, Word Searching, Structuring, Focusing, and Hesitative.

Prolongations associated with the word searching (WS) function tended to be longer as they suspend speech delivery for more time, which leads to acknowledging the search for and the selection of a specific lexical item as a more time-consuming process. In addition, these prolongations showed the lowest intensity and F0 values, aligning with prior research on hesitation markers in Italian discourse (Schettino et al., 2021a), in which lower pitch was found to characterize prolongations as well as other hesitations associated with Word Searching function. Prolongations associated with structuring (STR) function, involved in the organization of information, were characterized by higher duration as long as higher intensity and fundamental frequency. Higher F0 values also characterize prolongations with focusing (FOC) function, which, however, were also characterized by shorter mean duration and, on average, lower intensity values. Finally, when prolongations co-occurred with the Hesitative (HES) function, they exhibited lower duration, intensity as well as F0 values.

7. *Conclusions*

To conclude, the findings presented in this study suggest that the production of markedly prolonged segments corresponds to a reduction in articulatory effort,

which is consistent with lower intensity values, and a more periodic and stable voice characterized by lower jitter and shimmer values and higher HNR measurements.

Furthermore, the phonetic-prosodic features of these phenomena have been observed to correlate with functional differences. Results show that more time is used either for the retrieval of specific lexical items or for marking information structure. However, F0 values are found to distinguish items involved in word searching or generic planning (Word Searching and Hesitative) from those involved in discourse structuring (Structuring and Focusing), respectively lower and higher values. This supports the idea that prolongations play a role in marking discourse, with their acoustic realization reflecting different specific functions.

The findings reported in this work contribute, to a certain extent, to add pieces to the general picture in understanding the perception of markedness of disfluency-related lengthenings, making them systematically recognizable from other types of lengthening in speech. Further investigations may concern other factors that influence the perception of disfluent prolongations, such as speakers' speech rate, prolongations' position in the phrase, and the surrounding context, taking into consideration when these phenomena co-occur with other disfluency markers, like silent pauses, filled pauses, and repetitions. Also, the characterization of these phenomena may be supported by further investigations from a multimodal perspective.

References

Allwood, J., Nivre, J. & Ahse E. (1990). Speech management – on the on-written life of speech. In *Nordic Journal of Linguistics*, 13, 3-48. https://doi.org/10.1017/S0332586500002092

Altiparmak, A., Kuruoğlu, G. (2018). Gender and Speech Disfluency Production: a Psycholinguistic Analysis on Turkish Speakers. In *Psycholinguistics*, *24* (2), 114-143. https://doi.org/10.31470/2309-1797-2018-24-1-114-143

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. In *Journal of Statistical Software*, 67 (1), 1-48. https://doi.org/10.18637/jss.v067.i01

Betz, S. (2020). Hesitations in Spoken Dialogue Systems. (Doctoral dissertation). Universität Bielefeld.

Betz, S., Eklund, R. & Wagner, P. (2017). Prolongation in German. In *Proceedings of DiSS 2017 The 8th Workshop on Disfluency in Spontaneous Speech*, Stockholm, Sweden, 18-19 August 2017, 13-16.

Betz, S., Wagner, P. (2016). Disfluent lengthening in spontaneous speech. In *Elektronische Sprachsignalverarbeitung (ESSV) 2016*, Leipzig, Germany, 2-4 March 2016.

Boersma, P., Weenink, D. (2021). Praat: Doing phonetics by computer [computer program]

Crocco, C., Savy, R. (2003). Fenomeni di esitazione e dintorni: una rassegna bibliografica. In Crocco, C., Savy, R. & Cutugno, F. (a cura di), *API. Archivio di Parlato Italiano*, DVD.

Den, Y. (2003). Some strategies in prolonging speech segments in spontaneous speech. In Eklund, R. (Ed.), *Proceedings of the ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech (DiSS 2003)*, Göteborg, Sweden, 5-8 September 2003, 87-90.

Di Napoli, J. (2020). Filled pauses and prolongations in Roman Italian task-oriented dialogue. In *Proceedings of the Laughter and Other Non-Verbal Vocalisations Workshop*, Biefeld, 5 October 2020, 24-27.

Eklund, R. (2001). Prolongations: A dark horse in the disfluency stable. In *Proceedings of DiSS 2001 Disfluency in Spontaneous Speech*, Edinburgh, Scotland, UK, 29-31 August 2001, 5-8.

Eklund, R. (2004). Disfluency in Swedish human-human and human-machine travel booking dialogues. PhD Dissertation, Linköping University: Electronic Press.

Eklund, R., Shriberg, E. (1998). Crosslinguistic disfluency modelling: a comparative analysis of Swedish and American English human-human and human-machine dialogues. In *Proc. 5th International Conference on Spoken Language Processing (ICSLP 1998)*. https://doi.org/10.21437/ICSLP.1998-756

Gafos, A., Van Lieshout, P. (2020). Models and theories of speech production. In *Frontiers in Psychology*, 11, 1238.

Giannini, A. (2003). Hesitation phenomena in spontaneous Italian. In *Proceedings of the 15th International Congress of Phonetic Sciences* (ICPhS), 2653-2656.

Ginzburg, J., Fernández, R. & Schlangen, D. (2014). Disfluencies as intra-utterance dialogue moves. In *Semantics and Pragmatics*, 7 (9), 1-64. https://doi.org/10.3765/sp.7.9

Gósy, M., Eklund, R. (2017). Segment prolongation in Hungarian. The 8th Workshop on Disfluency in Spontaneous Speech (DiSS 2017), 29–32.

Gósy, M., Eklund, R. (2018). Language-Specific Patterns of Segment Prolongation in Hungarian. *In Phonetician*, *115*, 36-52.

Hamzah, R., Jamil, N. & Seman, N. (2012). Acoustical analysis of filled pause in Malay spontaneous speech. In *Computer Applications for Communication, Networking, and Digital Contents: International Conferences, FGCN and DCA 2012, Held as Part of the Future Generation Information Technology Conference, FGIT 2012, Gangneug, Korea, December 16-19, 2012. Proceedings* (pp. 251-259). Springer Berlin Heidelberg.

Hope, R.M. (2012). Rmisc: *Ryan Miscellaneous*. R package version 1.5.1, https://cran.r-project.org/web/packages/Rmisc

Jabeen, F., Wagner, P. (2023). Variability in hesitations in Punjabi semi-spontaneous narrative speech: An automatic clustering based analysis, In *Disfluency in Spontaneous Speech (DiSS) Workshop 2023*, ISCA, 71–75. https://doi.org/10.21437/DiSS.2023-15

Johnson, W. (1961). Measurements of oral reading and speaking rate and disfluency of adult male and female stutterers and nonstutterers. *Journal of Speech and Hearing Disorders. Monograph Supplement*, 1–20.

Kosmala, L. (2024). *Beyond disfluency: The interplay of speech, gesture, and interaction*. John Benjamins.

Kuruoglu, G., Altiparmak, A. (2015). A Psycholinguistic point of view on prolongations of the native speakers of Turkish. In *International Journal of Arts & Sciences*, *8*(3), 185.

Landis, J.R., Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Lee, T.-L., He, Y.-F., Huang, Y.-J., Tseng, S.-C. & Eklund, R. (2004). Prolongation in spontaneous Mandarin. In *Proceedings of Interspeech 2004*, Jeju Island, Korea, 4-8 October 2004, 2181-2184.

Lenth, R., Singmann, H., Love, J., Buerkner, P. & Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means. R package version, 1 (1), 1–97.

Levelt, W.J. (1983). Monitoring and self-repair in speech. In *Cognition*, 14(1), 41-104.

Levelt, W.J. (1989). *Speaking: From intention to articulation* (Vol. 1). Cambridge, MA: MIT Press.

Li, X., Ishi, C.T., Fu, C. & Hayashi, R. (2022). Prosodic and Voice Quality Analyses of Filled Pauses in Japanese Spontaneous Conversation by Chinese learners and Japanese Native Speakers. In *Proceedings of Speech Prosody 2022*, 550–554. https://doi.org/10.21437/SpeechProsody.2022-112

Lickley, R.J. (2017). Disfluency in typical and stuttered speech. In Bertini, C., Celata, C., Lenoci, G., Meluzzi, C. & Ricci, I. (Eds.), *Fattori sociali e biologici nella variazione fonetica*. Studi AISV 3, 373–387. Milano: Officinaventuno. https://doi.org/10.17469/O2103AISV000019

Lickley, R.J. (2015). Fluency and Disfluency. In *Redford, M.A. (Ed.), The handbook of speech production*. Chichester: John Wiley & Sons, 445-474. https://doi.org/10.1002/9781118584156.ch20

Lobanov, B. (1971). Classification of Russian vowels spoken by different listeners. In *Journal of the Acoustical Society of America*. 49. 606–608.

Maekawa, K., Mori, H. (2017). Comparison of Voice Quality between the Vowels in Filled Pauses and Ordinary Lexical Items. In *Journal of the Phonetic Society of Japan*, vol. 21, no. 3, 53-62.

Mayer, M. (1969). *Frog, ¿where are you?*. New York: Dial Press.

Moniz, H., Mata, A.I. & Céuviana, M.C. (2007). On filled-pauses and prolongations in European Portuguese. In *8th Annual Conference of the International Speech Communication Association*. Antwerp, Belgium, 27-31 August, 2820-2824 https://doi.org/10.21437/Interspeech.2007-695

Onslow, M. (2023). Stuttering and its treatment: Twelve lectures. Last retrieved October 20, 2024, from https://www.uts.edu.au/asrc/resources

R core team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna. https://www.R-project.org/

Sarro, G. (2023). The many ways to search for an Italian frog. The Manner encoding in an Italian corpus collected with Modokit. (Master Thesis). Università degli Studi dell'Aquila.

Schettino, L. (2022). The Role of Disfluencies in Italian Discourse. Modelling and Speech Synthesis Applications. (Doctoral dissertation). Università degli Studi di Salerno.

Schettino, L., Betz, S., Cutugno, F. & Wagner, P. (2021a). Pitch and Functional Characterization of Hesitation Phenomena in Italian Discourse. In *Proceedings of the 4th Phonetics and Phonology in Europe* (PaPE 2021).

Schettino, L., Betz, S., Cutugno, F. & Wagner, P. (2021b). Hesitations and individual variability in Italian tourist guides' speech. In *Speaker individuality in phonetics and speech sciences: Speech technology and forensic applications*, Studi AISV 8, 243–262. Milano: Officinaventuno.

Schettino, L., Campisi, E. & Origlia, A. (2024) Disfluenze bimodali. Forme, funzioni e pattern di fenomeni di disfluenza e gesti per la gestione del parlato di guide turistiche. In Cirillo, L., Nodari, R. (a cura di) Studi AItLA 18: *Contesti, pratiche e risorse della comunicazione multimodale*, 209–223. Milano: Officinaventuno.

Schettino, L., Eklund, R. (2023). Prolongation in Italian. In *Disfluency in Spontaneous Speech (DiSS) Workshop 2023*, ISCA, 81–85. https://doi.org/10.21437/DiSS.2023-17

Shriberg, E.E. (1994). Preliminaries to a theory of speech disfluencies (Doctoral dissertation). University of California.

Silber-varod, V., Gósy, M. & Eklund, R. (2019). Segment prolongation in Hebrew, in *Proceedings of DiSS 2019*. The 9th Workshop on Disfluency in Spontaneous Speech, ELTE Faculty of Humanities, 47–50. https://doi.org/10.21862/diss-09-013-silb-etal

Sloetjes, H., Wittenburg, P. (2008). Annotation by category-ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakesh, Marocco, 28-30 May 2008, 816-820.

Stepanova, S. (2007). Some features of filled hesitation pauses in spontaneous Russian. In: *Proceedings of ICPhS*.

Voghera, M., Mayhofer, V., Ricci, M., Rosi, F. & Sammarco, C. (2020). Il Modokit: un identikit modale delle produzioni parlate e scritte dalle medie al biennio. In Voghera, M., Maturi, P. & Rosi, F. (Eds.), *Orale e scritto, verbale e non verbale: la multimodalità nell'ora di lezione*, 227-245. Firenze: Cesati.

Voghera, M. (2017). *Dal parlato alla grammatica*. Roma: Carocci.

Williams, S. (2022). Disfluency and proficiency in second language speech production. Cham: Palgrave Macmillan.

Wingate, M.E. (1984). Fluency, disfluency, dysfluency, and stuttering. In *Journal of Fluency Disorders*, 9 (2), 163–168.

Zmarich, C. (2017). Stuttering and phonetic theory: An introduction. In Bertini, C., Celata, C., Lenoci, G., Meluzzi, C. & Ricci, I. (Eds.), *Social and biological factors in speech variation*, Studi AISV 3, 357–372. Milano: Officinaventuno.

DOI: 10.17469/O2112AISV000010

SIMONA SBRANNA, MICHELINA SAVINO, FLORENCE BAILLS,
MARTINE GRICE

Vocal feedback and eye gaze patterns in Italian task-based dyadic conversations

Previous research has provided evidence of the interaction between eye gaze and turn-taking, showing that turn yielding is generally signalled with mutual gaze, whereby the primary speaker assesses the interlocutor's availability, which is subsequently confirmed by the interlocutor, as the next or secondary speaker reciprocates the gaze. However, the relation between eye gaze and turn-regulating vocal feedback signals, indicating the intention to take the conversational floor or not, can be influenced by different experimental designs. In task-based conversations, turn-regulating vocal feedback reportedly occurs predominantly when speakers avert their gaze. In this study, we investigate the relation between turn-regulating vocal feedback and eye gaze patterns, extending the binary paradigm of averted versus mutual gaze to include unilateral gaze, which can be directed towards the interlocutor when feedback signals are followed by a turn. We also find by-dyad variability, which we attribute to the fact that interlocutors can use the vocal and visual channels of communication together or separately.

Keywords: eye gaze, vocal feedback, turn-taking, multimodality, interaction.

1. *Introduction*

While the investigation of human communication has traditionally been focused on spoken language, the centrality of its multimodal nature is nowadays well recognized in linguistic studies (Paulmann, Jessen & Kotz, 2009; Perniss, 2018). Linguistic multimodality refers to the integration of a variety of information conveyed through both aural and visual channels, encompassing prosody, gestures, facial expressions, and body movements, all of which are intertwined with linguistic expression. This property of communication is particularly evident in face-to-face dialogues, where vocal and visual elements are used by interlocutors to co-construct meaning in a structured manner (Rasenberg, Pouw, Özyürek, & Dingemanse, 2022). Due to its collaborative properties, dialogue has also been compared to a Complex Dynamic System (Hollenstein, 2013), since it emerges dynamically (i.e., it is not planned) and evolves through a process of co-regulation between interlocutors based on feedback loops (Hu, Chen, 2021 for a review; Stahl, 2016 for task-based dialogues). A fundamental cooperative mechanism for structuring interactions is turn-taking, wherein speakers alternate their speech in dialogue with minimal disruption.

In spoken communication, transitions between turns have been found to occur with brief pauses and minimal overlaps, and this phenomenon is consistently

evident across individuals, methodologies, and contexts (Levinson, Torreira, 2015). This mechanism is executed with exceptional temporal precision and flexibility (Schegloff, 2020), showing that interlocutors must be able to predict in advance the endpoint of their interlocutors' turn, and respond when vocal and/or visual turn-final cues are detected (Barthel, Meyer & Levinson, 2017), allowing a participant in a dialogue to shift from recipiency to speakership with high temporal accuracy. This impressive synchronization between conversational partners has sparked scholarly interest regarding the factors governing it, and numerous studies have demonstrated the multimodal nature of turn transitions (e.g., Zellers, House & Alexanderson, 2016; Holler, Kendrick, Levinson, 2018; Kendrick, Holler & Levinson, 2023). However, the roles of eye gaze, on the one hand (e.g., Degutyte, Astell, 2021), and vocal feedback with turn-alternating functions, on the other (e.g., Drummond, Hopper, 1993), have mainly been studied independently.

This study aims to advance the understanding of how visual and vocal channels interact to create the complex feedback loops through which turn-taking is managed in real time. To do so, we conduct an exploratory analysis of the interplay between eye gaze and turn-regulating vocal feedback in signalling turn alternation in dyadic conversations. Specifically, we analyse gaze pattern frequency and temporal dynamics at the moment when turn-regulating vocal feedback occurs, based on whether the speaker takes the floor, or refrains from doing so, after producing the feedback signal. Given the multimodal nature of turn-transition feedback, we argue that such an integrated approach is particularly suitable for better accounting for conversational dynamics. Particular attention is also devoted to inter-dyadic variability.

The paper is organised as follows: Section 2 provides a review of the literature on the interplay between eye gaze and turn-taking, vocal feedback and turn-taking, and subsequently, vocal feedback and eye gaze. We conclude this section by presenting our research goal. Section 3 delineates our dataset, data collection methodology, annotation process, and measurements for the analysis. In section 4, we show the results pertaining to gaze patterns and their temporal dynamics. The results and their implications in relation to existing knowledge are discussed in section 5, followed by the conclusion in section 6.

2. *Background*

2.1 Eye gaze and turn-taking

Studies on eye gaze and turn-taking have been conducted since the 1960s with the seminal work of Kendon (1967), which significantly influenced the understanding of eye gaze dynamics at the onset and offset of a speaker's turn. Through the analysis of spontaneous dialogues, Kendon observed that a substantial majority of utterances began with the speaker looking away from the listener; conversely, they concluded with the speaker directing their gaze towards the listener. According to Kendon, at the initiation of their speaking turn, participants usually avert their gaze from their

interlocutor, serving two primary functions: 1) mitigating cognitive processing as they recall information and formulate their speech and 2) signalling the intention to take the conversational floor. For these reasons, averting gaze may constitute a strategy to signal unavailability as a listener. These observations were corroborated by subsequent studies on dyadic interactions (Cummins, 2012; Duncan, 1972; Duncan, Fiske, 1977; Ho, Foulsham & Kingstone 2015; Oertel, Włodarczak, Edlund, Wagner & Gustafson, 2012). However, some studies have challenged these findings (Beattie, 1978; Rossano, 2012; Rutter, Stephenson, Ayling & White, 1978; Streeck, 2014), suggesting that gaze aversion during early utterance production merely functions to mitigate cognitive load rather than serving as a mechanism for regulating turn-taking as such.

Regarding turn-ending, Kendon (1967) suggested that the speaker directs their gaze towards the listener to ascertain their readiness to assume the role of the next speaker, thereby signalling turn-yielding. In contrast to his claim about turn beginnings, this proposal finds strong support in subsequent studies (Novick, Hansen & Ward, 1996; Rutter et al., 1978; Lerner, 2003; Jokinen, Nishida & Yamamoto, 2009; Jokinen, Furukawa, Nishida & Yamamoto, 2013; Ho et al., 2015; Brône, Oben, Jehoul, Vranjes, & Feyaerts, 2017; Auer, 2018; Blythe, Gardner, Mushin & Stirling, 2018; Streeck, 2014). Nevertheless, Rutter et al. (1978) observed that Kendon's predictions regarding changes in conversational turns are contingent upon mutual gaze between speaker and listener. In other words, for Kendon's predictions to be substantiated, not only one, but both participants must maintain visual contact with each other. Additionally, Novik et al. (1996) identified two primary patterns of gaze direction during turn-taking: mutual-break and mutual-hold. In the mutual-break pattern, which was observed to be the most prevalent, at the conclusion of an utterance, the speaker looks at the listener, resulting in a brief period of mutual gaze until the secondary speaker initiates speech and breaks the mutual gaze. Conversely, in the mutual-hold pattern, mutual gaze is maintained while the secondary speaker begins speaking, without immediately averting their gaze. Contrary to Kendon's findings, Beattie (1978) observed that utterances ended with more immediate turn switches in the absence of mutual gaze compared to when there was mutual gaze, suggesting that gaze may not serve in determining turn-taking. In line with this finding, Kendrick et al. (2023) found that turn transitions were sped up by the use of manual gestures rather than gaze behaviour.

It is important to note that variations in findings regarding turn boundaries and gaze behaviour may stem from differences in experimental design (Degutyte, Astell, 2021), including the choice of interaction types in the analysis (such as dyadic vs. triadic/multiparty interaction or question/answer sequences vs. other types of adjacency pairs) and the task being performed (task-oriented dialogue vs. free conversation). Moreover, the methodology employed can also affect results in terms of the level of granularity (eye tracking vs. video). Furthermore, numerous individual factors can influence gaze behaviour during conversation, including gender (Argyle, Dean, 1965; Myszka, 1975; Bissonnette, 1993), social status (Myszka, 1975;

Foulsham, Cheng, Tracy, Henrich & Kingstone, 2010), acquaintance status (Rutter et al., 1978; Bissonnette, 1993), and cultural background (Rossano et al., 2009). Thus, considering and comparing these variables can help clarify some contradictory statements found in the literature.

Despite some variability and different interpretations[1], overall, it still appears that either gaze towards the listener or mutual gaze might be associated with turn yielding, while averted gaze might be related to turn initiation.

2.2 Vocal feedback and turn-taking

Speakers' contribution to a conversation significantly depends on listeners' reactions to their speech, conveyed through feedback signals. These vocal and/or visual signals enhance fluency in social interactions (Amador-Moreno, McCarthy & O'Keeffe, 2013) by supporting the ongoing turn of the interlocutor, revealing listeners' attitude towards its content, and structuring conversations (Kraut, Lewis & Swezey, 1982; Sacks, Schegloff & Jefferson, 1974; Schegloff, 1982) by guiding floor management (Jefferson, 1983).

Vocal feedback signals have been largely investigated in spoken interaction and variously named as there is no consensus in the literature regarding their definition. Fries (1952) is recognized as one of the earliest to identify these "signals of attention", which do not interrupt the speaker's discourse in telephone conversations. Over time, various terms have been used to describe this phenomenon, such as "accompaniment signals" (Kendon, 1967), "receipt tokens" (Heritage, 1984), "minimal responses" (Fellegy, 1995), "reactive tokens" (Clancy, Thompson, Suzuki & Tao, 1996), "response tokens" (Gardner, 2001), "engaged listenership" (Lambertz, 2011), and "active listening responses" (Simon, 2018). Yngve (1970) introduced the concept of "backchannel communication" to distinguish the primary channel used by the speaker from that used by the listener (i.e., the back channel), to convey essential information without actively taking the floor. In other words, backchannels are described as vocal tokens expressing acknowledgment and understanding, thereby encouraging the current speaker to continue. As backchannels are generally lexical and/or non-lexical tokens such as 'yes' or 'yeah', okay', 'mh-mh', 'uh-huh', etc., Jefferson (1983) first suggested that the use of specific tokens could additionally cue whether the listener intends to transition from a listening to a speaking role after having signalled acknowledgment/understanding to the current speaker. She observed that 'mh-mh' generally shows Passive Recipiency (PR), indicating that the current speaker should continue speaking, while 'yeah' is more often used by the listener to signal Incipient Speakership (IS), that is, the intention of taking the floor (see also Drummond, Hopper, 1993). This pragmatic distinction between PR and IS was adopted in later studies addressing dialogue act classification (e.g.,

[1] Note that there are also different interpretations of eye gaze functions, such as the one proposed by Rossano, Brown and Levinson (2009) and Streeck (2014), who claimed that gaze does not support turn-taking, but rather the organization of complex actions, which can stretch over several turns.

Jurafsky, Shriberg, Fox & Curl, 1998), and in research investigating the lexical/non-lexical and intonational characterisation of PR and IS tokens in languages other than English in task-oriented dialogues, notably in Italian and German (Savino, 2010, 2011, 2014; Savino, Refice, 2013 for Italian; Spaniol, Janz, Wehrle, Vogeley & Grice, 2023, Janz, 2023, Wehrle, 2023 for German; Sbranna, Möking, Wehrle & Grice, 2022 and Sbranna, Wehrle & Grice, in press for native Italian and German, and in L2 German by Italian learners).

In this study, the same distinction is adopted between PR tokens (produced without the speaker taking the floor, i.e., backchannels as conceptualised by Yngve, 1970, or continuers by Schegloff, 1981), and IS tokens (produced to acknowledge the interlocutor's turn before initiating their own, facilitating a smooth transition of turns) for classifying backchannels produced by participants in dyadic game-based sessions, to investigate their interplay with gaze behaviour during interaction.

2.3 Vocal feedback and eye gaze

It has been proposed that speakers might provide certain cues to invite listeners to produce feedback (see "backchannel relevant spaces" by Heldner, Hjalmarsson & Edlund, 2013). These invitation signals, often investigated under the term "backchanneling inviting cues", can be conveyed through different cues, such as pausing, intonation, head movements, or gaze direction adjustments. However, there is evidence that listeners do not actually fill all backchannel relevance spaces, as they actively and intentionally select feedback positions to support speakers in structuring their speech (Heldner et al., 2013). Furthermore, evidence suggests that a single channel is insufficient to fully account for the timing of feedback (Hjalmarsson, Oertel, 2012), suggesting the complex and multimodal nature of these inviting cues.

Several studies examining the interaction between vocal feedback and eye gaze have demonstrated that the vocal and non-vocal feedback signals which are not followed by a turn and are only used to convey acknowledgment and understanding to the speaker (backchannels as defined by Yngve, 1970; or PR tokens as defined by Jefferson, 1983), tend to occur during mutual gaze between interlocutors (Kendon, 1967; Bavelas, Coates & Johnson, 2002; Eberhard, Nicholson, 2010; Cummins, 2012; Oertel et al., 2012). In particular, Bavelas et al. (2002) found that speakers actively seek a response at crucial moments in their speech by establishing eye contact with the listener, who reciprocates, leading to brief intervals of mutual gaze termed "gaze windows", which are interrupted by the listener who averts their gaze shortly afterwards. They noted that during these gaze windows, the listener was highly likely to react, such as with a vocal "mhm" or a nod, prompting the speaker to rapidly avert their gaze and resume speaking. The authors explain this behaviour by referring to the cooperative nature of interaction (Hall, 1995; Jacoby, Ochs, 1995), suggesting a cooperative relationship between the speaker's acts and the listener's responses, which are coordinated by eye gaze.

Mutual gaze is widely recognized as a strong indicator of backchannels, regardless of their vocal or visual nature (Ferré, Renaudier, 2017; Hjalmarsson, Oertel, 2012;

Poppe, Truong & Heylen, 2011). However, a distinction has been highlighted between vocal and visual feedback signals, with visual signals being more frequently observed in conjunction with mutual gaze compared to vocal ones (Ferré, Renaudier, 2017; Eberhard, Nicholson, 2010; Truong, Poppe, Kok, & Heylen, 2011). Additionally, gaze tends to be sustained more consistently throughout the entire duration of a visual backchannel (e.g., a head nod) compared to a verbal one (Ferré, Renaudier, 2017). These findings can be explained by the fact that the presence of vocal signals does not require visual contact. The strong correlation between being looked at and gestural backchannels, but not vocal ones, has further been motivated by the hypothesis that gaze may establish a mode of communication between interlocutors, whereby a gesture responds to another gesture. Indeed, in contexts where the speaker does not make eye contact with the listener, greater variability across gestural and vocal backchannels is observed (Bertrand, Ferré, Blache, Espesser & Rauzy, 2007). This hypothesis may modulate the strong expectation that vocal feedback signals tend to cooccur with mutual gaze, supported by the aforementioned findings, partially explaining the variability in results across studies.

Further information relative to the temporal details of the "gaze windows" comes from Ondáš et al. (2023), who examined the duration of the interval between the initiation of the gaze directed towards the listener and the subsequent backchannel in interviews. They found that in most cases, the moderator receives vocal feedback from their guest within 500 ms of initiating eye contact, while the second most frequent duration value ranged from 1500 ms to 2000 ms. The interval of 1500 ms to 2000 ms was the most probable duration when the guest receives feedback from the moderator.

Despite providing relevant evidence for the interplay between eye-gaze and vocal feedback signals in general, these studies defined feedback as acknowledgments (as in Yngve, 1970 or Schegloff, 1981), whereas less is known about the interplay between their turn-regulating function (as, by contrast, in Jefferson, 1983) and gaze patterns, which was not considered in the aforementioned studies.

2.4 The interplay of turn-regulating vocal feedback and eye gaze

To the best of our knowledge, there is only one recent study (Spaniol et al., 2023) conducted on four German dyads, which examined the interaction between gaze and turn-regulating vocal feedback, taking the PR- and IS-specific functions of backchannels into consideration, under mutual and averted gaze conditions. Participants' behaviour was compared in three different and subsequent communicative contexts: 1) casual conversation for mutual acquaintance, 2) engagement in a Tangram game, and 3) free discussion regarding the game. They found that during task-based interactions, participants predominantly produced vocal feedback while averting their gaze from their interlocutor, irrespective of whether it was followed by a change in turn. Interestingly, this finding contrasts with prior research claiming that feedback is elicited by a gaze window (Bavelas et al., 2002; Novik et al., 1996) and is in line with the theory that interlocutors might respect a mutual "gesture mode" of communication (Ferré, Renaudier, 2017;

Bertrand et al., 2007). Spaniol et al. attribute their result to the presence of a visual distractor during the game, that is, the materials they had to examine to accomplish the task. In the remaining two conversational contexts, PR feedback signals were more often associated with mutual gaze than to IS feedback signals, indicating a different distribution of turn-regulating backchannels and gaze behaviour patterns according to their function and to the context of the interaction.

However, in most previous studies reviewed, only two gaze patterns are analysed: directed vs. averted gaze by the backchannel producer in the Spaniol et al. study and mutual vs. averted by both speakers in the studies in section 2.3. In the former case, these binary categories do not capture the interactive dynamic of gaze established by both interlocutors, and in the latter case, the distinction between instances in which only one participant looks at the interlocutor from instances in which both participants avert their gaze. Consequently, a binary distinction risks losing information about the search for eye contact by each speaker, which might reveal that even in task-based settings speakers do seek eye contact despite a visual competitor.

Therefore, in the present paper, we aim to expand the exploration of the relation between eye gaze and turn-regulating vocal feedback beyond the binary gaze pattern distinction of mutual vs. averted gaze to include unilateral gaze. In particular, we aim to (1) assess the relation between a specific gaze pattern and the production of turn-regulating vocal feedback, and (2) examine the extent to which the duration of this specific gaze pattern expedites the production of the feedback utterance and the potential subsequent turn alternation.

3. *Method*

To address our research questions, we analysed the gaze pattern (mutual, averted, unilateral) of participants while producing vocal feedback expressions classified as PR or IS backchannels (i.e., followed or not by a turn change) by pairs interacting in a Tangram game-based dialogue.

3.1 Data

The analysed data consist of six audio-video recorded sessions including pairs of Italian participants involved in Tangram game-based dialogues (as described in Savino, Lapertosa & Refice, 2018).

The use of Tangram-based dialogues was intended to create a controlled setting in which we could assume a certain homogeneity in behaviour across speakers. By contrast, free conversation, might include personal topics with emotional components, potentially influencing eye gaze unpredictably across dyads (Adams & Kleck, 2005; Liang, Zou, Liang, Wu & Yan, 2021). Additionally, Tangram-based games also reflect real-world communication contexts involving a visual object as the topic of conversation (e.g., discussing food choices from a menu or choosing directions while looking at a navigation device, among many other examples) with the advantage of not involving any emotional component. Finally, this approach

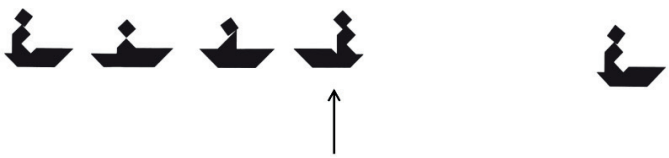
has already been used to study the interplay between eye gaze and turn alternation, providing some comparability (Spaniol et al., 2023). At the time of recording, participants were all students at the University of Bari (Italy) and shared the same geographic area of origin, gender, and acquaintance status, as they were all from the Bari metropolitan area, were female (aged 21–25), and knew each other as they were classmates. Speakers of each dyad sat at desks facing each other, separated by a panel which allowed eye contact while inhibiting the view of Tangram figures. Participants wore head-mounted microphones connected to a digital recorder, whereas camcorders were placed behind each participant, capturing their upper body and face frontally.

In each game round, participants were assigned roles as either Director or Matcher and given sets of Tangram figures accordingly. The Director received a set of four figures, in which one figure was marked with an arrow, and the Matcher received only one figure extracted from the Director's set. The objective of each round was to ascertain whether the marked figure on the Director's sheet and the figure on the Matcher's sheet were the same (an example of Director's and Matcher's sets is shown in Figure 1). Instructions prompted the Director to initially describe the marked figure; afterwards, participants were meant to interact collaboratively until reaching consensus about whether the two figures were the same or not. They verified their decision by lifting up their sheet and showing their figures to each other.

Each session includes 22 rounds of the Tangram game. Participants alternated their role as Director or Matcher in each round so that the distribution of role types was balanced between partners across the entire recording session.

We analysed rounds 1, 2, 9, 10, 21, and 22 of each recorded session. This selection was decided to account for the beginning (rounds 1-2), the medial (rounds 9-19), and the final part (rounds 21-22) of the recording sessions, with the aim of ascertaining the role of round progressions on the gaze patterns[2]. The total amount of analysed interactions corresponds to 45 minutes.

Figure 1 - *Example of a Tangram set of figures used in the Bari Italian game dialogues. The sheet on the left was given to the participant playing the Director role in that round, the sheet on the right to the participant playing the Matcher role. The aim of each round was to collaboratively ascertain whether the arrowed figure on the left matched that on the right*



[2] We also accounted for the influence of round progression on gaze and vocal feedback behaviour. However, the analysis did not reveal any consistent changing pattern as rounds progressed, instead we found a highly variable picture with dyad-specific preferences. This is why we show results pooled across rounds.

3.2 Annotations

The data contain transcripts of all game dialogues, where the beginning and end of each game round are reported, along with the participant's role (Director/Matcher) in that round. Multilevel annotations of the speech signal are also provided, which include game round intervals, as well as inter-pausal units (IPUs – speech intervals separated by at least 100 ms of silence), phonological words, and syllables produced by each of the two participants in each dyad. An additional level for vocal feedback annotation is also provided, as described in the following section.

3.2.1 Annotation of vocal feedback signals

Vocal feedback entailing lexical and non-lexical tokens conveying acknowledgment to the speaker are annotated, along with their function in regulating turns. As described in section 2.2, when the secondary speaker refrains from taking the floor after providing vocal feedback, allowing the primary speaker to continue, we categorised the feedback as PR. Conversely, when the secondary speaker takes the floor immediately after the production of vocal feedback, we annotated it as IS (Savino, 2010; 2011; 2014, also adopted by Sbranna et al., 2022; in press, and by Spaniol et al. 2023).

An example of both categories is provided in the following example from the data[3]:

Speaker A: allora la mia figura
(Eng.: so my figure)
Speaker B: sì **[vocal feedback-PR]**
(Eng.: yes)
Speaker A: ha la forma di un vaso e la parte superiore assomiglia ad un diamante
(Eng.: has a vase shape and the upper part looks like a diamond)
Speaker B: <m> **[vocal feedback-PR]**
(Eng.: <m>)
Speaker A: immagina che alla base ci sia un triangolo e sopra ci sia un diamante rove+ un diamante
(Eng.: Imagine that at the base there is a triangle, and on top, there is a diamond upsided+, a diamond)
Speaker B: <m> **[vocal feedback-IS]** da quante<ee> parti è composta?
(Eng.: <m> how many parts it is composed of?)

3.2.2 Annotation of eye gaze directions

We annotated gaze in ELAN (The Language Archive, 2023; Wittenburg, Brugman, Russel, Klassmann & Sloetjes, 2006). Following the methodologies established in prior research (Kendon, 1967; Beattie, 1978; Goodwin, 1980; Egbert, 1996; Novik et al., 1996; Jokinen et al., 2009; Streeck, 2014; Auer, 2018; Blythe et al., 2018), we annotated the intervals of directed gaze. These intervals denote periods during which participants consistently gaze in one direction without shifts, with interval boundaries marked by gaze transitions in different directions.

[3] The annotation convention "+" indicates an interruption. Vocalisations and nasalisations are indicated in angle brackets "< >".

We annotated gaze as directed towards "Task", "Addressee", or "Other". Gaze labelled as "Task" entailed the participant looking at their desk where the Tangram figures were placed, while "Addressee" involved the participant looking directly at their conversational partner. With "Other" we coded cases when participants gazed at the experimenter; however, since such cases were rather rare, they were not included in the analysis. Notably, participants showed very homogeneous gaze behaviour in that they either looked at the desk or directly at their partner.

We did not establish a minimum duration threshold for eye gaze annotation, but annotated all perceivable changes between gaze directions. However, since some previous studies used minimum duration thresholds (Beattie, 1979; Jokinen et al., 2009; Brône et al., 2017; Zima et al., 2019; Bavelas et al., 2002), we report the shortest duration of gaze direction intervals found in our data, which is 0.069 seconds, a value below most of the previously used thresholds (see Degutyte, Astell, 2021 for a review).

Twenty percent of the data were independently annotated by another linguist trained in eye gaze annotation. Inter-annotator reliability was assessed using Staccato in ELAN (Lücking, Ptock & Bergmann, 2011), a reliability measure of the temporal overlap between annotations based on Thomann's technique (2001). The result shows a mean degree of overlap of 70 %, indicating substantial reliability. No disagreements in category labels were found across annotators. The annotations by the first author were used for analysis.

3.3 Measurements

From the annotations, we derived the following categories and measurements for our analysis:

- *Mutual gaze:* when intervals of gaze to "Addressee" from both participants overlap, that is, they look at each other;
- *Averted gaze:* stands for overlapping intervals of gaze to "Task" from both participants, i.e., when both participants do not look at each other but at the task materials;
- *Unilateral gaze:* derives from one single interval of gaze to "Addressee", whereas the other speaker features gaze to "Task", i.e., one participant is looking at the other, while the interlocutor is looking at the task materials.

Given these three categories, which we will refer to as gaze patterns, we analysed: 1) the proportion of occurrences of PR and IS feedback signals under the different gaze pattern conditions, that is, at the moment when the feedback is uttered; 2) the length of this gaze pattern before vocal feedback is uttered, that is, how long a specific gaze pattern had already been established when the vocal feedback signal is produced. This time window was operationalised as a time interval (i.e., duration) calculated from the onset of the gaze pattern to the onset of the vocal feedback.

Given the exploratory nature of this study, we do not provide any inferential statistics. Instead, a detailed analysis at a by-dyad level is presented to account for

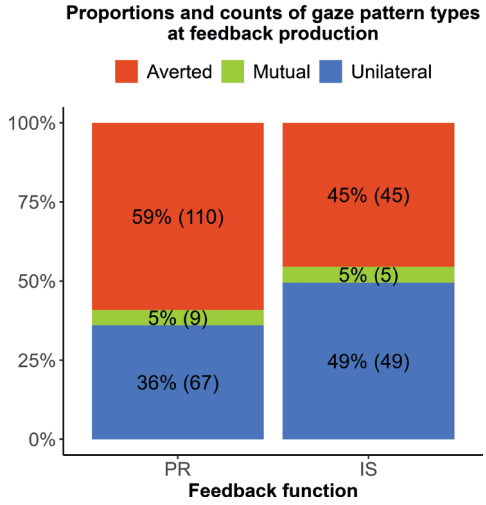
the individual conversational dynamic of each pair. The data tables and the code used to perform the analysis are openly available in the accompanying repository at https://osf.io/qzgwj/?view_only=270e19b9c36b4070ae41b6444f0220a9.

4. *Results*

4.1 Occurrences of gaze pattern types at feedback production

Figure 2 shows that both IS and PR vocal feedback are rarely produced in co-occurrence with mutual gaze (5 %). There is a prevalence of averted gaze co-occurring with PR feedback (59 %), in which both members of the dyad look at the task materials. However, speakers do not devote their visual attention exclusively to the task: unilateral gaze co-occurs with PR feedback in 36 % of cases and with IS vocal feedback in 49 % of cases, showing that speakers still use gaze to visually check the interlocutor across both feedback functions, prevalently before turn transitions.

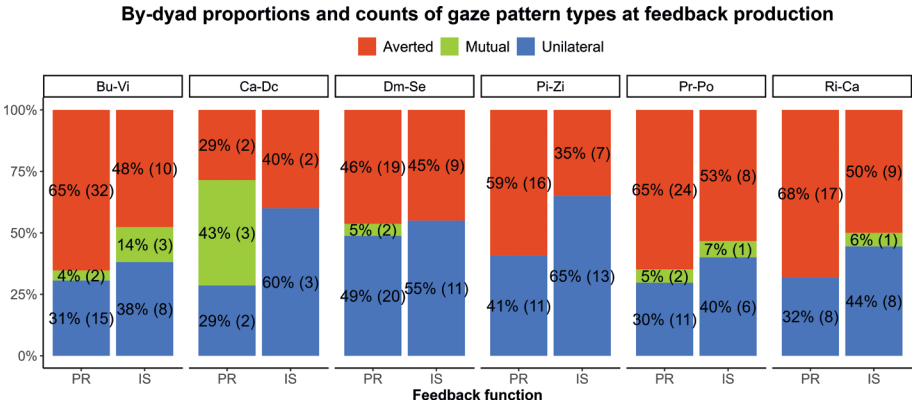
Figure 2 - *Proportions (and counts) of gaze pattern types when PR and IS vocal feedback signals were produced by all dyads in the Tangram game rounds*



A partially different view is shown in Figure 3, which displays the same proportions presented by dyad. Here, a higher percentage of unilateral gaze than averted gaze connected to the production of IS feedback is confirmed in three dyads only (Ca-Dc; Dm-Se; Pi-Zi). This suggests that dyad-specific variability is involved. It is worth noting that in these three dyads, no mutual gaze is present at all for IS, whereas the other three dyads (Bu-Vi; Pr-Po; Ri-Ca), with values of averted gaze higher than unilateral gaze, present a low percentage of mutual gaze. When summing the two values for mutual and unilateral gaze for the latter three dyads, results show a very similar distribution of "no gaze involved" vs "some gaze involved". As for PR feedback signals, the general trend is confirmed in four dyads out of six (Bu-Vi; Pi-Zi; Pr-Po;

Ri-Ca), that is a prevalence of averted gaze. However, while the specific percentage values are dyad-dependent, an overall trend can be identified: PR feedback shows a higher co-occurrence of averted gaze than IS, and IS shows a higher co-occurrence of unilateral gaze than PR.

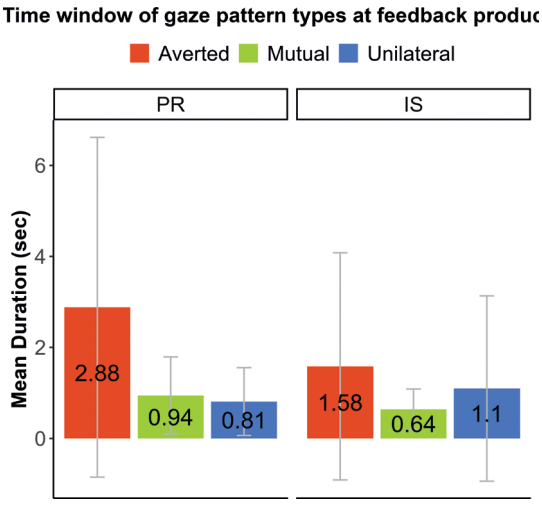
Figure 3 - *Proportions (and counts) of gaze pattern types at PR and IS vocal feedback productions, pooled by dyad*



4.2 Time window of gaze pattern types before feedback production

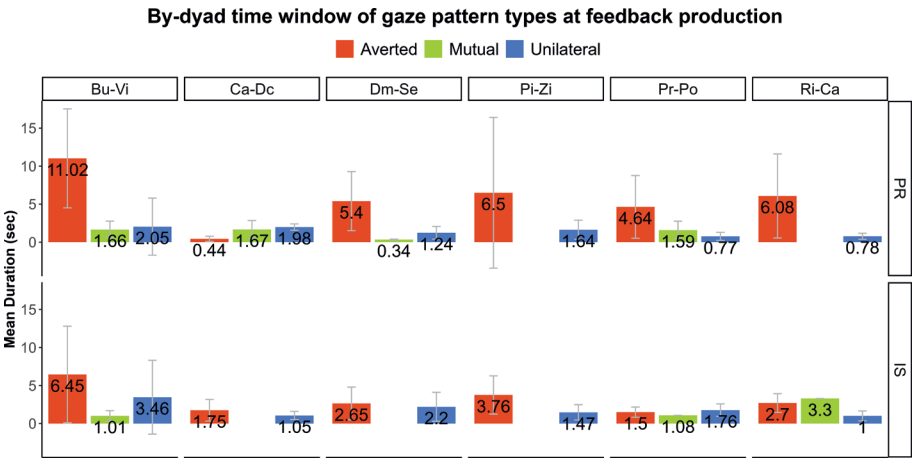
Figure 4 displays the time window for the gaze pattern preceding feedback production. It reveals that the duration of averted gaze up to the moment of feedback production is longer before PR feedback than IS feedback. Conversely, unilateral gaze is longer before IS feedback. The mutual gaze period is shorter before a turn initiation (that is, following an IS vocal feedback) than in cases in which feedback signals do not entail a change of speaker (PR vocal feedback). This shows that if a "gaze window" is activated and there is willingness to take the floor, this happens more quickly than when a speaker is only willing to acknowledge the other's speaker speech. However, given the very few data points of mutual gaze compared to averted and unilateral gaze (see raw numbers in Fig. 2), this observation should be considered only preliminary and requires further investigation.

Figure 4 - *Mean duration of gaze pattern window, for each gaze pattern types, when PR and IS vocal feedback are produced*



By-dyad values in Figure 5 show, again, some variability with respect to the observed duration of the gaze patterns preceding the different feedback functions, albeit less than for the occurrences of gaze pattern types (Fig. 3). Indeed, the averaged trends observed above hold in four dyads out of six (Bu-Vi; Dm-Se; Pr-Po; Ri-Ca, with the latter not presenting any instance of mutual gaze co-occurring with PR feedback, impeding a comparison across feedback functions). The other two dyads diverge from the general trends to varying extents. Ca-Dc shows shorter averted gaze before PR than IS feedback signals and longer unilateral gaze before PR than IS feedback signals, contradicting the general trend for both gaze patterns, while Pi-Zi differs from the trend only due to slightly longer unilateral gaze before PR than IS feedback signals. Importantly, looking at individual by-dyad values enables us to refine the observations on mutual gaze duration found in the averaged data. While averaged data (Fig. 4) display values under one second, by-dyad values are in all cases but one (Dm-Se for PR) above one second. Still, where data points are available for a comparison across feedback functions (Bu-Vi; Pr-Po), the statement that mutual gaze preceding PR feedback signals is longer than when a turn alternation (IS) is involved holds true.

Figure 5 - *Mean duration of gaze pattern window, for each gaze pattern types, when PR and IS vocal feedback are produced, pooled by dyad*



5. *Discussion*

Research question 1 assessed the relation between a specific gaze pattern and the production of turn-regulating vocal feedback. Our findings indicate that both averted and unilateral gaze were prevalent, with averted gaze being more frequently associated with PR, and unilateral gaze with IS. This somewhat unexpected result might be explained by the speakers' differential use of intonational cues for turn-taking purposes. Previous studies on backchannel intonation contours in Italian task-oriented dialogues (Savino, 2010, 2011, 2014; Sbranna et al., 2021; Sbranna et al., in press) have shown that continuers (PR vocal feedback) predominantly exhibit a rising intonation contour, whereas IS vocal feedback tokens are predominantly realised with a falling intonation. Therefore, the rising intonation characteristic of PR feedback production may be readily interpreted by interlocutors, thus reducing the need for additional visual confirmation. In contrast, the fact that IS vocal feedback co-occurred more frequently with unilateral gaze (than PR feedback) suggests that IS feedback may require additional visual confirmation to distinguish it from a possible falling PR vocal signal, and may serve as a request for joint attention to clarify the intention of a subsequent turn transition both vocally and visually.

We also observed that mutual gaze rarely co-occurs with vocal feedback signals, regardless of their turn-taking regulating function, confirming previous findings in a similar task-based study in German (Spaniol et al., 2023). This phenomenon might be attributed to the task-based setting and the presence of task materials, which necessitate visual attention to achieve the communicative goal of the conversation. It can be assumed that other interactional contexts – e.g. a free conversation without such a strong a visual competitor – might reveal different distributions of gaze direction and vocal feedback combinations than those emerging from our data.

On the other hand, other studies have also argued that mutual gaze is more closely associated with visual feedback in the form of head and hand gestures, rather than vocal feedback (Ferré, Renaudier, 2017; Bertrand et al., 2007). This might explain the low association between gaze with vocal feedback in our finding, although this hypothesis cannot be tested in the present study, as head and hand gestures are not analysed. Therefore, irrespective of the potential role of the task, this result is in line with previous findings, supporting the hypothesis that gestural acts are more readily followed by gestural responses.

Research question 2 aimed to examine whether the duration of a gaze pattern expedites the production of feedback utterance and potential subsequent turn alternation. Here, we observed longer averted gaze patterns preceding PR feedback production, whereas longer unilateral gaze patterns were observed preceding IS feedback production. One possible interpretation is that a prolonged averted gaze may signal a preference by the secondary speaker to maintain the listener role, visually indicating their unavailability for a turn transition and encouraging the primary speaker to continue. This situation may occur when one interlocutor describes their picture and the other participant is expected to listen carefully in order to complete the Tangram-game task. Conversely, when the secondary speaker intends to take the floor (i.e., when mostly longer unilateral gaze patterns occur), they may first look at the primary speaker to visually signal their intention. If this gaze is not reciprocated, resulting in a unilateral gaze, the secondary speaker eventually initiates the turn after a vocal feedback signal without having obtained visual attention. However, if the gaze is reciprocated, creating a period of mutual gaze, turn initiation after feedback occurs more rapidly than when no turn transition is intended after feedback (PR), possibly benefitting from visual contact as an additional channel to communicate the imminent turn alternation. This result contrasts with the findings of Kendrick et al. (2023), who observed that eye gaze does not speed up turn transitions.

Finally, these general trends appear to be relatively consistent across dyads, although the exact proportions and duration values exhibit substantial dyad-specific variability. This underscores the importance of exploring multimodal dynamics at the dyadic level to complement and refine interpretations based on averaged data. In essence, each interaction outcome is contingent upon the dynamics negotiated and co-constructed by the participants.

6. *Conclusion*

We explored the relationship between eye gaze patterns and turn-regulating vocal feedback (continuers/Passive Recipiency, turn-initiating feedback/Incipient Speakership) based on the occurrences and the time window of the gaze patterns, prior to and concurrent with the utterance of vocal feedback.

Our findings indicate that a binary classification of gaze patterns (mutual vs. averted) is not sufficient to capture the interplay between gaze and turn-regulating vocal feedback. In fact, despite averted gaze co-occurring with both types of vocal

feedback signals, unilateral gaze towards the interlocutor is found predominantly when a turn transition is imminent. We also observed that averted gaze is longer when preceding PR than when preceding IS. In contrast, unilateral gaze is longer before IS than before PR, suggesting that speakers may attempt to establish and wait for gaze contact before taking the turn, which they eventually indicate with a vocal feedback signal if their interlocutor fails to reciprocate their gaze. Finally, we found dyad-specific variability, as expected, due to the availability of two communication channels (vocal and visual), which can be used either in combination or independently. Moreover, it is worth remembering that each interaction can be regarded as a complex system resulting from the dynamics created by the participants in the conversation through a cycle of actions and responses, which inevitably results in unique detailed aspects beyond the more general common trends. Thus, we emphasise the importance of examining intra-dyadic details to obtain a more comprehensive understanding of interactional phenomena.

The application of this methodology to a larger sample of dyads and a comparative analysis with dialogue formats without visual competitors, will help further elucidate the phenomena underscored by these preliminary yet insightful results.

Acknowledgments

This research was funded by the Cluster Development Program Language Challenges, a funding line within the Excellent Research Support Program of the University of Cologne, and the German Research Foundation (DFG), grant number 281511265, SFB 1252 Prominence in Language.

Credit Author Statement

Simona Sbranna: Conceptualization, Data curation, Formal Analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing.

Michelina Savino: Data curation, Investigation, Methodology, Resources, Writing – review & editing.

Florence Bailles: Methodology, Writing – review & editing.

Martine Grice: Methodology, Supervision, Writing – review & editing.

References

Adams, R.B. Jr., Kleck, R.E. (2005). Effects of direct and averted gaze on the perception of facially communicated emotion. In *Emotion*, 5, 3-11. doi: 10.1037/1528-3542.5.1.3

Amador-Moreno, C.P., Mccarthy, M. & O'keeffe, A. (2013). Can English Provide a Framework for Spanish Response Tokens?. In Romero-Trillo, J. (Eds.), *Yearbook of Corpus Linguistics and Pragmatics 2013*, vol 1. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-6250-3_9

Argyle, M., Dean, J. (1965), Eye-contact, distance and affiliation. In *Sociometry*, 289-304, doi:102307/2786027

Auer, P. (2018). Gaze, addressee selection and turn-taking in three-party interaction. In Brône, G., Oben, B. (Eds.), *Eye-tracking in interaction: Studies on the role of eye gaze in dialogue*. John Benjamins, 197-232. http://doi.org/10.1075/ais.10.09aue

Barthel, M., Meyer, A.S. & Levinson, S.C. (2017). Next speakers plan their turn early and speak after turn-final "go-signals". In *Frontiers in psychology*, 8, 239095.

Bavelas, J.B., Coates, L. & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. In *Journal of Communication*, 52, 566-580. https://doi.org/10.1111/jcom.2002.52.issue-3

Beattie, G.W. (1978). Floor apportionment and gaze in conversational dyads. In *British Journal of Social and Clinical Psychology*, 17, 7–15. http://doi.org/10.1111/j.2044-8260.1978.tb00889.x

Beattie, G.W. (1979). Contextual constraints on the floor-apportionment function of speaker gaze in dyadic conversation. In *British Journal of Social and Clinical Psychology*, 18, 391-392.

Bertrand, R., Ferré, G., Blache, P., Espesser, R. & Rauzy, S. (2007). Backchannels revisited from a multimodal perspective. In *Auditory-visual Speech Processing*, 1-5.

Bissonnette, V.L. (1993). *Interdependence in dyadic gazing* [Doctoral dissertation, The University of Texas at Arlington]. The University of Texas at Arlington ProQuest Dissertations Publishing.

Blythe, J., Gardner, R., Mushin, I. & Stirling, L. (2018). Tools of engagement: selecting a next speaker in Australian aboriginal multiparty conversations. In *Research on Language and Social Interaction*, 51, 145-170. http://doi.org/10.1080/08351813.2018.1449441

Brône, G., Oben, B., Jehoul, A., Vranjes, J. & Feyaerts, K. (2017). Eye gaze and viewpoint in multimodal interaction management. In *Cognitive Linguistics*, 28, 449-483. http://doi.org/10.1515/cog-2016-0119

Clancy, P.M., Thompson, S.A., Suzuki, R. & Tao, H. (1996). The conversational use of reactive tokens in English, Japanese, and Mandarin. In *Journal of Pragmatics*, *26* (3), 355-387.

Cummins, F. (2012). Gaze and blinking in dyadic conversation: a study in coordinated behaviour among individuals. In *Language and Cognitive Processes*, 27, 1525-1549. http://doi.org/10.1080/01690965.2011.615220

Degutyte, Z., Astell, A. (2021). The role of eye gaze in regulating turn taking in conversations: A systematized review of methods and findings. In *Frontiers in Psychology*, 12, 616471. http://doi.org/10.3389/fpsyg.2021.616471

Drummond, K., Hopper, R. (1993). Back channels revisited: Acknowledgment tokens and speakership incipiency. In *Research on Language and Social Interaction*, 26, 157-177. http://doi.org/10.1207/s15327973rlsi2602_3

Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology*, 23, 283-292.

Duncan, S. (1973). Toward a Grammar for Dyadic Conversation. In *Semiotica*, 9(1), 29-46. https://doi.org/10.1515/semi.1973.9.1.29

Duncan, S., Fiske, D.W. (1977). *Face-to-face interaction: Research, methods, and theory*. L. Erlbaum Associates.

Eberhard, K.M., Nicholson, H. (2010). Coordination of understanding in face-to-face narrative dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*, 2051-2056. eScholarship, University of California. https://escholarship.org/uc/item/6bk8h630

Egbert, M.M. (1996). Context-sensitivity in conversation: Eye gaze and the German repair initiator bitte?. In *Language in Society*, 25(4), 587-612.

ELAN (Version 6.7) [Computer software]. (2023). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Fellegy, A.M. (1995). Patterns and functions of minimal response. In *American Speech International Journal of Educational Best Practices*, 2 (1), 186-199.

Ferré, G., Renaudier, S. (2017). Unimodal and Bimodal Backchannels in Conversational English. In *Proceedings of SemDial*, Saarbrücken, Germany, 27-37.

Foulsham, T., Cheng, J.T., Tracy, J.L., Henrich, J. & Kingstone, A. (2010). Gaze allocation in a dynamic situation: effects of social status and speaking. In *Cognition*, 117, 319–331. http://doi.org/10.1016/j.cognition.2010.09.003

Fries, C.C. (1952). *The structure of English*. London: Longmans, Green; Company.

Gardner, R. (2001). *When listeners talk: Response tokens and listener stance*. Amsterdam: John Benjamins Publishing Company.

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. Academic Press.

Hall, S. (1995). Review of the Observer 3.0. In *Journal of Child Psychology & Psychiatry*, 36,1495-1498.

Heldner, M., Hjalmarsson, A. & Edlund, J. (2013). Backchannel relevance spaces. In Asu, E.L., Lippus, P. (Eds.), *Nordic Prosody: Proceedings of the XIth Conference, Tartu 2012*. Peter Lang, 137-146. http://doi.org/10.3726/978-3-653-03047-1

Heritage, J. (1984). A change-of-state token and aspects of its sequential placement. In *Structures of Social Action: Studies in Conversation Analysis*, 299-345.

Hjalmarsson, A., Oertel, C. (2012). Gaze direction as a back-channel inviting cue in dialogue. In *IVA 2012 workshop on realtime conversational virtual agents* (Vol. 9). Citeseer.

Ho, S., Foulsham, T. & Kingstone, A. (2015). Speaking and listening with the eyes: gaze signaling during dyadic interactions. In *PLoS ONE., 10*, Article e0136905. http://doi.org/10.1371/journal.pone.0136905

Hollenstein, T. (2013). State Space Grids. In *State Space Grids*. Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-5007-8_2

Holler, J., Kendrick, K.H. & Levinson, S.C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. In *Psychonomic bulletin & review*, 25(5), 1900-1908. https://doi.org/10.3758/s13423-017-1363-z

Hu, L., Chen, G. (2021). Trajectories of Idea Emergence in Dialogic Collaborative Problem Solving: Toward a Complex Dynamic Systems Perspective. In *Frontiers in Psychology*, *12*, 735534. https://doi.org/10.3389/fpsyg.2021.735534

Jacoby, S., Ochs, E. (1995). Co-Construction: An Introduction. In *Research on Language and Social Interaction*, 28(3), 171-183. https://doi.org/10.1207/s15327973rlsi2803_1

Janz, A. (2022). Navigating common ground using feedback in conversation – A phonetic analysis. Cologne, Germany: University of Cologne. (MA thesis).

Jefferson, G. (1983). Two explorations of the organization of overlapping talk in conversation: Notes on some orderlinesses of overlap onset. In *Tilburg Papers in Language and Literature, 28*, 1–28. https://liso-archives.liso.ucsb.edu/Jefferson/Onset.pdf

Jokinen, K., Furukawa, H., Nishida, M. & Yamamoto, S. (2013). Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems*, 3, 12. http://doi.org/10.1145/2499474.2499481

Jokinen, K., Nishida, M. & Yamamoto, S. (2009). Eye-gaze experiments for conversation monitoring. In *Proceedings of the 3rd International Universal Communication Symposium*. ACM, 303-308. http://doi.org/10.1145/1667780.1667843

Jurafsky, D., Shriberg, E., Fox, B. & Curl, T. (1998). Lexical, prosodic, and syntactic cues for dialog acts. In Stede, M., Wanner, L. & Hovy, E. (Eds.), *Discourse Relations and Discourse Markers*. ACL Anthology, 114-120. https://aclanthology.org/W98-0319.pdf

Kendon, A. (1967). Some functions of gaze direction in social interaction. In *Acta Psychologica*, 26, 22-63. https://doi.org/10.1016/0001-6918(67)90005-4

Kendrick, K.H., Holler, J. & Levinson, S.C. (2023). Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. In *Philosophical Transaction of the Royal Society B, 378*, Article 20210473. http://doi.org/10.1098/rstb.2021.0473

Kraut, R.E., Lewis, S.H. & Swezey, L.W. (1982). Listener responsiveness and the coordination of conversation. In *Journal of Personality and Social Psychology*, 43, 718-731.

Lambertz, K. (2011). Back-channelling: The use of yeah and mm to portray engaged listenership. In *Griffith Working Papers in Pragmatics and Intercultural Communication*, 4, 11-18.

Lerner, G.H. (2003). Selecting next speaker: The context-sensitive operation of a context-free organization. In *Language in Society*, 32, 177-201. http://doi.org/10.1017/S004740450332202X

Levinson, S.C., Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. In *Frontiers in Psychology*, 6, 731. https://doi.org/10.3389/fpsyg.2015.00731

Liang, J., Zou, Q., Liang, Y., Wu, W. & Yan, J. (2021). Emotional Gaze: The Effects of Gaze Direction on the Perception of Facial Emotions. In *Frontiers in Psychology*, 12. https://doi.org/10.3389/fpsyg.2021.684357

Lücking, A., Ptock, S. & Bergmann, K. (2011). Assessing agreement on segmentations by means of staccato, the segmentation agreement calculator according to thomann. In *International gesture workshop*. Berlin, Heidelberg: Springer Berlin Heidelberg, 129-138. https://doi.org/10.1007/978-3-642-34182-3_12

Myszka, T.J. (1975). Situational and intrapersonal determinants of eye contact, direction of gaze aversion, smiling and other non-verbal behaviors during an interview [Doctoral dissertation, University of Windsor]. Electronic Thesis and Dissertations University of Windsor. https://scholar.uwindsor.ca/etd/3477

Novick, D., Hansen, B. & Ward, K. (1996). Coordinating turn-taking with gaze. In *Proceedings of Fourth International Conference on Spoken Language Processing* (ICSLP-96). IEEE Xplore, 1888-91. http://doi.org/10.1109/ICSLP.1996.608001

Oertel, C., Włodarczak, M., Edlund, J., Wagner, P. & Gustafson, J. (2012). Gaze patterns in turn-taking. In Sproat, R. (Ed.), *Proceedings of Interspeech 2012*. ISCA, 2246-2249. http://doi.org/10.21437/Interspeech.2012-132

Ondáš, S., Kiktová, E., Pleva, M. & Juhár, J. (2023). Analysis of Backchannel Inviting Cues in Dyadic Speech Communication. In *Electronics*, 12, 3705. https://doi.org/ 10.3390/electronics12173705

Paulmann, S., Jessen, S. & Kotz, S.A. (2009). Investigating the multimodal nature of human communication. In *Journal of Psychophysiology*, 23, 63-76. https://doi.org/10.1027/0269-8803.23.2.63

Perniss, P. (2018). Why we should study multimodal language. In *Frontiers in Psychology*, *9*, 342098. https://doi.org/10.3389/fpsyg.2018.01109

Poppe, R., Truong, K.P. & Heylen, D. (2011). Backchannels: Quantity, type and timing matters. In *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11*. Springer Berlin Heidelberg, 228-239.

Rasenberg, M., Pouw, W., Özyürek, A. & Dingemanse, M. (2022). The multimodal nature of communicative efficiency in social interaction. In *Scientific Report*, 12, Article 19111. https://doi.org/10.1038/s41598-022-22883-w

Rossano, F. (2012). Gaze behaviour in face-to-face interaction. PhD Dissertation, Radboud University, Nijmegen. https://hdl.handle.net/11858/00-001M-0000-000F-ED23-5

Rossano, F., Brown, P. & Levinson, S.C. (2009). Gaze, questioning and culture. In *Conversation Analysis, 27*, 187-249. http://doi.org/10.1017/CBO9780511635670.008

Rutter, D.R., Stephenson, G.M., Ayling, K. & White, P.A. (1978). The timing of looks in dyadic conversation. In *British Journal of Social and Clinical Psychology*, 17, 17-21. http://doi.org/10.1111/j.2044-8260.1978.tb00890.x

Sacks, H., Schegloff, E. & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. Lg. 50.696-735.

Savino, M. (2010). Prosodic strategies for backchanneling in Italian Map Task dialogues. In Botinis, A. (Ed.), *Proceedings of the ISCA Tutorial and Research Workshop on Experimental Linguistics (Exling 2010)*. ISCA, 157-160. http://doi.org/.36505/ExLing-2010/03/0040/000160

Savino, M. (2011). The intonation of backchannels in Italian task-oriented dialogues: cues to turn-taking dynamics, information status and speaker's attitude. In *Proceedings of the 5th Language and Technology Conference: "Human Language Technology as a Challenge for Computer Science and Linguistics*. Poznan (Poland), 25-27 November 2011, 370-374.

Savino, M. (2014). The intonation of backchannel tokens in Italian collaborative dialogues. In Vetulani, Z., Mariani, J. (Eds.), *Human Language Technology Challenges for Computer Science and Linguistics*, Springer LNAI Series n. 8387, 17-28. https://doi.org/10.1007/978-3-319-08958-4_3

Savino, M., Refice, M. (2013). Acknowledgement or reply? Prosodic features for disambiguating pragmatic functions of the Italian token 'sì'. In *Proceedings of the 7th International*

Conference on Speech Technology and Human-Computer Interaction (SpeD 2013). IEEE, 117-112. http://doi.org/10.1109/SpeD.2013.6682660

Savino, M., Lapertosa, L. & Refice, M. (2018). Seeing or not seeing your conversational partner: the influence of interaction modality on prosodic entrainment. In Karpov, A., Jokisch, O. & Potapova, R. (Eds.), *Speech and Computer, Proceedings of the 20th SPECOM International Conference*, Leipzig 18-22 September 2918, LNCS series n. 11096, Springer, 574-584. DOI: 10.1007/978-3-319-99579-3

Sbranna, S., Möking, E., Wehrle, S. & Grice, M. (2022). Backchannelling across languages: Rate, lexical choice and intonation in L1 Italian, L1 German and L2 German. In Frota, S., Vigário, M. (Eds.), *11th International Conference on Speech Prosody*. http://doi.org/10.21437/SpeechProsody.2022-149

Sbranna, S., Wehrle, S. & Grice, M. (in press). A multi-dimensional analysis of backchannels in L1 German, L1 Italian and L2 German. In *Language, Interaction, and Acquisition*. 15.2.

Schegloff, E.A. (1982). Discourse as an interactional achievement: Some uses of "uh huh" and other things that come between sentences. In Tannen, D. (Ed.), *Analyzing discourse: Text and talk, Georgetown University Roundtable on Languages and Linguistics* 1982, 71–93. Washington, DC: Georgetown University Press.

Schegloff, E.A. (2020). Interaction: The infrastructure for social institutions, the natural ecological niche for language, and the arena in which culture is enacted. In *Roots of human sociality*. Routledge, 70-96.

Simon, C. (2018). The functions of active listening responses. In *Behavioural Processes*, 157, 47–53.

Spaniol, M., Janz, A., Wehrle, S., Vogeley, K. & Grice, M. (2023). Multimodal signalling: The interplay of oral and visual feedback in conversation. In Skarnitzl, R., Volín, J. (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences*. Guarant International, 4110-4114.

Stahl, G. (2016). *Constructing dynamic triangles together: The development of mathematical group cognition*. Cambridge: Cambridge University Press.

Streeck, J. (2014). Mutual gaze and recognition. Revisiting Kendon's Gaze direction in two-person conversation. In Seyfeddinipur, M., Gullberg, M. (Eds.), *From Gesture in Conversation to Visible Action as Utterance*. John Benjamins, 35–55. http://doi.org/10.1075/z.188.03str

Thomann, B. (2001). Observation and judgment in psychology: Assessing agreement among markings of behavioral events. In *Behavior Research Methods, Instruments, & Computers*, 33, 339-348.

Truong, K.P., Poppe, R., Kok, I.D. & Heylen, D. (2011). A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. *Proc. Interspeech 2011*, 2973-2976, DOI: 10.21437/Interspeech.2011-744

Wehrle, S. (2023). *Conversation and intonation in autism: A multi-dimensional analysis*. Studies in Laboratory Phonology 14. Berlin: Language Science Press. DOI: 10.5281/zenodo.10069004

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*.

Yngve, V. (1970). On getting a word in edgewise. In *Papers from the sixth regional meeting, Chicago Linguistic Society*, 567-577. Chicago: Chicago Linguistic Society.

Zellers, M., House, D. & Alexanderson, S. (2016). Prosody and hand gesture at turn boundaries in Swedish. In *Speech Prosody*, 831-835.

Zima, E., Weiss, C. & Brône, G. (2019). Gaze and overlap resolution in triadic interactions. In *Journal of Pragmatics*, 140, 49-69.

DOI: 10.17469/O2112AISV000011

FEDERICA CAVICCHIO, MIRKO GRIMALDI

The Effect of Subtitles on Second Language Pronunciation

The increasing prevalence of video-on-demand services for watching movies and TV shows presents a unique opportunity for foreign language learning, thanks to easy access to original language content. When watching films in the original language, viewers typically can follow the audiovisual content with subtitles either in their language (L1) or the film's language (L2). Here, we present a pilot study that examines the impact of English (L2) subtitles, Italian (L1) subtitles, or no subtitles on L1 native Italian speakers' production of L2 English phonemes. Our findings reveal that subtitles in the L2 language (English) can improve the production of L2 phonemes. However, while L2 subtitles facilitated the improved pronunciation of the vowel /æ/, no significant enhancements were observed for /ʌ/. The study emphasizes the utility of incorporating reading and listening in L2 for effective language acquisition. It also points out the importance of considering L1 phonemic traits that may pose challenges to L2 acquisition.

Keywords: Audiovisual online content, Subtitles, L2 acquisition, L1-L2 interference.

1. *Introduction*

The critical period hypothesis (CPH, see, for example, Lennenberg, 1967; Pinker, 1984; Hartshorne, Tenenbaum & Pinker, 2018; MacWhinney, 1987) posits that there is a specific window in human development during which acquisition of the native language (L1) occurs most naturally and efficiently. This period is typically considered to be during early childhood. After this window closes, it is believed that the ability to learn a new language (L2) with native-like fluency significantly decreases. This notion is often juxtaposed with neuroplasticity (see, for example, Merzenich, Nahum, & Van Vleet, 2013), that is, the capacity of the brain to form new neural connections throughout an individual's life in response to new experiences and learning.

During the critical period, children's brains exhibit heightened neuroplasticity (e.g., Kuhl, Conboy, Padden, Nelson & Pruitt, 2005; Kuhl, Conboy, Coffey-Corina, Padden, Rivera-Gaxiola & Nelson, 2008; Morford, Mayberry, 2000), allowing for more efficient development of neural pathways responsible for language processing and production. This heightened plasticity enables children to assimilate new phonological systems more readily, particularly when exposed to an L2 through native speakers, even in the context of the classroom (see Cheour, Shestakova, Alku, Ceponiene & Näätänen, 2002; Shestakova, Huotilainen, Čeponien & Cheour, 2003; Peltola, Kuntola, Tamminen, Hämäläinen & Aaltonen, 2005).

Neuroplasticity allows for continuous learning and adaptation within the human brain, including acquiring an L2 at different ages (Stein, Kuntola, Tamminen, Hämäläinen & Aaltonen, 2006; Leagault, Grant, Fang & Li, 2019a; Leagault, Fang, Lan & Li, 2019b). Neuroplasticity underpins the processes through which adult learners can immerse themselves in an L2 environment and experience a range of linguistic inputs that challenge them to categorise and produce new L2 sounds, as described by Flege (1995), Flege, Bohn (2021) and Winkler, Kujala, Tiitinen, Sivonen, Alku, Lehtokoski, Czigler, Csépe, Ilmoniemi & Näätänen (1999). The ability to leverage neuroplasticity for L2 learning, even in adults, underlines the brain's capacity to reorganise and adjust to new linguistic contexts, even outside the critical period. Several factors might amplify L2 learning, including linguistic aptitude (Hu, Ackermann, Martin, Erb, Winkler & Reiterer, 2013; Chai, Berken, Barbeau, Soles, Callahan & Chen, 2016), along with the intensity (Thomson, Derwing, 2015) and quality (Zhang, Kuhl, Imada, Iverson, Pruitt, Stevens & Nemoto, 2009) of L2 learning. Conversely, when L2 acquisition occurs in a classroom environment where the teachers have a non-native accent, and the primary language of communication among students remains their L1, the immersion necessary for this neural activation is insufficient. In such environments, the exposure to native L2 phonetics and syntax is significantly reduced, and students may inadvertently adopt their non-native teachers' inaccurate pronunciations and rhythms. Furthermore, the prevalence of L1 during classroom activities may lead to a reliance on familiar linguistic structures, preventing the necessary neural adaptation and integration of L2 linguistic patterns (Grimaldi, Sisinni, Gili Fivela, Invitto, Resta & Alku, 2014).

Despite this inherent potential for neuroplastic adaptation to new linguistic environments, longitudinal studies on auditory plasticity have shown the challenges involved in L2 phonetic learning within the constraints of traditional classroom settings. These studies, such as those conducted by Grimaldi et al. (2014), Jost, Eberhard-Moscicka, Pleisch, Heusser, Brandeis, Zevin & Maurer (2015), Hisagi, Shafer, Miyagawa, Kotek, Sugawara & Pantazis (2016), Wottawa, Adda-Decker & Isel (2022), and Højlund, Horn, Sørensen, McGregor & Wallentin (2022), indicate that sustained exposure to an L2 in a classroom environment does not necessarily result in changes to auditory plasticity. These results suggest that, over time, the auditory system's sensitivity to the acoustic nuances of a new language may not be significantly enhanced in such settings. As posited by Piske (2007), one potential explanation for this finding is that the quality and intensity of L2 stimuli provided in classrooms are insufficient to reinvigorate the auditory system's sensitivity to the spectro-temporal characteristics unique to non-native languages. In a typical classroom setting, the auditory experience of L2 may be diluted by various factors: the presence of non-native accents, limited opportunity for immersive interaction, and the predominance of controlled, repetitive exercises over spontaneous language use. These findings have considerable implications for the formation of mental representations of L2 phonemes. The ability to create accurate mental models of

L2 phonemic structures is vital for successful L2 acquisition, but when the auditory input lacks richness and authenticity or when exposure to the language is fragmented or inconsistent, the establishment of clear and distinct phonemic representations may be impeded.

Classroom learning may challenge the development of L2 representations for several reasons. First, the frequency of exposure to L2 sounds may be limited to a few hours per week, which is insufficient compared to the constant stimulation provided in a naturalistic environment. Second, the diversity of L2 input in a classroom is often narrower than in real-world settings, offering fewer opportunities for learners to experience the full range of phonetic variations and contextual uses of the language. In summary, the dynamics of a classroom may not always assist the perceptual and learning processes essential for refining the perception and production of L2 phonemes. Regarding L2 sound acquisition, recent meta-analytic reviews (e.g., Sakai, Moorman, 2018; Rato, Oliveira, 2023) have robustly validated that the use of brief training sessions focused on L2 sound perception (e.g. High Variability Phonetic Training, HVPT) can enhance the identification and production of L2 phonemes. In light of these findings, we must rethink L2 teaching methodologies incorporating methods that can enhance auditory plasticity, such as immersive experiences and short acoustic training, and technologies that can simulate natural language exposure.

The need for innovative immersive learning strategies resonates with the growing use of video-on-demand services, which provide access to content in its original language accompanied by subtitles in either the learners' native language (L1) or the original language of the film (L2). In a seminal study, Mitterer, McQueen (2009) explored the effects of either L1 or L2 subtitles on Dutch L2 speakers' perception of Scots English words after watching the film "Trainspotting". Their findings suggested that exposure to L2 subtitles could enhance speech comprehension more than L1 subtitles. Subsequent studies have confirmed that using L2 subtitles significantly facilitates L2 acquisition. A study by Vulchanova, Aurstad, Kvitnes & Eshuis (2015) demonstrates that L2 subtitles are more effective than L1 subtitles in aiding the acquisition of L2 vocabulary. Frumuselu, De Maeyer, Donche & Colon Plana (2015) found that L2 subtitles contribute to a better long-term understanding of spoken L2. These results are likely due to the direct visual reinforcement of spoken words and the active engagement of both auditory and visual processing pathways, which enriches the language learning experience and aids the processing of language constructs. The simultaneous exposure to spoken and written forms of the target language enables learners to make connections between the phonetic sounds of words and their orthographic representations, fostering a more comprehensive grasp of the language as used in natural contexts (on the subject of L2 phono-lexical representation and phoneme/grapheme interference see Darcy, Holliday, 2019; and Bassetti, Cerni & Masterson, 2022, respectively).

Expanding on the effect on L2 perception of L2 subtitles, Wisniewska, Mora (2021) report that exposure to L2 subtitles not only can enhance L2 phoneme

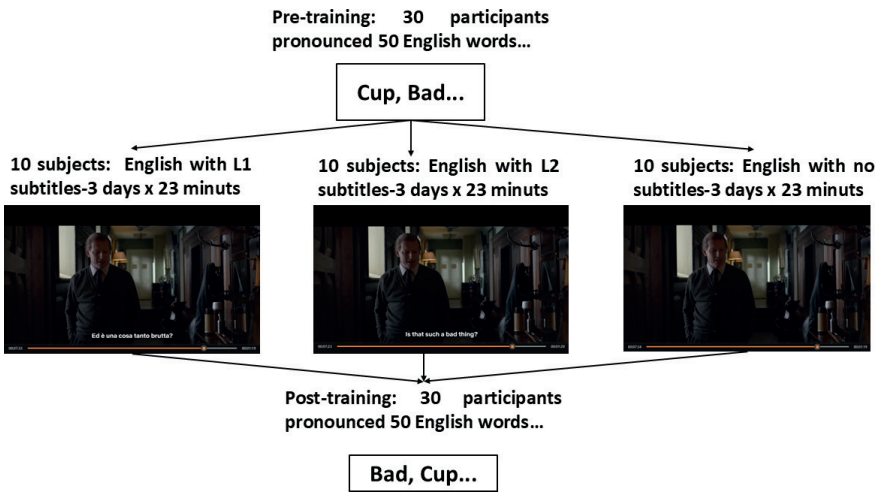
perception, effectively helping L2 learners to distinguish between similar words, but can also improve the learners' accent, as per native speakers' ratings of L2 speech productions. However, it is crucial to note that while subtitles may improve the perception of L2 pronunciation, they do not necessarily translate into the learners' ability to produce all the sounds accurately, as reported by Hutchinson, Dmitrieva (2022), who point out that exposure to L2 subtitles does not improve L2 vowel production, as per native speakers' ratings.

In summary, according to results from previous studies, L2 subtitles, while widely acknowledged for their efficacy in enhancing language comprehension and vocabulary acquisition, may not directly facilitate the accurate production of L2 phonemes. Our pilot study aims to test whether exposure to L2 subtitles can influence the pronunciation skills of native Salento Italian speakers learning English as L2. Our study specifically investigates the impact of different subtitle settings on the acquisition of British English phonemes that are commonly mispronounced by Italian speakers from Salento. We selected the phonemes /ʌ/ (as in "cup") and /æ/ (as in "bad"), which are often confused by Salento Italian speakers with the Italian vowels /a/ and /ɛ/, respectively (Escudero, Sisinni, & Grimaldi, 2014). English subtitles might provide auditory and visual reinforcement of the phonetic characteristics of English, potentially aiding learners in overcoming the influence of their L1 phonemes. By contrast, Italian subtitles might reinforce reliance on native pronunciation norms, whereas the no subtitles condition, should not help or hinder L2 pronunciation. Our method included pre- and post-tests to evaluate the accuracy of phoneme production before and after participants were exposed to one of the three subtitles conditions. Their L2 productions were compared to the productions of native English speakers. We were able to make this direct comparison because, unlike previous studies, we do not employ qualitative assessments by L1 speakers to assess the participants' L2 productions. Instead, to quantify the deviation of L2 speakers' vowel pronunciation from that of native English speakers, we utilised the Mahalanobis distance (Mahalanobis, 1936), a statistical measure useful in the context of measuring how similar or different a given observation is from a reference group, especially when the data is multidimensional. Unlike other metrics, such as the Euclidean distance, the Mahalanobis distance considers the correlations across multidimensional data. Therefore, the Mahalanobis distance is particularly suitable for analysing vowel existence spaces. Vowels are characterised by their formant frequencies, specifically the first and second formants (F1 and F2), which are multidimensional and interrelated. The Mahalanobis distance provides a robust method for measuring distances within this complex space. Consequently, it enabled us to precisely gauge how closely or not the vowel pronunciations of L2 speakers matched those of native English speakers, considering variations within each subtitles group. Unlike previous studies, our approach provides a quantitative, rather than qualitative, assessment of the L2 speakers' phonetic variation.

2. *Materials and Methods*

Thirty native Italian speakers from the Salento region (mean age 23.4 years, 23 women) participated in the study. They were all first-year students at the Faculty of Foreign Languages and Literatures, and their English level, according to CEFR, was B1. It should be noted that a five-vowel system characterises the local dialect in Salento (/i, ɛ, a, ɔ, u/), with noted variations as detailed by Grimaldi (2003) and Romano (2013). We selected two segments from the Netflix series "The Crown," specifically from Season 1, Episode 2 and Season 4, Episode 8, which together amounted to approximately 23 minutes of viewing. From these episodes, we selected twenty-five words containing the vowel/ʌ/ and twenty-five words containing the vowel /æ/ in stressed positions. The words were chosen for their similar frequency of usage, as reported in the *BNC XML*. Five female actresses and five male actors pronounced the words in the film segments, and the English variant spoken in "The Crown" is the Received Pronunciation (RP). In Fig. 1, we present a sketch of the experimental procedure.

Figure 1 - *Experimental Procedure Overview. The experimental protocol began with the pre-test. Participants recorded 50 words, 25 words containing the vowel /æ/ (as in "bad") and 25 words with the vowel /ʌ/ (as in "cup"). The words were selected from the film segments used later in the study. Participants were then assigned to one of three subtitle conditions and then re-tested in the post-test*



Participants recorded the words selected from the TV series segments in a soundproof booth. The words were displayed one at a time in written format via MS PowerPoint on a Lenovo portable PC. Participants articulated the words at their own pace. The recording sessions lasted 10 minutes on average and were conducted using Audacity software version 3.08, with a Trust 21674 Table Microphone. Subsequently, participants were divided into three groups: the first

group watched the film segments with L2 English subtitles, the second with L1 Italian subtitles, and the third without subtitles. Participants viewed the segments once daily for three consecutive days via the online experiment platform *Gorilla*. Following this exposure, participants returned to record the same words presented in the pre-test, which were presented to them in a counterbalanced order with respect to the initial recording.

3. *Data Analysis*

We extracted the F1 and F2 formant values of the /ʌ/ and /æ/ vowels articulated by the 30 L2 participants and 10 L1 speakers using DARtmouth Linguistic Automation (DARLA) software (Reddy, Stanford, 2015). To assess the vowels' production differences between L2 and L1 speakers, we computed the Mahalanobis distance (Mahalanobis, 1936) using package dplyr (Wickham, François, Henry, Müller & Vaughan, 2023) in R (version 4.2). We calculated the mean vectors and the covariance matrix for F1 and F2 values for each vowel and each speaker and then computed the Mahalanobis distance between L2 and L1 speakers. Smaller Mahalanobis distance values indicate a higher similarity between the vowel spaces of L2 speakers and L1 speakers, suggesting a closer approximation to native-like vowel production by the L2 speakers. Conversely, larger Mahalanobis distances point to significant discrepancies in vowel production between L2 and L1 speakers. By comparing the Mahalanobis distance values of the pre-test and post-test under different subtitle conditions, we can gauge whether a phonological adjustment may have occurred for L2 speakers due to exposure to different subtitle conditions.

4. *Results and Discussion*

Tab. 1 reports the statistical summary of Mahalanobis distances for the vowels /ʌ/ and /æ/ measured before (pre-test) and after (post-test) the participants were exposed to three different subtitle conditions. The summary includes the mean, maximum, minimum, and median Mahalanobis distance values. The results indicate a reduction in Mahalanobis distances for the vowel /æ/ following exposure to subtitles. Specifically, the mean Mahalanobis distance for /æ/ in the pre-test was 2.23; this distance decreased to 1.65 after exposure to L1 subtitles and an even lower mean value of 0.91 when subtitles were in the L2. In contrast, the condition with no subtitles showed no substantial change in the mean value of the Mahalanobis distance (mean=2.6). Conversely, the phoneme /ʌ/ exhibited no significant variation in Mahalanobis distances across different subtitle conditions, indicating that the efficacy of subtitle-based phonetic training may be vowel-specific. This finding suggests that the impact of subtitles on phonetic learning is influenced not only by the language of the subtitles but also by the phonemes involved, with L2 subtitles showing greater effectiveness for the vowel /æ/ but not for /ʌ/.

Table 1 - *Summary of Mahalanobis Distance Measures for Vowels /ʌ/ and /æ/ comparing values from the pre-test and post-test across different subtitle conditions. The mean Mahalanobis distance values lower in the post-test, indicating greater proximity to native speakers' vowels, are highlighted in bold print*

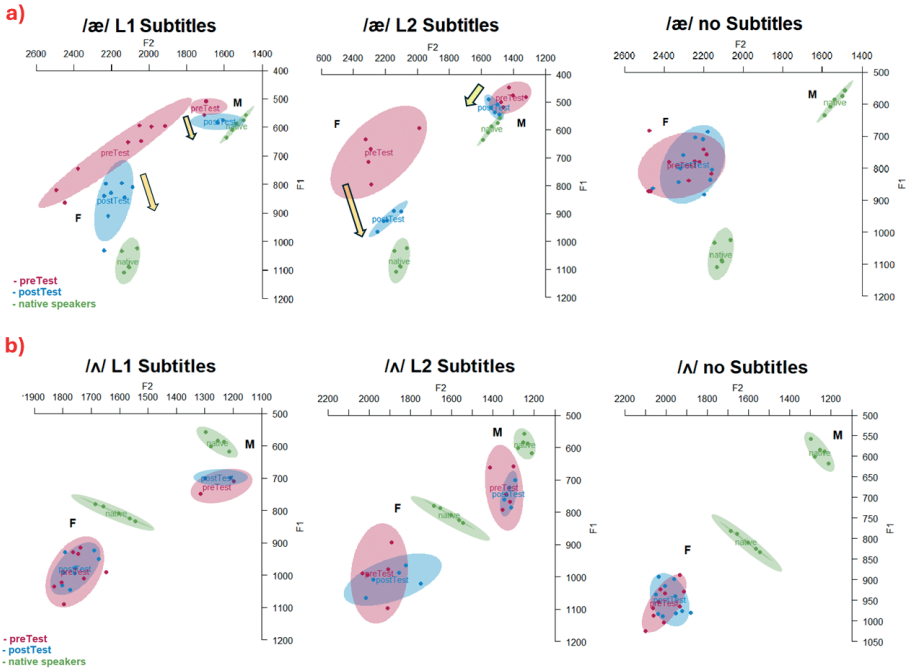
		Pre-test	*Post-test L2 Subtitles*	*Post-test L1 Subtitles*	*Post-test No Subtitles*
/ʌ/	max	9.48	6.93	6.6	11.8
	min	0.02	0.02	0.02	0.07
	mean	2.25	2.04	2.06	2.28
	median	1.97	1.6	2	2.32
/æ/	max	11.65	5.92	9.3	14.72
	min	0.02	0.004	0.009	0.08
	mean	2.23	*0.91*	*1.65*	2.6
	median	1.58	0.65	1.01	1.82

In Fig. 2, we present a visual assessment of participants' different subtitle conditions for vowel production. The figure presents the existence areas (F1xF2) for vowels /æ/ and /ʌ/, comparing data collected during the pre-test and post-test sessions. The analysis is differentiated by subtitle conditions (L1 subtitles, L2 subtitles, and no subtitles). Additionally, the figure delineates the vowel existence areas for both female and male speakers. Participants consistently undershoot (Flege, Schirru & MacKay, 2003) across conditions, that is, participants did not fully achieve the articulatory targets of the English vowels and consistently fell short of the precise phonetic qualities of the target English sounds. However, for the vowel /æ/, speakers in the L2 subtitles condition showed a notable shift in their vowel existence areas towards those typical of native speakers.

The findings of this pilot study show that the subtitles' role in L2 production is nuanced. Our results revealed that exposure to L2 subtitles improved the vowel /æ/production, as evidenced by the reduced Mahalanobis distances in the post-test data for the L2 subtitles' group. This improvement aligns with previous observation that English subtitles might facilitate the L2 learners' ability to differentiate and accurately produce non-native phonemes. In particular, our results are congruent with the findings of Wisniewska, Mora (2021), which suggested that L2 subtitles can enhance phoneme perception and even improve accent in L2 learners. Using the Mahalanobis distance, our study adds a quantitative dimension to Wisniewska and Mora (2021) qualitative observations, demonstrating that L2 subtitles' impact on L2 speech production is selective and varies by phoneme. As Hutchinson, Dmitrieva (2022) noted, exposure to L2 subtitles does not necessarily translate into an equivalent L2 speech production enhancement for all the L2 vowels. Indeed, in our study, the phoneme /ʌ/ showed no substantial changes in the Mahalanobis distances across any of the subtitle conditions, indicating a potential limitation in the impact of subtitles on the acquisition and production of that specific phonemes.

The "resistance to change" for /ʌ/ could depend on a stronger L1 interference for /ʌ/ and /æ/.

Figure 2 - *Existence areas (F1XF2) for vowels /æ/ (panel a) and /ʌ/ (panel b) in the pre-test and post-test by subtitle conditions (L1 subtitles, L2 subtitles, and no subtitles). Vowel existence areas are pinpointed for female (F) and male (M) speakers*



Previous studies by Cavicchio, Grimaldi (2022) have found that Salento Italian speakers consistently identify the vowel /ʌ/ less accurately than the vowel /ɒ/. In Salento Italian, a five-vowel system, lip rounding is not a distinctive feature, unlike frontness/backness and the height trait. Indeed, the trait [±rounded] in Salento Italian is redundant, as the vowels /ɔ, u/, compared to /i, ɛ/, are both [+back] and [+rounded], and the vowel /a/ is characterised solely by the [+low] trait: hence the traits [+low] and [-rounded] attributable to /a/ are predictable (Grimaldi, 2019). This difference could explain why acquiring /ʌ/by speakers of Salento Italian is challenging. The lack of rounding as a distinctive phonological feature can impede the perception and production of /ʌ/, leading to a tendency to substitute /ʌ/ with the nearest L1 vowels, /a/.

Finally, the lack of significant improvement in vowel production under the Italian subtitles and the no subtitles conditions reinforces that L2 language exposure is more beneficial than reliance on L1 norms or the absence of linguistic cues. This finding is consistent with prior research by Mitterer, McQueen (2009). It also supports the theoretical framework posited by Vulchanova, Aurstad, Kvitnes

& Eshuis (2015), which emphasises the value of direct L2 input for its acquisition as opposed to instructions on L2 in the learners' native language.

Our study has some obvious limitations. First, it is a pilot that involves a relatively small cohort of participants, which can significantly constrain the generalizability of the findings. Due to the small number of observations, we did not include test statistics in our analysis. This choice reflects the nature of our study, which was intended to explore potential trends rather than provide definitive conclusions. Moreover, the predominance of female participants may introduce a bias, i.e., the results may not adequately represent how male learners, or a more balanced mix of female and male speakers, would respond to subtitle exposure. Future studies should aim to recruit a larger and more diverse participant pool to enhance the representativeness and reliability of the findings and allow for statistical analyses.

5. *Conclusions*

This pilot study contributes to the ongoing collection of evidence on the effectiveness of subtitles in L2 phonetic training by offering quantitative evidence of their possible benefits and underlying possible limitations regarding L2 speech production. Our findings contribute to the piling of evidence on the role of subtitle exposure in L2 learning. However, the selective efficacy of subtitles in improving the production of certain L2 vowels highlights the complex interplay between phonological features in L2 language learning. Ultimately, our results show the importance of considering the learners' native language phonological system when designing pedagogical strategies for L2 acquisition. Future studies should expand the cohort of L2 learners tested to provide a deeper understanding of the role of subtitles in L2 learning and explore the possibility of combining different forms of L2 training to address the acquisition of the more difficult phonemes.

References

Bassetti, B., Cerni, T. & Masterson, J. (2022). The efficacy of grapheme-phoneme correspondence instruction in reducing the effect of orthographic forms on second language phonology. In *Applied Psycholinguistics*, 43(3), 683–705.

BNC XML Edition: *http://www.natcorp.ox.ac.uk/XMLedition/*

Cavicchio, F., Grimaldi, M. (2022). Enhancing L2 Speech Perception Using Auditory and Audiovisual Training. In AMLaP 28, University of York, 7-9 September 2022.

Chai, X.J., Berken, J.A., Barbeau, E.B., Soles, J., Callahan, M. & Chen, J.-K. (2016). Intrinsic functional connectivity in the adult brain and success in second-language learning. In *Journal of Neuroscience*, 36, 755–761.

Cheour, M., Shestakova, A., Alku, P., Ceponiene, R. & Näätänen, R. (2002). Mismatch negativity shows that 3–6-year-old children can learn to discriminate non-native speech sounds within two months. In *Neuroscience Letters*, 325(3), 187–190.

Darcy, I., Holliday, J.J. (2019). Teaching an old work new tricks: Phonological updates in the L2 mental lexicon. In Levis, J., Nagle, C. & Todey, E. (Eds.), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*, Ames, IA, September 2018, 10–26.

Escudero, P., Sisinni, B. & Grimaldi, M. (2014). The effect of vowel inventory and acoustic properties in Salento Italian learners of Southern British English vowels. In *Journal of the Acoustic Society of America*, 135(3), 1577–1584.

Flege, J.E. (1995). Second-language speech learning: Theory, findings, and problems. In Strange, W. (Ed.), *Speech perception and linguistic experience: Issue in cross-language research*. Timonium, MD: York, 229–273.

Flege, J.E., Schirru, C. & MacKay, I.R. (2003). Interaction between the native and second language phonetic subsystems. In *Speech Communication*, 40(4), 467–491.

Flege, J.E., Bohn, O-S. (2021). The Revised Speech Learning Model (SLM-r). In: Wayland, R. (Ed.). *Second Language Speech Learning: Theoretical and Empirical Progress*. Cambridge: Cambridge University Press, 3–83.

Frumuselu, A.D., De Maeyer, S., Donche, V. & Gutierrez Colon-Plana, M. (2015). Television series inside the EFL classroom: Bridging the gap between teaching and learning informal language through subtitles. In *Linguistics and Education*, 32, 107–117.

Gorilla Experiment Software: *www.gorilla.sc*

Grimaldi, M., Sisinni, B., Gili Fivela, B., Invitto, S., Resta, D. & Alku, P. (2014). Assimilation of L2 vowels to L1 phonemes governs L2 learning in adulthood: a behavioral and ERP study. In *Frontiers in Human Neuroscience*, 8, 279.

Grimaldi, M. (2003). *Nuove ricerche sul vocalismo tonico del Salento meridionale. Analisi acustica e trattamento fonologico dei dati*. Alessandria: Edizioni dell'Orso.

Grimaldi, M. (2019). *Il cervello fonologico*. Roma: Carocci Editore.

Hartshorne, J.K., Tenenbaum, J.B. & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. In *Cognition*, 177, 263–277.

Hisagi, M., Shafer, V.L., Miyagawa, S., Kotek, H., Sugawara, A. & Pantazis, D. (2016). Second-language learning effects on automaticity of speech processing of Japanese phonetic contrasts: An MEG study. In *Brain Research*, 1652, 111–118.

Højlund, A., Horn, N.T., Sørensen, S.D., McGregor, W.B. & Wallentin, M. (2022). Foreign language learning and the mismatch negativity (MMN): A longitudinal ERP study. In *Neuroimage: Reports*, 2(4), 100138.

Hu, X., Ackermann, H., Martin, J.A., Erb, M., Winkler, S. & Reiterer, S.M. (2013). Language aptitude for pronunciation in advanced second language (L2) Learners: Behavioural predictors and neural substrates. In *Brain and Language*, 127, 366–376.

Hutchinson, A.E., Dmitrieva, O. (2022). Exposure to speech via foreign film and its effects on non-native vowel production and perception. In *Journal of Phonetics*, 95, 101189.

Jost, L.B., Eberhard-Moscicka, A.K., Pleisch, G., Heusser, V., Brandeis, D., Zevin, J.D. & Maurer, U. (2015). Native and non-native speech sound processing and the neural mismatch responses: A longitudinal study on classroom-based foreign language learning. In *Neuropsychologia*, 72, 94–104.

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. In *Philosophical Transactions of the Royal Society of London B Biological Science*, 363, 979–1000.

Kuhl, P.K., Conboy, B.T., Padden, D., Nelson, T. & Pruitt, J. (2005). Early speech perception and later language development: implications for the "critical period". In *Language Learning Development*, 1, 237–264.

Legault, J., Grant, A., Fang, S.Y. & Li, P. (2019a). A longitudinal investigation of structural brain changes during second language learning. In *Brain and Language*, 197, 104661.

Legault, J., Fang, S.Y., Lan, Y.J. & Li, P. (2019b). Structural brain changes as a function of second language vocabulary training: Effects of learning context. In *Brain and Cognition*, 134, 90–102.

Lenneberg, E.H. (1967). *Biological foundations of language*. New York: Wiley.

Mahalanobis, P.C. (1936). On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India*, 2(1), 49–55.

MacWhinney, B. (1987). Toward a psycholinguistically plausible parser. In Thomason, S. (Ed.), *Proceedings of the Eastern States Conference on Linguistics*. Columbus, OH: Ohio State University, 12–32.

Merzenich, M.M, Nahum, M. & Van Vleet, T.M. (2013). Neuroplasticity: introduction. In *Progress in Brain Research*, 2(7), xxi-xxvi.

Mitterer, H., McQueen, J.M. (2009). Foreign Subtitles Help but Native-Language Subtitles Harm Foreign Speech Perception. In *PLoS ONE*, 4(11), e7785.

Morford, J, Mayberry, R. (2000). *A reexamination of "early exposure" and its implications for language acquisition by eye*. Lawrence Erlbaum Associates.

Pinker, S. (1984). *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.

Reddy, S., Stanford, J.N. (2015). Toward Completely Automated Vowel Extraction: Introducing DARLA. In *Linguistics Vanguard*, 1(1), 15–28.

Peltola, M.S., Kuntola, M., Tamminen, H., Hämäläinen, H. & Aaltonen, O. (2005). Early exposure to non-native language alters preattentive vowel discrimination. In *Neuroscience Letters*, 388(3), 121–125.

Piske, T. (2007). Implications of James E. Flege's research for the foreign language classroom. In Munro, M.J., Bohn, O-S. (Eds.), *Language experience in second language speech learning. In honor of James Emil Flege*. Amsterdam: John Benjamins, 301–314.

Rato, A., Oliveira, D. (2023). Assessing the robustness of L2 perceptual training: A closer look at generalisation and retention of learning. Second Language Pronunciation. In Alves, U.K., Alcantara de Albuquerque, J.I. (Eds.), *Different Approaches to Teaching and Training*, Berlin, Boston: De Gruyter Mouton, 369–396.

Romano, A. (2013). Il vocalismo del dialetto salentino di Galàtone: differenze di apertura metafonetiche, tracce isolate di romanzo comune o interferenze diasistematiche? In Romano, A., Spedicato, M. (Eds.), *Sub voce Sallentinitas. Studi in onore di p. Giovan Battista Mancarella*. Lecce: Edizioni del Grifo, 247–276.

Sakai, M., Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. In *Applied Psycholinguistics*, 39(1), 187–224.

Shestakova, A., Huotilainen, M., Čeponien, R. & Cheour, M. (2003). Event-related potentials associated with second language learning in children. In *Clinical Neurophysiology*, 114(8), 1507–1512.

Stein, M., Dierks, T., Brandeis, D., Wirth, M., Strik, W. & Koenig, T. (2006). Plasticity in the adult language system: a longitudinal electrophysiological study on second language learning. In *Neuroimage*, 33(2), 774–783.

Thomson, R.I., Derwing, T.M. (2015). The Effectiveness of L2 Pronunciation Instruction: A Narrative Review. In *Applied Linguistics*, 36(3), 326–344.

Vulchanova, M., Aurstad, L.M., Kvitnes, I.E. & Eshuis, H. (2015). As naturalistic as it gets: Subtitles in the English classroom in Norway. In *Frontiers in Psychology*, 5, 98076.

Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4, *https://github.com/tidyverse/dplyr*

Winkler, I., Kujala, T., Tiitinen, H., Sivonen, P., Alku, P., Lehtokoski, A., Czigler, I., Csépe, V., Ilmoniemi, R.J. & Näätänen, R. (1999). Brain responses reveal the learning of foreign language phonemes. In *Psychophysiology*, 36(5), 638–642.

Wisniewska, N., Mora, J.C. (2020). Can Captioned Video Benefit Second Language Pronunciation? In *Studies in Second Language Acquisition*, 42(3), 599–624.

Wottawa, J., Adda-Decker, M. & Isel, F. (2022). Neurophysiology of non-native sound discrimination: Evidence from German vowels and consonants in successive French–German bilinguals using an MMN oddball paradigm. In *Bilingualism: Language and Cognition*, 25, 137–147.

Zhang, Y., Kuhl, P.K., Imada, T., Iverson, P., Pruitt, J., Stevens, E.B. & Nemoto, I. (2009). Neural signatures of phonetic learning in adulthood: A magnetoencephalography study. In *Neuroimage*, 46, 226–240.

DOI: 10.17469/O2112AISV000012

MARTA MAFFIA, VINCENZO VACCHIANO, ANNA DE MEO

Descrivere la competenza ritmica in lingua straniera: uno studio sull'italiano di apprendenti anglofoni

Describing Rhythmic Competence in a Foreign Language: A Study on the Italian Spoken by English-Speaking Learners

Despite the increased interest toward suprasegmental aspects in foreign/second language acquisition, a shared «acquisitional prosody» is still lacking, especially for rhythm, with studies often reporting controversial and inconsistent findings. The aim of the study is to observe the rhythmic competence in Italian as FL spoken by English-speaking learners with different levels of proficiency. A corpus of read speech, both in the native and non-native languages, was collected from 5 Italian native speakers and 12 English-speaking learners of Italian (from A1 to B2 level of CEFR). Recordings were spectroacoustically analysed and manually segmented into vocalic and consonantal intervals, allowing the calculation of different rhythmic metrics. Results highlight a varied contribution of metrics in describing the rhythm of interlanguage and differentiating learners' speech on the basis of proficiency level. In this complex picture, vowel duration appears to play a key role, both when considered as a percentage of total utterance duration and in rate-normalized measures.

Keywords: rhythmic competence, speech rhythm, foreign language acquisition, Italian as FL, rhythmic metrics.

1. *Introduzione*[1]

Il crescente interesse verso gli aspetti ritmico-prosodici nell'acquisizione di lingue seconde e straniere è testimoniato da una ricca recente letteratura (tra gli altri, Trofimovich, Baker, 2006; Delais-Roussarie, Avanzi & Herment, 2015), anche in relazione all'apprendimento della lingua italiana (Pettorino, De Meo, 2012; Chini, 2015). Tale interesse è dimostrato, inoltre, dall'aggiornamento dei descrittori del Quadro Comune Europeo di Riferimento per le Lingue, con riferimento proprio ai tratti prosodici.

Nella prima versione del QCER, infatti, sebbene gli aspetti soprasegmentali dell'eloquio fossero elencati nella sezione dedicata alla competenza fonologica del parlante/apprendente (Council of Europe, 2002: 116-117), non erano presenti,

[1] Benché lo studio sia frutto di ricerche condotte congiuntamente dagli autori del testo, ai fini della valutazione scientifica, a VV è da attribuirsi la redazione della prima stesura dei §§ 2, 3.1, 3.2; a MM la redazione dei §§ 1, 3.3, 3.4, 4; a ADM la supervisione scientifica e la revisione finale del testo. Le conclusioni sono comuni.

tuttavia, descrittori specifici dedicati all'acquisizione di intonazione e ritmo, ma esclusivamente una scala "globale" relativa al controllo fonologico.

Il rapporto ufficiale sul processo di revisione dei descrittori fonologici del Quadro, nell'identificare i punti di debolezza di tale scala, la definisce "*unrealistic*" e "[...] *not consistent as it mixes such diverse factors as stress/intonation, pronunciation, accent and intelligibility without providing clear indication of progression in any of these factors specifically*" (Piccardo, 2016: 9).

Nel report, quindi, si afferma la necessità di una revisione che tenga conto dei risultati delle più recenti ricerche scientifiche nel campo dell'acquisizione degli aspetti fonetico-fonologici delle lingue e che possa portare alla definizione di scale sia globali sia analitiche. Allo stesso tempo, si evidenzia chiaramente uno "scollamento" tra teoria e pratica, ossia la difficoltà di "tradurre" nei termini dell'educazione linguistica i risultati di studi specialistici, come quelli di fonetica acustica.

> An extensive, growing literature on L2 speech has been published in journals that focus on speech production and perception, for example, Journal of the Acoustical Society of America, Journal of Phonetics, and Language and Speech. Yet this work is rarely cited or interpreted in teacher oriented publications (Derwing, Munro, 2005: 382).

I lavori presi in considerazione nella rassegna della letteratura riportata nel rapporto, infatti, si concentrano per lo più su aspetti percettivi, legati alla intelligibilità/comprensibilità del parlato in lingua seconda/straniera e al concetto di accento (Piccardo, 2006: 10-15).

I nuovi descrittori proposti nel Volume Complementare del QCER in relazione ai tratti prosodici sono riportati nella Tab. 1.

Tabella 1 - *Descrittori della competenza prosodica (Consiglio d'Europa, 2021: 147)*

	Tratti prosodici
C2	È in grado di utilizzare in modo appropriato ed efficace i tratti prosodici (ad es. l'accento, il ritmo e l'intonazione) per trasmettere fini sfumature di significato (ad es. per differenziare e valorizzare).
C1	È in grado di produrre un discorso parlato fluente e intelligibile con qualche errore occasionale negli accenti, nel ritmo e/o nell'intonazione che non compromette né l'intelligibilità né l'efficacia. È in grado di variare l'intonazione e collocare correttamente l'accento per esprimere precisamente ciò che intende dire.
B2	È in grado di utilizzare i tratti prosodici (ad es. l'accento, l'intonazione, il ritmo) per far passare il messaggio che intende trasmettere, anche se con qualche influenza proveniente dall'altra o dalle altre lingue che parla.
B1	È in grado di trasmettere il suo messaggio in modo intelligibile malgrado una forte influenza sull'accento, l'intonazione e/o il ritmo proveniente dall'altra o dalle altre lingue che parla.
A2	È in grado di utilizzare in modo intelligibile i tratti prosodici di parole e di espressioni quotidiane, malgrado una forte influenza dell'altra o delle altre lingue che parla sull'accento, l'intonazione e/o il ritmo. I tratti prosodici (ad es. l'accento tonico) di parole familiari e quotidiane e di enunciati semplici sono adeguati.

	Tratti prosodici
A1	È in grado di utilizzare in modo intelligibile i tratti prosodici di un repertorio limitato di parole e semplici espressioni, malgrado una fortissima influenza dell'altra o delle altre lingue che parla sull'accento, il ritmo e/o l'intonazione; il suo interlocutore deve essere collaborativo.

Si nota immediatamente come nella nuova scala si faccia sempre riferimento ad accento, ritmo e intonazione, senza che siano considerati aspetti specifici di questi tre parametri. Se, come in tutti i descrittori del QCER, la prospettiva è quella dell'efficacia comunicativa e dell'uso funzionale (e intelligibile) degli aspetti soprasegmentali dell'eloquio, grande risalto è dato ai fenomeni di interferenza dalla lingua materna e dalle altre lingue conosciute dall'apprendente.

Appare ancora lontana, quindi, la condivisione di una "prosodia acquisizionale", ossia la definizione di specifiche sequenze acquisizionali per ciascuno dei tratti prosodici, che riescano a integrare i risultati degli studi sperimentali condotti su diverse lingue, come quelli riportati nel prossimo paragrafo in relazione al ritmo[2].

2. *Il ritmo delle lingue seconde/straniere*

Studi condotti su neonati in famiglie monolingui e bilingui utilizzando la procedura del monitoraggio della suzione non nutritiva hanno dimostrato come il ritmo della propria L1 (o delle proprie L1) sia riconosciuto molto presto, già nei primissimi giorni di vita (Mehler, Jusczyk, Lambertz, Halsted, Bertoncini & Amiel-Tison, 1988; Ramus, 2000; Byers-Heinlein, Burns & Werker, 2010). Al tempo stesso, il ritmo di una lingua seconda o straniera, insieme ad altre caratteristiche prosodiche e fonetico-fonologiche dell'eloquio, sembra essere un aspetto piuttosto difficile da acquisire, in particolare nei casi di bilinguismo consecutivo/tardivo (Costamagna, Giannini, 2003; De Meo, Pettorino & Vitale, 2015): la struttura ritmico-temporale del parlato è, infatti, uno dei fattori che contribuiscono alla percezione di quello che viene comunemente definito "accento straniero".

Gli studi finora condotti sul ritmo delle lingue seconde/straniere hanno assunto principalmente una prospettiva contrastiva e hanno cercato di verificare l'influenza degli schemi ritmico-prosodici della L1 sulla L2, e la possibilità di applicazione ai fenomeni soprasegmentali di modelli teorici di natura acquisizionale elaborati originariamente in relazione agli aspetti segmentali del parlato: lo *Speech Learning Model* di Flege (nella versione aggiornata in Flege, Bohn, 2021), il *Perceptual Assimilation Model* di Best (Best, Tyler, 2007), l'*Ontology, Phylogeny Model* di Major (2001), la meno recente *Markedness Differential Hypothesis* di Eckman (1977)[3].

[2] Si veda Mennen (2015) per un tentativo di formulare un modello per l'apprendimento dell'intonazione in lingua seconda, LILt (*L2 Intonation Learning theory*).

[3] Per una recente rassegna degli studi sul ritmo del parlato di soggetti bilingui, si veda Matticchio, 2020.

I risultati di tali studi sono spesso molto eterogenei, controversi e a volte di difficile interpretazione, tanto da mettere in dubbio l'efficacia dei tradizionali strumenti di analisi ritmica (le metriche) nella descrizione delle varietà di apprendimento (Gut, 2012)[4].

Alcune recenti ricerche si sono concentrate sul contatto tra lingue descritte come appartenenti a due classi ritmiche diverse: Carter (2005), ad esempio, ha analizzato il parlato in inglese di immigrati messicani adulti del North Carolina, osservando una accentuata variabilità tra i parlanti ispanofoni nei valori delle metriche (in particolare, nPVI-V) e mettendola in relazione a diverse caratteristiche sociolinguistiche di ciascun soggetto; anche White e Mattys (2007) hanno osservato inglese e spagnolo L1 e L2, riscontrando negli apprendenti caratteristiche ritmiche "a cavallo" tra le due lingue e definendo in particolare la metrica %V come resistente alle variazioni di velocità ed efficace nella discriminazione tra le classi ritmiche ipotizzate (p. 520); Grenon e White (2008) hanno invece condotto uno studio su inglese e giapponese, coinvolgendo apprendenti avanzati di entrambe le lingue e constatando l'efficacia delle metriche %V, VarcoV e PVI-C nel rappresentare le differenze e, soprattutto, le somiglianze tra parlato nativo e non nativo.

Le indagini sperimentali che hanno preso in considerazione processi di apprendimento che coinvolgessero lingue tradizionalmente descritte come appartenenti alla stessa classe ritmica hanno spesso riscontrato difficoltà nella distinzione tra parlato nativo e non nativo attraverso le metriche, giustificata con possibili fenomeni di transfer positivo (White, Mattys, 2007; Gut, 2012).

Sono pochi, e principalmente sull'inglese L2, gli studi che hanno cercato di descrivere traiettorie di sviluppo della competenza ritmica in lingua seconda/straniera, prendendo in considerazione parlanti con diversi livelli di competenza linguistico-comunicativa. Anche in questo caso, i risultati appaiono controversi: in alcuni casi le metriche non sono risultate efficaci nel distinguere tra apprendenti principianti, intermedi e avanzati (Guilbault, 2002; Jang, 2008; Gut, 2009); in altri casi, invece, sono emerse alcune differenze, come nello studio di Stockmal, Markus & Bond (2005) condotto su apprendenti lettoni di russo, che si distinguono per livello di competenza rispetto ai valori di ΔC e PVI-C, o in quello di Tortel e Hirst (2010) che, nell'analisi del parlato di apprendenti francofoni di inglese L2, hanno comparato adulti e studenti universitari, riscontrando la possibilità di distinguere tra le produzioni in interlingua, sulla base di diverse metriche (soprattutto ΔC e VarcoC).

Lo sviluppo della competenza ritmica in una lingua seconda/straniera appare quindi un fenomeno complesso, condizionato da fattori di natura linguistica (da quelli macro, come gli schemi ritmici della lingua di partenza e di quella di arrivo, a quelli micro, come lo stile di parlato considerato), legati alle caratteristiche dei parlanti (il livello di competenza e il contesto di apprendimento) ma anche a questioni extralinguistiche, come, ad esempio, agli atteggiamenti nei confronti della

[4] Ulrike Gut riprende ed estende anche al caso delle lingue seconde/straniere le osservazioni e le critiche di Amalia Arvaniti (2012) rispetto all'utilità delle metriche tradizionali, di natura quantitativa e basate sul calcolo di durate, nel distinguere le lingue naturali in classi ritmiche e nel tenere conto dell'enorme variabilità esistente già all'interno di ciascuna lingua.

lingua target (si veda, a tal proposito, lo studio di Gabriel, Stahnke & Thulke del 2014 su apprendenti tedeschi di francese con il cinese mandarino come *heritage language*) o alla percezione di dominanza di una lingua sull'altra (Henriksen, 2016).

2.1 Studi sull'italiano L2/LS

Le ricerche che hanno assunto come oggetto di interesse lo sviluppo della competenza ritmica in apprendenti adulti di italiano come lingua seconda o straniera non sono numerose.

Tra i primi, lo studio condotto da Schmid e Dellwo (2012) su parlanti nativi, non-nativi e bilingui di tedesco e di italiano ha previsto l'applicazione di diverse metriche ritmiche, con risultati talvolta a favore di un'ipotesi "nativa", secondo cui i bilingui sono in grado di padroneggiare due diversi modelli ritmici, paragonabili a quelli dei parlanti nativi monolingui, talvolta di una posizione "intermedia", che vede i bilingui "occupare uno spazio ritmico" a metà strada tra le due lingue del proprio repertorio. Con riferimento alle caratteristiche ritmiche del parlato non-nativo, si è notato in questo studio come i tedescofoni non abbiano trasferito nell'italiano il valore di variabilità delle durate vocaliche della propria lingua materna, tradizionalmente inserita nel gruppo delle lingue isoaccentuali, ma abbiano piuttosto realizzato "un ritmo quasi più 'sillabico' degli stessi locutori italofoni" (p. 168-169).

Romito e Tarasi (2012) hanno proposto, invece, un confronto di natura ritmico-prosodica tra il parlato in italiano L1 di calabresi e quello in L2 di soggetti polacchi, romeni, cinesi e albanesi, riportando anche in questo caso risultati controversi rispetto all'emersione di fenomeni di transfer ritmico rilevati attraverso l'uso di diverse metriche e presentando interessanti considerazioni di natura glottodidattica.

Vitale e De Meo (2017) hanno analizzato il parlato letto di apprendenti di italiano con un livello avanzato di competenza e con diverse lingue materne (spagnolo, inglese e giapponese), applicando le metriche %V e VtoV (cfr. § 3.3) ed evidenziando un "movimento" verso l'italiano, ossia un adattamento dello schema ritmico del proprio eloquio rispetto a quello della lingua target, solo nei casi in cui L1 e L2 non appartenessero allo stesso gruppo ritmico. Tali risultati sono confermati da quelli di uno studio successivo (De Meo, Vitale, 2018), in cui sono esaminati gli aspetti ritmico-temporali nel parlato di apprendenti avanzati di italiano L2 con 8 lingue materne (cinese, estone, giapponese, inglese, portoghese, russo, spagnolo, tedesco). Anche in questo caso, i soggetti con L1 e L2 appartenenti a uno stesso gruppo ritmico non sono risultati in grado di percepire differenze tra i due codici, rimanendo ancorati alla struttura ritmica della lingua materna, in accordo con quanto è descritto per gli aspetti segmentali nello *Speech Learning Model* di Flege (Flege, Bohn, 2021).

Lo studio di Budeanu, De Meo & Pettorino (2020) ha, invece, considerato l'italiano L2 di apprendenti romeni suddivisi in due gruppi sulla base del livello di scolarizzazione pregressa (diplomati e laureati). L'analisi ha mostrato un significativo aumento della %V in entrambi i gruppi (ma in particolare in quello dei diplomati) nell'eloquio in lingua seconda, sebbene il romeno, come l'italiano, sia

tradizionalmente riconosciuto come una lingua isosillabica (e, quindi, tale aumento non possa essere legato a un fenomeno di transfer).

Un'indagine condotta su apprendenti anglofoni di italiano con percorsi di acquisizione molto eterogenei è quella di Mairano, Mois, De Iacovo & Romano (2018). Anche in questo studio si riscontrano nel parlato dei soggetti non nativi valori significativamente più alti della metrica %V in confronto sia all'inglese L1 sia all'italiano nativo di un gruppo di controllo. Tale dato è spiegato dagli autori come correlato al rallentamento dell'eloquio nell'interlingua e a fenomeni di *overshooting* e ipercorrezione a livello prosodico (già evidenziati da White e Mattys, 2007). Il gruppo degli apprendenti si distingue inoltre dai parlanti nativi sulla base delle metriche nPVI-V e VarcoV, mentre presenta valori simili ai nativi anglofoni rispetto ai fenomeni di durata consonantica.

3. *Lo studio*

L'obiettivo del presente studio è quello di provare a descrivere la competenza ritmica, ossia l'abilità di gestire in maniera appropriata ed efficace la struttura ritmica di una lingua, nell'italiano LS di apprendenti anglofoni con diversi livelli di competenza comunicativa nella lingua straniera. Nello specifico, lo studio intende rispondere a due domande di ricerca: le "tradizionali" metriche ritmiche sono efficaci nel distinguere le produzioni degli apprendenti da quelle dei parlanti nativi? Riescono a differenziare tali produzioni in base al livello di competenza nella lingua straniera e a tracciare un percorso di sviluppo della competenza ritmica, in cui si assumono come punto di partenza le caratteristiche della lingua materna e come punto di arrivo quelle del parlato nativo nella lingua target?

3.1 I partecipanti

Al fine di provare a rispondere alle domande di ricerca, sono stati coinvolti nello studio 17 soggetti:

- cinque parlanti nativi di italiano di area campana (4 donne e 1 uomo);
- 12 apprendenti anglofoni di italiano LS (5 donne e 7 uomini), provenienti da USA e Regno Unito, suddivisi in quattro gruppi, ciascuno composto da tre partecipanti, sulla base del livello di competenza linguistico-comunicativa nella lingua straniera (A1, A2, B1 e B2 del Quadro Comune Europeo di Riferimento per le lingue).

I gruppi di parlanti nativi e non nativi di italiano, con la media di 29 anni, risultano comparabili per età (test di Wilcoxon-Mann Whitney per campioni indipendenti, p=.36).

Il reclutamento degli apprendenti di italiano LS è avvenuto tramite una piattaforma online dedicata all'educazione linguistica, attraverso cui uno degli autori di questo contributo tiene regolarmente lezioni di italiano a stranieri. Il loro livello di padronanza dell'italiano è stato definito grazie alla somministrazione di un test standardizzato di valutazione delle competenze in entrata, messo a disposizione

dalla suddetta piattaforma. Il docente di lingua straniera ha successivamente avuto modo di confermare o ridiscutere gli esiti del test.

Al momento della raccolta dei dati, infatti, tutti i partecipanti anglofoni erano inseriti in un percorso guidato online di formazione linguistica e hanno dichiarato di essere spinti nell'apprendimento dell'italiano LS principalmente da motivazioni di tipo culturale e integrativo[5]. Tutti hanno affermato, inoltre, di aver trascorso in passato un periodo di soggiorno in Italia di lunghezza variabile (da due settimane a sei anni).

La scelta di coinvolgere parlanti nativi di italiano di area campana è stata dettata dalla necessità di assumere come "modello di italofonia" la varietà cui gli apprendenti fossero maggiormente esposti al momento dello studio, ovvero quella del loro docente di lingua.

3.2 La raccolta dei dati e il dataset

A ciascun partecipante è stato chiesto di leggere a voce alta un testo nella lingua materna e, nel caso degli apprendenti, nella lingua straniera. Il testo, in Appendice, di circa 350 sillabe in italiano e 230 in inglese, già utilizzato in studi precedenti (Maffia, De Micco, Pettorino, Siciliano, Tessitore & De Meo, 2021; Maffia, 2023), è stato costruito (e tradotto) per essere molto semplice dal punto di vista lessicale, morfosintattico e dei contenuti proposti (le abitudini culinarie del presente e del passato). Tale scelta è stata dettata dalla necessità di evitare interruzioni o esitazioni legate alla non comprensione di quanto letto, per quanto possibile.

Durante le regolari lezioni online di italiano LS, i partecipanti sono stati invitati a leggere nel modo più naturale possibile, al volume e alla velocità per loro più congeniali, e a registrare la propria lettura in un ambiente silenzioso.

Nella Tab. 2 sono riportate alcune informazioni quantitative sui dati raccolti.

Tabella 2 - *Descrizione del dataset*

	n. di campioni	*durata media (s)*	*durata tot (s)*
en	12	68.2	818
itaLS	12	122	1464
ita	5	73.4	367
tot	**29**	**91.3**	**2649**

3.3 L'analisi ritmica

Ciascun campione di parlato in L1 o in LS è stato acusticamente analizzato con il software Praat (versione 6.2.06 - Boersma, Weenink, 2022) e segmentato manualmente in intervalli vocalici (V) e consonantici (C), seguendo una procedura di etichettatura

[5] È da specificare che il corso di lingua italiana frequentato dai partecipanti a questo studio non ha previsto l'uso di tecniche glottodidattiche specificamente mirate allo sviluppo della competenza ritmica in lingua seconda.

già applicata in studi precedenti (Maffia et al., 2021). Anche le pause silenti e gli eventuali fenomeni di disfluenza sono stati individuati e annotati (con X e DIS, rispettivamente), sebbene non siano stati presi in considerazione nella misurazione delle metriche ritmiche. In totale, sono stati etichettati circa 16.000 intervalli V/C.

Esempi di segmentazione sono riportati nelle Fig. 1 e 2.

Figura 1 - *Spettrogramma annotato dell'enunciato in italiano "secondo diversi studi"*[6]

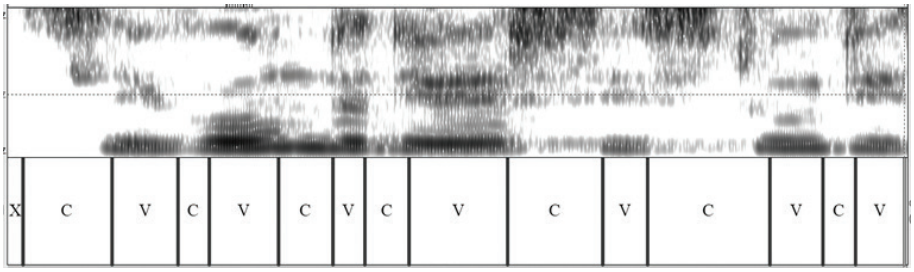
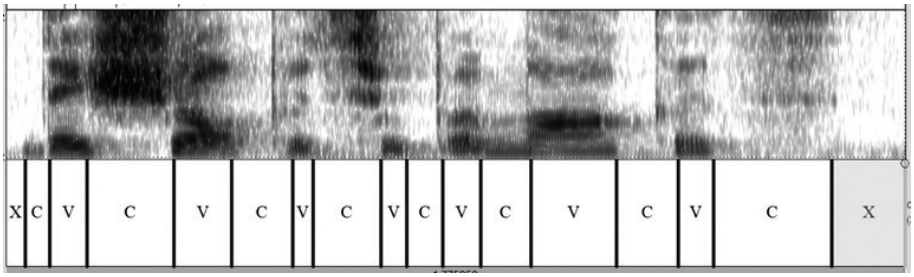


Figura 2 - *Spettrogramma annotato dell'enunciato in inglese "they shop at the supermarket"*



Allo scopo di ottenere una descrizione delle caratteristiche ritmico-temporali del parlato in lingua materna e straniera, si è scelto di misurare alcune tra le metriche tradizionalmente più utilizzate nella letteratura sul tema, riportate in Tab. 3.

Tabella 3 - *Definizione delle metriche ritmiche utilizzate nell'analisi del parlato in L1 e LS*

Metrica	*Definizione*	*Riferimenti*
%V	Percentuale vocalica. Indica in che percentuale di durata l'enunciato è costituito da intervalli vocalici. Valori più alti di %V sono associati in letteratura a una maggiore isosillabicità e alla percezione di un parlato più "legato".	Ramus, Nespor & Mehler, 1999; Pettorino, Maffia, Pellegrino, Vitale & De Meo, 2013

[6] Si noterà l'assenza in Fig. 1 e 2 di una trascrizione fonetica dei singoli segmenti. Questa scelta è funzionale a chiarire la modalità in cui è stata condotta, attraverso l'ispezione dello spettrogramma e della forma d'onda, la segmentazione del segnale acustico. Tale operazione, non preceduta da una fase di trascrizione, non si è basata su una forma fonologica "attesa" bensì esclusivamente sulla forma fonica effettivamente realizzata dai parlanti e sulla distinzione tra intervalli consonantici e intervalli vocalici/pienamente sonori, al di là della specifica configurazione articolatoria dei diversi suoni.

Metrica	*Definizione*	*Riferimenti*
VtoV	*Vowel to Vowel*. Misura la durata media degli intervalli tra un punto di inizio vocalico e il successivo. Il VtoV è sensibile alla complessità fonotattica degli intervalli vocalici e consonantici in una data lingua (a lingue isoaccentuali corrispondono in media valori più alti di VtoV) e fornisce informazioni sulla velocità di articolazione (valori maggiori corrispondono a un parlato più lento).	Pettorino et al., 2013
ΔC	Deviazione standard degli intervalli consonantici. Restituisce il grado di variabilità della durata delle porzioni consonantiche. Lingue isoaccentuali sono associate in letterature a valori maggiori di ΔC.	Ramus et al., 1999
VarcoV	Coefficiente di varianza degli intervalli vocalici. È misurato dividendo la deviazione standard della durata degli intervalli vocalici per la durata media degli stessi intervalli (x 100). Il coefficiente permette di ottenere una normalizzazione rispetto alla velocità di articolazione del parlante.	Dellwo, Wagner, 2003
Varco C	Coefficiente di varianza degli intervalli consonantici. È misurato come la corrispondente metrica vocalica.	Dellwo, Wagner, 2003
nPVI-V	Indice normalizzato di variabilità a coppie degli intervalli vocalici. È misurato dividendo la media delle differenze di durata tra intervalli vocalici adiacenti per la loro somma (x100). Il coefficiente permette di ottenere una normalizzazione rispetto alla velocità di articolazione del parlante. Lingue isoaccentuali sono associate in letteratura a valori maggiori di nPVI-V.	Grabe, Low, 2002
rPVI-C	Indice non normalizzato di variabilità a coppie degli intervalli vocalici. Il coefficiente di normalizzazione non è applicato, perché ritenuto superfluo nel caso delle consonanti. Lingue isoaccentuali sono associate in letteratura a valori maggiori di rPVI-C.	Grabe, Low, 2002

La misurazione delle metriche è stata effettuata grazie all'applicazione di uno script di *Praat* e del software Correlatore (Mairano, Romano, 2010).

3.4 L'analisi statistica

Per verificare, anche da un punto di vista statistico, se le variazioni nei valori delle metriche ritmiche fossero legate alle differenze tra i diversi codici linguistici presenti nel dataset e per tenere sotto controllo la variabilità individuale nei dati, per ciascuna metrica è stato costruito un modello lineare a effetti misti, assumendo come variabile indipendente la lingua parlata (en, itaLS, ita), come variabile dipendente il valore di ciascuna metrica ritmica e come effetto random i parlanti (pacchetto 'lme4', Bates, Mächler, Bolker & Walker, 2015).

Per testare, invece, l'impatto dei diversi livelli di competenza in italiano come lingua straniera sui valori delle metriche ritmiche, sono stati costruiti modelli lineari multipli (pacchetto 'lmtest', Hothorn, Zeileis, Farebrother & Cummins, 2022), assumendo i quattro livelli (A1, A2, B1, B2) come variabile esplicativa e il valore di ciascuna metrica come variabile dipendente.

Le informazioni biografiche sul sesso e l'età dei partecipanti sono state incluse come variabili indipendenti in ciascuno dei modelli costruiti.

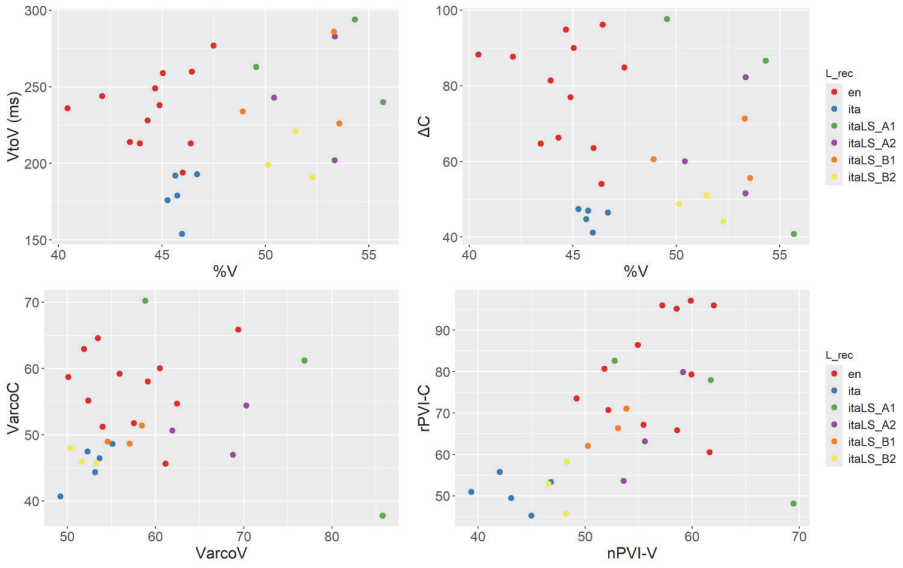
Test del rapporto di verosimiglianza (*likelihood ratio comparisons*) sono stati, inoltre, applicati per escludere le variabili indipendenti non significative e ottenere i modelli che meglio riuscissero a rappresentare la variabilità dei dati.

Sono stati, infine, utilizzati post-hoc test con correzione di Bonferroni per i confronti a coppie (pacchetto 'emmeans', Lenth, Singmann, Love, Buerkner & Herve, 2018). La soglia di significatività, come di consueto nelle scienze umane, è stata fissata a .05. Le analisi statistiche sono state condotte utilizzando il software R (versione 4.3.3 - R Core Team, 2022).

4. *Risultati*

In Fig. 3 sono riportati i valori delle metriche per ciascun parlante e ciascuna lingua, rappresentati nei quattro modelli più utilizzati in letteratura: %V/VtoV (Pettorino et al., 2013); %V/ΔC (Ramus et al., 1999); VarcoV/VarcoC (Dellwo, Wagner, 2003); nPVI-V/rPVI-C (Grabe, Low, 2002).

Figura 3 - *Valori delle metriche ritmiche per ciascun parlante, per lingua e livello di competenza nella LS*



Osservando i grafici, è possibile notare come in tutti i modelli le aree dei punti relativi all'italiano e all'inglese di parlanti nativi siano abbastanza differenziate, occupando le posizioni "attese" e ampiamente documentate (ad es. in Ramus et al., 1999; Mairano, Romano, 2010; Roseano, 2020) per lingue tradizionalmente descritte come appartenenti a due classi ritmiche diverse (isosillabica l'italiano, isoaccentuale l'inglese). Si osserva altresì un discreto margine di variabilità intra-gruppo, in particolare per la lingua inglese.

I dati relativi al gruppo di apprendenti, invece, si collocano in maniera diversa nello spazio dei grafici, e rispetto alle due lingue native, a seconda del modello utilizzato: nei primi due (%V/VtoV e %V/ΔC), il "blocco" di parlato in lingua straniera si distingue dal parlato nativo, sia in italiano sia in inglese, sull'asse orizzontale, presentando valori maggiori di %V; negli altri due modelli (VarcoV/VarcoC e nPVI-V/rPVI-C), in cui sono riportate misure normalizzate per velocità di articolazione, le aree occupate dai punti relativi agli apprendenti si sovrappongono maggiormente a quelle di entrambe le lingue native. Nel modello in cui sono utilizzate le metriche proposte da Grabe e Low (Fig. 6, in basso a destra), in particolare, i dati degli apprendenti sembrano occupare in maniera piuttosto definita una posizione intermedia rispetto a quelli dei parlanti nativi di italiano e inglese, che presentano rispettivamente i valori più bassi e più alti sia di nPVI-V, sia di rPVI-C.

Sebbene in nessuno dei grafici vi sia una chiara distribuzione dei punti relativi agli apprendenti sulla base del livello di competenza in italiano LS, emergono tuttavia delle tendenze di "progressione" dall'area occupata dall'inglese, dove si collocano soprattutto i valori dei gruppi A1 e A2, verso quella dell'italiano nativo, dove si concentrano maggiormente i gruppi B1 e B2.

4.1 Metriche ritmiche e lingua parlata

Per confermare o confutare la validità di tali osservazioni e verificarle anche dal punto di vista statistico, si prenderanno di seguito in esame le metriche, considerate singolarmente e non in combinazione.

Nella Tab. 4 si riportano i valori medi e la deviazione standard di ciascuna metrica per lingua parlata (inglese, italiano LS e italiano).

Tabella 4 - *Valori delle metriche ritmiche nel dataset per lingua parlata (media e deviazione standard)*

	en *media ± dev. st.*	*itaLS* *media ± dev. st.*	*ita* *media ± dev. st.*
%V	44.5 ± 1.9	52.1 ± 2.1	45.8 ± 0.5
VtoV	235 ± 24	240 ± 35	179 ± 16
ΔC	79.1 ± 13.8	62.6 ± 18.1	45.4 ± 2.5
VarcoV	57.3 ± 5.5	62.3 ± 10.9	52.7 ± 2.1
Varco C	57.3 ± 5.9	50.8 ± 8.2	45.5 ± 3.1
nPVI-V	56.7 ± 4.1	54.4 ± 6.4	43.2 ± 2.8
rPVI-C	80.7 ± 13.3	63.5 ± 12.4	50.9 ± 4

È evidente come, al variare del codice linguistico, vi sia una variazione nei valori delle metriche, con l'eloquio in lingua straniera che presenta spesso una maggiore deviazione standard rispetto alle lingue native.

Tabella 5 - *Output dei test a coppie sui modelli lineari a effetto misto per ciascuna metrica*

		Est.	*SE*	*t*	*p*
%V	en - ita	-1.56	0.9	-1.63	.34
	en - itaLS	-7.52	0.54	-13.7	<.0001***
	ita - itaLS	-5.96	0.96	-6.1	<.0001***
VtoV	en - ita	58.75	14.69	4	.001**
	en - itaLS	-4.18	9.84	-0.4	1
	ita - itaLS	-62.92	14.83	-4.24	.0007***
ΔC	en - ita	34.4	7.43	4.6	.0002***
	en - itaLS	17.1	4.7	3.6	.008**
	ita - itaLS	-17.3	7.5	-2.3	.09
VarcoV	en - ita	2.91	4.06	0.71	1
	en - itaLS	-4.97	2.81	-1.77	.29
	ita - itaLS	-7.88	4.10	-1.92	.19
Varco C	en - ita	11.65	3.31	3.5	.004**
	en - itaLS	6.86	2.27	3.01	.02*
	ita - itaLS	-4.79	3.35	-1.43	.48
nPVI-V	en - ita	12.6	2.66	4.76	.0002***
	en - itaLS	2.47	1.67	1.48	.48
	ita - itaLS	-10.19	2.68	-3.8	.002**
rPVI-C	en - ita	31.7	5.97	5.31	<.0001***
	en - itaLS	17.5	3.5	5	.0005***
	ita - itaLS	-14.2	6.04	-2.35	.08

I risultati dei test a coppie applicati ai modelli lineari a effetto misto (riportati in Tab. 5) confermano quanto precedentemente osservato in relazione alle due lingue materne: nei dati considerati, i valori di quasi tutte le metriche differiscono in maniera statisticamente significativa tra inglese e italiano, tranne che per %V e VarcoV.

Il parlato in italiano LS, invece, si distingue significativamente dall'inglese nativo sulla base della %V, del ΔC, di VarcoC e di rPVI-C. Rispetto a quanto accade nell'inglese, lingua materna, il parlato degli apprendenti in lingua straniera è quindi caratterizzato da un aumento della durata relativa dei suoni vocalici e da una più ridotta variabilità nella durata degli intervalli consonantici, sia se considerati sull'intero enunciato, sia se considerati a coppie.

Posto a confronto, invece, con l'italiano di parlanti nativi, quello degli apprendenti si distingue in maniera statisticamente significativa sulla base della %V, del VtoV e di nPVI-V. L'eloquio nella lingua straniera, quindi, risulta nel complesso più lento, come atteso nel parlato non-nativo (e ampiamente documentato nella varietà basiche in italiano L2, come in Pellegrino, 2012), con intervalli vocalici che presentano una

durata più variabile, quando osservati a coppie, e complessivamente maggiore in termini percentuali, rispetto a quanto accade nell'italiano del gruppo di controllo.

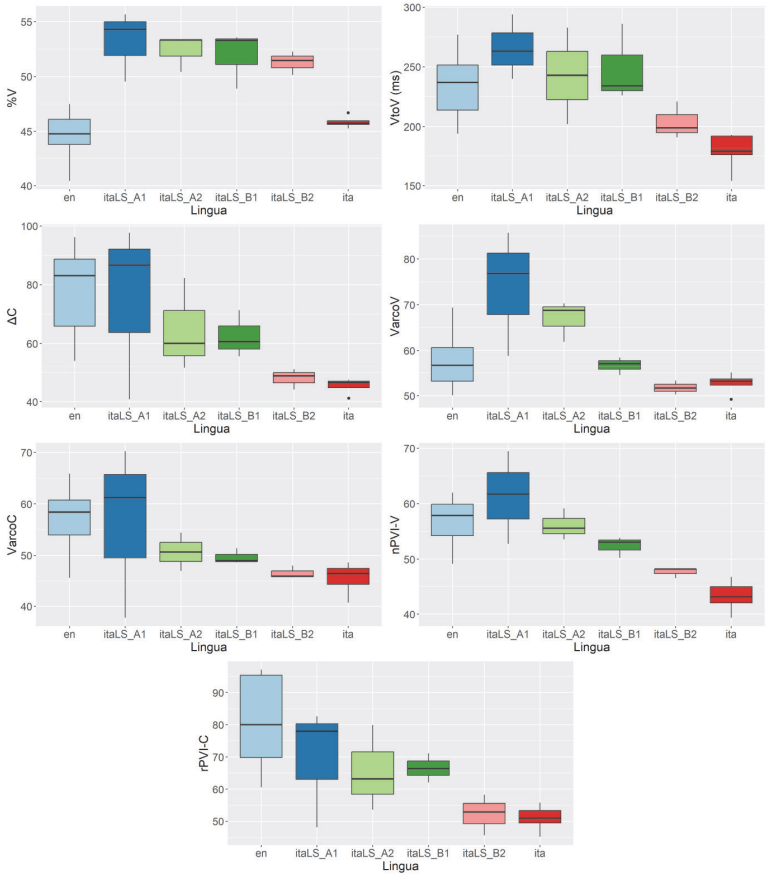
La metrica VarcoV è l'unica la cui variazione non è condizionata in maniera statisticamente significativa dalla lingua parlata nei dati analizzati.

Le variabili biografiche, ossia il sesso e l'età dei partecipanti, infine, non risultano mai esplicative delle variazioni dei valori delle metriche in modo statisticamente significativo, in nessuno dei modelli applicati.

4.2 Metriche ritmiche e livelli di competenza in italiano LS

I grafici della Fig. 4, in cui i valori di ciascuna metrica sono riportati in diagrammi a scatole, suddivisi per lingua parlata e con la distinzione tra i diversi livelli di competenza linguistica in italiano LS, offrono un quadro più dettagliato e permettono di verificare la presenza/assenza di un andamento progressivo dello sviluppo della competenza ritmica, assumendo come riferimenti la lingua di partenza, l'inglese, e quella di arrivo, l'italiano.

Figura 4 - *Valori di ciascuna metrica ritmica nel dataset considerato, per lingua parlata e livello di competenza in italiano LS*



È evidente in tutti i grafici la tendenza a una progressiva riduzione del valore medio delle metriche e anche della variabilità dei dati all'aumentare della competenza linguistica nella lingua straniera, con una conseguente approssimazione al modello italofono nativo. Meno evidente risulta, invece, la vicinanza dei valori dei gruppi con una competenza più bassa (A1 e A2) a quelli dell'inglese, in particolare nel caso delle metriche che prendono in considerazione in vario modo le durate degli intervalli vocalici.

Allo scopo di osservare i dati nel dettaglio, nella Tab. 6 sono riportati i valori medi e la deviazione standard delle diverse metriche ritmiche con riferimento ai quattro gruppi di apprendenti, suddivisi per livello di competenza in italiano LS.

Tabella 6 - *Valori delle metriche ritmiche per livello di competenza in italiano LS (media e deviazione standard)*

	itaLS_A1 *media ± dev. st.*	*itaLS_A2* *media ± dev. st.*	*itaLS_B1* *media ± dev. st.*	*itaLS_B2* *media ± dev. st.*
%V	53.2 ± 3.2	52.3 ± 1.7	51.9 ± 2.6	51.2 ± 1.1
VtoV	266 ± 27	243 ± 40	249 ± 32	204 ± 15
ΔC	75.1 ± 30.1	64.7 ± 15.8	62.5 ± 8	47.9 ± 3.5
VarcoV	73.8 ± 13.7	66.9 ± 4.5	56.7 ± 1.9	51.7 ± 1.5
Varco C	56.4 ± 16.7	50.7 ± 3.7	49.7 ± 1.5	46.6 ± 1.2
nPVI-V	61.3 ± 8.3	56.1 ± 2.8	52.4 ± 1.9	47.7 ± 0.9
rPVI-C	69.6 ± 18.7	65.6 ± 13.2	66.5 ± 4.5	52.2 ± 6.3

Sebbene, come già osservato nei grafici in Fig. 4, in quasi tutti i casi si osservi un interessante andamento "a scalare" e una diminuzione della deviazione standard con l'avanzare della competenza in LS, gli esiti dell'applicazione dei modelli lineari a ciascuna metrica riportano risultati statisticamente significativi solo per VarcoV e nPVI-V.

Dai risultati dei test appaiati, inoltre, emerge che, sia per VarcoV sia per nPVI-V, esiste una differenza statisticamente poco significativa solo tra i due livelli di competenza più "estremi" nei dati analizzati, A1 e B2 (VarcoV: Est=22.05, SE=5.9, t=3.7, p= .04; nPVI-V: Est=13.64, SE=3.7, t=3.7, p= .04).

Anche in questo caso, il sesso e l'età dei parlanti non risultano variabili significative nei modelli statistici.

5. *Discussione e conclusioni*

L'analisi del parlato letto di apprendenti anglofoni di italiano LS e di parlanti nativi di italiano e di inglese ha avuto l'obiettivo di provare a descrivere un possibile percorso di sviluppo della competenza ritmica in lingua straniera, da un livello elementare a un livello intermedio (da A1 a B2 del QCER), attraverso la misurazione delle metriche ritmiche più utilizzate in letteratura.

In relazione alla prima domanda di ricerca (cfr. § 2), i risultati dell'analisi evidenziano innanzitutto come le metriche siano state efficaci nel cogliere una differenza tra le due lingue native del dataset considerato: l'italiano, lingua tradizionalmente definita isosillabica, presenta, rispetto all'inglese, valori medi più alti di %V (sebbene la differenza non sia statisticamente significativa) e una minore variabilità nella durata sia degli intervalli vocalici sia di quelli consonantici; al contrario, il ritmo della lingua inglese, definita isoaccentuale, appare condizionato dalla presenza di più marcati fenomeni di riduzione delle vocali atone e di nessi consonantici più complessi (e quindi più variabili in durata)[7].

Nel parlato in lingua straniera, in analogia con studi precedenti (White, Mattys, 2007; Grenon, White, 2008; Schmid, Dellwo, 2012; Mairano et al., 2018; Budeanu et al., 2020), si osservano valori significativamente più alti di %V, probabilmente legati a fenomeni di ipercorrettismo prosodico. C'è da chiedersi se la generale tendenza degli apprendenti alla *overarticulation* nell'interlingua (Barry, 2007), qualsiasi sia la lingua di partenza, non possa determinare, sul piano ritmico, un fenomeno sistematico dal punto di vista acquisizionale, ossia l'aumento in percentuale della durata degli intervalli vocalici. Tale aumento sembra non essere condizionato (solo) dall'attestato rallentamento dell'eloquio e dalla riduzione della velocità di articolazione nell'interlingua nei dati considerati (*Pearson's correlation* = .240, p = .43).

Se si considerano, invece, i valori delle metriche che normalizzano la variazione di velocità nel parlato, i dati relativi agli apprendenti sembrano situarsi in una posizione intermedia, tra quelli dell'inglese L1 (da cui si differenziano soprattutto rispetto alla variabilità di durata degli intervalli consonantici) e quelli dell'italiano nativo (rispetto ai quali presentano un maggiore rallentamento dell'eloquio e una più accentuata variabilità nella durata delle vocali).

Rispetto alla seconda domanda di ricerca (cfr. § 2), sono proprio le metriche normalizzate VarcoV e nPVI-V le uniche a cogliere una differenza anche statisticamente significativa tra i diversi gruppi di apprendenti, sulla base del livello di competenza comunicativa nella LS.

Osservando però più nel dettaglio i grafici riportati nella Fig. 4, benché siano riscontrabili per tutte le metriche un progressivo avvicinamento degli apprendenti ai valori dell'italiano nativo e una riduzione della variabilità nella distribuzione dei dati, meno evidente risulta, invece, la congruenza dei valori relativi ai soggetti di livello A1 (e A2) a quelli dell'inglese, in particolare proprio in relazione a VarcoV e nPVI-V. Questo dato sembra avvalorare l'ipotesi del ruolo chiave giocato dalla durata degli intervalli vocalici nella caratterizzazione delle varietà basiche, forse indipendentemente dalle caratteristiche ritmiche della lingua materna degli apprendenti.

I risultati di questo elaborato sono sicuramente da confermare/confutare con ulteriori indagini. È, inoltre, ipotizzabile che un dataset più ampio di quello analizzato

[7] Per una raccolta di riflessioni sulla definizione della nozione di sillaba e sulla possibilità di descriverne la struttura, nonché di studi sperimentali condotti su diverse lingue, si veda Russo (2015).

in questa sede possa permettere di ottenere risultati statisticamente più significativi anche per altre metriche, in relazione allo sviluppo della competenza ritmica.

In conclusione, riprendendo le riflessioni proposte nell'introduzione a questo studio, appare importante ricordare che i valori delle metriche ritmiche misurati su un dato enunciato danno conto di diversi fenomeni fonotattici caratteristici delle lingue e delle interlingue. Nella prospettiva della linguistica acquisizionale e della didattica, piuttosto che ragionare su una generica e astratta nozione di ritmo, sarebbe essenziale descrivere il modo in cui gli apprendenti imparano (o possono imparare) a gestire efficacemente e in maniera consapevole i parametri di durata vocalica e consonantica e i fenomeni di prominenza che determinano il ritmo del parlato in una qualunque lingua target (Barry, 2007), forse ridimensionando il ruolo dato nel Quadro all'influenza delle altre lingue del repertorio sull'interlingua.

Si ribadiscono, infine, la necessità di ottenere nel QCER descrittori condivisi, specifici per ciascun tratto prosodico e al tempo stesso generalizzabili a lingue diverse, e la possibilità di dotarli di una maggiore concretezza attraverso il riferimento agli studi (anche) di fonetica acustica.

Ringraziamenti

Questo studio si sviluppa all'interno del progetto STRADD - *Speech Technology for [L2] Rhythm, Affect and Disease Detection*, finanziato dal Dipartimento di Studi Letterari, Linguistici e Comparati dell'Università di Napoli L'Orientale.

Si ringraziano, inoltre, i revisori anonimi che, con le loro indicazioni, hanno contribuito senz'altro a migliorare questo contributo.

Appendice

Testi per il compito di lettura in italiano e in inglese:

Secondo diversi studi 60 anni fa mangiavamo in modo diverso.
Mangiavamo molto pane e molta pasta, poca carne.
Generalmente il pesce si mangiava una volta alla settimana.
Mangiavamo abbastanza frutta di stagione.
Si accompagnava tutto con un po' di vino.
Di solito a pranzo tutta la famiglia sedeva a tavola, si parlava delle cose fatte in giornata.
Si faceva la spesa tutti i giorni.
Le persone compravano prodotti freschi nei mercati o in negozi piccoli.
Si spendeva molto ma si mangiava bene.
Oggi la cena è diventata il momento in cui tutta la famiglia è a tavola.
Le persone parlano ma spesso stanno anche al telefono o vedono la tivù.
Di solito si fa la spesa nei supermercati, si comprano alimenti che provengono da diversi paesi del mondo.

Si spende meno e si comprano prodotti surgelati o cibi già pronti.
Il tempo dedicato alla cucina è poco.
60 anni fa si mangiava meglio e le persone non si preoccupavano della linea come oggi.
Oggi si mangia in modo disordinato ma la vita si è allungata.

Several research studies have revealed that our eating habits were different sixty years ago.
We used to eat a lot of bread, a lot of pasta, little meat.
Fish would usually be consumed once a week.
We would eat an adequate quantity of seasonal fruit.
All this was accompanied by a small glass of wine.
The whole family would usually sit together for lunch and chat about how they had spent the day.
People used to go shopping every day.
They would buy fresh food at the market or in small shops.
One would spend much money and eat well.
Nowadays it is at dinner time that the whole family seat together at the table.
They chat but often they use their phones or watch television.
They shop at supermarkets, buy food which comes from all over the world.
They spend less and buy frozen or pre-cooked food.
They spend little time cooking.
Sixty years ago, people ate better and did not worry about their shape as we do today.
In the present time we eat untidily but we live longer lives.

Riferimenti bibliografici

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. In *Journal of Phonetics*, 40(3), 351-373.

Barry, W.J. (2007). Rhythm as an L2 problem. How prosodic is it?. In Trouvain J., Gut, U. (Eds.), *Non-native Prosody: Bridging the Gap between Research and Teaching*. Berlin, New York: De Gruyter Mouton, 97–120.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. In *J. Stat. Softw.* 67(1), 1-48. http://dx.doi.org/10.18637/jss.v067.i01

Best, C., Tyler, M. (2007). Nonnative and Second-Language Speech Perception: Commonalities and Complementarities. In Bohn, O-S., Munro, M. (Eds.), *Language Experience in Second Language Speech Learning: in Honor of James Emil Flege*. Amsterdam: Benjamins, 13-33.

Boersma, P., Weenink, D. (2022). PRAAT: doing phonetics by computer. [Computer program] Version 6.2.06. https://www.praat.org

Budeanu, A., De Meo, A. & Pettorino, M. (2020). Caratteristiche fonetiche dell'italiano di romeni in Calabria. Riflessioni su ritmo e lunghezza vocalica e consonantica. In Romito, L. (a cura di), *La variazione linguistica in condizioni di contatto: contesti*

acquisizionali, lingue, dialetti e minoranze in Italia e nel mondo, Studi AISV 7. Milano: Officinaventuno, 233-243.

Byers-Heinlein, K., Burns, T.C. & Werker, J.F. (2010). The roots of bilingualism in newborns. In *Psychological science*, 21(3), 343-348.

Carter, P., M. (2005). Quantifying rhythmic differences between Spanish, English and Hispanic English. In Gess R., Rubin, E.J., *Theoretical and Experimental Approaches to Romance Linguistics: Selected papers from the 34th Linguistic Symposium on Romance Languages (LSRL)*, Salt Lake City, USA, March 2004, 63-75.

Chini, M. (Ed.) (2015). *Il parlato in (italiano) L2: aspetti pragmatici e prosodici.* Milano: Franco Angeli.

Consiglio d'Europa (2021). *Quadro comune europeo di riferimento per le lingue: apprendimento, insegnamento, valutazione. Volume complementare*. Trad. italiana, a cura di Monica Barsi, M, Lugarini E. & Cardinaletti, A., di Council of Europe (2020). *CEFR Companion volume*, Strasbourg: Council of Europe Publishing.

Costamagna, L., Giannini, S. (2003). *La fonologia dell'interlingua. Principi e metodi di analisi*. Milano: Franco Angeli Edizioni.

Council of Europe (2002). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Strasbourg: Council of Europe Publishing.

De Meo, A., Pettorino, M. & Vitale, M. (2015). Parli nativo? I correlati acustici dell'accento "nativo" in italiano: uno studio acustico-percettivo. In Bruno, C., Casini, S., Gallina, F. & Siebetcheu, R. (Eds.), *Atti XLVI Congresso Internazionale SLI "Plurilinguismo/Sintassi"*. Roma: Bulzoni, 191-206.

De Meo, A., Vitale, M. (2018). Tipologia ritmica e apprendimento di una seconda lingua. In Brincat, G., Caruana, S. (Eds.). *Tipologia e "dintorni". Il Metodo tipologico alla intersezione di piani d'analisi*. Roma: Bulzoni, 45-62.

Delais-Roussarie, E., Avanzi, M. & Herment, S. (Eds.) (2015). *Prosody and Languages in Contact: L2 Acquisition, Attrition, Languages in Multilingual Situations*. Berlin: Springer.

Dellwo, V., Wagner, P. (2003). Relations between language rhythm and speech rate. In *Proceedings of the 15th International Congress of Phonetics Science. International Congress of Phonetics Science*, Barcellona, Spain, 3-9 August 2003, 471-474.

Derwing, T.M., Munro, M.J. (2005). Second language accent and pronunciation teaching: A research-based approach. In *TESOL Quarterly*, 39(3), 379-398.

Eckman, F.R. (1977). Markedness and the contrastive analysis hypothesis. In *Language Learning*, 27, 315-330.

Flege, J.E., Bohn, O-S. (2021). The Revised Speech Learning Model (SLM-r). In Wayland R. (Ed.), *Second Language Speech Learning: Theoretical and Empirical Progress*. Cambridge: Cambridge University Press, 3-83.

Gabriel, C., Stahnke J. & Thulke J. (2014). On the acquisition of French speech rhythm in a multilingual classroom: Evidence from linguistic and extra-linguistic data. In *SHSWeb of Conferences* 8, EDP Sciences, 1267-1283.

Grabe, E., Low, E.L. (2002). Durational variability in speech and the Rhythm Class Hypothesis. In Gussenhoven, C., Warner, N. (Eds.), *Laboratory Phonology 7*. Berlin-New York: De Gruyter Mouton, 515-546.

Grenon, I., White, L. (2008). Acquiring rhythm: A comparison of L1 and L2 speakers of Canadian English and Japanese. In *Proceedings of the 32nd annual Boston University Conference on Language Development*, Boston, USA, 2-4 November 2007, 155-166.

Guilbault, C. (2002). *The Acquisition of French Rhythm by Second Language Learners.* PhD thesis, University of Alberta.

Gut, U. (2009). *Non-native Speech: a Corpus-based Analysis of Phonological and Phonetic Properties of L2 English and German*. Frankfurt: Peter Lang.

Gut, U. (2012). Rhythm in L2 speech. In *Speech and Language Technology*, 14/15, 83-94.

Henriksen, N. (2016). Convergence effects in Spanish-English bilingual rhythm. In *Proceedings of Speech Prosody 2016*, Boston, USA, 31 May - 3 June 2016, 721-725.

Hothorn, T., Zeileis, A., Farebrother, R.W. & Cummins, C. (2022). *lmtest: Testing Linear Regression Models*. https://CRAN.R-project.org/package=lmtest

Jang, T.-Y. (2008). Speech Rhythm Metrics for Automatic Scoring of English Speech by Korean EFL Learners. In *Malsori Speech Sounds The Korean Society of Phonetic Sciences and Speech Technology*, 66, 41-59.

Lenth, R., Singmann, H., Love, J., Buerkner & P., Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means. In *R package version* 1(1), 1-97.

Maffia, M. (2023). Il ritmo del Parkinson in lingua straniera: uno studio pilota. In Dovetto, F. (Ed.), *Lingua e patologia. Parole dentro parole fuori.* Roma: Aracne, 527-538.

Maffia, M., De Micco, R., Pettorino, M., Siciliano, M., Tessitore, A. & De Meo, A. (2021). Speech Rhythm Variation in Early-Stage Parkinson's Disease: A Study on Different Speaking Tasks. In *Frontiers in Psychology*, 12, 2021, 668291.

Mairano, P., Mois, M., De Iacovo, V. & Romano, A. (2018). Acquisizione di fenomeni temporali e ritmici dell'italiano: analisi di produzioni di apprendenti anglofoni di italiano L2. In *RiCognizioni - Rivista di lingue, letterature e culture moderne*, 5.10, 121-136.

Mairano, P., Romano, A. (2010). Un confronto tra diverse metriche ritmiche usando Correlatore. In Schmid, S., Schwarzenbach, M. & Studer, D. *La dimensione temporale del parlato*. Torriana: EDK Editore, 79-100.

Major R.C. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology*. Mahwah, NJ: Lawrence Erlbaum Associates.

Matticchio, I. (2020). Le metriche ritmiche applicate allo studio del parlato bilingue. Stato dell'arte e implicazioni per possibili studi sul contatto slavo-romanzo nell'Alpe-Adria. In *Colloquium: New Philologies*, 5(2), 71-104.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J. & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. In *Cognition* 29, 143-178.

Mennen, I. (2015). Beyond Segments: Towards A L2 Intonation Learning Theory. In Delais-Roussarie, E., Avanzi, M. & Herment, S. (Eds.), *Prosody and Languages in Contact: L2 Acquisition, Attrition, Languages in Multilingual Situations*. Berlin: Springer, 171-188.

Pellegrino, E. (2012), The perception of foreign accent and speech. Segmental and suprasegmental features affecting degree of foreign accent in Italian L2. In Mello, H., Pettorino, M., Raso, T. (Eds.), Proceeding of the VIIth GSCP International Conference – Speech and Corpora, Firenze: Firenze University Press, 261-267.

Pettorino, M., De Meo, A. (2012), *Prosodic and Rhythmic Aspects of L2 Acquisition: The Case of Italian*, Cambridge: Cambridge Scholars Publishing.

Pettorino, M., Maffia, M., Pellegrino, E., Vitale, M. & De Meo, A. (2013). VtoV: a perceptual cue for rhythm identification. In Mertens, P., Simon, A.C. (Eds.), *Proceedings of the Prosody-Discourse Interface Conference IDP-2013*, Leuven: Belgium, 11-13 September 2013, 101-106.

Piccardo, E. (2016), *Phonological Scale Revision Process Report*, Education Policy Division, Council of Europe. https://rm.coe.int/168073fff9. Ultimo accesso 31 maggio 2024.

R Core Team (2022). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/

Ramus, F. (2002). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. In *Annu. Rev. Lang. Acquis*. 2. 10.1075/arla.2.05ram.

Ramus, F., Nespor, M. & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. In *Cognition*, 73(3), 265-292.

Romito, L., Tarasi, A. (2012), A Rhythmic-Prosodic Analysis of Italian L1 and L2. In Pettorino, M., De Meo, A. (Eds.), *Prosodic and Rhythmic Aspects of L2 Acquisition: The Case of Italian*, Cambridge: Cambridge Scholars Publishing, 137-152.

Roseano, P. (2020), Il ritmo linguistico del ladino dolomitico: studio acustico del badiotto. In *Ladinia*, 40, 351-373.

Russo, D. (2015). *The Notion of Syllable across History, Theories and Analysis*. Cambridge: Cambridge Scholars Publishing.

Schmid, S., Dellwo, V. (2012). Caratteristiche temporali del parlato italiano e tedesco: un confronto tra parlanti nativi, bilingui e non-nativi. In: Falcone, M., Paoloni, A. (Eds.), *La voce nelle applicazioni*. Roma: Bulzoni, 159-174.

Stockmal, V., Markus, D. & Bond, D. (2005). Measures of Native and Non-Native Rhythm in a Quantity Language. In L*anguage and Speech*, 48, 55-63.

Tortel, A., Hirst, D. (2010). Rhythm metrics and the production of English L1/L2. In *Proceedings of Speech Prosody 2010*, Chicago, USA. 10-14 May 2010, paper 959.

Trofimovich, P., Baker, W. (2006). Learning Second-language Suprasegmentals: Effect of L2 Experience on Prosody and Fluency Characteristics of L2 Speech. In *Studies in Second Language Acquisition*, 28, 1-30.

Vitale, M., De Meo, A. (2017). Rhythmic Differences and Second Language Acquisition. An Italian-based study. In Álvarez López, L., Bardel, C., Gudmundson, A. (Eds.), *Multilingualism and acquisition*, Frankfurt am Main: Peter Lang, 109-122.

White, L., Mattys, S., L. (2007). Calibrating rhythm: First language and Second language studies. In *Journal of Phonetics*, 35, 501-522.

DOI: 10.17469/O2112AISV000013

MATTEO GAY, CLAUDIA ROBERTA COMBEI

Whose Voice Speaks Volumes? The Problem with Gender Identification from Speech[1]

This study investigates the topic of voice-based gender identification, focusing on aspects that concern accuracy and ethical implications. One aim is to review the literature available in order to understand what shapes the expectations of gender as revealed through voice. Then, an experimental investigation is conducted to examine the impact of age on voice-based gender classification. Our evaluation of a Multilayer Perceptron neural network model using MFCC features indicates 0.90 overall accuracy in binary gender classification. This seemingly high accuracy score would mask a substantial 10% misclassification rate in real-world applications. When considering age, accuracy drops further, with varying scores across the seven groups (0.64-0.88). Generally, our results indicate age-related variability in model performance, highlighting limitations and ethical concerns in generalizing across age groups. As concerns societal implications, our literature review and experiments suggest that greater awareness and diversity-informed approaches are needed in the design, development, and marketing of speech technologies.

Keywords: speech technology, voice, gender, bias, ethics.

1. *Introduction*

The recent advancements in voice-based technologies have been possible due to significant progress in the fields of speech and natural language processing (NLP), natural language understanding (NLU), and natural language generation (NLG). It has been reported that more complex deep learning architectures, such as recurrent neural networks (RNNs) and end-to-end training techniques, can improve the accuracy of automatic speech recognition (ASR), particularly as regards adverse conditions (Baby et al., 2022; Chen et al., 2023a).

Nevertheless, even advanced ASR systems and well-trained acoustic models can still struggle under challenging conditions (e.g., noise, non-native speech, etc.). Inspired by the fact that human listeners use contextual cues to infer meaning and reduce reliance on auditory input, Chen et al. (2023b) propose an open-source benchmark that uses large language models (LLMs) for ASR error correction, leveraging N-best decoding hypotheses to improve speech recognition accuracy.

[1] This paper is the result of close and continuous collaboration between the two authors, who are equally responsible for its content. For the purposes of Italian academia, C.R. Combei wrote § 1 (Introduction), § 2 (The topic), and § 5 (Conclusions), while M. Gay wrote § 3 (Data and methods) and § 4 (Results).

The study of Yu et al. (2024) goes in a similar direction, as it compares different ways of connecting LLMs to speech encoders for ASR. The authors claim that LLMs that use the Q-Former method may achieve better accuracy, even for out-of-domain speech. Additionally, they propose a segment-level Q-Former for handling longer speech to further improve accuracy, even though it appears that excessively long inputs increase the likelihood of hallucinating LLMs.

Other areas of research, such as intent recognition for conversational Artificial Intelligence (AI) chatbots (Chandrakala et al., 2024) or dialogue modelling and management (Fernández-Rodicio et al., 2020), have also contributed to the advancement of speech technologies, allowing, for instance, voice assistants to engage in conversations that are more natural and aware of the context. The combination of the aforementioned techniques has facilitated the development of commercial applications for everyday usage, such as real-time speech translation (Bentivogli et al., 2021), accessibility tools for visually impaired users (Norkhalid et al., 2020), and other voice-controlled interfaces.

These recent advancements in speech technologies, along with the possibility to use voice commands to interact with electronic devices (e.g., smartphones, computers, virtual assistants, in-vehicle infotainment, etc.) have profoundly transformed the way we communicate and access information. In line with what has been said above, voice-based applications have evolved well beyond basic commands (e.g., recognizing dictated numbers) to now being able to perform complex and sophisticated tasks (e.g., robust speech recognition under adverse conditions).

Voice has long been used also in biometric technologies, particularly within security systems designed for access control and user authentication (Fong, 2012). Since biometric voice authentication relies on the speaker's unique characteristics, it is featured in commercial applications for speaker verification that allow to unlock personal devices and access home banking, or smart home systems through vocal commands (Meng et al., 2020). Moreover, phone surveillance operations, border security devices, and forensic applications have been known to make use of voice classification techniques to categorize speakers based on attributes like gender[2], ethnicity, and age (Jain, Ross, 2015; Dumbrava, 2021).

Furthermore, customer relationship management systems have recently evolved to incorporate voice-based demographic classification for highly targeted advertising and marketing strategies (Aravinda et al., 2019). Consider, for instance, a scenario where a person is either scrolling through a social media application on their phone or using voice commands in a search engine application. Both applications could employ voice classification technology through the user's microphone (assuming the user has given prior permission for microphone access).

[2] To align with the terminology used in the corpus documentation used in our study (Burkhardt et al., 2010, 2023), we consistently refer to this variable as 'gender'. Besides ensuring coherence with the literature on the topic at hand, this terminological choice facilitates consistency throughout our analyses. The meaning of the term 'gender' is detailed in § 2.1.

The speech classification technology could analyse the voice interacting with the device, and based on that evaluation, assume the gender of the user. Then, this information could be used to customize the advertisements displayed to the user. For example, a person whose voice is classified by the system as male could receive advertisements for beard grooming products, while a person classified as female could receive advertisements for lipstick. Finally, it has been suggested that voice assistants on commercial devices are now able to tailor their responses based on the perceived gender of the user, inferred from voice characteristics and other cues derived from their behaviour (Alnuaim et al., 2022).

Similar to automated systems, humans also engage in the practice of inferring gender from the vocal characteristics of the speakers they hear and/or talk to. According to Dunkerson and Weiler (2023), listeners tend to assign a (binary) gender identity to a speaker exclusively on the basis of voice traits (e.g., when engaging in a telephone conversation with an unfamiliar interlocutor, one usually decides whether to address them as "Mr." or "Ms.", solely based on auditory cues). Nonetheless, Sutton (2020) warns that gender is not inherent in voice (listeners assign gender to voice) and that it is actually constructed through a number of variables (see also § 2.1).

As a matter of fact, the issue of assuming a person's gender, whether by humans or algorithms, has been a topic of intense debate in recent years. One significant issue reported in the literature is misgendering, which according to Fosch-Villaronga et al. (2021: 1) "reinforces gender stereotypes, accentuates gender binarism, undermines privacy and autonomy, and may cause feelings of rejection, impacting people's self-esteem, confidence, and authenticity". Their research examines the consequences of misgendering and highlights how algorithmic bias[3] can inadvertently violate privacy and perpetuate social biases and prejudices regarding gender identity.

Our study goes in a similar direction, and as anticipated throughout this section, it aims to critically address the concerns regarding automatic systems tasked with classifying speakers into gender categories based on their voice. To this end, we undertake a two-pronged approach. First, looking at the literature on the topic at hand, we examine the constructs that inform our inferences about gender as revealed through voice; in our exploration of the scholarly debate, we also highlight the potential issues of these assumptions in real-world situations. Second, we complement our theoretical arguments with an empirical investigation on gender recognition from speech by exploring the impact of the age variable on the detection process.

In addition to this introduction, the paper contains four other sections organized as follows: § 2 details the relationship between gender and voice, as

[3] We would like to stress the fact that bias in technology should not be relegated only to the issue of algorithmic bias. In fact, a significant portion of bias originates from naturally occurring data produced by human beings that are used to train and develop NLP technologies (Schramowski et al., 2022). These data inherently reflect the stereotypes, prejudices, and discriminatory practices that exist in the real world.

well as the motivations and objectives of our study; § 3 describes the data and the methods; § 4 presents the results of our analysis; § 5 concludes the paper.

2. *The topic*

This work deals with a topic that resonates with the growing attention towards ethical considerations in language technologies. A recent position paper by Sarker (2024) emphasizes the potential of AI and LLMs, while at the same time stressing the need to promote awareness and responsible practices in their integration in society. The paper highlights the importance of ethics, trustworthiness, and fairness in the development and use of these technologies, given their increasing presence in real-world applications. A relevant issue addressed is algorithmic bias, which can result in unfair or discriminatory outcomes, as well as in the spread and consolidation of harmful stereotypes and prejudices (Blodgett et al., 2020). Moreover, as highlighted by a large body of research (see, inter alia, Wu et al., 2022; Singh et al., 2023), the opaque nature of these systems makes it difficult to understand the mechanisms behind their outputs. In order to address these challenges, Sarker (2024) argues for the promotion of accountability, transparency, and interdisciplinary collaboration among stakeholders. Our study aligns with the existing literature on the ethical implications of developing algorithms for everyday tasks, as we advocate for continued debate and research on the topic to ensure a responsible deployment of these technologies.

The timeliness of this type of research is further demonstrated by the recent legislative developments regarding AI in the European Union. In fact, the European Parliament and Council negotiators reached a preliminary agreement on the Artificial Intelligence Act (AI Act)[4] on December 8, 2023, that was formally adopted on March 13th, 2024. The AI Act establishes a regulatory framework that assigns responsibilities to both developers and users of AI systems, with the level of accountability contingent upon the perceived risk associated with the technology. These regulations outline that some AI technologies and applications are entirely banned. They include biometric categorization systems that rely on sensitive data, unauthorized large-scale collection of facial recognition data from public sources (e.g., internet, CCTV footage, etc.), the use of AI for emotion recognition in workplaces and schools, social scoring systems, predictive policing based on user profiling, and AI designed to employ cognitive behavioral manipulation techniques targeting individuals or vulnerable groups (e.g., voice-activated toys that encourage dangerous behavior in children).

Voice technology encounters high expectations from users, whether they are individuals, private companies, or public institutions and organizations. However, what users might not know is that each decision of these automated

[4] The text adopted as the AI Act is available here: https://www.europarl.europa.eu/doceo/document/TA-9-2024-03-13-TOC_EN.html (accessed on May 30th, 2024).

systems consists of various intermediate steps that bring in elements of statistical probability and chance. The widespread idea that AI technologies, including voice-based systems, are infallible presents some problems, as it undermines critical scrutiny of their outputs in real-world scenarios and perpetuates undue trust. In fact, Kidd and Birhane (2023: 1222) warn that "[o]verhyped, unrealistic, and exaggerated capabilities permeate how generative AI models are presented" and this, of course, contributes to the "misconception that these models exceed human-level reasoning".

Our paper is motivated by the need to address the gap between the perceived infallibility of these technologies and the reality of their limitations, inaccuracies, biases, and ethical concerns.

2.1 Gender and voice

Commercial applications provide the statistically most probable outcome for speech recognition, verification, and classification tasks, but the factors influencing their outputs remain unknow to the public, due to the lack of transparency in the (commercial) model development and to the complexities associated with its interpretability. Moreover, the results depend on the quality and quantity of the training data employed for the construction of the model.

At the same time, the outputs of voice-based technologies are influenced by the developer's preconceived expectations regarding human voices and their interaction with demographic factors, such as gender and age (e.g., how an adult female voice is expected to differ from an adolescent male voice). As shown by previous research (see, inter alia, Fant, 1966; Nordström, 1977), these assumptions are typically supported by anatomical differences (e.g., size and shape of the vocal tract and vocal folds).

Traditionally, the phonetic differences between female and male speakers have been investigated in terms of the static acoustic and perceptual consequences of various articulatory dimensions, indicating, among others, an average difference of approximately 20% in vocal tract length between females and males (Simpson, 2001). Specifically, the pioneering work by Fant (1966, 1975) identifies larger laryngeal cavities and proportionally longer pharynges in male speakers as compared to female speakers.

In more recent times, the study of Smith et al. (2007) synthesises a range of vowels from recordings of four different speakers (an adult female, an adult male, a young boy, and a young girl) to assess whether gender and age influence the listeners' perception, after the voices are equated for glottal-pulse rate and vocal-tract length. The results show that listeners manage to distinguish children from adults by their voices, but they are unable to differentiate between male and female voices. A recent work by Zhang (2021) investigates the contribution of vocal fold length, thickness, and depth to differentiate between male and female voice production. The results indicate that the differences in voice production between

adult females, adult males, and children can be explained by differences in length and thickness, while the effect of vocal fold depth is small.

That being said, speakers and listeners use their voice to shape and contest the societal understandings of masculinity, femininity, and sexuality (Meyerhoff, Ehrlich, 2019). In fact, according to the social constructionist theories, gender emerges not as a fixed, inherent, or biological quality, but as a dynamic construct that is perpetually shaped and enacted in our daily interactions, both with ourselves and with others (Butler, 1999). Previous research has shown that the standards for effectively performing gender differs across historical and cultural contexts (Schilt, Westbrook, 2009). According to Dunkerson and Weiler (2023: 2), the Western paradigm fundamentally associates the 'correct' gender performance with binary genitalia distinctions (for female and male) and heterosexuality, constituting the so called 'cisheteronormativity'. In ordinary social interactions, however, gender attribution is more likely to be an inference based on a culturally constructed set of cues and behaviours associated with masculinity and femininity (West, Zimmerman, 1987).

The developers of speech technology likely base their projections regarding gender and voice on the two factors discussed so far. First, the culturally ingrained perceptions of masculinity and femininity, encompassing notions of gender-specific behaviours. Second, acoustic and auditory traits originating, for instance, from biological differences in the vocal tract and vocal folds. Besides these two elements, the corpora used for training purposes in the development of voice-based applications play a crucial role, as gender biases or misrepresentations in the training data have a significant impact on the performance of the systems developed. All things considered, insufficient and poorly representative data together with the developers' preconceived expectations have problematic consequences on the functioning of voice applications, resulting in various issues, such as misidentification and misgendering, as well as in the reinforcement of gender biases.

Various external variables can, in fact, alter the characteristics of one's voice. Smoking, for instance, may render the voice hoarse (Tafiadis et al., 2017), while alcohol consumption may result in slurred speech (Pisoni, Martin, 2023). Altered voice productions are also caused by dysphonia in individuals experiencing conditions that affect the vocal tract, such as the cold, the flu, the Parkinson's and the Alzheimer's disease (Contreras et al., 2023).

Recent research has also demonstrated that acoustic features, particularly fundamental frequency (F0), which is commonly used in voice-based gender classification tasks, are often influenced by conversational, contextual, and psycho-social factors. For example, a paper by Hughes and Puts (2021: 1) shows that speakers often lower their voice pitch in order "to exert dominance, display status, and compete with rivals". Then, Bradshaw et al.'s (2022) study indicates that F0 variability is heightened during initial conversational turns, such as greetings, likely serving to signal identity, mark boundaries, or attract attention.

The assumption that algorithms can accurately identify a person's gender based on their physical characteristics, such as the voice, is fraught with significant risks. As highlighted by Hamidi et al. (2018), diverse gender identities, including transgender and nonbinary individuals, have expressed profound apprehensions regarding the use of automated recognition and verification systems. The primary concern of these individuals regards the possibility that language technologies misgender or misunderstand them. In the online article "Queer Eye for AI: Risks and limitations of artificial intelligence for the sexual and gender diverse community" published on May 26th, 2023, for Open Global Rights (hosted by the Center for Human Rights and Global Justice and the Future of Rights Program at New York University School of Law)[5], Ilia Savelev illustrates the following problematic scenario:

> Imagine calling your bank to resolve an issue, but the customer service representative suddenly accuses you of fraud and threatens to suspend your bank account if you don't invite the "real" you. Why? Because their artificial intelligence (AI) voice recognition system concluded that your voice was not "male" enough to match their records. This is a universal experience among transgender individuals, including myself, and is just one of the significant risks that AI poses to the LGBTQI+ community.

In fact, the research of Ovalle et al. (2023) reveals a dominance of binary gender norms reflected by the AI models and stresses the need for further investigation into the manifestation of transgender and non-binary invisibility and bias within language technology.

A different example of misgendering based on physical traits is provided by Peynircioğlu et al. (2017) in their study regarding the McGurk effect in gender identification. Their results show that while participants accurately identified gender 100% of the time in conditions where video and audio were matched, the accuracy dropped to 31% in conditions where video and audio were mismatched. This finding indicates that visual gender cues can override auditory gender identification, revealing a strong non-speech McGurk effect.

2.2 From theory to practice

The elements discussed in the previous sections highlight the need for scholarly research that can guide the development of more inclusive speech technologies able to represent the complexity and diversity of human voices. In particular, interdisciplinary studies – with insights from gender studies, computer science, sociology, psychology, and linguistics – are needed to address the concerns surrounding speech classification systems.

To contribute to this debate, we conducted an empirical investigation on gender recognition from speech by exploring the effect of age on the identification process. Our hypothesis is that the age variable has a negative effect on the performance

[5] The article is available here: https://www.openglobalrights.org/risks-limitations-artificial-intelligence-sexual-gender-diverse-community/ (accessed on May 30th, 2024).

of the gender classification task, thereby diminishing the expected reliability and efficacy of this type of application.

This paper uses a practical real-world example where the factors within the data can significantly impact the performance of a speech classification system. Thus, we seek to demonstrate how the presence of an experimental variable (i.e., age) can influence the accuracy in identifying the classes of the target variable (i.e., gender). By doing so, our goal is to raise awareness of the potential for errors in these systems, discussing, at the same time the need for, a more prudent, inclusive, and informed approach to their design, development, and use.

3. *Data and methods*

Our study evaluates one of the prevailing baseline technologies in automatic speech classification, namely the application of neural network architectures based on Mel Frequency Cepstral Coefficients features. The corpus under examination is derived from the Agender corpus (Burkhard et al., 2010) for the German language, a well-established and extensive telephonic speech dataset built for age and gender classification, originally released for the INTERSPEECH 2010 Paralinguistic Challenge. The Agender corpus comprises 65,241 individual utterances from 945 native German speakers, categorized into four age classes. These age classes, except for children, are further divided based on the gender variable, resulting in seven groups: CHILD (7-14 years old, one class), YOUNG (15-24 years old, two classes), ADULT (25-54 years, two classes), and SENIOR (55-80 years, two classes). Each speaker delivered 18 speech items in up to six different sessions.

Since we are not engaging in the replication of the challenge, neither we aim to achieve high performance scores and compete with previous results, we decided to randomly split the train/development corpus to create the train/test sets (85% and 15%). Special attention was given to ensuring that the same speaker did not appear in both sets to eliminate any possibility of voice-specific pre-knowledge bias. As a result, our corpus includes the seven classes displayed in Tab. 1: CHILD (106 speakers, 6,802 instances, not divided by gender), YOUTH (99 females, 7,360 instances; 88 males, 6,189 instances), ADULT (113 females, 7,934 instances; 107 males, 6,929 instances), and SENIOR (133 females, 8,485 instances; 134 males, 9,375 instances). Thus, the final dataset for our experiment included a total of 53,074 instances from 770 speakers.

Table 1 - *Dataset description: f and m abbreviate female and male, x represents children w/o gender discrimination. The last two columns represent the number of speakers/instances per partition (Train and Test).*

Class	*Group*	*Age*	*Gender*	*#Train*	*#Test*
1	**C**hild	7-14	*x*	415/5,965	57/839
2	**Y**outh	15-24	*f*	393/6,151	79/1,209

3	**Y**outh	15-24	*m*	343/5,130	73/1,059
4	**A**dult	25-54	*f*	472/6,857	77/1,077
5	**A**dult	25-54	*m*	403/5,869	73/1,033
6	**S**enior	55-80	*f*	494/7,105	96/1,380
7	**S**enior	55-80	*m*	551/7,986	87/1,227

The distinctiveness of this corpus lies in the fact that the original audio files were recorded using standard cell phones (GSM) and landline connections at a sampling rate of 8,000 Hz in 8-bit ALAW format, introducing substantial disturbances and noise into the acoustic spectrum of the voice recordings. The training and development sets of the Agender corpus are publicly accessible and can be downloaded from the BAS CLARIN repository[6] of the Bavarian Archive for Speech Signals at Ludwig-Maximilians-Universität Munich. The files are provided in .raw format, so for the purposes of our analysis we converted them to .wav.

As mentioned in § 1, in the domain of speech processing, there exists a significant area of research known as speech classification. Unlike ASR, the primary objective of speech classification is not to obtain textual transcriptions of utterances but rather to assign human speech production to predefined classes related primarily to the speaker. These classes encompass attributes such as age range, gender, presence of clinical disorders, emotional states, speaker intentions (e.g., asking a question, making a statement, expressing uncertainty), and geographical provenance.

When approaching a classification problem from a computational point of view, typically, the primary focus is to optimize the classifier algorithm. However, before reaching the mapping function, an essential and non-negligible step is the selection and extraction of the most informative features from the input data, in order to reduce the space of the problem. These features (numerically encoded to feed the classification function) capture different aspects of the speech signal, containing valuable information for distinguishing between different classes. While the importance of specific features may vary depending on the classification task and the nature of the data, several key features have been widely recognized as crucial in speech classification, such as Mel Frequency Cepstral Coefficients (MFCCs), Shifted Delta Cepstral Coefficients (SDCC), Linear Prediction Cepstral Coefficients (LPCC), among low-level spectral features, spectral centroid, pitch-related features (mainly F0), as well as high-level features such as jitter and shimmer, commonly used in speech disorders and emotion detection (Teixeira et al., 2013).

In this study, we trained a vanilla Multilayer Perceptron (two hidden layers, 80 and 40 nodes) with a binary classifier head using the Scikit-learn MLPClassifier implementation on Python 3.9, employing only MFCC features. The model was trained on pre-processed audio data, which included noise reduction, silence trimming, and pre-emphasis techniques. Since our goal was to identify a possible

[6] The Agender corpus is available here: https://clarin.phonetik.uni-muenchen.de/BASRepository/index.php?target=Public/Corpora/aGender/aGender.1.php (accessed on May 30th, 2024).

source of misclassification (i.e., misgendering), not to build a robust and high-performing model[7], we opted for this simple yet fundamental architecture to maintain low complexity and higher interpretability.

4. *Results*

In this section, we present the evaluation results for the Multilayer Perceptron classifier in two specific tasks: overall gender prediction (G) and gender prediction within specific age classes (G-in-A). We report as measures F1 metric and accuracy.

The best score, overall accuracy and F1 of 0.90 (Tab. 2), was achieved in the binary G classification (excluding the first class, consisting of children aged 7 to 14 years old, for which gender information is not provided in the dataset). Table 2 shows the relevant confusion matrix for the results of this task.

Table 2 - *Confusion Matrix, gender binary classification task*

	Male	*Female*
Male	3187	291
Female	394	3006

While these metrics might suggest that the model performs well in distinguishing between male and female speakers, the 10% of the samples that were incorrectly classified represent a substantial number of individuals being misgendered in such a simple task, highlighting a critical issue that pure numerical performance metrics may overlook.

Notably, we observed an improvement (+0.05) in classification accuracy after applying the previously described preprocessing techniques (see § 3), which appears to be a crucial initial step when dealing with a noisy corpus like a telephone recorded speech dataset. It is also worth mentioning that the model exhibited a slight preference for predicting male gender. The performance on pre-processed audio file confirms the fact that MFCC features alone have, indeed, a relevant role in gender detection.

However, the most interesting results achieved are those produced by the experiment on G-in-A classification (Tab. 3 and Tab. 4).

Table 3 - *Results for Gender in selected Age Class experiment (Y = Young, 15-24 yrs; A= Adult, 25-54 yrs; S = Senior, 55-80 yrs; M = Male, F = Female.)*

	YF + YM	*AF + AM*	*SF + SM*
F1 score	0.92	0.84	0.80

[7] For more complex models addressing the gender classification task, including experiments on the Agender corpus, see Qawaqneh et al. (2017) and Burkhardt et al. (2023).

In particular, we attempted to train the model on gender recognition within specific age ranges, and we discovered that our SK-Model yielded better results in correctly detecting gender in younger speakers compared to seniors, with demonstrating median performance in the adults' class (Tab. 3).

As a demonstration of the utility and precision of MFCCs in suggesting possible correlations between speech, vocal tract characteristics, gender, and age, we also attempted to predict gender within specific age class pairings that were constructed for this purpose (Tab. 4). In this case, we included children as a separate class, since the Agender Corpus does not provide gender information for them. We observed that discrimination was generally easier between children and male speakers than between children and female speakers.

Table 4 - *Results for Child (C) vs Male/Female across different age classes.*

	C + YF	*C + YM*	*C + AF*	*C + AM*	*C + SF*	*C + SM*
F1 score	**0.64**	0.86	**0.76**	0.87	**0.79**	0.88

Specifically, the accuracy in predicting the values Child or Male gender remained consistently around 0.86-0.88 across all age classes (Young, Adult, Senior), whereas the model performed better in predicting Child or Female gender with older age classes than with younger (voice characteristics more similar to those of children), ranging from 0.64 to 0.79.

To substantiate these observations, we conducted bootstrap resampling analyses to evaluate the statistical significance of the differences in macro-averaged F1 scores between binary classifiers. When comparing C + YF to C + SF, the 95% Confidence Interval (CI) for the mean difference in F1 scores, derived from 10,000 bootstrap samples, was [0.071933, 0.126724]. This result allows us to reject the null hypothesis of no difference with high confidence. Similarly, significant differences were observed when comparing C + YF to C + AF (95% CI: [0.114848, 0.167898]) and C + AF to C + SF (95% CI: [-0.065162, -0.017584]), as detailed in Fig. 1, 2, and 3, in the Appendix.

Additionally, similar significant results were obtained when analyzing differences across gender groups within same age classes (e.g., Fig. 4, in the Appendix). Moreover, the slight improvement in distinguishing between males and children with increasing age classes did not reach statistical significance (95% CI: [-0.0028, 0.0355]), as shown in Fig. 5 and 6, in the Appendix. As reported by the literature cited in § 2.1, these findings may be linked to the physiological evolution of the vocal tract and vocal folds across different ages and the distinct physical characteristics of females and males.

Our findings reveal that the accuracy of the model varies greatly across different age groups, indicating that age acts as a confounding variable in voice-based gender classification. This variability in performance suggests a possible limitation in the capability of the model to generalize across the lifespan of an individual. We would like to stress that this issue extends beyond the influence of age on model

performance; it also raises significant ethical concerns related to misclassification in systems using architectures similar to the one we tested. This is particularly problematic for under-represented groups that may be more susceptible to the negative effects of bias and misgendering.

5. *Conclusions*

This paper addressed the topic of voice-based gender identification that, in very recent times, has started to receive attention in the scholarly debate, particularly as regards the issue of misgendering. First, we examined the topic, by reviewing the existing literature, to outline the elements that determine the way we shape our expectations about gender and voice. Second, we performed an experiment to investigate the impact of age on the task of gender classification from speech. We hypothesized that the age variable would affect the performance of gender identification, thereby reducing the accuracy of the model.

To this end, we tested one of the prevailing baseline technologies in automatic speech classification, namely the deployment of Multilayer Perceptron neural network architectures based on MFCC features. The model achieved an overall accuracy and F1 score of 0.90 in binary gender classification (the class of children aged 7 to 14 were excluded from this analysis). Even though these metrics may suggest robust performance in recognizing male and female speakers, the 10% error rate would represent a substantial number of misgendering instances, in a real-world situation.

As hypothesized, when predicting all seven classes, thus co-varied by the age factor, the accuracy decreased. Male speakers and children were identified with an accuracy from 0.86 to 0.88 across all age groups (young, adult, senior), while female speakers and children displayed a wider range of accuracy, namely from 0.64 to 0.79, with older age groups being more accurately classified than younger ones.

In general, our results showed that the accuracy of the model varied across different age groups, highlighting the role of age in gender classification from speech. This variability not only affects model performance but also reflects the limitations of the systems that are based on models similar to the one we tested. At the same time, it suggests that the topic needs attention from both researchers and the industry, for a more cautious, diversity-aware approach to the design, development, and marketing of speech technology. Switching to the societal perspective, we believe that studies like the one presented here are needed to raise awareness about the potential errors, inaccuracies, and biases in voice-based applications.

Appendix

Figure 1

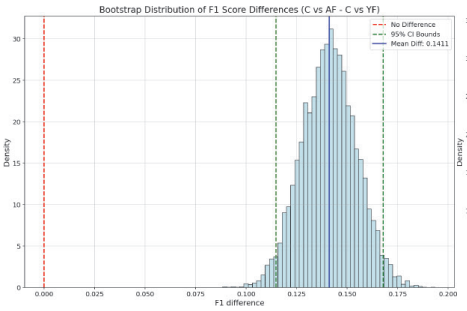


Figure 2

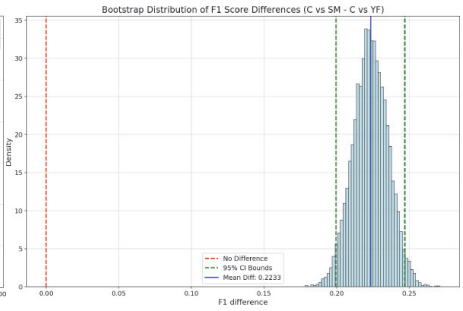


Figure 3

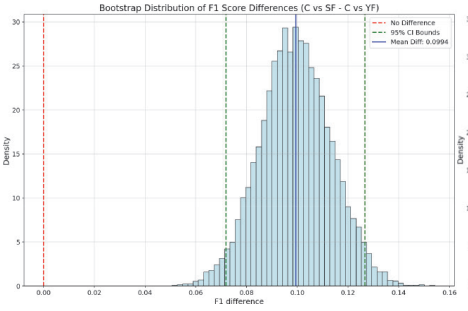


Figure 4

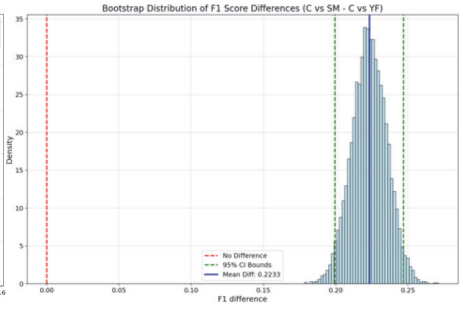


Figure 5

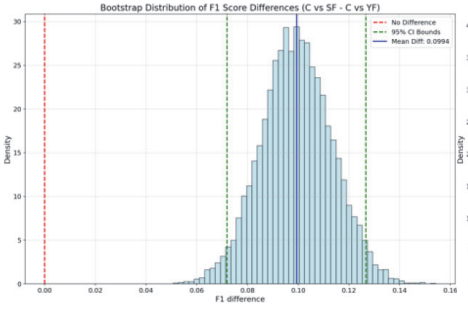
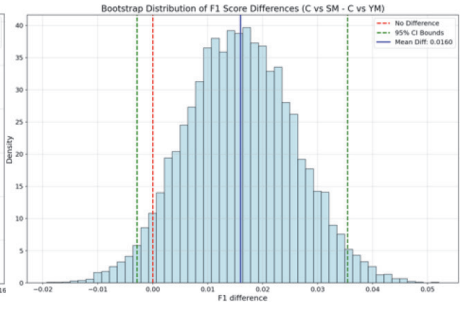


Figure 6



Acknowledgements

The first draft of this paper was written while C.R. Combei was engaged as a researcher under the initiative PON Ricerca e Innovazione 2014-2020 – Linea Innovazione (D.M. 1062/2021) at the University of Pavia.

References

Alnuaim, A.A., Zakariah, M., Shashidhar, C., Hatamleh, W.A., Tarazi, H., Shukla, P.K. & Ratna, R. (2022). Speaker gender recognition based on deep neural networks and ResNet50. In *Wireless Communications and Mobile Computing*, 2022(Special Issue), 4444388. https://doi.org/10.1155/2022/4444388.

Aravinda, D., Arpitha, N. & Mamatha, M. (2019). Gender Voice Classification using Deep Learning Convolutional Neural Networks. In *Journal of Critical Reviews*, 6(6), 2595-2601.

Baby, D., D'Alterio, P. & Mendelev, V. (2022). Incremental learning for RNN-Transducer based speech recognition models. In *Proceedings of Interspeech 2022*, Incheon, Korea, 18-22 September 2022, 71-75. https://doi.org/10.21437/Interspeech.2022-10795.

Bentivogli, L., Cettolo, M., Gaido, M., Karakanta, A., Martinelli, A., Negri, M. & Turchi, M. (2021). Cascade versus direct speech translation: Do the differences still make a difference? In Zong, C., Xia, F., Li, W. & Navigli, R. (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, Bangkok, Thailand, 1-6 August 2021, 2873-2887. https://doi:10.18653/v1/2021.acl-long.224.

Blodgett, S.L., Barocas, S., Daumé III, H. & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In Jurafsky, D., Chai, J., Schluter, N. & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, Online, 5-10 July 2020, 5454-5476. https://doi.org/10.18653/v1/2020.acl-main.485.

Bradshaw, L., Chodroff, E., Jaeger, L. & Dellwo, V. (2022). Fundamental frequency variability over time in telephone interactions. In *Proceedings of Interspeech 2022*, Incheon, Korea, 18-22 September 2022, 101-105. https://doi.org/10.21437/Interspeech.2022-10669.

Burkhardt, F., Eckert, M., Johannsen, W. & Stegmann, J. (2010). A Database of Age and Gender Annotated Telephone Speech. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10)*, La Valetta, Malta, 19-21 May 2010, 1562-1565.

Burkhardt, F., Wagner, J., Wierstorf, H., Eyben, F. & Schuller, B. (2023). Speech-based age and gender prediction with transformers. In *Proceedings of the 15th ITG Conference on Speech Communication*, Aachen, Germany, 20-22 September 2023, 1-5. https://doi.org/10.30420/456164008.

Butler, J. (1999). *Gender Trouble: Feminism and the Subversion of Identity*. Tenth Anniversary Edition. London/ New York: Routledge.

Chandrakala, C.B., Bhardwaj, R. & Pujari, C. (2024). An intent recognition pipeline for conversational AI. In *International Journal of Information Technology*, 16, 731-743. https://doi.org/10.1007/s41870-023-01642-8.

Chen, Z.-C., Yang, C.-H.H., Li, B., Zhang, Y., Chen, N., Chang, S.-Y., Prabhavalkar, R., Lee, H.-Y. & Sainath, T. (2023a). How to estimate model transferability of pre-trained speech models? In *Proceedings of Interspeech 2023*, Dublin, Ireland, 20-24 August 2023, 456-460. https://doi.org/10.21437/Interspeech.2023-1079.

Chen, C., Hu, Y., Yang, C.-H.H., Siniscalchi, S.M., Chen, P.-Y. & Chng, E.-S. (2023b). HyPoradise: An open baseline for generative speech recognition with large

language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M. & Levine, S. (Eds.), *Advances in Neural Information Processing Systems*, New Orleans, USA, 10-16 December 2023, Vol. 36, 31665-31688.

Contreras, R.C., Viana, M.S., Fonseca, E.S., Dos Santos, F.L., Zanin, R.B. & Guido, R.C. (2023). An Experimental Analysis on Multicepstral Projection Representation Strategies for Dysphonia Detection. In *Sensors*, 23(11), 5196. https://doi.org/10.3390/s23115196.

Dumbrava, C. (2021). Artificial intelligence at EU borders: Overview of applications and key issues. In *European Parliamentary Research Service*, 21(12), 690706.

Dunkerson, A., Weiler, A. (2023). "Oh! Based on voice, assigned female at birth": Transmasculine voices and gender construction. In *Canadian Graduate Journal of Sociology and Criminology*, 6(1), 1-18. https://doi.org/10.15353/cgjsc-rcessc.v6i1.5012.

Fant, G. (1966). A note on vocal tract size factors and non-uniform F-pattern scalings. In *STL-QPSR*, 4, 22–30.

Fant, G. (1975). Non-uniform vowel normalization. In *STL-QPSR*, 2(3), 1-19.

Fernández-Rodicio, E., Castro-González, Á., Alonso-Martín, F., Maroto-Gómez, M. & Salichs, M.Á. (2020). Modelling multimodal dialogues for social robots using communicative acts. In *Sensors*, 20(12), 3440. https://doi.org/10.3390/s20123440.

Fong, S. (2012). Using hierarchical time series clustering algorithm and wavelet classifier for biometric voice classification. In *Journal of biomedicine & biotechnology*, 2012(1), 215019. https://doi:10.1155/2012/215019.

Fosch-Villaronga, E., Poulsen, A., Søraa, R.A. & Custers, B.H.M. (2021). A little bird told me your gender: Gender inferences in social media. In *Information Processing & Management*, 58(3), 102541. https://doi.org/10.1016/j.ipm.2021.102541.

Hamidi, F., Scheuerman, M.K. & Branham, S.M. (2018). Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In Mandryk, R., Hancock, M. (Eds.), *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*, Montreal, QC, Canada, 21-26 April 2018, Article No. 8, 1-13.

Hughes, S.M., Puts, D.A. (2021). Vocal modulation in human mating and competition. In *Philosophical Transactions of the Royal Society B: Biological Sciences,* 376(1840), 1-10, 20200388. https://doi.org/10.1098/rstb.2020.0388

Jain, A.K., Ross, A. (2015). Bridging the gap: from biometrics to forensics. In *Philosophical Transactions of the Royal Society of London*. Series B, Biological Sciences, 370(1674), 20140254.

Kidd, C., Birhane, A. (2023). How AI can distort human beliefs. In *Science*, 380(6650), 1222–1223. https://doi.org/10.1126/science.adi0248.

Meng, Z., Bin Altaf, M.U. & Juang, B.-H. (2020). Active voice authentication. In *Digital Signal Processing*, 101, 102672. https://doi.org/10.1016/j.dsp.2020.102672.

Meyerhoff, M., Ehrlich, S. (2019). Language, gender, and sexuality. In *Annual Review of Linguistics*, 5, 455-475. https://doi.org/10.1146/annurev-linguistics-052418-094326.

Nordström, P.-E. (1977). Female and infant vocal tracts simulated from male area functions. In *Journal of Phonetics*, 5(1), 81-92. https://doi.org/10.1016/S0095-4470(19)31115-5.

Norkhalid, A.M., Faudzi, M.A., Ghapar, A.A. & Rahim, F.A. (2020). Mobile application: Mobile assistance for visually impaired people - Speech Interface System (SIS). In *Proceeding of the 8th International Conference on Information Technology and Multimedia (ICIMU)*, Selangor, Malaysia, 24-26 August 2020, 329-333. https://doi:10.1109/ICIMU49871.2020.9243450.

Ovalle, A., Goyal, P., Dhamala, J., Jaggers, Z., Chang, K.-W., Galstyan, A., Zemel, R. & Gupta, R. (2023). "I'm fully who I am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, Chicago, USA, 12-15 June 2023, 1246–1266. https://doi.org/10.1145/3593013.3594078.

Peynircioğlu, Z.F., Brent, W., Tatz, J.R. & Wyatt, J. (2017). McGurk effect in gender identification: Vision trumps audition in voice judgments. In *The Journal of General Psychology*, 144(1), 59-68. https://doi.org/10.1080/00221309.2016.1258388.

Pisoni, D.B., Martin, C.S. (1989). Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analyses. In *Alcoholism: Clinical and Experimental Research*, 13(4), 577-587. https://doi.org/10.1111/j.1530-0277.1989.tb00381.x.

Qawaqneh, Z., Abu Mallouh, A. & Barkana, B.D. (2017). Deep neural network framework and transformed MFCCs for speaker's age and gender classification. In *Knowledge based system*, 115, 5-14. https://doi.org/10.1016/j.knosys.2016.10.008.

Sarker, I.H. (2024). LLM potentiality and awareness: A position paper from the perspective of trustworthy and responsible AI modeling. In *Discover Artificial Intelligence*, 4(40). https://doi.org/10.1007/s44163-024-00129-0.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C.A. & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. In *Nature Machine Intelligence*, 4(4), 258–268. https://doi.org/10.1038/s42256-022-00458-8.

Schilt, K., Westbrook, L. (2009). Doing Gender, Doing Heteronormativity: "Gender Normals," Transgender People, and the Social Maintenance of Heterosexuality. In *Gender & Society*, 23(4), 440-464. https://doi.org/10.1177/0891243209340034.

Simpson, A.P. (2001). Dynamic consequences of differences in male and female vocal tract dimensions. In *Journal of the Acoustical Society of America*, 109(5), 2153–2164. https://doi.org/10.1121/1.1356020.

Singh, C., Askari, A., Caruana, R. & Gao, J. (2023). Augmenting interpretable models with large language models during training. In *Nature Communications*, 14, 7913, 1-11. https://doi.org/10.1038/s41467-023-43713-1.

Smith, D.R.R., Walters, T.C. & Patterson, R.D. (2007). Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled. In *Journal of the Acoustical Society of America*, 122(6), 3628–3639. https://doi.org/10.1121/1.2799507.

Sutton, S.J. (2020). Gender ambiguous, not genderless: Designing gender in voice user interfaces (VUIs) with sensitivity. *In Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20)*, Bilbao, Spain, 22-24 July 2020, Article No. 11, 1–8. https://doi.org/10.1145/3405755.3406123.

Tafiadis, D., Toki, E.I., Miller, K.J. & Ziavra, N. (2017). Effects of early smoking habits on young adult female voices in Greece. In *Journal of Voice*, 31(6), 728-732. https://doi.org/10.1016/j.jvoice.2017.03.012.

Teixeira, J.P., Oliveira, C. & Lopes, C. (2013). Vocal acoustic analysis – Jitter, shimmer and HNR parameters. In *Procedia Technology*, 9, 1112-1122. https://doi.org/10.1016/j.protcy.2013.12.124.

West, C., Zimmerman, D.H. (1987). Doing gender. In *Gender & Society*, 1(2), 125-151. https://doi.org/10.1177/0891243287001002002.

Wu, T., Terry, M. & Cai, C.J. (2022). AI Chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '22)*, New Orleans, USA, 29 April-5 May 2022, Article No. 385, 1-22. https://doi.org/10.1145/3491102.3517582.

Yu, W., Tang, C., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z. & Zhang, C. (2024). Connecting speech encoder and large language model for ASR. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2024),* Seoul, Korea, 14-19 April 2024, 12637-12641. https://doi.org/10.1109/ICASSP48485.2024.10445874.

Zhang, Z. (2021). Contribution of laryngeal size to differences between male and female voice production. In *Journal of the Acoustical Society of America*, 150(6), 4511–4521. https://doi.org/10.1121/10.0009033.

DOI: 10.17469/O2112AISV000014

SARA PICCIAU, DOMENICO DE CRISTOFARO, ALESSANDRO VIETTI

Phonological patterns in the predictions of a syllable-based end-to-end ASR system

This paper explores the role of syllables in automatic speech recognition (ASR) systems, focusing on its linguistic implications. Traditionally, ASR has focused on processing speech at the segmental level, but recent research suggests the importance of syllabic processing for robust recognition. We trained a neural ASR model to recognize phonological syllables and conducted a linguistic analysis on its output. Our objective was to observe how various factors, such as syllable token frequency, lexical accent position, syllable type, and parts of speech, influence the neural representation of syllables. To achieve this, we developed a fine-grained linguistic annotation system to overcome the limitations of quantitative metrics like Word Error Rate. By applying Multiple Correspondence Analysis, we identified patterns of association between the neural network's output behavior and linguistic features of speech. Specifically, the study demonstrates that the network compensates for low-frequency syllables through substitution strategies, particularly with absent tokens, which are complex syllables often occurring in proper nouns, common nouns, or numerals. Unstressed high-frequency tokens, such as subordinating conjunctions and determiners, tend toward deletion, while mid-frequency syllables with simple structures (CV) achieve optimal recognition, indicating the network's ability to reflect natural language processing patterns based on token frequency and syllabic complexity. Our findings provide insights into the role of syllables in ASR and contribute to ongoing research in this field.

Keywords: syllable, ASR, pretrained neural models, discrete speech units.

1. *Introduction*

The syllable plays a central role in the spoken word recognition process. In a view that sees the traditional 'segmental' and 'suprasegmental' planes as integrated levels of an interdependent prosodic system, the syllable plays a pivotal role in helping the listener extract lexical and syntactic structure from acoustic information (Beckman, 1996; Hawkins, Smith, 2001). Specifically, the syllable constitutes the linguistic unit in which information useful for speech segmentation, rhythmic patterns, and lexical access is encoded (McQueen, Dilley, 2020).

In automatic speech recognition research, the basic unit of processing has traditionally been the segment. However, the use of the syllable, or phonetic units of similar size, has often been called upon as an alternative signal processing strategy that would apply to a larger time scale than the segment, so as to be more robust in the face of variability across speakers and speech styles (Greenberg 1999; Morgan, Bourlard & Hermansky, 2004; Coro, Massoli, Origlia & Cutugno, 2021). Similarly, in the most

recent research in ASR, the role of the syllable in processing is tested within Transformer-based neural models both through their use as tokens, instead of phonemes and subwords, for recognition of written words, and by trying to trace their role in the internal speech representations of the neural network (Anoop, Ramakrishnan, 2023; Cho, Mohamed, Li, Black & Anumanchipalli, 2024; Vitale, Cutugno, Origlia & Coro, 2024; Shon, Kim, Hsu, Sridhar, Watanabe & Livescu, 2024).

In this paper, we trained a neural model for ASR to process and recognize phonological syllables and assemble them into words. The purpose of this study is to conduct a linguistic analysis on the output of syllabic processing of a neural speech recognition system. A large acoustic model was therefore fine-tuned to map the speech signal onto a phonological transcription that was segmented into syllables and words (see § 2.2).

The primary objective of the linguistic analysis is to observe variation in the final neural representation of syllables based on the effect of syllable token frequency, lexical accent position, syllable type, and parts of speech (to which the syllable belongs). The effects of syllable-based phonetic restructuring of word boundaries, although present and identified in the analysis, will not be the subject of this study.

To identify the different types of "behavior" of the neural network in rescanning syllables and words, a fine-grained linguistic annotation system was developed, a system that was necessary to overcome the opacity of purely quantitative metrics such as Word-Error-Rate or, in our case, Token-Error-Rate. By doing so, it is possible to more precisely "count" prediction types and associate them with linguistic features of speech. This was performed using Multiple Correspondence Analysis, an exploratory multivariate statistic method that revealed patterns of association between the output behavior of the neural network and a set of linguistic factors. Although conducted with different computational methods and on a phonetically annotated dataset, the results of our exploratory analysis are compared with those of Schettino, Vitale & Cutugno (2023).

2. *Methodology*

The workflow for this research project was structured as follows: first, we collected and automatically transcribed the data for the training. Next, we designed a rule-based syllabification algorithm and integrated it in the tokenizer used to fine-tune the ASR model. Following the training, we tested the model and extracted a sample of 300 predicted sentences to conduct a qualitative linguistic analysis. We then categorized discrepancies between predictions and references at both token and word level. Subsequently, we generated a comprehensive database, including detailed information such as token composition and word and token frequency, using a semi-automatic annotation approach. Lastly, we treated deviation events as dependent variables and examined their relation to the features characterizing each token.

2.1 Preparation of the data

2.1.1 Dataset

The data used to fine-tune and test the model has been collected from the Italian dataset of the Common Voice 13.0 corpus (Ardila, Branson, Davis, Henretty, Kohler, Meyer, Morais, Saunders, Tyers & Weber, 2020). It consists of approximately 30 hours of read speech, with each audio file having a maximum length of 7 seconds. The prompts for reading are gathered from Wikipedia and other public domain resources, resulting in the presence of several low frequency words and proper nouns. Furthermore, the dataset is crowd-sourced, meaning that it is robust against inter-speaker and diatopic variation. Aiming to a qualitative analysis centered on phonological aspects, we opted to train and evaluate the model using the phonemic transcriptions of the sentences instead of their original orthographic format. We used the WebMAUS Basic tool provided by the Bavarian Archive for Speech Signal (Schiel, 2015) to transcribe our dataset into X-SAMPA. Before feeding the model with the training data, we ensured the quality of the transcriptions by manually correcting acronyms, ordinal numbers, foreign proper nouns, and other few elements that caused a discrepancy between the original graphemic sentence and its automatic transcription.

2.1.2 Tokenization

While several syllabification algorithms are already available for processing audio signal (Cutugno, Passaro & Petrillo, 2001; Cutugno, Origlia & Schettino, 2018) and text (Iacoponi, Savy, 2011; Bigi, Petroni, 2014), we chose to implement our own tokenization algorithm on textual data. Proceeding this way, we facilitated the integration of the syllabifier in the tokenizer class *Wav2Vec2PhonemeCTCTokenizer* required by the model to perform the training (Transformers, 2023). Moreover, we obtained the output in an optimal format, thus avoiding the need for additional processing of the predictions before conducting the linguistic analysis. Our rule-based syllabification algorithm acts according to the Maximal Onset Principle (Kahn, 1976) and the Sonority Sequencing Principle (Clements, 1990). To streamline data processing and reduce the complexity of managing exceptions, we opted to treat /s/ + plosive clusters and geminates as tautosyllabic onsets, despite acknowledging the lack of phonological grounding in this approach (Marotta, Vanelli, 2021). This computational simplification allowed for a more efficient handling of syllable structures within the model.

(1) aspettare → a.spe.tta.re

In tokenized sentences, blank spaces separate tokens, while pipe symbols separate words. This structure allows us to observe how the model performs word parsing while retaining the information concerning the recognition of individual tokens.

(2) lo studente aspetta → lo | stu dEn te | a spe tta

2.2 Training setup

The initial model used for the experiment is a pretrained Transformer-based model implemented by Microsoft and named WavLM-Large (Chen, Wang, Chen, Wu, Liu, Chen, Li, Kanda, Yoshioka, Xiao, Wu, Zhou, Ren, Qian, Qian, Zeng, Yu & Wei, 2022). This model was trained on 94k hours of English raw audio data. Since its training is based on pure audio signal, its capabilities extend beyond conventional speech recognition tasks, encompassing various other speech-related applications, such as speaker identification. Its proficiency in tasks beyond speech recognition underscores its capacity to extract not only spoken words but also valuable additional knowledge, such as for example speaker-related information. The fine-tuning process for this model included more than just training it with Italian data; it also involved replacing the existing tokenizer. As previously described, the tokenizer relied on a rule-based syllabification algorithm, which encompasses a set of categories, or what can be better defined as syllable-based vocabulary. Following a principle of economy, the vocabulary of tokens consists of the 487 highest frequency syllables and the phonemic inventory of Italian, reaching a total of 517. This approach ensures the transcription of a broad range of possible utterances.

2.3 Linguistic analysis

2.3.1 Categorization of prediction types

The analysis of transcriptions generated by speech recognition systems has been long instrumental in refining the accuracy of these models (Meripo, Nimshi, Konam & Sandeep, 2022; Palmerini, Savy, 2014; Perero-Codosero, Espinoza-Cuadros & Hernández-Gómez, 2022). This process involves examining the transcription to identify areas for improvement within the model. However, transcription error analysis is often conducted on models that transcribe the audio signal into string of characters, or graphemes. As a result, some errors may not be readily traceable to the actual content of the audio. When manually examining the transcriptions generated by our trained ASR system, we noticed discrepancies that could more likely have been attributed to the phonetic organization of connected speech rather than to a deficiency in the training process of the ASR system.

(3) per i brasiliani → pe i | bra zi lja ni

In the context of this study, labeling the deviations found in predictions as errors is misleading. We found it more suitable to categorize them as prediction types instead. In this category the occurrence of unmarked events, that is, when tokens were correctly transcribed by the model, is also included.

We defined prediction types to serve as prediction tags (PT) to categorize the result of the comparison between individual tokens, for both predicted (*Pt*) and reference (*Rt*) units. The classification is based on a system of ordered suffixes that denotes the type of phenomenon at different levels of granularity, as follows:

1. word-level event (if applies);
2. token movement direction (if applies);
3. operation/equality;
4. token or word level.

Word-level events are defined as instances where the misplacement of Pt affects canonical word boundaries. There are three possible situations: the merging of two or more words (*mer*); a single word split in two or more parts (*div*); lastly, tokens that are moved within the boundaries of an adjacent word (*mv*). The direction of the shift in the latter case is marked with the suffixes *forw* and *back*. At the core of each PT is the type of operation, namely substitution (*sub*), deletion (*del*) and insertion (*ins*). If no operation is detected, the equality tag (*eq*) is assigned. Additionally, the suffix syl is added when the event involves a token but not an entire word. On the other hand, if the syllable token is also a word, the suffix word is added to the tag. The detailed list of PT is provided in Appendix 1, while an example of annotation is illustrated in example 4 and Tab. 1.

(4) *Rs*: kom prEn de |i | kwa ttro | sti li | ka nO ni tSi |

Ps: kom prEn de | kwa ttro | sti lli ka | a nO ni tSi

Table 1 - *Prediction tags associated to prediction and reference token*

PT	*Pt*	*Rt*
eq_syl	kom	kom
eq_syl	prEn	prEn
eq_syl	de	de
del_syl_word	-	i
eq_syl	kwa	kwa
eq_syl	ttro	ttro
eq_syl	sti	sti
sub_syl	lli	li
mv_back_eq_syl	ka	ka
ins_syl	a	-
eq_syl	nO	nO
eq_syl	ni	ni
eq_syl	tSi	tSi

2.3.2 Creation of the database

The next step of our work aimed to create a detailed database focused on the single *Pt* and *Rt*, with the corresponding PT serving as a pivotal point. With a total of 300 sentences consisting of approximately 5900 tokens, we opted for a semi-automatic

approach to assign PT; manual revision was anyway necessary to ensure the correct tagging in cases of ambiguity.

To facilitate the computation of the comparison between *Pt* and *Rt*, we deleted the blank spaces serving as segment separators in the tokens that were not to be found in the training vocabulary. To obtain tokens without blanks, reference sentences were syllabified with our algorithm, while the predictions required a different approach as some recognized tokens lacked a potential nucleus for syllable construction. To handle these cases, we developed a specific algorithm to directly assemble predicted tokens instead of relying on syllabification. In instances where the algorithm's output was ambiguous, we manually adjusted the information to ensure that the segments were reconducted to the correct tokens, as showed in example 5:

(5) Algorithm output → 'zvva'

After manual adjustments → (*Pw*) 'zv va' ↔ (*Rw*) 'zve va'

We then proceeded with the design of the master algorithm to achieve three main goals:

1. identify correspondence between predicted and reference words (*Pw* and *Rw,* respectively) in the prediction and reference sentences (*Ps* and *Rs,* respectively) despite the mismatches caused by marked prediction types;
2. track the tokens that were misplaced during the recognition and assigned within non-canonical word boundaries;
3. operate the comparison between *Pt* and *Rt* and assign a *PT* label to each, based on detected prediction types.

The algorithm relies on a dynamic index (*current_i*), which retrieves the to-be-compared words *Pw* and *Rw* from *Ps* and *Rs*. This index changes based on the boolean values of flags that are activated when prediction types are detected during the comparison of the previous pair of *Pw* and *Rw*.

Within each loop, *current_i* is initialized, and a pair of *Ps* and *Rs* is split into its *Pw* and *Rw* words.

Table 2 - *Items compared by the algorithm at a sentence, word, and token level*

Ps	['al ko ni', 'ne i', 'swO i li vi', 'sO no', 'sta ti', 'tra do tti']
Rs	['al ku ni', 'de i', 'swO i', 'li bri', 'sO no', 'sta ti', 'tra do tti']
Pw	['al', 'ko', 'ni']
Rw	['al', 'ku', 'ni']
Pt	'al'
Rt	'al--

While the dynamic index is less than the length of the longer sentence among *Ps* and *Rs*, the pair of words to be compared token-wise is retrieved by indexing the sentences with *current_i*. To establish whether there is a correspondence between *Pw* and *Rw* and, consequently, recognize them as matching and comparable elements, we first consider the length of the sentences and then their similarity at a word level.

This is determined in two ways: firstly, by checking in *Pw* the presence of tokens contained in *Rw*. Secondly, by calculating the Levenshtein distance between the concatenated tokens of the list as a single string and comparing it to a threshold that varies in function of the token's length. Two strings are similar if the score is above 0.75 when token's length is equal or shorter than 3 characters or is above 0.50 if it's higher.

The comparison between *Pw* and *Rw* proceeds as follows.

1. If length(*Pw*) == length(*Rw*), we compare *Pt* and *Rt* pairwise to check for equality or similarity and we assign the tags *eq_syl* or *sub_syl* respectively.
2. If length(*Pw*) != length(*Rw*), the algorithm searches the correct correspondence between the words by retrieving the adjacent *Pw* and *Rw* indexing *Ps* and *Rs* with *current_i*; then, it attempts to classify the mismatch by identifying one of the following events based on a set of defined conditions exemplified below:
   a. word merging: len(*Pw*) > len(*Rw*), deviance(len(*Pw*)) = len(*Rs*[*current_i*+1]), deviance(*Pw*) in *Rs*[*current_i*+*1*], *Pw* in *Rs*[*current_i*+*1*], *Ps*[*current_i*+1] in *Rs*[*current_i*+2];
   b. token insertion: len(*Pw*) > len(*Rw*), deviance(len(*Pw*)) != len(*Rs*[*current_i*+1]), *Pw* != *Rs*[*current_i*+1], *Ps*[*current_i*+1] != *Rs*[*current_i*+2];
   c. token movement backwards: len(*Pw*) > len(*Rw*), len(*Ps*[*current_i*+1]) < len(*Rs*[*current_i*+1]), *Pw* ~ *Rw*;
   d. token movement forwards: len(*Pw*) > len(*Rw*), len(*Ps*[*current_i*+1]) > len(*Rs*[*current_i*+1]), deviance(len(*Pw*)) = (len(*Rs*[*current_i*+1]) - len(*Ps*[*current_i*+1]));
   e. word division: len(*Pw*) < len(*Rw*), *Ps*[*current_i*+1] ~ *Rw*;
   f. token deletion: len(*Pw*) < len(*Rw*), *Pw* ~ *Rw, Pw* != *Rs*[*current_i*+1].
3. Once the event is detected, the related flag is activated, and the analysis proceeds at a token level with the comparison between each *Pt* and *Rt* in *Pw* and *Rw*.
4. The resulting labels are added to the label list. If none of the conditions set is met or ambiguities arise, the label "manual annotation" is appended to notify the need for manual annotation of the involved sentence.
5. Based on the flag activated in the previous loop, the *Pw* and *Rw* to compare next are defined as follows:
   – no flag activated → *Ps*[*current_i*], *Rs*[*current_i*];
   – word division → *Ps*[*current_i*], *Rs*[*current_i* -1];
   – word merging → *Ps*[*current_i* -1], *Rs*[*current_i*];
   – word deletion → *Ps*[*current_i*-1], *Rs*[*current_i*];
   – word insertion → *Ps*[*current_i*], *Rs*[*current_i*-1].

Due to the high quality of the recognition performed by the model, most of the tags consisted in "eq_syl" and were assigned correctly, meaning that the correspondence between the words and tokens was found smoothly.

A challenging case was represented by assimilations, due to the fact that one of the tokens involved was not deleted or moved within the word but merged in another token. This represents a problem for the similarity function, since one

of the two Rt appears to be missing and is tagged with the "del" PT, despite the fact that it became part of an adjacent Pt during the recognition and, therefore, it has been recognized. In such cases, even though it may appear misleading from a linguistic point of view, the "del" PT it is functional for the algorithm: in fact, to ensure the computational integrity of the final data frame, it remains necessary to assign PT equal to the token count.

(6) *Ps*: non | ra ddZun se | ma i | tSi | pr**e** ttSe ttsjo na li

Rs: non | ra ddZun se | ma i | tSi fre | **e** ttSe ttsjo na li

PT: ['eq_syl'], ['eq_syl', 'eq_syl', 'eq_syl'], ['eq_syl', 'eq_syl'], ['mv_back_eq_syl'], ['aggl_sub_syl', '**aggl_del_syl**', 'aggl_eq_syl', 'aggl_eq_syl', 'aggl_eq_syl', 'aggl_eq_syl']

Given the exploratory nature of our research, we conducted a manual revision and correction of the provisional label list, despite recognizing that this approach is not optimal. During this check, we decided to discard few predictions due to the poor quality of the recognition, that made it impossible to assign PT as exemplified below:

(7) *Ps*: il | pa ti ko la | kom pi kka i | me re tSe

Rs: in | par ti ko la re | lo | im pe JJa va | l | i dE a | del | su o | mo nu men to | fu ne bre |

Lastly, we added detailed information in correspondence of each row and obtained a database with the following structure:

- "filename": ID of the file within the Common Voice dataset;
- "sentence": original sentence transcription in graphemic format as extracted from Common Voice;
- "tokenized_P": predicted sentence tokenized into phonological syllables (model's output format);
- "tokenized_R": predicted sentence tokenized into phonological syllables;
- "deviation": the presence or absence of a marked event in correspondence of the predicted token when compared to the reference token;
- "prediction type": type of recognition (see § 2.3.1 and Appendix 1);
- "token_P": predicted token (*Pt*);
- "token_R": reference token (*Rt*);
- "word": reference word in graphemes;
- "number_of_tokens": number of tokens in the reference sentence;
- "number of words": number of words in the reference sentence;
- "freq_tok_P": relative frequency of the predicted token calculated within the whole dataset used to train the model;
- "tok_type_P": syllabic type of the predicted token;
- "freq_tok_R": relative frequency of the reference token calculated within the whole datasets used to train the model;
- "tok_type_R": syllabic type of the reference token;
- "in_vocab_P": presence or absence of the predicted token in the training vocabulary;

- "in_vocab_R": presence or absence of the reference token in the training vocabulary;
- "stress_R": presence or absence of phonological stress in the reference token, extracted with G2P (Reichel, Kisler, 2014; Goslin, Galluzzi & Romani, 2014; Spinelli, Sulpizio & Burani, 2017);
- "pos_tags": part of speech tag of the reference word, assigned with spacy library (Honnibal, Montani, 2017);
- "freq_word_relative": frequency of reference words calculated on the entire dataset;
- "abs_freq_word_R": absolute frequency of reference words calculated on the entire dataset.

3. *Data analysis and results*

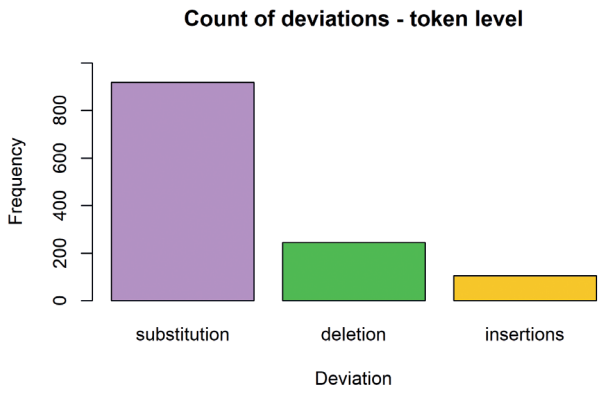
To perform the analysis on our prediction database we first examined the distribution of prediction types. In the second part of the analysis, we considered the patterns of association between prediction types, token frequency, presence in the training vocabulary, and lexical stress using Multiple Correspondence Analysis (MCA).

The quality of the recognition achieved by the syllable-based fine-tuned ASR model is reflected in the number of deviations found in the predictions: only 28 % of the tokens were affected by some marked recognition types, meaning that most tokens were recognized correctly and, therefore, "eq_syl" is the most frequent category (72 %).

The detailed distribution of marked prediction types is displayed in the following figures. Our system of structured labeling allows us to focus separately on the phenomena that are traceable to the token level and on those affecting the sentence structure through word boundary misplacement.

If we consider the high-level recognition labels that mark the type of operation at a token level (Fig. 1), it emerges that substitution is the most frequent operation, followed by deletion and insertion. This indicates that most tokens that were not recognized precisely are still to be found in the transcription hypothesized by the model. On the other hand, the deletion and insertion of tokens (including those which also correspond to entire words, like prepositions, determiners or some auxiliary verb forms) represent a more consistent discrepancy in the recognition. It should be noted that the use of automatically generated phonological transcriptions as a reference causes an increment in the number of substitutions due to the speech variability that characterizes the corpus.

Figure 1 - *Count of deviations at a token level*



Shifting our focus to the labels assigned to prediction types impacting canonical word boundaries, Fig. 2 shows the distribution of the operation/equality tags within the three categories. It can be observed that merging is the most frequent process with 401 tokens belonging to the words involved. Divided words follow with 206 occurrences, while the movement of single tokens to adjacent words is less frequent with a total of 48 instances. Additionally, the movement label applies to single tokens, unlike the other categories that consider all the tokens within the affected words. Examining the distribution of equality, it is evident that the tokens involved in merged and divided words were mostly recognized correctly, while substitution is the second most performed operation. It is interesting to notice that token deletion happens more often in merged words, while the amount of token insertion is proportionally higher in words that undergo division. Regarding tokens affected by movement, the distribution of equal and substituted tokens is nearly identical. Deletions do not apply since moved tokens cannot be missing from the prediction due to their intrinsic nature. As a result, insertions are not relevant either, because inserted tokens were not present in the reference and cannot be considered moved.

Figure 2 - *Count of deviations at a word level*

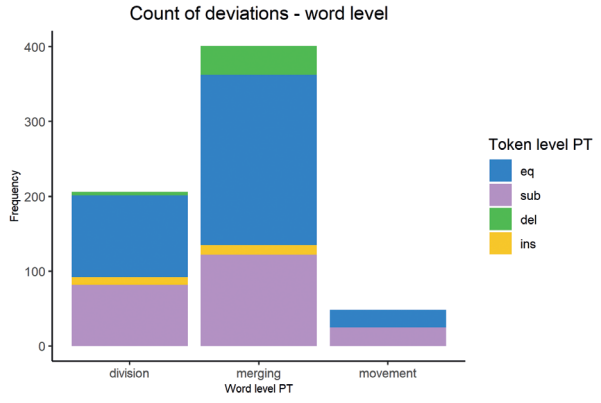
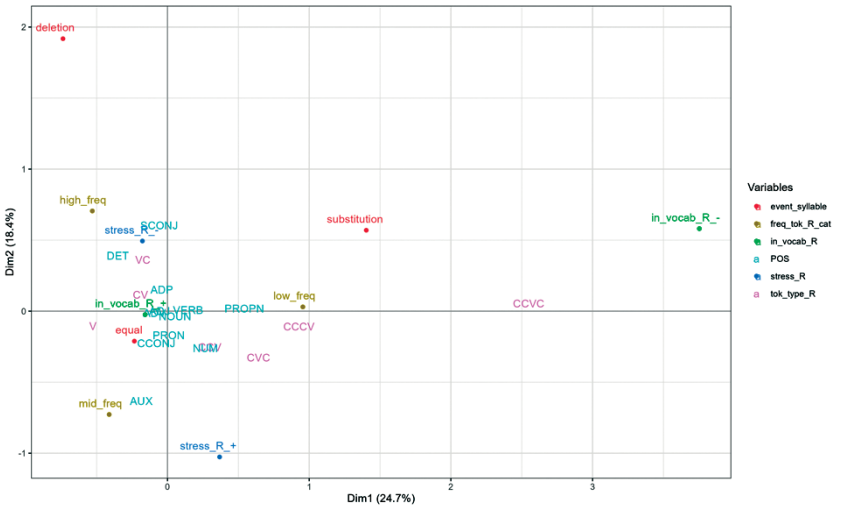


Fig. 3 shows the results of Multiple Correspondence Analysis (MCA), a multivariate statistical technique performed using *FactoMinerR* R package. As mentioned previously, in the analysis we observed the emergence of patterns of association between prediction types (event_syllable), token relative frequency (freq_tok_R_cat), token presence in the training vocabulary (in_vocab_R) and lexical stress (stress_R). In order to facilitate the analysis, the relative frequency of reference tokens was discretized in three categories using quantiles. Part of speech tag (POS) and syllable type (tok_type_R) were not included in the computation of the low-dimensional space but are projected onto it a posteriori, as supplementary variables, to contribute to the linguistic interpretation of the analysis. Since our focus is on reference tokens and their attributes, insertion (which is the least frequent operation) is not relevant in this context. The most complex syllable types (e.g. CCVCC or CCCVC as in 'edward' and 'mainstream') were also excluded from the analysis given their low frequency (less than 10 observations) and since, in most cases, they correspond to proper nouns in English.

MCA is a dimensionality reduction technique for categorical variables based on variance maximization, as in Principal Component Analysis (PCA). The dimensions on which observations are projected are those that represent the greatest variance in the original data (see Fig. 3: Dim1 = 24.7 %, Dim2 = 18.4 %), hence their meaning is derived from the actual distribution of variable values on the new plane. On the left side of the plot, we can observe the distribution of the tokens learned by the model during the training phase. It is interesting to notice in the top section how unstressed high-frequency tokens (greater than 2.23 %), consisting mostly in subordinating conjunctions and determiners, are related to deletion. In the bottom-left section we find the mid-frequency items (range: 0.5 % - 2.23 %). These tokens are characterized by a simple syllabic structure (CV) and tend to be recognized correctly. The tokens that were absent in the training vocabulary are on the right side of the MCA chart. Since these tokens are characterized by a more complex syllabic structure, they are naturally less frequent in the dataset. To compensate for this lack of information, the model adopts substitution as a strategy. The tokens are part (or, in some cases, represent) mostly proper and common nouns, as well as numerals.

Figure 3 - *Multiple Correspondence Analysis (MCA)*



4. *Discussion and future work*

Using a neural network to recognize syllables seems to show how the phonetic structure of the speech signal changes. In our analysis, the phonetic elements that undergo the deletion process are elements that usually disappear in connected and spontaneous speech, which was not expected since the training and test data consisted of read speech. The reconstructed words mainly involve deletion phenomena and known regularities in connected speech, such as element frequencies and certain parts of speech (POS) (Bybee, 2006). An innovative aspect of our study is the role of frequency, which shows up in three ways. For low-frequency syllables, the network compensates for their absence by substituting them. At medium frequency, the network works best, and it optimizes its recognition processing using high frequency syllables, as shown in other studies on spontaneous speech (Schettino, Vitale & Cutugno, 2023). This reveals the network ability to adapt based on how often syllables occur, demonstrating its potential to mimic natural language patterns.

To support linguistic analysis, we have developed a new annotation schema that could lead to a clearer metric than the traditional Word Error Rate (WER) from a linguistic standpoint. This annotation schema can be further developed into a metric. However, the algorithm currently lacks accuracy in some cases. Therefore, we aim to improve it by generalizing it to account for all instances of predictions.

One major limitation of our study is that we did not perform a phonetic analysis of the syllable directly from the acoustic signal, so its actual phonetic structure as well as the resyllabification process were not observed on the phonetic level. Additionally, the number of syllables used in our vocabulary (tokenizer) was limited, thus restricting the generalizability of the results.

Future research must therefore expand primarily in two areas: (1) the analysis of the role of token frequency, as information available to the neural model, in relation to the variables observed here, through the construction of balanced and representative vocabulary; (2) providing a phonetic representation of the syllable grounded in acoustic information that would allow the neural model's behavior to be evaluated not as a "deviation" from phonological expectations, but as a faithful processing of the acoustic characteristics of the speech signal.

In conclusion, we believe that our study shows, albeit in a preliminary way, how complex computational tools such as modern neural networks can be used by linguists as models to simulate and test linguistically relevant hypotheses.

References

Anoop, C.S., Ramakrishnan, A.G. (2023). Suitability of syllable-based modeling units for end-to-end speech recognition in Sanskrit and other Indian languages. In *Expert Systems with Applications,* 220, 119722. https://doi.org/10.1016/j.eswa.2023.119722

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M. & Weber, G. (2020). *Common Voice: A Massively-Multilingual Speech Corpus*. arXiv. https://doi.org/10.48550/arXiv.1912.06670

Beckman, M.E. (1996). The Parsing of Prosody. In *Language and Cognitive Processes*, 11(1–2), 17–68. https://doi.org/10.1080/016909696387213

Bigi, B., Petrone, C. (2014). A generic tool for the automatic syllabification of Italian. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014.* Pisa, Pisa University Press, 73-77.

Bybee, J. (2006). From Usage to Grammar: The Mind's Response to Repetition. In *Language*, 82, 711-733. 10.1353/lan.2006.0186

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Zeng, M., Yu, X. & Wei, F. (2022). WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. In *IEEE Journal of Selected Topics in Signal Processing*, 16, 1-14. 10.1109/JSTSP.2022.3188113

Cho, C.J., Mohamed, A., Li, S.-W., Black, A.W. & Anumanchipalli, G.K. (2024). SD-HuBERT: Sentence-Level Self-Distillation Induces Syllabic Organization in HuBERT. *arXiv*. Retrieved from http://arxiv.org/abs/2310.10803

Clements, G.N. (1990). The role of the sonority cycle in core syllabification. In Kingston, J. & Beckman, M.E. (Eds.), *Papers in Laboratory Phonology: Volume 1: Between the Grammar and Physics of Speech.* 1. Cambridge University Press, 283-333. https://doi.org/10.1017/CBO9780511627736.017

Coro, G., Massoli, F.V., Origlia, A. & Cutugno, F. (2021). Psycho-acoustics inspired automatic speech recognition. In *Computers & Electrical Engineering*, 93, 107238. https://doi.org/10.1016/j.compeleceng.2021.107238

Cutugno, F., Origlia, A. & Schettino, V. (2018). Syllable structure, automatic syllabification and reduction phenomena. In Cangemi, F., Clayards, M., Niebuhr, O.,

Schuppler, B. & Zellers, M. (Eds.), *Rethinking Reduction: Interdisciplinary Perspectives on Conditions, Mechanisms, and Domains for Phonetic Variation*. Berlin/Boston: De Gruyter Mouton, 205-242. https://doi.org/10.1515/9783110524178-007

Cutugno, F., Passaro, G. & Petrillo, M. (2001). Sillabificazione fonologica e sillabificazione fonetica. In Albano Leoni, F., Sornicola, R., Stenta Krosbakken, E. & Stromboli, C. (Eds.), *Dati empirici e teorie linguistiche, Atti del XXXIII, Congresso della Società di Linguistica Italiana*. Roma: Bulzoni, 205-232.

Goslin J., Galluzzi C. & Romani, C. (2024). PhonItalia: a phonological lexicon for Italian. *Behav Res Methods*, 46(3), 872-86. doi: 10.3758/s13428-013-0400-8. PMID: 24092524.

Greenberg, S. (1999). Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. In *Speech Communication*, 29, 159-176.

Hawkins, S., Smith, R. (2001). Polysp: A polysystemic, phonetically-rich approach to speech understanding. In *Italian Journal of Linguistics*, 13, 99-189.

Honnibal, M., Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. [Python Package] Version 3.8.1. https://spacy.io/

Iacoponi, L., Savy, R. (2011). Sylli: Automatic Phonological Syllabification for Italian. [Computer Program] Version 0.9.8. https://sylli.sourceforge.net/

Kahn, D. (1976). Syllable-based generalizations in English phonology. Ph.D. dissertation, Massachusetts Institute of Technology. https://dspace.mit.edu/handle/1721.1/16397

Marotta, G., Vanelli, L. (2021). *Fonologia e prosodia dell'italiano*. Roma: Carocci Editore.

Mcqueen, J.M., Dilley, L. (2020). Prosody and Spoken-Word Recognition. In Gussenhoven, C., Chen, A. (Eds.), *The Oxford Handbook of Language Prosody*. Oxford University Press, 508-521. https://doi.org/10.1093/oxfordhb/9780198832232.013.33

Meripo, N., Konam, S. (2022). ASR Error Detection via Audio-Transcript entailment. 10.48550/arXiv.2207.10849

Morgan, N., Bourlard, H. & Hermansky, H. (2004). Automatic Speech Recognition: An Auditory Perspective. In Greenberg, S., Ainsworth, W.A., Popper, A.N., Fay, R.R. (Eds.), *Speech processing in the auditory system*. New York: Springer, 309-338.

Palmerini, M., Savy, R. (2014). Gli errori di un sistema di riconoscimento automatico del parlato: analisi linguistica e primi risultati di una ricerca interdisciplinare. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014*. Pisa: Pisa University Press, 281-285.

Perero-Codosero, J.M., Espinoza-Cuadros, F.M. & Hernández-Gómez, L.A. (2022). A Comparison of Hybrid and End-to-End ASR Systems for the IberSpeech-RTVE 2020 Speech-to-Text Transcription Challenge. In *Applied Sciences*, 12*(*2), 903. https://doi.org/10.3390/app12020903

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

Sasindran, Z., Yelchuri, H. & Prabhakar, T.V. (2024) SeMaScore: A new evaluation metric for automatic speech recognition tasks, arXiv.org. Available at: https://arxiv.org/abs/2401.07506

Schettino, L., Vitale, V.N. & Cutugno, F. (2023). Syllabic Reduction in Italian Connected Speech: Towards the Integration of Linguistic and Computational Approaches. In *Proceedings of the 20th International Congress of Phonetic Sciences*, 2039-2043. Prague: Guarant International.

Schiel, F. (2015). A Statistical Model for Predicting Pronunciation. In *Proceedings of the ICPhS 2015*, Glasgow, UK.

Shon, S., Kim, K., Hsu, Y.-T., Sridhar, P., Watanabe, S., & Livescu, K. (2024). DiscreteSLU: A Large Language Model with Self-Supervised Discrete Speech Units for Spoken Language Understanding. ArXiv. doi: 10.48550/arXiv.2406.09345

Spinelli, G., Sulpizio, S. & Burani, C. (2017). Q2Stress: A database for multiple cues to stress assignment in Italian. In *Behavior Research Methods*, 49(6), 2113-2126. doi: 10.3758/s13428-016-0845-7. PMID: 28039679.

Transformers. (2023). Tokenization Wav2Vec2 Phoneme (v4.35.2) [Computer software]. Hugging Face. https://github.com/huggingface/transformers/blob/v4.35.2/src/transformers/models/wav2vec2_phoneme/tokenization_wav2vec2_phoneme.py#L94

Vitale, V.N., Cutugno, F., Origlia, A. & Coro, G. (2024). Exploring emergent syllables in end-to-end automatic speech recognizers through model explainability technique. In *Neural Computing and Applications*, 36(12), 6875-6901. https://doi.org/10.1007/s00521-024-09435-1

Appendix

Appendix 1 - *Prediction tags*

Label	*Prediction*	*Reference*
eq_syl	do po | al ku ni |	do po | al ku ni |
sub_syl	mO do | **ve** tSo |	mO do | de tSo |
ins_syl	i | lo ro | a bi ta **tta** |	i | lo ro | a bi t a t |
del_syl	kom ple ta men te | sO **-** |	kom ple ta men te | so **lo** |
sub_syl_word	**kon** | E | di ven ta to |	non | E | di ven ta to |
ins_syl_word	te | **i** |	ti |
del_syl_word	so pra ttu tto | **-** | ma ssa ka tSe ts |	so pra ttu tto | **in** | ma ssa tS u se tts |
mv_eq_forw_syl	o ri dZi | **ni** mi ti ke |	o ri dZi ni | mi ti ke |
mv_sub_forw_syl	E stre | **ro** u ma no |	E sse re | u ma no |
mv_eq_back_syl	da **ve** | tra te |	da | ve tra te |
mv_sub_back_syl	tu tta vi a **no** |	tu tta vi a | non |
div_eq_syl	**a** | pu ddZa | da	a ppo ddZa ta |
div_sub_syl	a | pu ddZa | **da**	a ppo ddZa ta |
div_ins_syl	fra | zi **i** |	fra zi |
mer_eq_syl	**kwa ttro** po sti |	kwa ttro | po sti |
mer_sub_syl	**sE** | la u re a to |	si | E | la u re a to |
mer_ins_syl	pu kwe stE **ro** no | kO lle	kwe s t E r mo | ko lle |
mer_del_syl	fi nO **-** tto |	fi no | ad | O tto |

DOI: 10.17469/O2112AISV000015

FRANCESCA FONTANELLA, TOMMASO BALSEMIN, BARBARA GILI FIVELA, PASCAL PERRIER, CHRISTOPHE SAVARIAUX, GRAZIANO TISATO, CLAUDIO ZMARICH

L'organizzazione temporale delle geminate dell'italiano: uno studio di modellizzazione tramite la Fonologia Articolatoria

Modelling the temporal organization of geminates in Italian: A study in Articulatory Phonology

Italian geminates are still a matter of debate about their phonological status and temporal organization with adjacent vowels. Although kinematic analyses can clarify what strategies are used to achieve length contrast, articulatory studies on Italian are scarce. The present work adopts the theoretical and methodological approach of *Articulatory Phonology* (Browman, Goldstein, 1989) to investigate the articulatory mechanisms involved in the gestural organization of Italian geminates, relying on the data of temporal intervals of consonantal versus vowel gestures, interpreted according to Gafos' (2002) formalization.
The main objective is to test two alternative models of gestural synchronization, the "*Combined Vowel and Consonant*", proposed by *Articulatory Phonology*, and the "*Vowel-to-Vowel*", proposed by Öhman (1967). For this purpose, a "virtual geminate" was constructed by doubling the articulatory closure interval derived from the singleton and superimposing it on some contextual vowel *landmarks* (Gafos, 2002). The results of the different synchronizations of the virtual geminate according to the two models are then be compared with those of the real geminate.

Keywords: bilabial geminates, Italian, Articulatory Phonology, synchronization VC(C)V.

1. *Introduzione*

Nel panorama internazionale, le geminate sono ancora oggi oggetto di discussione in ambito fonetico-fonologico e l'italiano è spesso citato come esempio nella letteratura internazionale per la pervasività della geminazione che lo caratterizza.

Una prima questione concerne lo statuto fonologico e vede contrapporsi due antitetiche ipotesi esplicative, una bisegmentale e eterosillabica (che Loporcaro, 1990, riconduce a Trubetzkoj, 1939) e una monosegmentale e tautosillabica (che Loporcaro, 1990, riconduce a Martinet, 1975). La prima proposta descrive le geminate come due segmenti identici posti l'uno in coda della sillaba precedente e l'altro in attacco della sillaba successiva. La seconda ipotesi vede le geminate come un unico segmento a livello fonologico, avente tratto [+lungo] e sillabificato interamente nell'*onset* della seconda sillaba.

Nel tentativo di supportare l'una o l'altra ipotesi, molti autori si sono serviti di "fatti" interni alla lingua, tra i quali l'osservazione delle regolarità distribuzionali e la derivazione dal latino. Tuttavia, i "fatti" di natura "esterna", relativi all'analisi acustica o cinematica, quantitativi e oggettivi, sono stati poco indagati, tanto più in italiano. In particolar modo, le analisi cinematiche potrebbero contribuire a spiegare quale delle quattro strategie per la realizzazione del contrasto di lunghezza descritte da Cho (2002) è quella realizzata preferenzialmente dal parlante italiano, e se ci sono varianti individuali o regionali. Collegata a ciò vi è poi la questione inerente al tipo di sincronizzazione con le vocali adiacenti.

L'analisi acustica ha aiutato a fare luce sui correlati della geminazione, il principale dei quali è la maggiore durata della fase di tenuta articolatoria. L'italiano non fa eccezione: diversi studi (Esposito, Di Benedetto, 1999; Mattei, Di Benedetto, 2001; Stevens, Hajek, 2004; Payne, 2005; Zmarich, Gili Fivela, 2005; Zmarich et al., 2006; Zmarich et al., 2007; Dipino, Celata, 2018; Celata, Meluzzi & Bertini, 2022) hanno mostrato che le geminate italiane sono lunghe circa il doppio della loro controparte scempia.

Invece, l'accorciamento della vocale precedente (V_1) il segmento raddoppiato, seppur noto (Esposito, Di Benedetto, 1999; Payne, 2005; Zmarich, Gili Fivela, 2005; Zmarich et al., 2006; Dipino, Celata, 2018; Mairano, De Iacovo, 2019; Di Benedetto, De Nardis, 2021; Celata, Meluzzi & Bertini, 2022), risulta più debole, in quanto influenzato da fattori quali l'accento (Bertinetto, 1981; Mairano, De Iacovo, 2019).

Circa l'analisi cinematica, lo studio di riferimento sul contrasto scempia-geminata è quello di Smith (1995), che si propone di verificare le ipotesi della Fonologia Articolatoria (Browman, Goldstein, 1989). Dato che il nostro lavoro utilizza lo stesso approccio teorico-metodologico, prima di descrivere i vari studi articolatori, si riporteranno i suoi principali assunti.

2. *Il contrasto scempia-geminata nella Fonologia Articolatoria*

2.1 Lo status delle geminate e la struttura della sillaba

La Fonologia Articolatoria (FA) è una teoria nata attorno agli anni '80, con i lavori pionieristici di E. Saltzman, P. Rubin, C. Browman, L. Goldstein (Browman, Goldstein, 1989).

Secondo FA, gli atomi di base della pianificazione fonologica e della produzione fonetica sono i gesti, azioni primitive degli articolatori presenti nel tratto vocale. Essi fungono sia da unità di contrasto che da unità d'azione (Browman, Goldstein, 1989), superando la tradizionale dicotomia tra fonologia e fonetica. Parte integrante di FA è il *Task Dynamic System* (TADA), esteso dal movimento degli arti, per il quale era stato originariamente concepito, a quello degli articolatori del parlato (Saltzman, 1986). TADA fornisce un meccanismo matematico per modellare il movimento dell'articolatore nello spazio e nel tempo, specificando come le coordinate spaziali del compito per realizzare un determinato bersaglio articolatorio (per es. la chiusura

labiale) vengano mappate sulle coordinate del tratto vocale e dell'articolatore, anche detto effettore (Pouplier, 2020). Così, ogni gesto è specificato da un aspetto spaziale intrinseco (i target delle variabili di tratto), ma anche da caratteristiche temporali intrinseche (es. *stiffness*, vedi oltre). Tali caratteristiche fanno sì che i gesti possano sovrapporsi (Fowler, 1980; Browman, Goldstein, 1989), ancorandosi in punti preferenziali della curva cinematica, denominati "punti di riferimento gestuale" (*landmark*), che sono: *onset* (o), l'inizio del gesto; *target* (t), l'istante in cui viene raggiunto il target articolatorio; *c-center* (cc), punto centrale tra raggiungimento del target e allontanamento dal target, ovvero il centro della fase di tenuta gestuale o "*plateau*"; *release* (r), momento in cui inizia l'allontanamento; *release offset* (roff), la fine del controllo attivo del gesto da parte dell'articolatore. Nel presente studio si farà riferimento al metodo di allineamento dei *landmark* secondo Gafos (2002).

Altro punto cardine di FA è la coordinazione inter-gestuale a livello della struttura sillabica, in quanto consonanti iniziali e finali si comportano differentemente. Quelle in attacco mostrano il cosiddetto effetto *c-center* (Browman, Goldstein, 1988): in inglese, ad esempio, l'aggiunta di ulteriori consonanti a inizio sillaba provoca un riaggiustamento della sincronizzazione dell'*onset*, tale che i gesti consonantici in attacco si coordinano con la vocale come fossero un blocco unitario. Lo stesso effetto non si manifesta nelle consonanti in coda: la prima consonante in coda ha una relazione così stabile con la vocale precedente, da non essere alterata dall'ipotetica aggiunta di ulteriori consonanti. La diversità di comportamento di *onset* e coda sillabici rappresenterebbe la manifestazione delle due fondamentali relazioni di fase (*phasing*) tra consonanti e vocali: "in-fase" e in "anti-fase" (Hall, 2010). La fisica dei sistemi biologici (Turvey, 1990) assume che il movimento di un gesto sia rappresentabile tramite il concetto di ciclo, che è assimilato a un angolo giro (360°). Nella relazione "in-fase" (o "fase zero", con angolo di 0°) due gesti iniziano nello stesso istante. Nella sincronizzazione in "anti-fase" (cfr. *out-of-phase*), invece, il secondo gesto inizia quando il primo è a metà del tragitto, dando luogo a una temporizzazione spostata di 180° (come tamburellare le dita in alternanza). Circa la sincronizzazione consonante-vocale, si assume che la consonante in *onset* sia in una relazione di "fase" con la vocale, mentre la consonante in coda sia in una relazione di "anti-fase" con la vocale. Le consonanti, invece, affinché conservino la loro udibilità, devono coordinarsi tra loro in "anti-fase" (Browman, Goldstein, 1990).

Riassumendo, le relazioni di coordinazione sono tre: 1) C-V, relazione tra ciascun *onset* consonantico e la vocale successiva e che prevede che il *c-center* del gesto consonantico si sincronizzi con l'*onset* del gesto vocalico; 2) C-C, coordinazione tra le consonanti, sincronizzate in "anti-fase" tra loro (Browman, Goldstein, 2000)[1];

[1] Autori come Gafos, tuttavia, assumono che la coordinazione C-C sia linguo-specifica e non esista una relazione di default in tal senso (Gafos, 2002). In generale, le lingue presentano due modalità di transizione tra consonanti: le cosiddette "*close transition*" e "*open transition*", le quali si rifletterebbero in una differente coordinazione. Nella "*close transition*" la costrizione articolatoria della seconda consonante si forma prima che la fase di tenuta per la prima consonante sia rilasciata (Catford, 1988; citato in Gafos, 2002). Nella "*open transition*", invece, tra le consonanti facenti parte del nesso, emerge un vo-

3) V-C, relazione in anti-fase tra la vocale e la prima consonante post-vocalica occupante la coda sillabica. Le ipotetiche ulteriori consonanti poste in coda non presentano alcun tipo di coordinazione con la vocale, ma solo con la consonante che la precede (Browman, Goldstein, 1988).

Tuttavia, alcuni studi successivi (cfr. Marin, 2013, per il rumeno; Pastätter, Pouplier, 2015, per il polacco) evidenziano come la sincronizzazione C-V non è spiegabile in modo uniforme mediante la teoria di organizzazione della sillaba. Infatti, la coordinazione del *c-center* con la vocale dipende dalla resistenza alla coarticolazione, che varia in base allo specifico segmento consonantico presente in attacco. Tale peculiarità non è specificata da Browman, Goldstein (2000).

Fatte tali premesse, due gesti che a parità di ampiezza presentano una diversa durata, possono essere regolati da meccanismi che agiscono sulle caratteristiche intragestuali (es. *stiffness*, o costante di rigidità) o sulle relazioni intergestuali di *phasing*.

2.2 Indagini cinematiche relative al contrasto scempia geminata

Lo studio di Smith (1995) è considerato il capostipite delle successive indagini sulle strategie cinematiche implicate nella geminazione italiana. La ricercatrice confronta la geminazione dell'italiano con quella del giapponese, analizzando parole con struttura mVC(C)V (con V= [a,i] e C=[p, t, m, n]), anche se poi l'analisi si limita alle parole con struttura miC(C)a. I risultati mostrano che l'italiano si conforma al modello "*Vowel-to-Vowel timing*" (*V-to-V*) di Öhman (1967; citato in Smith, 1995), mentre il giapponese obbedisce al "*combined vowel and consonant timing*" (*V-to-C*) di Browman, Goldstein (1990). Il modello *V-to-V*, che Smith (1995) associa a lingue *stress-timed* (come l'inglese) e *syllable-timed* (come l'italiano), assume che l'articolazione delle sillabe si basi sulla produzione delle vocali interpretate come successione di cicli articolatori. Le consonanti, invece, sono articolate separatamente come modulazione del ciclo articolatorio delle vocali. Dunque, le consonanti sono irrilevanti per la sequenziazione vocalica che, quindi, non dovrebbe essere influenzata né dalla durata consonantica né da loro altre caratteristiche (Smith, 1995). Contrariamente, nel modello *V-to-C*, che Smith attribuisce alle lingue *mora-timed* (come il giapponese), vocali e consonanti sono reciprocamente coordinate. Pertanto, anche le consonanti possono influenzare la temporizzazione vocalica, cosicché una maggiore durata della consonante mediana corrisponderebbe a un allungamento da vocale a vocale (Smith, 1995). Nella geminazione italiana il ciclo vocalico sarebbe mantenuto costante grazie a un'anticipazione del gesto di chiusura consonantica a livello di V_1 e un ritardo nel gesto di rilascio della stessa consonante a livello di V_2, dunque grazie a una maggiore sovrapposizione. Inoltre, Smith (1995) evidenzia come le geminate siano caratterizzate da una maggiore durata del gesto di chiusura e della fase di tenuta articolatoria: la prima dovuta a una minore rigidità (*stiffness*); la seconda conseguente a un periodo di attivazione più lungo. Infine,

coide di transizione, il quale permette un periodo di transizione tra i due gesti utile affinché entrambi siano udibili da parte dell'ascoltatore.

Smith (1995) rileva un solo picco di ampiezza nelle geminate, corrispondente al target, come nelle scempie, sottolineando con questo la problematicità di un modello bisegmentale/bigestuale.

Successivamente, Zmarich, Gili Fivela (2005) hanno registrato parole e pseudo-parole con consonanti alternativamente scempie e geminate ([f, v, p, b, m]) e struttura CaCCa, e stimoli contenenti nessi consonantici (/mf/, /mb/ e /mp/), prodotti da due parlanti originari dell'Italia settentrionale; i risultati cinematici, per i target prodotti a velocità normale e focalizzazione ampia, mostrano differenze significative tra scempie e geminate: le geminate somigliano ai nessi consonantici, soprattutto a livello del gesto di apertura del labbro inferiore, il quale presenta durata e ampiezza maggiori, una durata maggiore per il raggiungimento del picco di velocità (*time-to-peak*) e minore rigidità (*stiffness*). Tale similarità ha condotto Zmarich, Gili Fivela (2005) a ipotizzare l'eterosillabicità delle geminate. La differenza maggiore riscontrata tra scempie e geminate, soprattutto a livello del gesto di apertura labiale verso la seconda vocale, ha confutato l'ipotesi di Smith (1995), secondo la quale l'occlusione labiale è rappresentata da un gesto in cui le fasi di chiusura e apertura sono simili. Gli studiosi, inoltre, confermano parzialmente l'appartenenza dell'italiano al modello *V-to-V* di Öhman (1967), suggerendo invece come la sincronizzazione temporale sia soggetta a variazione interindividuale.

Lo stesso anno, Gili Fivela, Zmarich (2005) hanno analizzato gli stimoli di Zmarich, Gili Fivela (2005) prodotti a velocità d'eloquio rapida e nelle condizioni prosodiche "non in focus" e in "focus contrastivo". I dati con focalizzazione contrastiva confermano quanto evidenziato da Zmarich, Gili Fivela (2005): le geminate assomigliano ai nessi consonantici eterosillabici, soprattutto circa le caratteristiche del gesto di apertura del labbro inferiore. Con l'aumento della velocità d'eloquio, le differenze cinematiche, a differenza di quelle acustiche, vengono preservate e ciò conduce a ipotizzare che i correlati cinematici della geminazione siano più stabili.

Zmarich et al. (2006) hanno studiato, mediante l'*Electromagnetic Midsagittal Articulometer* 2D (EMA; Hoole, Nguyen, 1999), parole e pseudo-parole di struttura ˈCVC(C)V contrastanti per lunghezza (con V=[a] e C=occlusive sonore dentali, velari e laterali alveolari), e parole e non parole contenenti gruppi consonantici (del tipo /ld/ o /dl/), prodotti da due locutrici di italiano settentrionale. Le analisi, parimenti a quanto rilevato in Zmarich, Gili Fivela (2005), mostrano che le geminate sono modellate da due gesti, chiusura e apertura, e che somigliano maggiormente ai gruppi consonantici eterosillabici in termini di maggiore durata, maggiore ampiezza, minore rigidità (*stiffness*), maggiore tempo di raggiungimento del picco di velocità (*time-to-peak*), cioè minore accelerazione, maggiore velocità nel gesto di chiusura e minore velocità del gesto di apertura. La somiglianza è maggiore a livello del gesto di apertura rispetto a quello di chiusura, evidenziando la rilevanza di tale gesto nel contrasto di lunghezza consonantica.

Un'indagine elettropalatografica (EPG) di Payne (2006) su coppie minime scempia-geminata con struttura "aC(C)a" prodotte da una locutrice romana, ha evidenziato una conformazione più piatta della superficie della lingua nelle

geminate e più concava nelle scempie. Tali evidenze supportano le ipotesi di Payne (2005), secondo cui il contrasto di lunghezza in italiano non è dettato da una mera differenza di durata temporale della fase di occlusione, ma anche da una differenza in termini spaziali.

Zmarich et al. (2007) e Gili Fivela et al. (2007) hanno analizzato gli stimoli di Zmarich, Gili Fivela (2005) per la consonante /m/ in contesto /i/-/a/. Alla velocità di eloquio normale, il ciclo vocalico varia significativamente in base alla lunghezza consonantica, diversamente dal modello *V-to-V* e similmente a quello *V-to-C*, pur mantenendo l'anticipazione del gesto consonantico, coerente col modello *V-to-V*, trovata anche da Smith (1992). Gli studiosi concludono che le geminate italiane sembrano conformarsi a un "modello ibrido", composto nella prima parte dal *V-to-V* e nella seconda parte dal *V-to-C*.

Uno studio con EMA 2D di Zmarich et al. (2011) sulle strategie articolatorie per la realizzazione del contrasto di lunghezza consonantica nella coppia minima "*mima/mimma*", confuta l'ipotesi del modello *V-to-V* sia alla velocità di elocuzione normale che a quella rapida: il ciclo vocalico varia significativamente tra scempie e geminate. Tuttavia, la predizione di Smith (1995) circa l'anticipazione del gesto consonantico della geminata nella vocale precedente riceve una nuova conferma. Alla velocità d'eloquio rapida le differenze tra scempie e geminate si riducono.

Indagini relative alle geminate sono state svolte anche utilizzando altri metodi e strumenti di misurazione. Hagedorn, Proctor & Goldstein (2011), mediante rtMRI, hanno indagato coppie minime scempia-geminata, costituite da occlusive, affricate e sonoranti, prodotte da un locutore di Roma. I risultati mostrano una differenza significativa nel picco di velocità massima (*peak intensity*) nelle geminate occlusive bilabiali, coronali laterali-nasali-occlusive; nelle occlusive coronali, tuttavia, la differenza non è significativa. Circa la cinematica gestuale, l'intera durata della consonante (dall'inizio al rilascio) è maggiore nelle geminate.

Dipino, Celata (2018) hanno effettuato un'ispezione dei profili sagittali della lingua per le occlusive alveolari, registrando cinque locutori della varietà di toscano tramite UTI (cfr. Stone, 2005). Coerentemente con Zmarich et al. (2007), le geminate sono prodotte con un gesto linguale più ampio.

Celata, Meluzzi & Bertini (2022) hanno confrontato per mezzo dell'UTI i gruppi consonantici tautosillabici ed eterosillabici italiani per comprendere se le caratteristiche temporali e cinematiche proprie del contrasto scempia-geminata valevano anche per il contrasto tra due tipi di gruppi consonantici, partendo dal presupposto che in entrambi i casi ci fosse un'opposizione tra sillaba aperta e chiusa. I due effetti cinematici della geminazione, ovvero l'aumento della velocità di chiusura e l'aumento dell'anticipazione temporale dell'inizio del gesto di chiusura nella vocale precedente, sono stati confermati in tutti i tipi di consonanti e in tutti i contesti all'interno del contrasto scempia-geminata, ma non in quello inerente alle due tipologie di gruppi consonantici. Anche il correlato acustico corrispondente all'accorciamento della vocale precedente la consonante in coda sillabica si è mostrato meno pregnante nel contrasto tra le due tipologie di gruppi consonantici.

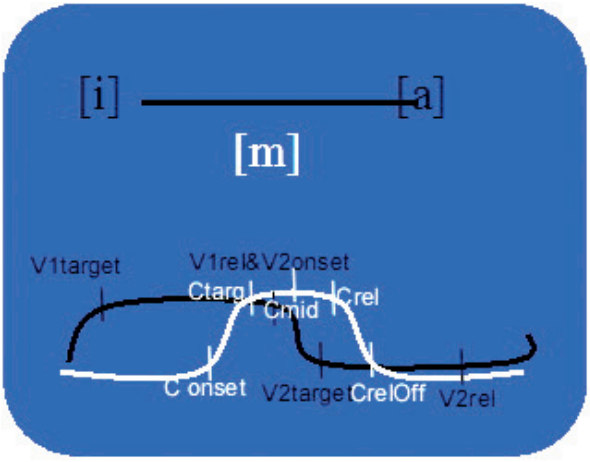
Tali risultati hanno evidenziato come le caratteristiche cinematiche e temporali che contraddistinguono il contrasto scempia-geminata sono proprie del contrasto in sé e non della specifica struttura sillabica.

Infine, Burroni et al. (2024) hanno indagato le proprietà cinematiche delle geminate italiane mediante EMA (3D, Carstens AG501), utilizzando pseudo-parole bisillabiche "VC(:)V" (con C= [p], [m] e V=[i], [a]) a 5 diverse velocità di elocuzione. Lo studio ha confermato la maggiore durata nella fase di chiusura per le geminate. Circa le proprietà cinematiche, le geminate sono prodotte, a prescindere dalla velocità di elocuzione, con movimenti articolatori più ampi, velocità massima leggermente minore e target più estremi (come in Löfqvist, 2005). Burroni et al. sostengono che le geminate non siano meramente delle scempie con intervalli di attivazione più lunghi (cfr. Gafos e Goldstein, 2012), bensì una specifica categoria gestuale. Una prova di ciò deriva anche dalla differente temporizzazione delle geminate con le vocali adiacenti: viene confermata l'anticipazione del gesto di chiusura della consonante in caso di geminazione rispetto a V_1 (come in Smith, 1995, e come in tutti gli autori fin qui considerati); tuttavia, l'intervallo vocalico V_1-V_2, differentemente dal modello *V-to-V*, risulta minore in caso di interposizione di una geminata. Inoltre, gli studiosi mostrano come anche il gesto per V_2 inizia prima se è presente una consonante raddoppiata, differentemente da quanto emerso in Smith (1995). Tali contraddizioni, rispetto a studi citati in precedenza, possono derivare da differenti scelte nel posizionamento dei vari *landmark*.

3. *Obiettivi e ipotesi*

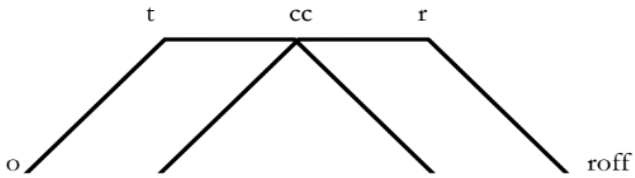
Il presente studio si pone diversi obiettivi inerenti alla sincronizzazione gestuale delle consonanti scempie (VCV) e geminate (VCCV), utilizzando i dati raccolti e parzialmente elaborati da Zmarich et al. (2011). I dati ivi trattati sono esclusivamente di tipo temporale: altre variabili cinematiche (ampiezza, *stiffness*, *time-to-peak* etc.) non sono state indagate. I dati analizzati riguardano la velocità d'eloquio "normale" (N, *normal*). Prima di ragionare sugli intervalli, è stata modificata la Fig.1 che compariva in Zmarich et al. (2011), adottando la terminologia dei *landmark* di Gafos (2002).

Figura 1 - *Landmark della traiettoria cinematica di "mim(m)a" secondo la formalizzazione di Gafos (2002) (adattata da Zmarich et al., 2011). Traiettoria linguale (nella dimensione verticale) per la prima e seconda vocale (nero); apertura labiale (LA) per la consonante (bianco). Si noti che "Cmid"="Ccent"*



Si noti come la traiettoria del dorso linguale per le vocali (in nero, Fig. 1) non possieda tutti i *landmark* della formalizzazione di Gafos (2002, Fig. 2). Inoltre, si osservi come il rilascio della prima vocale (*V1rel*) coincide con l'inizio della seconda vocale (*V2ons*).

Figura 2 - *Sovrapposizione di due curve gestuali in "close transition" come da formalizzazione di Gafos (2002)*



Il primo obiettivo del lavoro è quello di testare due modelli alternativi di sincronizzazione gestuale, il "*combined vowel and consonant timing*" (*V-to-C*, Browman, Goldstein, 1990) e il "*Vowel-to-Vowel timing*" (*V-to-V*, Öhman, 1967).

Di seguito riportiamo gli intervalli ritenuti cruciali per il confronto, nonché le nostre predizioni a riguardo, aiutandoci con le Figg. 3 e 4:

– *V1targ_Cons* (indicato con 1 nelle Figg. 3 e 4) di VCCV è diverso (≠) dallo stesso intervallo in VCV per entrambi i modelli. Per il *V-to-C* tale intervallo dovrebbe variare in caso di geminazione, in quanto l'aggiunta di ulteriori consonanti modifica la sincronizzazione vocalica. Infatti, Browman, Goldstein (1990) prevedono che in VCV la consonante sia sincronizzata con la seconda vocale, mentre in VC1C2V, C1 è sincronizzata con la prima vocale. Invece,

per il modello *V-to-V*, l'intervallo *V1targ_Cons* differisce tra VCV e VCCV, in quanto mantenendo il ciclo vocalico inalterato, (cioè mantenendo invariato il valore temporale corrispondente a V1onset, si assiste ad un'anticipazione del gesto consonantico per consentire la sua espansione, cioè Ctarget si avvicina a V1).

– *V1targ_Ccent* (indicato con 2 nelle Figg. 3 e 4) di VCCV è diverso (≠) dallo stesso intervallo in VCV secondo *V-to-C*, in quanto risente della durata consonantica e della sincronizzazione in "anti-fase" della prima parte della consonante. Nella scempia, invece, la prima vocale non ha alcuna relazione di fase con la vocale precedente. Nel modello *V-to-V* questo intervallo si mantiene uguale (=).
– *Ccent_V2rel* (indicato col 3 nelle Figg. 3 e 4) di VCCV è simile (=) allo stesso intervallo in VCV secondo *V-to-C*, in quanto lo spostamento verso destra di *Creloff*, secondo Browman, Goldstein (1990) è compensato parzialmente dallo spostamento di V2 sempre nella stessa direzione. Per motivi diversi questo intervallo resta invariato (=) anche nel modello *V-to-V*.

Figura 3 - *Schema degli intervalli temporali indagati per verificare il modello di coordinazione V-to-V di Öhman (1967) applicato a "mim(m)a". La curva nera rappresenta il movimento del dorso della lingua da [i] ad [a]; la linea bianca simula il gesto dell'Apertura Labiale per la realizzazione della consonante [m]/[mm]. Vedi testo per la spiegazione sopra=singola; sotto=geminata*

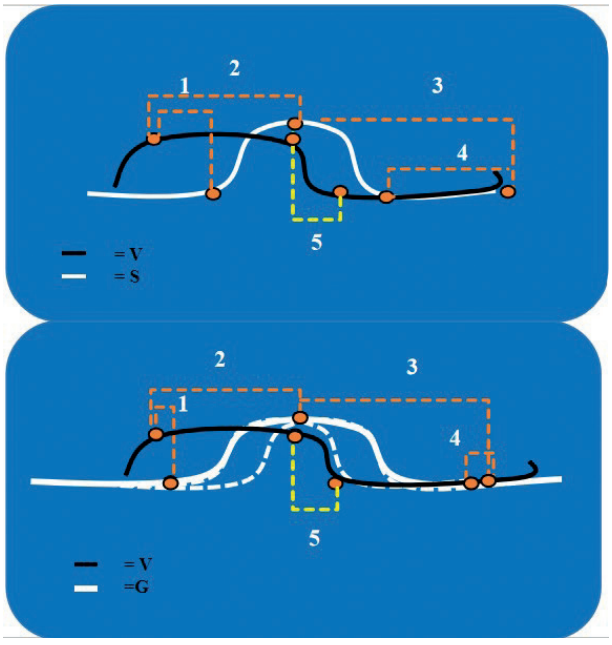
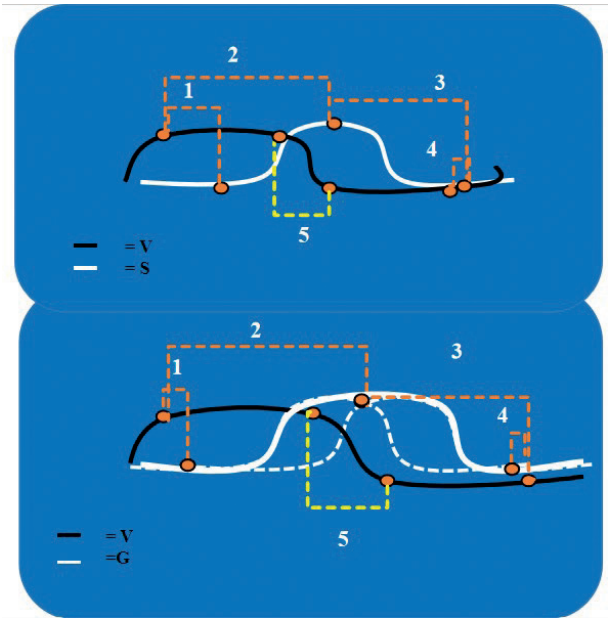


Figura 4 - *Schema degli intervalli temporali indagati per verificare il modello di coordinazione V-to-C di Browman e Goldstein (1990) applicato a "mim(m)a". La curva nera rappresenta il movimento del dorso della lingua da [i] ad [a]; la linea bianca simula il gesto dell'Apertura Labiale per la realizzazione della consonante [m]/[mm]. Vedi testo per la spiegazione sopra=singola; sotto=geminata*



– *Creloff_V2rel* di VCCV (indicato col 4 nelle Figg. 3 e 4) è simile (=) allo stesso intervallo in VCV secondo Browman, Goldstein (1990), in quanto V2 subirà un ritardo gestuale causato dalla maggiore lunghezza consonantica, compensata dallo spostamento a destra di V2. Invece, nel caso del modello di Öhman (1967), il suddetto intervallo sarà diverso (≠) dallo stesso intervallo in VCV. Ciò si spiega col fatto che lo spostamento a destra di *Crel* non influisce su *V2rel*, che rimane fisso.
– *V1rel_V2targ* (indicato col 5 nelle Figg. 3 e 4) non era stato calcolato da Zmarich et al. (2011). L'intervallo rappresenta la transizione dalla fine di V1 all'inizio di V2. Basandosi su studi precedenti di Zmarich et al. (2005, 2007, 2011), si ipotizza che *V1rel_V2targ* in VCCV sia (≠) dallo stesso intervallo in VCV secondo *V-to-C*, ovvero maggiore nel primo caso, in quanto risente della lunghezza consonantica; contrariamente, secondo il modello di Öhman (1967) tale intervallo, relativo al ciclo vocalico, rimarrà invariato tra scempia e geminata.

Un ulteriore obiettivo riguarda lo studio dei singoli intervalli consonantici ricavabili dai 4 *landmark* del ciclo consonantico proposti da Gafos (2002, Fig. 2), ossia *Cons_Ctarg*; *Ctarg_Crel* e *Crel_Creloff* (Fig. 1). Tale analisi non è funzionale alla verifica di alcun modello, ma è finalizzata a comprendere se il contrasto di lunghezza in italiano sia motivato dal prolungamento di un unico intervallo temporale, per

esempio la fase di tenuta dell'occlusione, che è il correlato acustico pregnante per la geminazione. Si potrebbe ipotizzare che il prolungamento dell'intervallo di tenuta sia ricavato tramite un'anticipazione del gesto consonantico nella geminazione, nonché un ritardo del suo rilascio, i cui effetti acustici dovrebbero consistere in una riduzione della durata delle due vocali contestuali come predetto dal modello di Öhman (1967), richiamato da Smith (1995) e da Zmarich et al. (2007).

Un'ultima modalità di verifica e confronto dei modelli cinematici V-to-V e V-to-C è consistita nella creazione di una "geminata virtuale" (GV, si veda § 4.3).

4. *Metodi*

4.1 Corpus e partecipanti

Il corpus sperimentale è stato mutuato da Zmarich et al. (2005) ed è formato dalla coppia minima scempia-geminata "*mima/mimma*". Le parole target erano inserite entro la frase cornice "*richiama mim(m)a malamente*". Le frasi erano state randomizzate e presentate a gruppi di tre su pannelli posti a circa due metri di distanza dai soggetti. Erano state previste dieci ripetizioni per frase a velocità normale e altre dieci a velocità sostenuta, sebbene nel presente studio vengano riportati solo i risultati inerenti alla velocità normale.

I quattro soggetti sono gli stessi dell'esperimento descritto da Zmarich et al. (2005), Gili Fivela et al. (2007) e Zmarich et al. (2011). La locutrice (AG) appartiene a una varietà italiana del nord-est; due soggetti, una femmina (BG) e un maschio (MP), parlano un italiano regionale del nord-ovest, mentre (FC) proviene da una varietà del centro Italia. Tutti i locutori si erano dichiarati esenti da qualunque disturbo di linguaggio e/o voce, presente o pregresso.

4.2 Acquisizione dei materiali e misure

Il segnale audio è stato registrato con due dispositivi distinti, uno integrato all'hardware di EMA (vedi oltre) e l'altro esterno. Il segnale analizzato è quello acquisito col dispositivo esterno, di maggior qualità (44 kHz e 16 bit), costituito da un DAT collegato a un microfono professionale. È stato utilizzato il software PRAAT (https://www.fon.hum.uva.nl/praat/) per l'analisi del segnale acustico e per la segmentazione manuale delle parole entro i loro confini. Nello specifico, la sequenza VC(C)V è stata segmentata e misurata mediante lo studio del sonogramma a banda larga. Nel caso delle geminate non si sono posti confini acustici interni alla sequenza consonantica.

Per la raccolta dei dati cinematici, è stato utilizzato l'*Electro Magnetic Articulograph* o EMA AG200 (© Carstens), nella versione 2D all'epoca in uso presso il laboratorio ICP-CNRF (ora GIPSA lab) a Grenoble, in Francia. I sensori erano stati posizionati, tramite l'uso di colla tissutale, nelle seguenti posizioni (con riferimento al piano medio-sagittale della testa): due di essi, utili per compensare i movimenti della testa, erano stati collocati sul ponte nasale (lo spazio tra le

sopracciglia) e sul prolabio (solco sotto-nasale); due erano stati posti al centro del labbro inferiore e superiore; quattro erano stati posizionati sulla superficie linguale a distanza l'uno dall'altro di circa un centimetro, partendo da un centimetro subito dopo l'apice linguale e proseguendo verso la radice.

Per le consonanti bilabiali si sono utilizzate le misurazioni dei sensori incollati al centro delle labbra, mediante i quali è stata rilevata l'apertura labiale (*Lip Aperture*), ovvero la distanza verticale. Per ottenere dati sulla cinematica delle vocali, si è fatto ricorso al sensore posto sul dorso della lingua, posizionato a circa 3.5 cm dall'apice della lingua.

Le misurazioni cinematiche sono state eseguite in una fase successiva alla segmentazione e all'etichettatura del segnale acustico e sono state allineate a queste ultime. Per l'elaborazione dei dati cinematici si è impiegato il software *Interface* (cfr. Tisato, Cosi, Sommavilla, Zmarich, 2007). Similmente a Smith (1995), per ottenere l'inizio e la fine di ogni gesto si sono utilizzati i punti in cui la velocità dell'articolatore in esame rispettivamente raggiungeva o diminuiva del 10 % rispetto al suo picco massimo.

Alle misure acustiche e cinematiche, Zmarich et al. (2011) avevano affiancato anche un test di percezione, volto a verificare se i locutori dello studio avessero prodotto adeguatamente le consonanti geminate. Infatti, in alcune varietà italiane, nella fattispecie settentrionali, il contrasto scempia-geminata è segnalato con una durata di occlusione minore (degeminazione settentrionale) e percettivamente può risultare indebolito o assente. Con il software *Perceval* (https://blog.bitergia.com/2018/05/01/perceval-software-project-data-at-your-will/) cinque ascoltatori del nord Italia e cinque del sud Italia avevano ascoltato le tracce audio contenenti le frasi cornice e avevano giudicato se le parole target al loro interno fossero geminate o meno. Le produzioni qui analizzate sono quelle risultate statisticamente prototipiche di ciascuna delle due categorie di lunghezza, e tra loro contrastanti (per i dettagli della verifica percettiva, si rimanda agli articoli citati).

La procedura sperimentale effettuata ex novo, univoca per tutti gli obiettivi sopra citati, è consistita nel calcolo di intervalli temporali ritenuti utili alle verifiche prefissate. Gli intervalli temporali sono stati ricavati sottraendo o sommando i vari *landmark* gestuali già calcolati da Zmarich et al. (2011), e altri ricavati da *landmark* "virtuali", espressamente ricavati per questo lavoro. Gli intervalli ritenuti utili alle verifiche inerenti al confronto tra i modelli *V-to-V* e *V-to-C* si sono rivelati essere: *V1targ_Cons*, *V1targ_Ccent*, *Ccent_V2rel*, *Creloff_V2rel*, *V1rel_V2targ*. Gli intervalli utilizzati per testare il loro ruolo nel contrasto scempia-geminata sono stati: *Cons_Ctarg*; *Ctarg_Crel* e *Crel_Creloff*. Questi intervalli sono poi stati organizzati in un database *Excel*, utili a effettuare le successive analisi statistiche.

4.3 Costruzione della "geminata virtuale"

Nella parte che segue descriveremo le modalità con cui abbiamo costruito la cosiddetta "geminata virtuale" (GV). In primis, si è raddoppiata la curva della scempia partendo dai dati reali ottenuti da Zmarich et al. (2011), sovrapponendo

successivamente le due curve ottenute in modalità *close transition*. Tale scelta si attiene a Gafos (2002), il quale critica l'idea della FA inerente a rapporti di fase consonantica di tipo universale, assumendo che questi siano linguo-specifici. Nel caso dell'italiano non vi sono prove a favore di una transizione aperta in caso di geminazione (consideriamo qui come occasionale la presenza di due *bursts*, indici acustici del rilascio di un'occlusione, riscontrati da Di Benedetto et al., 2021 nelle loro registrazioni). Per ottenere la sovrapposizione consonantica in "*close transition*", il *C1rel* è stato allineato con il *C2targ*, creando il *CcenterGV* della geminata virtuale (*cc* nella Fig. 2).

Dapprima si sono ricavati i vari *landmark* di GV: l'*onset* e il *target* non sono stati ricavati ex novo, in quanto corrispondenti a quelli della consonante scempia "reale"; anche per il *CcenterGV* non è stata necessaria alcuna derivazione, in quanto corrispondeva al *Crel* della scempia; il *CrelGV* in *close transition* è stato computato partendo dal *Crel* della scempia e aggiungendovi l'intervallo *Ctarg_Crel* della stessa; per ottenere il *CreloffGV* in *close transition*; infine, l'intervallo *Crel_Creloff* della scempia è stato aggiunto al *CrelGV*.

Successivamente, l'organizzazione temporale della GV è stata costruita in modo tale da simulare l'organizzazione temporale di ciascuno dei due modelli di sincronizzazione, *V-to-V* e *V-to-C*, al fine di confrontare statisticamente i dati simulati di alcuni intervalli critici con quelli ottenuti dalle geminate effettivamente prodotte (cfr. Zmarich et al., 2011), e verificare così l'adeguatezza dei modelli. A tale scopo si sono allineati i *landmark* della GV ottenuta a quelli delle vocali contestuali alla consonante scempia.

A questo scopo abbiamo utilizzato il modello *V-to-V* in due modi. Nel primo modo, lavorando sulle matrici di *Excel*, l'espansione del ciclo consonantico della geminata rispetto alla scempia è stata simulata facendo coincidere il *CcentGV* al *Ccent* della scempia, ed espandendo i *landmark* di GV a sinistra e a destra di un intervallo temporale corrispondente a metà della durata di occlusione articolatoria *Ctarget_Crel* della scempia. In termini pratici questo significa che abbiamo sottratto il valore temporale di tale intervallo ai *landmark* consonantici a sinistra del *CcentGV* (anticipando così la loro occorrenza temporale rispetto ai corrispondenti *landmark* della scempia), ottenendo *ConsGV* e *CtargGV*. Parallelamente, abbiamo aggiunto il valore temporale di tale intervallo a tutti i *landmark* a destra del *CcentGV* (posticipando temporalmente la loro occorrenza rispetto ai corrispondenti *landmark* della scempia), ottenendo *CrelGV* e *CreloffGV*. Particolare importante, abbiamo lasciato l'organizzazione temporale del ciclo vocalico inalterata, e non abbiamo fatto riferimento ad alcun *landmark* della vocale per simulare l'organizzazione temporale di GV.

Con la seconda modalità di simulazione di GV in base al modello *V-to-V*, assumiamo che il *Ccent* sia ancorato al *V2onset*, adottando la proposta di sincronizzazione che Browman, Goldstein (1988, 2000) avanzano per la sincronizzazione della consonante iniziale nella sillaba CV. Pur riconoscendo che si tratta di una forzatura, può essere però un'ipotesi di lavoro interessante nel senso

che questo tipo di sincronizzazione "cala nella realtà" la proposta fonologica di una consonante lunga sillabificata in *onset* (V.CCV), avanzata per primo da Martinet (1975). Pertanto, si è fatto corrispondere il *CcenterGV* (coincidente col *Crel* della scempia) al *V2ons* del ciclo vocalico VCV; successivamente è stata calcolata la differenza tra il *Ccenter* originario e il nuovo *CcenterGV* (ovvero, *V2ons*), al fine di sottrarre questo intervallo a ogni *landmark* della scempia e ottenere i nuovi punti della GV. In questo modo, togliendo la differenza (*Ccenter-CcentGV*) si sono ottenuti il *CtargGV*, *CrelGV*, *CreloffGV*. Poiché il *Ccent* della scempia occorreva diversi ms dopo il *V2onset*, dove è stato ancorato il *Ccent* della GV, l'effetto prodotto è consistito in uno spostamento globale a sinistra dell'intera configurazione della GV.

Di seguito si è testato il modello *V-to-C*, tenendo conto dei vari rapporti di fase, e partendo dal presupposto che, secondo quest'ultimo, la quantità consonantica modifica il ciclo vocalico (allungandolo). Dunque, sono state simulate le diverse sincronizzazioni tra consonante e vocali contestuali: il *CtargGV* è stato allineato al *V1rel* della scempia, riproducendo così la temporizzazione in "anti-fase" tra la consonante in coda sillabica e la vocale precedente; successivamente, è stata calcolata la differenza tra il *Ctarget* della scempia e il *CtargGV*; tale differenza, poi, è stata sottratta ai *landmark* della scempia per ottenere gli ancoraggi della GV (*ConsGV*, *CcentGV*, *CrelGV* e *CreloffGV*); successivamente, si è fatto coincidere il *V2onsetGV* con il *CcenterGV*, simulando la sincronizzazione "in-fase" tra la consonante in *onset* e la vocale successiva, nel senso che *CcenterGV* impone che *V2onsetGV* sincronizzi con esso; di seguito si è calcolata la differenza tra il *V2onsGV* e il *V2onset* della scempia, in modo che anche i *landmark* successivi della seconda vocale fossero spostati a destra, risentendo della lunghezza consonantica; si sono quindi calcolati i *landmark V2targGV* e *V2relGV*, sottraendo la differenza di cui sopra ai *landmark V2targ* e *V2rel* della scempia.

Con un'ultima simulazione abbiamo tentato di riprodurre l'ipotesi avanzata da Zmarich et al. (2007) e Gili Fivela et al. (2007), secondo cui le geminate italiane si conformano a un "modello ibrido", che possa spiegare sia la riduzione della durata temporale che sia ha in V_1 di V_1CCV_2, sia il sostanziale mantenimento in V_2. Per questo motivo, la serie di sincronizzazioni è la stessa del modello *V-to-C* appena descritto, salvo distaccarsene coll'imporre a *V2ons* di sincronizzare non più con *Ccent*, ma con *Crel*. L'effetto sostanziale è quello di spostare la durata del ciclo di V_2 a destra, sottraendolo meglio alla sovrapposizione consonantica.

4.4. Analisi statistica

Le analisi statistiche descrittive sono state condotte mediante il software *Jamovi* (https://www.jamovi.org/, versione 2.2.5), importando le tabelle direttamente da *Excel*. L'analisi statistica inferenziale è stata fatta precedere dal test di normalità "*Shapiro-Wilk*", che ha rivelato come alcune misure non fossero distribuite in modo normale. È stato dunque utilizzato il test non parametrico *Mann-Whitney*. La variabile dipendente considerata è stata sempre la "geminazione" (GEMINATION), mentre

le variabili dipendenti, di tipo continuo, sono rappresentate dalle misure temporali (puntuali o sotto forma di intervalli). Inoltre, sono stati utilizzati i filtri di *Jamovi*, per produrre statistiche divise per soggetto (AG, MP, FC, BG).

5. *Risultati*

5.1 Ruolo degli intervalli consonantici nella geminazione

In Tab. 1 si riportano le medie dei singoli intervalli consonantici indagati per comprendere il loro ruolo nella geminazione, ossia *Cons_Ctarg*, *Ctarg_Crel* e *Crel_Creloff*. Per tutti e quattro i soggetti, la totalità degli intervalli mostra una differenza significativa tra consonante geminata e scempia alla velocità d'elocuzione "normale".

Tabella 1 - *Media intervalli consonantici principali nei singoli soggetti. Unità di misura ms; G=geminates; S= singletons. *=coppie p<.05. Test Mann-Whitney U*

	AG		*MP*		*FC*		*BG*	
Time Intervals	*G*	*S*	*G*	*S*	*G*	*S*	*G*	*S*
Cons_Ctarg	111*	89*	90*	80*	102*	89*	122*	91*
Ctarg_Crel	42*	17*	51*	17*	45*	15*	60*	14*
Crel_Creloff	92*	68*	87*	74*	101*	76*	99*	78*

Questi dati smentiscono l'ipotesi che la maggior durata delle geminate rispetto alle scempie dell'italiano sia dovuta solo all'aumento della fase di tenuta dell'occlusione.

5.2 Analisi dei dati sperimentali "reali" per gli intervalli "critici"

Le medie degli intervalli da noi scelti perché in grado di rivelare, come una cartina al tornasole, le conseguenze diverse delle sincronizzazioni previste dai diversi modelli, suddivise per soggetto sono sintetizzati nella Tab. 2, che evidenzia anche se la differenza tra scempie e geminate all'interno di ciascun soggetto è significativa (p<.05; mediante asterisco).

Tabella 2 - *Media dei soggetti per gli intervalli temporali (arrotondati a ms; G = geminates; S = Singletons. *=la differenza nelle coppie è significativa per p<.05. Test Mann-Whitney U*

	AG		*MP*		*FC*		*BG*	
Time Intervals	*G*	*S*	*G*	*S*	*G*	*S*	*G*	*S*
V1targ_Cons	3	21	47*	65*	70*	90*	45*	81*
V1targ_Ccent	137*	115*	166	153	193	188	183	179
Ccent_V2rel	130	139	110*	92*	141*	119*	153*	96*
Creloff_V2rel	65	64	8	10*	15*	38	13	11
V1rel_V2targ	210*	159*	167*	119*	208*	151*	207*	132*

I risultati in questa tabella dipingono una situazione in cui, per la maggioranza dei soggetti (almeno tre su quattro), l'accorciamento della prima vocale è ottenuto tramite l'anticipazione dell'inizio consonantico (*V1targ_Cons*) ma senza anticipare il baricentro della tenuta dell'occlusione (*V1targ_Ccent*). Il mantenimento della durata della seconda vocale è realizzato tramite una posticipazione della fine della seconda vocale (*Ccent_V2rel*) e da un mantenimento dell'intervallo tra la fine della consonante e la fine della seconda vocale (*Creloff_V2rel*). Si assiste inoltre a un allungamento della durata della transizione tra la fine della prima vocale e il raggiungimento del target articolatorio della seconda (*V1rel_V2targ*).

5.3 Confronto con i modelli cinematici a partire dai dati reali

La Tab. 3 mostra se le predizioni fatte da ciascun modello per ogni intervallo al variare della lunghezza consonantica (=: non varia; ≠: varia) si conformano o meno ai risultati ottenuti (=: non varia; ≠: varia).

L'intervallo *V1targ_Cons* si mostra differente tra VCV e VCCV in 3 soggetti su 4, come predetto da tutti i modelli.

V1targ_Ccent resta invariato in 3 soggetti su 4, diversamente dalle predizioni dei modelli *VtoC* e *Hybrid*. L'intervallo *Ccent_V2rel*, invece, è risultato significativamente diverso in 3 soggetti su 4, anche qui divergendo dalle predizioni del modello *VtoC*.

Creloff_V2rel si è rivelato simile tra VCV e VCCV in 3 soggetti su 4, in conformità alle predizioni del modello *VtoC* e *Hybrid*. Infine, *V1rel_V2targ* si è mostrato significativamente diverso tra scempia e geminata in tutti i soggetti, conformemente alle predizioni dei modelli *VtoC* e *Hybrid*.

Tabella 3 - *Risposta dei soggetti alle predizioni dei modelli VtoV or VtoC (=: VCV non significativamente diverso da VCCV; ≠: VCV significativamente diverso da VCCV; p<.05)*

	VtoV	*VtoVmod*	*VtoC*	*Hybrid*	*AG*	*MP*	*FC*	*BG*
V1targ_Cons	≠	≠	≠	≠	=	≠	≠	≠
V1targ_Ccent	=	=	≠	≠	≠	=	=	=
Ccent_V2rel	≠	≠	=	≠	=	≠	≠	≠
Creloff_V2rel	≠	≠	=	=	=	=	≠	=
V1rel_V2targ	=	=	≠	≠	≠	≠	≠	≠

5.4 Confronto tra i dati ottenuti dalla "Geminata Reale (GR)" e i modelli cinematici simulati con la "Geminata Virtuale (GV)"

Nelle tabelle che seguono, mostreremo i risultati del confronto tra i dati ricavati dalle simulazioni delle conseguenze sull'organizzazione temporale predette da ciascuno dei quattro modelli di organizzazione temporale della GV e i dati ottenuti per la geminata reale, per ciascuno dei 5 intervalli critici. Ovviamente, la logica sottesa a questo confronto è che se la simulazione relativa a un certo modello non produce dati temporali significativamente diversi per l'intervallo in questione rispetto ai dati realmente ottenuto, allora il modello risulta confermato.

La Tab. 4 mostra i valori temporali (ms) ottenuti con la simulazione del modello *V-to-V* (GV) affiancati ai dati misurati dalla produzione reale (GR), soggetto per soggetto per ciascuno dei 5 intervalli.

Tabella 4 - *Medie dei soggetti per gli intervalli temporali critici delle geminate "reali" (GR) e virtuali (GV), simulate in base al modello V-to-V (arrotondati a ms; G = geminates; S = Singletons. *=la differenza nelle coppie è significativa per p<.05. Test Mann-Whitney U*

VtoV	*AG*		*MP*		*FC*		*BG*	
Time Intervals	*GR*	*GV*	*GR*	*GV*	*GR*	*GV*	*GR*	*GV*
V1targ_Cons	3	8	47*	56*	70*	83*	45*	74*
V1targ_Ccent	137*	113*	166	154	193	187	183	179
Ccent_V2rel	130	140	110*	92*	141	121	153*	96*
Creloff_V2rel	65	55	8	1	15	30	13	4
V1rel_V2targ	210*	159*	167*	119*	208*	151*	207*	132*

La Tab. 5 esibisce i valori temporali (ms) ottenuti con la simulazione del modello *V-to-Vmod*, che è stato modificato tramite l'allineamento forzato del centro della tenuta articolatoria (*Ccent*) con l'inizio del movimento verso la seconda vocale (*V2ons*). I dati simulati (GV) sono affiancati ai dati misurati dalla produzione reale (GV), soggetto per soggetto per ciascuno dei 5 intervalli.

Tabella 5 - *Medie dei soggetti per gli intervalli temporali critici delle geminate "reali" (GR) e virtuali (GV), simulate in base al modello V-to-V modificato (arrotondati a ms; G = geminates; S = Singletons. *=la differenza nelle coppie è significativa per p<.05. Test Mann-Whitney U*

VtoVmod	*AG*		*MP*		*FC*		*BG*	
Time Intervals	*GR*	*GV*	*GR*	*GV*	*GR*	*GV*	*GR*	*GV*
V1targ_Cons	3*	-41*	47*	19*	70*	38*	45	29
V1targ_Ccent	137*	69*	166*	116*	193*	142*	183*	134*
Ccent_V2rel	130*	188*	110*	130*	141	166	153*	141*
Creloff_V2rel	65*	104*	8*	39*	15*	76*	13*	49*
V1rel_V2targ	210*	159*	167*	119*	208*	151*	207*	132*

La Tab. 6 mostra i valori temporali (ms) ottenuti con la simulazione del modello *V-to-C*. I dati simulati (GV) sono affiancati ai dati misurati dalla produzione reale (GR), soggetto per soggetto per ciascuno dei 5 intervalli.

Tabella 6 - *Medie dei soggetti per gli intervalli temporali critici delle geminate "reali" (GR) e virtuali (GV), simulate in base al modello V-to-C (arrotondati a ms; G = geminates; S = Singletons. *=la differenza nelle coppie è significativa per p<.05. Test Mann-Whitney U*

VtoC	*AG*		*MP*		*FC*		*BG*	
Time Intervals	*GR*	*GV*	*GR*	*GV*	*GR*	*GV*	*GR*	*GV*
V1targ_Cons	3 *	-23 *	47 *	36 *	70	53	45	43
V1targ_Ccent	137 *	83 *	166 *	133 *	193 *	157 *	183 *	148 *
Ccent_V2rel	130 *	188 *	110 *	130 *	141	166	153 *	141 *
Creloff_V2rel	65 *	104 *	8 *	39 *	15 *	76 *	13 *	49 *
V1rel_V2targ	210 *	177 *	167 *	136 *	208 *	165 *	207 *	146 *

La Tab. 7 esibisce i valori temporali (ms) ottenuti con la simulazione del modello "*ibrido*". I dati simulati (GV) sono affiancati ai dati misurati dalla produzione reale (GR), soggetto per soggetto per ciascuno dei 5 intervalli.

Tabella 7 - *Medie dei soggetti per gli intervalli temporali critici delle geminate "reali" (GR) e virtuali (GV), simulate in base al modello Hybrid (arrotondati a ms; G = geminates; S = Singletons. *=la differenza nelle coppie è significativa per p<.05. Test Mann-Whitney U*

Hybrid	*AG*		*MP*		*FC*		*BG*	
Time Intervals	*GR*	*GV*	*GR*	*GV*	*GR*	*GV*	*GR*	*GV*
V1targ_Cons	3 *	-23 *	47 *	36 *	70	53	45	43
V1targ_Ccent	137 *	83 *	166 *	133 *	193 *	157 *	183 *	148 *
Ccent_V2rel	130 *	205 *	110 *	147 *	141 *	181 *	153	155
Creloff_V2rel	65 *	120 *	8 *	56 *	15 *	90 *	13 *	63 *
V1rel_V2targ	210	193	167 *	153 *	208	180	207 *	160 *

La Tab. 8 indica per quali soggetti la simulazione in base a ciascun modello ha prodotto un valore temporale non significativamente diverso dai valori reali.

Tabella 8 - *Soggetti la cui geminata virtuale non ha una durata statisticamente diversa dalla geminata reale*

	VtoV	*VtoVmod*	*VtoC*	*Hybrid*
V1targ_Cons	AG	BG	FC,BG	FC,BG
V1targ_Ccent	MP,FC,BG	-	-	-
Ccent_V2rel	AG,FC	FC	FC	BG
Creloff_V2rel	AG,MP,FC,BG	-	-	-
V1rel_V2targ	-	-	-	AG,FC

Come è facilmente osservabile, la simulazione di maggior successo è quella relativa al modello *V-to-V*, in base al quale un semplice allungamento simmetrico a destra e

a sinistra del *Ccenter* della consonante scempia basta a rendere conto dell'intervallo tra il raggiungimento del target articolatorio della prima vocale e il centro della fase di tenuta articolatoria, di quello tra quest'ultimo e il rilascio della consonante (almeno per due soggetti), e, per tutti i soggetti, dell'intervallo relativo allo "spazio vitale" della seconda vocale, liberata dalla sovrapposizione con la consonante. L'unico intervallo non riprodotto è quello relativo alla durata della transizione tra la posizione articolatoria della prima vocale e quella della seconda, che per il modello *V-to-V* deve rimanere necessariamente invariato tra VCV e VCCV (la variazione interessando il solo ciclo consonantico), ma questo era un dato atteso dalle rilevazioni reali.

Un'altra osservazione interessante di tipo generale è l'estrema idiosincraticità delle risposte. All'interno di ciascun modello, salvo il *V-to-V*, i soggetti che producono una simulazione di successo per un intervallo non sono mai gli stessi per gli altri intervalli (con l'eccezione, forse di FC).

Da notare anche il fallimento della simulazione con il modello *V-to-C*.

6. *Discussione e conclusioni*

È nostra opinione che i dati che abbiamo presentato gettino una nuova luce sulla natura delle consonanti geminate italiane e propongano un metodo originale per lo studio dell'organizzazione temporale intra- e inter-gestuale come quello della geminata virtuale. Ricordiamo brevemente che l'impostazione di questo studio si è basata sull'individuazione di alcuni intervalli temporali che mettono in relazione dei *landmark*, nella terminologia di Gafos (2002), che non sono altro che punti di ancoraggio nel ciclo articolatorio per la sincronizzazione temporale di gesti consonantici e vocalici di tipo sia intra-sillabico (come la relazione C-V) che inter-sillabico (come la relazione V-C nel bisillabo VCV), sia solo consonantici (C-C in VCCV) che solo vocalici (V-V in V(C)CV). Gli intervalli individuati costituiscono dei veri banchi di prova del funzionamento di alcuni modelli di organizzazione temporale intra-e inter-sillabica presenti in letteratura. I modelli analizzati sono il modello *V-to-V* originalmente proposto da Öhman (1967), lo stesso modello modificato tramite una sincronizzazione del centro del ciclo consonantico "in fase" con l'inizio della seconda vocale (*V-to-Vmod*), il modello *V-to-C* di Browman, Goldstein (1989, 2000) e un modello ibrido da noi proposto sulla base dei risultati ottenuti da Zmarich et al. (2007) e Gili Fivela et al. (2007).

Riassumendo, ci si è posti tre obiettivi: testare i modelli *V-to-V* e *V-to-C*, basandosi su intervalli inerenti a dati reali; verificare il ruolo dei singoli intervalli consonantici nel contrasto di lunghezza consonantica; verificare il funzionamento dei modelli di temporizzazione *V-to-V* e *V-to-C* mediante la costruzione di una geminata virtuale.

L'analisi dei dati cinematici "reali" prodotti dai soggetti nell'esperimento di Zmarich et al. (2011) ha decretato che, per la maggioranza dei soggetti (almeno tre su quattro), l'accorciamento della prima vocale è ottenuto tramite l'anticipazione dell'inizio consonantico (*V1targ_Cons*) ma senza anticipare il baricentro della

tenuta dell'occlusione (*V1targ_Ccent*). Il mantenimento della durata della seconda vocale è realizzato tramite una posticipazione della fine della seconda vocale (*Ccent_V2rel*) e da un mantenimento dell'intervallo tra la fine della consonante e la fine della seconda vocale (*Creloff_V2rel*). Si assiste, inoltre, a un allungamento della durata della transizione tra la fine della prima vocale e il raggiungimento del target articolatorio della seconda (*V1rel_V2targ*). Vale la pena notare che l'allungamento di questo intervallo è l'unica variazione tra singola e geminata che è condivisa da tutti i soggetti.

Un risultato interessante, e mai presentato prima, riguarda la verifica dell'intervallo di transizione vocalica *V1rel_V2targ*: il risultato per quest'ultimo intervallo, mai indagato nei precedenti contributi di Zmarich, Gili Fileva e colleghi, confuta l'ipotesi avanzata da Smith (1995), secondo la quale la geminazione italiana si conforma senza riserve al modello *V-to-V* di Öhman (1967). Infatti, diversamente da quanto predetto dalla studiosa, la fase di transizione vocalica del ciclo vocalico *V1rel_V2targ* si è mostrata significativamente differente per tutti i soggetti, da quella presente nella scempia. Il risultato sostanzia e spiega quanto rilevato da Zmarich et al. (2011) per la maggior durata del ciclo vocalico *V1targ_V2rel* (*V1ons-V2off*, nella terminologia di Zmarich e colleghi), nel senso che questo allungamento, non previsto dal modello *V-to-V*, è dovuto principalmente all'aumento della durata della transizione vocalica (anche questo non previsto dal modello).

Circa il secondo obiettivo, si è evidenziato come nella geminazione sia coinvolto un aumento di durata di ciascun singolo intervallo consonantico (*Cons_Ctarg*, *Ctarg_Crel* e *Crel_Creloff*). In primo luogo, tale dato confuta l'idea comune che la geminazione sia causata da un mero allungamento della sola fase di tenuta articolatoria (*Ctarg_Crel*). Inoltre, sebbene lo studio si focalizzi puramente sulle durate temporali, questo risultato potrebbe portare a ipotizzare che il mutamento da geminata a scempia, in italiano, possa essere spiegato dal cosiddetto *change by shrinking* (Cho, 2006), oppure dal cosiddetto *change by stiffness* (Cho, 2006). In altri termini, nel primo caso la scempia, rispetto alla geminata, presenterebbe una modifica di tutti i parametri dell'equazione massa-molla (*stiffness* e ampiezza), che causerebbero una maggiore durata della geminata (ricordiamo che nell'equazione massa-molla la durata non costituisce un parametro (da pianificare) ma è un (sotto) prodotto implicito dello svolgimento del processo descritto dall'equazione). Nel secondo caso, la geminata, rispetto alla scempia, presenterebbe una modifica della sola *stiffness*. Per avvalorare tale predizione, tuttavia, sarebbe auspicabile effettuare nuove misurazioni, volte a determinare il ruolo delle altre variabili cinematiche come l'ampiezza, la velocità massima *time-to-peak velocity* (che indica la proporzione della fase di accelerazione su quella di decelerazione), che sono attualmente in fase di studio.

Una parte importante e metodologicamente nuova di questo studio ha riguardato la simulazione delle organizzazioni temporali dei gesti cinematici per consonanti e vocali nel bisillabo VCCV in base ai quattro modelli e il loro confronto con i dati delle geminate "reali". Anche qui siamo partiti dalla formalizzazione

di Gafos (2002) che modella la geminata in *close transition*, come quella italiana, raddoppiando la curva della scempia ricavata dai dati reali ottenuti da Zmarich et al. (2011) e sovrapponendola ai profili cinematici del ciclo vocalico V-V dello stimolo *singleton*. Il confronto ha messo in luce che, nonostante i problemi del modello *V-to-V* esposti in precedenza, quest'ultimo ha approssimato con più successo il tipo di sincronizzazione che compare nelle geminate "reali".

In conclusione, non va nascosto che la consonante labiale qui indagata presenta un ridotto grado di complessità poiché la consonante e le vocali contestuali sono realizzate da articolatori diversi (rispettivamente, labbra e dorso della lingua) i cui movimenti sono reciprocamente indipendenti, e che altri tipi di consonanti linguali aumenterebbero il grado di complessità e di non linearità della sincronizzazione tra consonanti e vocali, ma proprio per questo motivo l'indipendenza articolatoria garantisce un ambiente ideale per lo studio di un tipo di sincronizzazione "pura".

Bibliografia

Bertinetto, P.M. (1981). Strutture Prosodiche dell'Italiano, Studi di grammatica italiana, *Accademia della Crusca*, Firenze.

Browman, C., Goldstein, L. (1988). Some notes on syllable structure in Articulatory Phonology. In *Phonetica*, 45(2-4), 140-155.

Browman, C., Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, 6, 201-251.

Browman, C., Goldstein, L. (1990). Tiers in Articulatory Phonology, with some implications for casual speech. In Kingston, J., Beckman, M.E. (Eds.), *Papers in Laboratory Phonology I: Between the grammar and physics of speech*. Cambridge: Cambridge University Press, 340-376.

Browman, C., Goldstein, L. (1992). Articulatory Phonology: An Overview. In *Phonetica*, 49, 155-180.

Browman, C., Goldstein, L. (2000). Competing constraints on intergestural coordination and self-organization of phonological structures. In *Bulletin de la Communication Parlée*, 5, 25-34.

Burroni, F., Maspong, S., Benker, N., Hoole, P. & Kirby, J. (2024). Spatiotemporal Features of Bilabial Geminate and Singleton Consonants in Italian. *Proceedings of the 13th International Seminar of Speech Production, 13-17 May 2024 Autrans FR*. 189-192.

Celata, C., Meluzzi, C. & Bertini, C. (2022). Acoustic and Kinematic Correlates of Heterosyllabicity in Different Phonological Contexts. In *Language and Speech*, 65(3), 755-780.

Cho, T. (2002). *The Effects of Prosody on Articulation in English*. New York, London: Routledge.

Di Benedetto, M.G., De Nardis, L. (2021a). "Gemination in Italian: The affricate and fricative case," *Speech Commun. (in press)*; arXiv:2005.06959.

Dipino, D., Celata, C. (2018). An UTI study of alveolar stops in Italian. In Vietti, A., Spreafico, L., Mereu, D. & Galatà, V. (Eds.), *Il parlato nel contesto naturale. Speech in the natural context*. Milano: Officinaventuno, 41-53.

Esposito, A., Di Benedetto, M.G. (1999). Acoustical and perceptual study of gemination in Italian stops. In *Journal of The Acoustical Society of America*, 106, 2051-2062.

Fowler, C. (1980). Coarticulation and theories of extrinsic timing. In *Journal of Phonetics*, 8, 113-133.

Gafos, A. (2002). A grammar of gestural coordination. In *Natural Language and Linguistic Theory*, 20(2), 269-337.

Gafos, A., Goldstein, L. (2012). Articulatory representation and organization. In Cohn, A., Fourgeron, C. & Huffman, E.M. (Eds.), *Oxford Handbook of Laboratory Phonology*. Oxford: Oxford University Press, 220-231.

Gili Fivela, B., Zmarich, C. (2005). Italian Geminates under Speech Rate and Focalization Changes: Kinematics, Acoustic, and Perception Data. *InterSpeech 2005*, Lisbon, 2897-2900.

Gili Fivela, B., Zmarich, C., Perrier, P., Savariaux, C. & Tisato, G. (2007). Acoustic and kinematic correlates of phonological length contrast in Italian consonants. In *Atti dell'International Conference of Phonetic Sciences (ICPhS'07)*, ISBN:978-3-9811535-0-7.

Hagedorn, C., Proctor, M. & Goldstein, L. (2011). Automatic Analysis of Singleton and Geminate Consonant Articulation Using Real-time Magnetic Resonance Imaging. In *Journal of the Acoustical Society of America*, 130(4), EL251-EL257.

Hall, N. (2010). Articulatory Phonology. In *Language and Linguistics*, 4(9), 818-830.

Krause, P.A., Kawamoto, A.H. (2020). On the timing and coordination of articulatory movements: Historical perspectives and current theoretical challenges. In *Language and Linguistics*, 14(6), e12373.

Lahiri, A., Hankamer, J. (1988). The timing of geminate consonants. In *Journal of Phonetics*, 16, 327-338.

Loporcaro, M. (1990). On the analysis of geminates in Standard Italian and Italian dialects. In Hurch, B., Rhodes, R.A. (Eds.), *Natural Phonology. The State of the Art*. Berlin, New York: Mouton de Gruyter, 153-188.

Löfqvist, A. (2005). Lip kinematics in long and short stop and fricative consonants. In *Journal of the Acoustical Society of America*, 117, 858-878.

Mairano P., De Iacovo, V. (2019). Gemination in Northern versus Central and Southern Varieties of Italian: A Corpus-based Investigation. In *Language and Speech*, 63(3): 608-634.

Marin, S. (2013). Organization of Complex Onsets and Codas in Romanian: A gestural approach. In *Journal of Phonetics*, 41, 211-227.

Marotta, G., Vanelli, L. (2021). *Fonologia e prosodia dell'italiano*. Roma: Carocci.

Mattei, M., Di Benedetto, M.G. (2000). Acoustic analysis of singleton and geminate nasals in Italian. *Web-SLS The European Journal of Language and Speech, 1-11. Retrieved from* http://wrangler.essex.ac.uk/web-sls

Payne, E.M. (2005). Phonetic variation in consonant gemination. In *Journal of the International Phonetic Association*, 35, 153-189.

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

Pastätter, M., Pouplier, M. (2015). Onset-vowel timing as a function of coarticulation resistance: Evidence from articulatory data. *Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow*.

Pouplier, M. (2020). Articulatory phonology. In *Oxford research encyclopedia of linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.745

Saltzman, E.L. (1986). Task dynamic coordination of the speech articulators: a preliminary model. Generation and modulation of action patterns, *Experimental Brain Research*, Series 15, 129.144.

Smith, C.L. (1995). Prosodic patterns in the coordination of vowel and consonant gestures. In Connell, B., Arvaniti, A. (Eds.), *Phonology and Phonetic Evidence. Papers in Laboratory Phonology IV*. Cambridge: Cambridge University Press, 205-222.

Stevens, M., Hajek, J. (2005). Raddoppiamento sintattico (RS) and word-medial gemination in Italian. In *Theoretical and Experimental Approaches to Romance Linguistics: Selected Papers from the 34th Linguistic Symposium on Romance Languages (LSRL)*, Salt Lake City, March 2004 (Vol. 272, p. 257-271). Amsterdam, Netherlands: John Benjamins.

Turvey, M.T. (1990). Coordination. In *American Psychologist*, 45, 938-953.

Zmarich, C., Gili Fivela, B. (2005). Consonanti scempie e geminate in italiano. Studio cinematico e percettivo dell'articolazione bilabiale e labiodentale. In Cosi, P. (Ed.), *La misura dei parametri. Aspetti tecnologici ed implicazioni nei modelli linguistici. Atti del I Convegno Nazionale AISV – Associazione Italiana di Scienze della Voce, Padova*, 2-4 dicembre 2004. Torriana (RN), EDK Editore, 429-448.

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. (a cura di), *Atti del III Convegno Nazionale dell'Associazione Italiana di Scienze della Voce (AISV), Trento, 29-30/11-1/12/2006*, EDK Editore srl, Torriana (RN), 151-163, 2006.

Zmarich, C., Gili Fivela, B., Perrier, P., Savariaux, C. & Tisato, G. (2007). 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 – Associazione Italiana di Scienze della Voce, Povo (Trento), 29 novembre – 1 dicembre 2006*. Torriana (RN), EDK Editore, 151-163.

Zmarich, C., Gili Fivela, B., Perrier, P., Savariaux, C. & Tisato, G. (2011). Speech Timing Organization for the Phonological Length Contrast in Italian Consonants. In Cosi, P., De Mori, R., Di Fabbrizio, G. & Pieraccini, R. (Eds.), *Proceedings of Interspeech 2011: 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, August 27-31, 2011. 401-404.

Autori

Amalia Arvaniti – Centre for Language Studies, Radboud University, Paesi Bassi
amalia.arvaniti@ru.nl; https://orcid.org/0000-0002-1689-1931

Florence Baills – Universitat de Lleida
florence.baills@udl.cat; https://orcid.org/0000-0001-6743-9008

Tommaso Balsemin – Dipartimento di studi linguistici e letterari, Università di Padova, Italia
tommaso.balsemin@unipd.it

Silvia Calamai – Dipartimento di Filologia e Critica delle Letterature Antiche e Moderne, Università degli Studi di Siena, Italia
silvia.calamai@unisi.it; https://orcid.org/0000-0002-6585-2576

Federica Cavicchio – CRIL, Università del Salento, Italia
federica.cavicchio@unisalento.it; https://orcid.org/0000-0002-2795-8396

Claudia Roberta Combei – Dipartimento di Studi Letterari, Filosofici e di Storia dell'Arte, Università di Roma "Tor Vergata", Italia
claudia.roberta.combei@uniroma2.it; https://orcid.org/0000-0003-1884-8205

Emanuela Cresti – Università di Firenze
elicresti@gmail.com; https://orcid.org/0000-0002-3018-6120

Alice Crochiquia – LIAAC/LAEL – PUC-SP, São Paulo, Brazil; Department of Linguistics – Stockholm University, Stockholm, Sweden
arcrochiquia@gmail.com; https://orcid.org/0000-0002-2344-5390

Domenico De Cristofaro – Facoltà di Scienze della Formazione, Libera Università di Bolzano, Italia
domenico.decristofaro@unibz.it

Valentina De Iacovo – Università di Torino
valentina.deiacovo@unito.it; https://orcid.org/0000-0002-3913-2727

Anna De Meo – Dipartimento di Studi Letterari, Linguistici e Comparati, Università di Napoli L'Orientale, Italia
ademeo@unior.it; https://orcid.org/0000-0002-8596-5041

Bianca Maria De Paolis – Dipartimento di Lingue e Letterature Straniere e Culture Moderne, Università di Torino
biancamaria.depaolis@unito.it; https://orcid.org/0000-0001-7725-9617

Simon Devauchelle – Université Paris-Saclay, Francia
simon.devauchelle@universite-paris-saclay.fr; https://orcid.org/0009-0009-0956-405X

David Doukhan – Institut National de l'Audiovisuel (INA), Francia
ddoukhan@ina.fr; https://orcid.org/0000-0002-1645-7334

Anders Eriksson – Department of Linguistics, Stockholm University, Svezia
anders.eriksson@ling.su.se; https://orcid.org/0000-0002-6844-4834

Francesca Fontanella – Ricercatrice indipendente
francesca.fontanella9@gmail.com

Emanuela Gallo – Centro Interdisciplinare di Ricerca UrbanEco, Università degli Studi di Napoli Federico II, Italia
ems.gallo@gmail.com

Matteo Gay – Dipartimento di Studi Umanistici, Università di Pavia, Italia
matteo.gay01@universitadipavia.it; https://orcid.org/0009-0003-6066-4036

Barbara Gili Fivela – Dipartimento di Studi Umanistici, Università del Salento, Italia
barbara.gilifivela@unisalento.it

Martine Grice – University of Cologne
martine.grice@uni-koeln.de; https://orcid.org/0000-0003-4973-4059

Mirko Grimaldi – CRIL, Università del Salento, Italia
mirko.grimaldi@unisalento.it; https://orcid.org/0000-0002-0940-3645

Stella Gryllia – Centre for Language Studies, Radboud University, Paesi Bassi
stella.gryllia@ru.nl; https://orcid.org/0000-0002-8930-5833

Glenda Gurrado – Università degli Studi di Bari Aldo Moro, Italia
glenda.gurrado@uniba.it

Jonathan Harrington – Institute for Phonetics and Speech Processing, Ludwig-Maximilians, Universität München, Germania
jmh@phonetik.uni-muenchen.de; https://orcid.org/0000-0002-9035-0949

Na Hu – Centre for Language Studies, Radboud University, Paesi Bassi
na.hu@ru.nl; https://orcid.org/0000-0002-2941-2100

Sandra Madureira – Pontifical Catholic University of São Paulo, Brasile
sandra.madureira.liaac@gmail.com; https://orcid.org/0000-0001-8263-053X

Marta Maffia – Dipartimento di Studi Letterari, Linguistici e Comparati, Università di Napoli L'Orientale, Italia
mmaffia@unior.it; https://orcid.org/0000-0002-4913-374X

Daniela Mereu – Università di Torino
daniela.mereu@unito.it; https://orcid.org/0009-0005-1571-0587

Massimo Moneglia – Università di Firenze
massimo.moneglia@unifi.it; https://orcid.org/0000-0002-7125-8653

Lucas Ondel Yang – Université Paris Saclay, CNRS, LISN, Francia
lucas.ondel@cnrs.fr; https://orcid.org/0000-0003-4512-0471

Riccardo Orrico – Centre for Language Studies, Radboud University, Paesi Bassi
riccardo.orrico@ru.nl; https://orcid.org/0000-0001-9260-7210

Pascal Perrier – Pôle Parole et Cognition [GIPSA-PPC], Università Grenoble Alpes, Francia
pascal.perrier@grenoble-inp.fr

Duccio Piccardi – Dipartimento di Studi Umanistici, Università degli Studi di Urbino Carlo Bo, Italia
duccio.piccardi@uniurb.it; https://orcid.org/0000-0002-4985-4360

Sara Picciau – Facoltà di Scienze della Formazione, Libera Università di Bolzano, Italia
sara.picciau@unibz.it

Albert Rilliard – Université Paris Saclay, CNRS, LISN, Francia
albert.rilliard@lisn.fr; https://orcid.org/0000-0001-6490-2386

Josiane Riverin-Coutlée – Institute for Phonetics and Speech Processing, Ludwig-Maximilians, Universität München, Germania
josiane.riverin@phonetik.uni-muenchen.de; https://orcid.org/0000-0001-9131-9217

Christophe Savariaux – Pôle Parole et Cognition [GIPSA-PPC], Università Grenoble Alpes, Francia
christophe.savariaux@gipsa-lab.grenoble-inp.fr

Michelina Savino – University of Bari, Italia
michelina.savino@uniba.it; https://orcid.org/0000-0001-8794-3113

Simona Sbranna – University of Cologne
s.sbranna@outlook.com; https://orcid.org/0000-0001-6915-7047

Loredana Schettino – Facoltà di Scienze della Formazione, Libera Università di Bolzano, Italia
lschettino@unibz.com; https://orcid.org/0000-0002-3788-3754

Patrizia Sorianello – Dipartimento di Ricerca e Innovazione Umanistica, Università degli Studi di Bari Aldo Moro, Italia
patrizia.sorianello@uniba.it

Graziano Tisato – Associato di ricerca, CNR-ISTC, Italia
graziano@tisato.it

Vincenzo Vacchiano – Dipartimento di Studi Letterari, Linguistici e Comparati, Università di Napoli L'Orientale, Italia
vincenzo.android.sia@gmail.com

Alessandro Vietti – Facoltà di Scienze della Formazione, Libera Università di Bolzano, Italia
alessandro.vietti@unibz.it; https://orcid.org/0000-0002-4166-540X

Giovanni Vinciguerra – Università degli Studi di Bari Aldo Moro, Italia
giovanni.vinciguerra@uniba.it

Claudio Zmarich – Ricercatore Senior, CNR-ISTC, Italia
claudio.zmarich@cnr.it

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.

Valentina De Iacovo è ricercatrice di tipo A in Glottologia e linguistica presso il Dip. di Lingue e LS e CM dell'Università di Torino. Tra i suoi ambiti di ricerca ci sono lo studio fonetico e prosodico della variazione dialettale e regionale in Italia, dell'italiano come L2 e lingua ereditaria. È inoltre responsabile della base degli archivi vocali del sito del LFSAG.

Bianca Maria De Paolis è assegnista di ricerca e docente a contratto presso l'Università di Torino, dove insegna Linguistica Generale. Ha conseguito il dottorato in cotutela tra l'Università di Torino e l'Université Paris 8, con una tesi sulle variazioni prosodiche e sintattiche legate alla struttura informativa in italiano e francese L1 e L2. Attualmente la sua linea di ricerca prosegue nell'ambito della prosodia e dell'acquisizione delle lingue, anche grazie a un soggiorno come Junior Fellow presso il Collaborative Research Center "Prominence in Language" dell'Università di Colonia, previsto per il 2025.

Daniela Mereu ha conseguito il dottorato di ricerca in Linguistica presso le Università di Bergamo e Pavia. In seguito, ha lavorato come assegnista di ricerca presso la Libera Università di Bolzano. Attualmente è ricercatrice all'Università di Torino, dove insegna Linguistica Generale. I suoi interessi di ricerca vertono principalmente su temi legati alla sociolinguistica, alla fonetica e alla dialettologia, con una particolare attenzione nei confronti del sardo e dell'italiano. È autrice del volume *Il sardo parlato a Cagliari. Una ricerca sociofonetica*, pubblicato da FrancoAngeli nel 2019.

AISV - Associazione Italiana Scienze della Voce
sito: www.aisv.it
email: aisv@aisv.it| redazione@aisv.it
ISBN: 978-88-97657-73-6

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