DOI: 10.17469/O2106AISV000013

VIOLETTA CATALDO, LOREDANA SCHETTINO, RENATA SAVY,
ISABELLA POGGI, ANTONIO ORIGLIA, ALESSANDRO ANSANI,
ISORA SESSA, ALESSANDRA CHIERA

Phonetic and functional features of pauses, and concurrent gestures, in tourist guides' speech[1]

This study falls into the bigger framework of the CHROME project, addressing the definition and testing of a methodology of collecting, analyzing and modeling multimodal data for the design of virtual agents serving in museums. The paper analyses three tourist guides' speech, focusing on silent pauses and voiced pauses (filled pauses and segmental prolongations). In this regard, a description of phonetic-acoustic and functional features of pauses and a classification of concomitant gestures have been performed. Results show a) speakers' idiosyncratic linguistic behaviours; b) a clear distinction between silent pauses, mainly used for grammatical and intentional reasons, and voiced pauses that instead occur as ungrammatical and hesitation devices; c) such a distinction is confirmed by concomitant gestures, they are semantically loaded in silent pauses and semantically empty in voiced pauses.

Key words: Speech, Disfluencies, Pauses, Gestures, Virtual Agent.

1. *Theoretical background*

1.1 Disfluencies and pauses

Spontaneous human speech shows a widespread occurrence of a number of heterogeneous phenomena, which interrupt, suspend and/or delay the speech flow in the production of the intended message, affecting its fluency. For this reason, such phenomena are generally referred to as disfluencies.

Several studies, belonging to different investigation fields, have concerned disfluencies since the 1950s. Such studies start from different theoretical approaches, ranging from phonetics, speech pathology, psycholinguistics to automatic speech processing (Crocco, Savy, 2003); the coexistence of several theoretical perspectives have caused studies to use and provide different terminologies[2] and a variety of descriptions and classifications (Lickley, 2015)[3]. Despite this situation, some

[1] Authors' responsibilities – University of Salerno, speech analysis: *Violetta Cataldo*, linguistic analysis, writing; *Loredana Schettino*, linguistic analysis, related work; *Renata Savy*, study concept, supervision. University of Naples "Federico II": *Antonio Origlia*, automatic data analysis. Roma Tre University, gesture analysis: *Isabella Poggi*, writing, supervision; *Alessandro Ansani*, writing, data analysis; *Isora Sessa*, gesture coding, discussion; *Alessandra Chiera*, gesture coding, related work.

[2] Among others, "disturbances" (Mahl, 1956), "hesitations" (Maclay, Osgood, 1959), "hesitation phenomena" (Blankenship, Kay, 1964), "dysfluencies" (Johnson, 1961).

[3] A broad and highly regarded overview of disfluencies is provided by Eklund (2004).

agreement has been reached, as reference is made to Levelt's description of the general structure of disfluencies[4] (Levelt, 1983) and his model of speech production (Levelt, 1989), comprising the "double perceptual loop" theory of self-monitoring. Accordingly, speakers' perception of both the own produced speech (external loop) and speech plan (internal loop) enables them to detect and repair a problem before it is articulated. Hence, disfluencies in general trace back to either the external monitoring stage or the internal planning stage.

Disfluency phenomena expressly characterize oral texts as they are closely related to the speech modality; in fact, conditions of low degree of message design by the speaker, simultaneity of planning and production processes and regular concurrence of signals' emission and reception are likely to undermine the speakers' flow of speech (Voghera, 2017).

Fundamentally, the basic distinction regards two main categories[5]:

- **Hesitations**; phenomena of hesitations show a certain degree of speakers' uncertainty and are used as means for taking time to plan what follows in the message. Generally, this kind of disfluencies does not affect the verbal sequence of speakers' productions; instead, a range of phonetic cues, including silent pauses, filled pauses and prolongations, are employed.
- **Repairs**; speakers need to retrace something that has been already said and change it. It can happen for different reasons, all revealing a need for correction of errors that have occurred somewhere in the planning process. In this case, a change in the verbal sequence of the utterance occurs, since the speaker acts on something already uttered, repeating, replacing, adding, removing part of it; indeed, successful repair strategies consist of repetitions, substitutions, insertions, deletions.

Disfluency occurrences belonging to the hesitation category usually involve the temporary suspension of flowing speech (Lickley, 2015), which means that, in most cases, speakers hesitate by means of elements which enable a short-term delay to the production process; therefore, pauses are commonly employed as hesitation devices (Maclay, Osgood, 1959; O'Shaughnessy, 1992). Although pauses are part of disfluencies, as they affect to some extent the speech flow, a number of studies claim that these interruption phenomena need to be investigated from a "positive" approach. Starting by Chafe (1980), hesitations produced in spontaneous speech communication have not been considered anymore as merely disturbances or "errors" of human speech production and researchers in this field have increasingly acknowledged their role in dialogue. In fact, pauses represent the natural outcome of the condition of on-line planning process in human spontaneous speech (Giannini, 2003)

[4] On such account, Shriberg (1994) developed her own highly influential account on disfluencies, whose structure may consist of the following regions: Reparandum (RM), Interruption Point (IP), Editing Phase (EP), Repair (RR).

[5] Such distinction has been considered in a number of studies, to recall some of them: "stalls" (silent pauses, filled pauses, prospective repeats, syllabic prolongations) and "repairs" (false starts, retrospective repeats or bridging) in Hieke (1981); "forward looking disfluencies" and "backward looking disfluencies" in Ginzburg, Fernandez & Schlangen (2014); "hesitations" and "repairs" in Lickley (2015); "*disfluenze fonetiche*" (phonetic disfluencies) and "*disfluenze testuali*" (textual disfluencies) in Voghera (2017).

and a possible solution to in-time discourse planning (Clark, 2002). Using pauses, speakers manage to avoid producing performance errors and to achieve greater well-formedness in their speech (Hieke, 1981). More specifically, the insertion of such elements allows the speaker to gain extra time to retrieve content and provide the listener with valuable meta-information about the ongoing speech (Betz, Wagner & Voße, 2016; Betz, Carlmeyer, Wagner & Wrede, 2018).

In the literature, pauses have been classified and termed differently, according to different research approaches and criteria. The most recurring classification concerns unfilled (or silent) pauses and filled pauses.

Unfilled pauses consist in occurrences of silence of unusual length in the spontaneous speech flow (Maclay, Osgood, 1959); from the articulatory point of view, silent pauses can be produced together with other phenomena, comprising inspiration, swallowing, any laryngo-phonatory reflex, or a silent expiration (Zellner, 1994).

Filled pauses are usually subdivided into "unlexicalized" and "lexicalized" filled pauses; the former, e.g. "ah" or "uh", have been defined as seemingly meaningless words (Gabrea, O'Shaughnessy, 2000), as they are independent nonverbal elements, realized as vocalizations and/or nasalizations. Furthermore, several studies tend to include in this category phenomena of segmental prolongation (Giannini, 2003), while others consider them as a specific form of unfilled pauses (non-phonemic lengthening of phonemes; Maclay, Osgood, 1959). The category of lexicalized filled pauses refers to pauses that have lexical form, thus including different cases found in a number of languages, such as discourse markers (e.g. "you know"), repetitions, false starts (Zellner, 1994).

Researchers on conversational speech synthesis, looking for a model for synthetic disfluencies, considered "pauses" as one of the three micro-structural elements that may constitute disfluencies (Betz, Wagner & Schlangen, 2015): pre-disfluent syllable lengthening (L); cut-offs leading to word fragments (F); silent or filled pauses (P). According to their study, most speech disfluent occurrences are expressed via a combination of the above mentioned elements.

From a more closely phonetic perspective, researchers have expressed particular interest in identifying both segmental and suprasegmental features of pauses, especially of those vocalizations considered as filled pauses.

As regards segmental content, several studies concerning English spontaneous speech claim that filled pauses are typically produced as a steady mid-central vowel close to *schwa* (Maclay, Osgood, 1959; O'Shaughnessy, 1992; 1993; Gabrea, O'Shaughnessy, 2000; Schriberg, 2001).

In the literature, different studies have investigated the prosodic cues of filled pauses. Duration is one of the most noteworthy, as it succeeds in distinguishing filled pauses' segments from other similar vowels contained in unstressed words (such as "a" and "the") in spontaneous speech; in the first context, such vowels are regularly longer than in the second one (Shriberg, 2001). Additionally, specific intonational features have been identified as peculiar to filled pauses and the relationship with their prosodic surrounding (Shriberg, Lickley, 1993); filled pauses

have been described as showing a low fundamental frequency F_0 compared to the adjacent context and displaying an ongoing F_0 fall (O'Shaughnessy, 1992; Gabrea, O'Shaughnessy, 2000).

Researchers investigating filled pauses, have also observed segmental prolongations. They were found to be often concurrent phenomena preceding filled pauses (Betz, Wagner, 2016); moreover, both phenomena, in contrast to other disfluency types, show features of vocalization and duration as means for expressing hesitation (Eklund, 2004).

In general, segmental prolongations that can be considered as pauses occur in disfluent contexts; in this way, it is possible to distinguish disfluent prolongations from other cases of lengthening, due to accentuation or in utterance-final or prepausal position. Recent studies have shown that the preferred segmental targets for disfluent prolongations are long vocalic nuclei and sonorant codas (Betz, Eklund & Wagner, 2017). On the intonation level, hesitant prolongations tend to show a specific flat pitch contour compared to non-disfluent prolongations, which are realized with a higher pitch range (for example, due to accentuation; Betz et al., 2017). As regards duration, segmental prolongations are commonly shorter than filled pauses (Eklund 2001; 2004; Betz et al., 2017).

1.2 Gestures

Communication entails a complex interplay between speech and gesture. The study of their relationship is characterized by a debate between two competing hypotheses: according to the Lexical Retrieval Hypothesis (e.g., Krauss, Hadar, 1999; Krauss, Chen & Gottesman, 2000; Morsella, Krauss, 2005) they fulfil different functions, with gestures mainly supporting the message encoding or having a compensatory role; a second hypothesis ascribes to gestures functions similar to those of speech (e.g., Kita, Özyürek, 2003; Kendon, 2004; McNeill, 2005). Support to the first view is given by a more frequent occurrence of gestures in disfluencies, their preparing language in infants (Liszkowski, 2008) and their helping word retrieval (Pine, Bird & Kirk, 2007).

Conversely, some models observe that gestures occur more often in absence of disfluencies (Christenfeld, Schachter & Bilous, 1991) and their strokes co-occur with prosodic peaks (Nobe, 2000), suggesting that speech and gesture are integrated systems playing similar pragmatic functions. For example, the Information Packaging Hypothesis (e.g., Alibali, Kita & Young, 2000; Kita, 2000) holds that gesture and speech are intertwined from early conceptual elaboration, and gestures provide help for thinking and then speaking; indeed, people gesture more frequently when facing a high conceptualization load (Kita, Davies, 2009; Kita, Alibali & Chu, 2017). Other proposals, inspired by McNeill (1992, 2005), argue that speech and gesture form a unique system where the propositional and the mental-image aspects of thought are linked together: they have a similar pattern of development in childhood (e.g., Capirci, Volterra, 2008) and different languages (Kita, 2009), they are synchronized in a semantic harmony in both production and comprehension (e.g.,

Kendon, 2004; Holler, Schubotz, Kelly, Hagoort, Schuetze & Özyürek, 2014), and are similarly affected by neurocognitive impairments (Duncan, Pedelty, 2007).

Since assumptions about the timing and functions of gesture in pauses (a crucial topic of this work) are opposite in the two frameworks above, their plausibility may be tested exploring the relationship between gesture production and disfluencies.

In the few studies exploring gestures accompanying speech pauses (e.g., Esposito, Marinaro, 2007; Stam, Tellier, 2017; Graziano, Gullberg, 2018; Krauss et al., 2000; Morsella, Krauss, 2005), data are inconsistent. As to their temporal relation, for some authors gestures occur just before or at the same time as disfluencies (Ragsdale, Fry Silvia, 1982), while for others speech and gesture are interrupted simultaneously (Mayberry, Jacques, 2000) or gesture stops before speech stops (Seyfeddinipur, 2006). Some find that disfluencies are specifically synchronized with gestures holds, i.e., the momentary suspension of movement, in both children and adults (Cibulka, 2016; Esposito, Marinaro, 2007): like speech pauses may be involved in the processes necessary to repair a problem in speaking, holds too may signal the activation processes to handle the same problem and re-plan a new message.

As for the function of gestures in pauses, along with a production-oriented function of lexical retrieval, an interactive role of managing turn-taking (Mondada, 2007) and a comprehension-oriented function of adding useful information for the interlocutor mainly in asymmetrical interactions (i.e., doctor-patient) are highlighted (Stam, Tellier, 2017). Graziano and Gullberg (2018) provide further evidence that speech and gesture form an integrated system by examining adult native speakers of two languages and language learners: gestures occur more in fluent than disfluent speech and, in the rare cases of strokes found in pauses, they belong not only to referential but also pragmatic gestures with an interactive function. Moreover, all participants tend to suspend or hold gestures in disfluency, showing that when speech stops, so does gesture (Yasinnik, Shattuck-Hufnagel & Veilleux, 2005).

2. *Object of the study*

The present study is part of the CHROME project – Cultural Heritage Resources Orienting Multimodal Experience, which is aimed at defining and testing a methodology of multimodal data collection, analysis and modeling for the development of a Virtual Agent (VA); such a VA should be able to serve in museums and present cultural sites using an "anthropomorphic" human-machine dialog system. The purpose of the research focuses on this last feature, providing an accurate description and modeling of a number of specific features of verbal behavior, which contribute to speech naturalness and adequacy to its situational context. In fact, VAs usually speak in a fully fluent way, giving the interlocutors a feeling of artificiality and distance; actually, VAs are expected to emulate human-human communication, which, however, shows a series of phonetic and prosodic characteristics that cannot be neglected. In this perspective, phenomena of hypospecification, coarticulation and disfluency need to be investigated on the positive side, looking for characteristics

and regularities, which can be effectively implemented in "text to speech" (TTS) synthesis in order to improve its performances. In the framework of the CHROME project, this work aims at: a) describing and analyzing a set of speech disfluency phenomena occurring in tourist guides' speech, focusing on three different types of pauses: silent pauses, filled pauses, vocalized filled pauses; b) researching patterns and regularities of such phenomena to be implemented in TTS system; c) investigating the existing relationship between pauses and concurrent gestures.

3. *Methodology*

3.1 Corpus and dataset

The present research focuses on a limited dataset drawn from the whole corpus collected for the CHROME project (Origlia, Savy, Poggi, Cutugno, Alfano, D'Errico, Vincze & Cataldo, 2018). The corpus consists of audiovisual recordings of guided tours led in Italian at the San Martino Charterhouse in Naples. Three female expert tourist guides accompany small groups of visitors in four guided tours, each of approximately an hour, for a total amount of about 3h30' of speech for each guide. The guided tours take place in six "points of interest" (POIs) of the Charterhouse: *pronaos*, great cloister, parlor, chapter hall, wooden choir, treasure hall[6]. The dataset of this study takes into account one POI of one visit led by each of the three expert guides, amounting to 36'88" of speech; the first POI, namely the *pronaos*, represents the opening moment of the guided tour, which starts at the doorway to the church of the Charterhouse.

From a linguistic point of view, the conditions set by the corpus identify a particular kind of oral texts, typical of tourist guides' oral performances. The guides' speech is characterised by a) a high degree of discourse planning, in fact, tourist guides make regular use of descriptive texts, which are partially pre-structured and repetitive; b) high selective attention of the participants, because of the hierarchical relationship between the guide and the audience, due to the guide's professional and linguistic competence in the topic; c) a resulting low degree of interlocutors' dialogic interaction and participation in the discourse construction; d) a close integration between verbal and non-verbal elements, due to the spatial context which, for example, makes it necessary to use spatial, verbal and gestural, deixis. Given the above features, it can be assumed that this study deals with semi-spontaneous and semi-monological speech.

3.2 Method

3.2.1 Disfluency annotation system

The three selected audiovisual recordings have been annotated on different linguistic levels.

[6] Detailed data collection protocols are provided by Origlia et al. (2018).

Firstly, orthographic transcription and phonetic and syllabic annotation have been carried out (see Origlia et al., 2018) using the software Praat (Boersma, Weenink, 2018).

As concerns disfluency phenomena, an *ad hoc* encoding system has been adopted, based on the disfluency modeling works of Hieke (1981), Shriberg (1994) and Lickley (1998). The system consists of four annotation tiers, which, although parallel to each other, refer to different occurrence domains. For this reason, disfluency phenomena have been annotated using the ELAN software (2018), which allows to carry out a multilevel annotation. A comprehensive description of the disfluencies' four domains and their relative annotation tiers follows.

- **Disfluency Type;** this first annotation level specifies the category of the occurring disfluent phenomenon. A finite number of categories has been selected, in the attempt to cover all the possible occurrences:
  - *Fresh Start*. It involves cases of false starts, when the speaker stops and rephrases a new utterance with no morphosyntactic and/or semantic relationships to the previous one.
  - *Repeated Start*. In the utterance, the speaker stops and exactly repeats something s/he has already uttered; such category deals with repetitions of single words, fragments of words, as well as whole utterances.
  - *Edited Start*. This category is divided into two subcategories; in both cases the speaker rephrases part of the string, either adding one or more elements (Addition subcategory), or substituting an element with another syntactically and/or semantically equivalent one (Substitution subcategory).
  - *Hesitative Start*. In this case, the speaker shows and produces a hesitation without repeating, substituting or abandoning any part of the utterance in production.
- **Disfluency Function;** this level provides information about the pragmatic function performed by each disfluent phenomenon. Such a pragmatic description specifically refers to Hieke's classification (1981) into retrospective and prospective disfluency phenomena. Accordingly, each disfluency type is associated with either a retrospective function, which plays a corrective role when the speaker "corrects" something s/he has already uttered, or a prospective function, which plays a control role for the speaker's production to avoid errors.
- **Disfluency Model;** this annotation level describes the occurrence model of disfluencies, based on Shriberg (1994) and Lickley (1998). Different regions are identified and annotated, namely:
  - *Reparandum*, the region where the speaker encounters difficulties and that will be later "repaired";
  - *Interruption Point*, a temporary suspension, where the speaker stops as s/he realizes to have difficulties;
  - *Interregnum*, a moment of transition where the speaker shows his/her hesitation;
  - *Repair*, the region that "repairs" to the Reparandum;

  - *Original Utterance*, the region that precedes the Interregnum in the case of the disfluencies that have no Reparandum (such as Hesitative Start);
  - *Continuation*, the region that follows the Interregnum in the case of the disfluencies that have no Repair (again, in the case of Hesitative Starts).
- **Disfluency Components;** based on the regions' identification in the previous level, here information about the inner phonic and/or linguistic components of such regions are provided.

A summary table for the disfluencies' annotation levels and the relative categories and labels is provided below (Table 1).

Table 1 - *Disfluency annotation system: levels of annotation and categories*

Levels of annotation	Categories
Disfluency Type	Fresh Start, Repeated Start, Edited Start Addition, Edited Start Substitution, Hesitative Start
Disfluency Function	Retrospective function, Prospective function
Disfluency Model	Reparandum, Interruption Point, Interregnum, Repair, Original Utterance, Continuation
Disfluency Component	Word, Word Fragment, Filled Pause, Vocalized Filled Pause, Discourse Marker Filled Pause, Silent Pause

3.2.2 Classification and analysis parameters of pauses

The present research primarily focuses on the analysis of pauses, classified as follows: a) Silent Pauses (SPs), pauses of silence; b) Filled Pauses (FPs), regarded as vocalizations and/or nasalizations, i.e. "eeh", "ehm", "mhh"; c) Vocalized Filled Pauses (VFPs), resulting from word-final segmental prolongation of lexical elements.

Pauses have been investigated from both a form and function perspective, in order to provide both a phonetic-acoustic description and a functional classification.

As concerns the phonetic-acoustic description, pauses' analysis has been conducted according to the following parameters:

- *Duration*; measured in milliseconds (ms).
- *Segmental content*; pauses' phonetic realizations and their characteristics have been investigated, considering their vowel quality and phenomena of diphthongization, triphthongization, nasalization, devocalization.
- *Pitch profile*; it provides a description of pauses' intonation patterns, classified as rising, falling, flat or valley. Additionally, their relative range level has been calculated in Herz (Hz) and semitones (ST). Both parameters refer to pauses compared to their cotextual sequence.

The occurrences of the three pause types have been extracted and classified according to the following primary functions:

- *Physiological function* (PHYS): it follows the "respiratory" function of Viola and Madureira (2008), which reflects the speaker's physiological need to pause and take a breath;
- *Demarcative function* (DEM): demarcative pauses play a grammatical role in structuring the discourse at different linguistic levels, such as intonation, syntax, information structure (Swerts, 1998);
- *Programmative function* (PROG): these pauses show the speaker's on-line process of planning and his/her difficulty in retrieving specific lexical elements (see Schnadt, Corley, 2006; Hartsuiker, Notebaert, 2009);
- *Hesitative function* (HES): pauses are widely employed to express the speaker's uncertainty over the general planning of the content as well as the form of his/her message ("unintended hesitation pauses", O'Shaughnessy 1992; "means of hesitation", Eklund 2001);
- *Strategic-rhetorical function* (STR-R): the speaker makes use of pauses in a partially conscious and deliberate way, in order to attract the audience's attention (Betz et al., 2016), introduce or underline key concepts (Duez, 1997), give particular emphasis on specific words (Strangert, 2003).

3.2.3 Gesture annotation system

To assess the relationships between gestures and pauses, among annotation systems proposed by previous works (see for instance Kong, Law, Kwanm Lai & Lam, 2015) the guides' gestures concomitant to pauses were annotated in terms of the following categories (Table 2), partly drawn from Colletta, Kunene, Venouil, Kaufmann & Simon (2009), but adapted according to Poggi (2007).

- **Gesture functions:** each gesture was classified in terms of these categories: *deictic*; extended index, thumb or the whole hand point at some place in the physical context where the referent of discourse presently is or can be connected to; e.g., while saying "*come dicee... la, laa... dicitura stessa del museo*" (as said by theeee... very wording of the museum), during the VFP *"laa..."*, the guide *moves both hands, with palms up, downward and rightward*, pointing at words written on the Museum entrance.
  - *Iconic;* the shape or movements of the hand(s) imitates the shape or movements of the referent. During a pause, before saying "*si prolunga*" (it is prolonged), the guide *moves her left hand, palm down, with almost closed fingers from right to left* as if picking and dragging something in a long fluctuating line, to represent something going through a long path.
  - *Metaphoric*; an iconic gesture refers to some abstract concept, or some inference is required to go from the bare imitation of a shape or movement to the intended meaning. While saying "*il certosino doveva preservare la sua vita isolata, contemplativa*" (the Chartusian had to preserve his isolated, contemplative life), the guide *pulls back her hands with palms forward*, to represent someone's withdrawing from something, which metaphorically means to retreat from life in the outside world.

  - *Coded*; this category includes not only symbolic gestures, those with a codified meaning and a shared verbal translation in a given culture (like *thumb and index making a ring* for "ok"), but also other gestures, without a straightforward verbal paraphrase, which yet do convey a shared meaning (Müller, 2004; Kendon, 2004; Nobili, 2019). The guide, while saying *"perché abbiamo... a cuore la conservazione del pavimento"* (because we care about the conservation of the floor) during the disfluency makes the *Palm Up Open Hands* gesture, which means "this is self-evident".
  - *Beat*; rhythmic gestures, with hands generally with a up-down movement, that emphasize a word in a sentence or a syllable in a word in order to highlight its importance, or scan the rhythmic structure of words for clearer articulation.
  - *Manipulator* (Ekman, Friesen, 1969); hands smoothing or rubbing between themselves or with other parts of the speaker's body, generally aimed at a reassuring self-contact, hence only indirectly being a cue to embarrassment or discomfort;
  - *Idle*: a "non-gesture", with hands not at rest.
- **Gesture phases**: following Kendon (1980) and Kita (1990), the phases of the gestures were annotated as:
  - *Preparation*; when the hand starts from its resting position;
  - *Stroke*: when it reaches its farthest point from the resting position, after which it starts to go back to it;
  - *Hold*: the hands remain on the stroke point before going back to the resting position;
  - *Chain*: the hand, after reaching the stroke point, starts retracting but then repeats its run to the stroke and back;
  - *Return*: after reaching the farthest point the hand goes back to the resting position.
- **Gesture meaning**: for communicative gestures (e.g., not idles), a verbal paraphrase of it is annotated: e.g. *Palm Up Open Hand* = "this is self-evident".
- **Synchrony**: based on the meaning attributed to the gesture, its semantic relationship is annotated with the co-occurring words or pause, with respect to which it can be anticipating, following, or synchronous. When the guide says *"come dicee... la, laa... dicitura stessa del museo"* (as said by theeee... very wording of the museum), although referring to the wording (*dicitura*), her deictic gesture falls during the VFP following "laa...", thus anticipating the referent it points at.

Table 2 - *Gesture annotation system: levels of annotation and categories*

Levels of annotation	Categories
Gesture Function (RHGF)	Deictic, Iconic, Metaphoric, Coded, Beat, Manipulator, Idle
Gesture Phases (RHGP)	Preparation, Stroke, Hold, Chain, Return
Gesture Meaning (Meaning)	"....."
Synchrony (RHGS)	Anticipating, following, Synchronous

The annotations were performed by two expert judges for pauses (Cohen's *k* of 0,7) and two for gestures (Cohen's *k* of 0,7).

4. *Results*

4.1 Linguistic analysis

The following section presents the research results, which emerged from the speech analysis. Firstly, general data on pauses' occurrence (Table 3) and incidence in the total speech of the dataset (36'88" of speech) are reported.

On a total of 384 pauses, occurrences of SPs exceed those of both FPs and VFPs, which present respectively 103 and 101 occurrences.

As regards incidence data, pauses, independently of the type of pause, register a per-word rate of 0,07 and a per-minute rate of 10,4, meaning that speakers produce a pause per about 14 words. More specifically, SPs register a per-word rate of 0,03 (a pause per about 30 words), FPs of 0,02 (a pause per about 53 words), and VFPs of 0,02 (a pause per about 54 words).

Table 3 - *Number of pauses' occurrences per pause type (SP, FP, VFP) and per speaker (G01: first guide; G02: second guide; G03: third guide)*

	G01	G02	G03	tot
SP	129	28	23	180
FP	82	14	7	103
VFP	72	27	2	101
tot	283	69	32	**384**

4.1.1 Idiosyncratic linguistic behaviour

Once reported the overall occurrence and incidence data of the three speakers of the dataset, it seems particularly noteworthy to examine speakers' individual speech in greater detail. In fact, it emerged that the three guides adopt different linguistic

behaviors, as it can be seen from the different number of pauses uttered by each speaker (Table 3).

The first guide (G01) makes use of a higher number of pauses compared with the other two guides. She reports a per-word rate of 0,19 and a per-minute rate of 24,5, i.e., a pause per only about 5 words. Differently, the second guide (G02) produces pauses four times less often than G01, with a per-word rate of 0,03 (a pause per about 33 words) and a per-minute rate of 5,1. Ultimately, pauses in the third guide's (G03) speech register a definitely lower incidence: a per-word rate of 0,02 and a per-minute rate of 2,7.

4.1.2 Phonetic-acoustic features

Firstly, duration values of the three types of pauses are presented (Table 4).

Table 4 - *Duration values (ms) per pause type (SP, FP, VFP) and per speaker (G01, G02, G03)*

	G01		G02		G03		all speakers	
	mean	st.dev.	mean	st.dev.	mean	st.dev.	mean	st.dev.
SP	371	0,41	750	0,88	285	0,19	469	0,49
FP	486	0,33	250	0,15	360	0,23	365	0,23
VFP	253	0,12	260	0,08	93	0	202	0,07

SPs are the type of pauses which have the longest mean duration (469 ms); SPs show quite high deviation standard values among the three speakers; this can be explained by the fact that very variable lengths have been found, as SPs can be stretched in time without appearing detrimental for the conversation. FPs show a mean duration of 365 ms, with lower values of intra- and inter-speaker standard deviation. VFPs have a mean duration of 202 ms, resulting in a less variable value (standard deviation of only 0,07). This seems to happen for two main reasons. On the one hand, G02 and G03 realize very few VFP occurrences compared to G01; consequently, duration values appear more homogenous, as they specifically refer to VFP occurrences in almost one speaker's speech, having an effect on the overall standard deviation value. On the other hand, it seems that VFPs, unlike SPs and FPs, cannot be excessively stretched in time, without risking to appear deviant for the conversation (d'Urso, Zammuner, 1990). The overall duration values of our study follow the same pattern found by Eklund (1999), namely VFPs < FPs < SPs[7].

The parameter of segmental content is obviously considered only for the two types of voiced pauses, namely FPs and VFPs (Table 5).

[7] In Eklund (1999): PRs < FPs < UPs (prolongations < filled pauses < unfilled pauses).

Table 5 - *Segmental content per pause type (FP, VFP)*

	FP		VFP	
	n.occ.	% occ.	n.occ.	% occ.
vowel	77	75%	95	94%
consonant	11	11%	1	1%
vowel + consonant	15	15%	1	1%
consonant + vowel	0	0%	4	4%
TOT	103	100%	101	100%

FPs and VFPs occur with almost the same number of occurrences in the dataset. With regard to their segmental content, both types of pauses are realized with an entirely vocalic content (75% of FPs and 94% of VFPs).

More specifically, more than half of FP occurrences (62%) is realized by a *schwa* (in a few cases, a feature of nasalization has been found, 8 cases). The other cases of fully vocalic FPs are realized by alternatives of mid-frontal vowels, namely [e] or [ɛ], or diphthongs in which one of the two vowels is a *schwa* (however, it happens in very few cases, 7 occurrences). As to other segmental realizations of FPs, all consonant cases present the nasal [m] (11 occurrences), whereas the cases of vowel-consonant sequence are realized by a *schwa* followed by the nasal [m] (15 occurrences). No cases of FPs expressed by a consonant-vowel sequence have been found.

Concerning voice quality, it is worth noticing that 67% of FPs is realized with creaky phonation, regardless of FP segmental content. This seems to be a regular feature of FPs (see also Schriberg, 2001), probably caused by a lesser articulatory effort by the speaker in producing non-lexical elements.

Detailed data on VFP segmental content are reported below. As mentioned, the vast majority of VFPs has vocalic content. As opposed to FPs, VFPs occurrences present a wider range of vocalic alternatives, in order of frequency: 27 cases of [a]; 16 cases of [ɛ]; 14 cases of [ɔ]; 10 cases of [i]; 4 cases of [e]; 1 case of [o]; 1 case of [u]; no cases of [ə]. Moreover, 21 occurrences of diphthongs and one occurrence of triphthong have been found. Such a heterogeneity can be explained considering that VFPs are vocal prolongations of the word-final sound; as Italian syllables show a consonant-vowel (CV) canonical structure (Berruto, Cerruti, 2011), VFPs mainly deal with prolongations of different vowels. Other possible VFP realizations concern consonant lengthening: such prolongations are realized producing a final *schwa* in order to lengthen the consonant, except for nasal prolongations (1 occurrence of prolongation of [m]). As an outcome of coarticulation, 37 vocalic VFPs out of 95 are affected by nasalization due to their nasal acoustic context; for instance:

tradizione<ee> napoletana

[tradit,tsjonẽ:napole'tana]

As to intonation features of voiced pauses, fundamental frequency (F_0) profile and range level have been considered. Table 6 shows the occurrences of voiced pauses

according to their F_0 profiles. One occurrence of both FP and VFP, totally realized with creaky voice, did not allow to clearly identify the actual intonation pattern; such cases are ignored in the analysis and reported as "not categorized".

Table 6 - *Pitch profiles per pause type (FP, VFP)*

	flat	rising	falling	valley	n.c.	tot
FP	43	20	8	31	1	103
VFP	68	7	11	14	1	101

As it can be observed, the two types of pauses show similar pattern behaviors: the greater part is realized with a flat pitch profile with respect to the preceding and following strings; in particular, VFPs register a higher number of flat occurrences (FPs: 43 cases; VFPs: 68 cases). Moreover, in both cases, few occurrences of rising and falling profiles have been found. Despite such similarities, what emerges is that FPs show a considerable amount of valley profiles (31 occurrences). In these cases, the pauses' intonation pattern is steadily overall flat, but the pitch is lower than the global F_0 pattern of its co-text.

Such a difference might be explained by the pauses' intrinsic characteristics. The majority of steady and flat F_0 profiles of VFPs suggests that such pauses, being word prolongations, are naturally performed as continuations of the lexical elements they lengthen on the segmental level. Conversely, FPs are not integral part of words, and behave as separate elements; hence, FPs are more likely to be produced with freestanding realizations, often showing a tonal trough relative to the co-text and appearing more embedded in it at the intonation level.

Range level values seem to confirm the difference between the two types of pauses (Table 7). In fact, VFPs show a mean F_0 value of 189,4 Hz, only 0,8 ST lower than their mean co-text (198,9 Hz) against the 1,4 ST of FPs. Such a higher difference reflects the number of valley pitch profiles. In this regard, a statistical analysis of range values of FPs and VFPs has been carried out; although it did not result significant, it indicated a general tendency towards a lower and more embedded F_0 for the FPs.

Table 7 - *F0 values of FPs and VFPs (per speaker: G01, G02, G03) and of their relative co-texts*

	FPs				VFPs			
	G01	G02	G03	mean	G01	G02	G03	mean
pause F_0 (Hz)	196,2	196,6	184,6	192,5	194,0	192,6	181,5	189,4
cotext F_0 (Hz)	209,9	211,3	206,7	209,3	199,7	201,9	195,0	198,9
difference (Hz)	13,7	14,7	22,1	16,8	5,7	9,4	13,5	9,5
difference (ST)	1,2	1,2	2,0	1,4	0,5	0,8	1,2	0,8

4.1.3 Functions

The functional classification aims at identifying the primary functions (described in § 3.2.2) performed by each pause occurrence. It is important to highlight that pauses have been considered as multifunctional elements, in fact, each pause occurrence is likely to simultaneously carry out more than one function. This allows for cases of overlapping of two or more functions fulfilled by a single pause. By implication, Table 8 shows the actual number of functions' occurrences per pause type.

Table 8 - *Occurrence number of functions per pause type*

	PHYS	DEM	PROG	HES	STR-R
SP	9	95	0	22	114
FP	0	18	52	99	3
VFP	0	11	66	99	8

An overall look at the pauses' functional behavior suggests that the three pause types can be considered as two macro categories: in fact, SPs systematically behave differently from voiced pauses, namely both FPs and VFPs.

Firstly, SPs are the only pause type to perform a PHYS function, fulfilling the speaker's respiratory need (9 occurrences). In this regard, it has to be noted that cases of breaths have not been taken into account as SP occurrences. On the contrary, no SP occurrence performing a PROG function has been found; indeed, such a function seems to be properly performed by voiced pauses. What appears quite remarkable data is that SPs show a considerable number of DEM e STR-R functions (95 and 114 occurrences, respectively). Pauses performing DEM functions assume the grammatical meaning of sectioning utterances at different linguistic levels; the majority of DEM SPs (65 out of 95) serves simultaneously as grammatical boundaries of constituents of syntactic structure, intonation units, and information structure. The following example shows a SP occurrence performing a DEM function according to the three linguistic levels mentioned above:

> *la Certosa di San Martino qui a Napoli ha almeno due anime* <SP> *una racconta la storia dei certosini.*
>
> the San Martino Charterhouse here in Naples has two souls <SP> one tells the story of the Carthusian monks.

SPs result to be extensively used for strategic goals (63%).

As already said, voiced pauses exhibit a similar behavior, as FPs and VFPs tend to share the same distribution of functions: no occurrences of PHYS function, rare occurrences of STR-R function (3 and 8 cases respectively), few occurrences of DEM function (18 and 11 respectively), common occurrences of PROG function (52 and 66 respectively) and a considerable number of HES occurrences (99 in both cases). More specifically, almost all occurrences of DEM and PROG functions overlap with the most performed HES function. Such a situation might be explained by the

fact that both types of voiced pauses are widely used by the speakers as a means of expressing their indecisiveness about the general message to convey. Both FPs and VFPs register almost the totality of HES function, although showing high multi-functionality.

What is particularly worth noticing is the total overlapping between PROG and HES functions; more specifically, all the pauses carrying out a PROG function have been at the same time considered as performing a HES function. In general, HES pauses reflect the speaker's planning phase and the possible difficulties in dealing with general planning demands. In almost half of these cases, corresponding to PROG pauses, such a planning phase concerns expressly lexical elements. In this perspective, the occurrence of PROG pauses reveal the ongoing of a Word Searching (WS) process: the speaker makes use of pauses when she is struggling with lexical retrieval. WS phenomenon is more likely to occur when what the speaker is trying to retrieve are lexical elements with lower frequency (Lickley, 2015). The non-linguistic task requested to the speakers of the present study affects in this regard their lexical choices: the three tourist guides resort to a great number of technical terms belonging primarily to architectural, artistic, religious specialized languages in order to describe the Charterhouse and its history[8]. Such technical terms are lexical items, which show lower frequency than others, including features of lower name agreement and familiarity in the communicative exchange between the speaker and her audience. These characteristics entail a greater deal of effort for accessing to these words in both production and perception. The following example shows the presence of a FP just preceding the technical term *cappella* ("chapel"):

poi c'è anche una <FP> *cappella dedicata a San Giovanni Battista*

furthermore, there is a <FP> chapel consecrated to St. John the Baptist

The FP performs simultaneously a HES function in a broad sense and a PROG function in a narrow sense, working as a WS tool.

4.2 Correlation

4.2.1 Correlation between phonetic-acoustic features and functions of pauses

This part of results mainly concerns FPs and the functional distinction between mere HES FPs and HES FPs working as WS devices (WS FPs). This is supported by the parameters taken into account for the phonetic-acoustic analysis; the parameters of duration and range level seem the most interesting. It should be considered that statistical analysis concerned data relative to the three guides; however, most

[8] An analysis of the lexicon of the whole oral corpus collected in the CHROME project has been carried out (Senigalliesi, Sparano, Schettino, Savy, Dell'Orletta, Lubello & Basile, 2018). The lexicon of the oral texts results to be composed as follows, according to the GRADIT (De Mauro, 1999): 41% of *lessico fondamentale*, 16% of *alto uso*, 4% of *alta disponibilità*, 39% of *tecnicismi* (technical terms). More than half of the technical terms belong to the architectural field.

FP occurrences were produced by G01 (§ 4.1.1). Hence, results regarding the phonetic-functional correlation mainly trace back to one guide's linguistic behavior.

As concerns the first parameter, duration values correlate with the function carried out by FPs. WS FPs are twice longer than HES FPs (mean values: 590 ms and 300 ms, respectively); such a difference results to have statistical significance ($p < 0.001$)[9].

With regard to range level, FPs show a number of flat or valley intonation realizations and low fundamental frequency, thus occurring as tonal troughs within the neighboring speech (§ 4.2.1). On closer inspection, the range level parameter emphasizes the functional distinction. In fact, WS FPs are realized with even lower range values compared to their co-text (WS FPs: 1,6 ST vs. HES FPs: 0,9 ST). By implication, FPs carrying out a HES function tend to preserve the mean range level of the surrounding strings; on the other hand, WS FPs are more embedded within the co-text.

4.2.2 Frequency of gestures during pauses

Concerning the gestures co-occurring with pauses, we only considered the pauses longer than .20 sec. The remaining pauses were 223: 101 SPs (45.29%), 57 VFPs (25.56%), and 65 FPs (29.15%). A chi-square test for the distribution of gesture types among the pause types is significant for $p < .10$ (Table 9, 10).

Table 9 - *Chi square test (gestures types per pause type)*

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	18,656a	12	,097
Likelihood Ratio	21,337	12	,046
N of Valid Cases	223		

a. 8 cells (38,1%) have expected count less than 5. The minimum expected count is ,77.

Table 10 - *Distribution of gestures among pauses*

		Coded	Deictic	Iconic	Metaphoric	Beat	Manipulator	Idle
SP	n.	32	10	5	2	14	6	32
	%	31,68	9,9	4,95	1,98	13,86	5,94	31,68
VFP	n.	20	2	1,98	0	4	7	24
	%	35,09	3,51	13,86	0	7,02	12,28	42,11
FP	n.	20	3	5,94	1	6	12	23
	%	30,77	4,92	31,68	1,54	9,23	18,46	35,38

As already found by Graziano and Gullberg (2018), "when speech stops, gesture stops"; our first robust result is that the idle gestures are the most frequent in all

[9] Linear regression: dependent variable=duration; independent variable: function (2 levels: WS, HES). FuncWS_t-value: 5.21; p-value < 0.001.

three pause types, with the highest value in correspondence with VFPs (42.11%). Moreover, a clear difference emerges between the frequency patterns of gesture categories in SPs, on the one side, vs. FPs and VFPs on the other. The respective patterns are represented in Table 11.

Table 11 - *Ranking of gesture functions in pauses*

	1	2	3	4	5	6
SP	idle & coded	beat	deictic	manipulator	iconic	metaphoric
VFP	idle	coded	manipulator	beat	deictic	
FP	idle	coded	manipulator	beat	deictic	metaphoric

While *idles* and *coded* have the same frequency in SPs, in both VFPs and FPs *idles* are slightly more frequent than *coded*, while *manipulators* are more frequent than *beats* and *deictics* in Voiced Pauses (FPs and VFPs), and the opposite is the case in Silent pauses.

Starting from gesture categories, the percentage of *coded* gestures is almost similar across pause types; *deictics* mostly occur in correspondence with SPs, probably because they are usually performed immediately after uttering a locative adverb or a deictic pronoun, in silence, by pointing to a given object. *Iconic* gestures only show during SPs, perhaps because in these cases the speaker helps herself to recall a specific word to describe an item by imitating the item shape. As for *metaphoric* gestures, the occurrences are too poor to be commented upon. *Beat* gestures are much more frequent in SPs than FPs and VFPs.

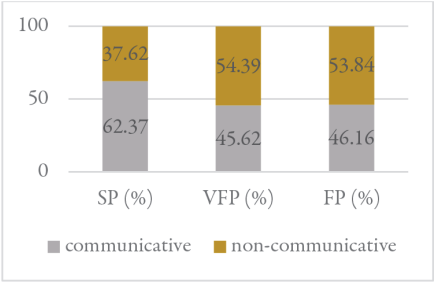
The remaining two categories show an inverse pattern of frequency: *manipulators* are more frequent in FPs, followed by VFPs, maybe because they are performed, at a low level of consciousness, in order to self-assure oneself; and *idles* are more frequent in VFPs than FPs, only finally followed by SPs.

How can we explain these opposite patterns for the first five categories as opposed to the last two? As argued in Origlia, Savy, Cataldo, Schettino, Ansani, Sessa, Chiera & Poggi (2019), these categories differ as to their communicative import. Four of them – *deictic, iconic, metaphoric, coded* – are definitely communicative, since in performing them the speaker has a conscious intention to convey some meaning, while the last two – *idle* and *manipulator* – are non-communicative: an *idle* is a non-gesture, simply resting hands, while a *manipulator* may sometimes provide information (embarrassment, anxiety) to an observer, but the Sender does not have the goal for it to leak out: so it may be informative but not communicative. Finally, a *beat* may be communicative if governed by a specific goal of emphasizing something, but sometimes it simply accompanies the rhythm of speech by synchronous body movements, thus making it easier for the Speaker to impress it the right temporal structure.

Based on a chi squared test, the distribution of communicative and non-communicative gestures across pauses is $c^2(2, N = 223) = 6.03, p = .049$ (Graph 1). Such

a distinction between communicative and non-communicative movements might account for the opposite pattern of Silent versus Voiced Pauses, with the communicative ones more often co-occurring with the former, and the non-communicative with the latter. Silent Pauses thus look as a moment of higher communicativeness, while Voiced ones as mainly of use for the Sender, less Addressee-oriented.

Graph 1 - *Communicative and Non-communicative gestures across pauses*



4.2.3 Relation between gestures and pauses functions

Three more analyses of HES, STR-R and PROG (WS) can deepen the relationship between pauses' functions and co-occurrent gestures' communicativeness (Table 12). Three 2-sided Fisher's Exact tests were run using a 2 x 2 contingency table. One for *HES x Communicativeness* (p = .007) shows that communicative gestures are less frequent, while non-communicative ones more frequent during hesitative pauses; one for *STR-R x Communicativeness* (p = .015) shows an opposite distribution: communicative gestures co-occur more with Strategic-Rhetorical pauses, non-communicative ones with non-strategic pauses. For *WS x Communicativeness* (p = .029), like with HES pauses, no-WS pauses present a higher number of communicative gestures, whereas those during WS pauses in the vast majority do not have a communicative function.

Table 12 - *Distribution of communicative and non-communicative gestures across pauses' functions*

	No-HES	**HES**	**No-STR-R**	**STR-R**	**No-WS**	**WS**
communicative	61,24%	42,55%	48,85%	69,39%	57.74%	40.00%
non-communicative	38,76%	57,45%	51,15%	30,61%	42.26%	60.00%

4.2.4 Amount of Movement (AoM)

To assess if the occurrence of pauses is associated with stops in hand movements, we computed the Amount of Movement (AoM) in the presenter's speech through automatic tracking of her hands: frame-by-frame changes in hands positions provide estimate of hands activity. In correspondence with pauses, two groups were considered: AoM equal or higher than a fixed threshold (0.2) at the beginning of FPs/VFPs vs. AoM lower than the threshold at the beginning of the FPs/VFPs. For each group, the rate of change was computed as the difference between starting AoM

and ending AoM, representing how much AoM changed during the occurrence of a pause. Since the Shapiro test confirmed the normality of the two distributions for Group 2 only, the Wilcoxon rank-sum test was used for the comparison. Group 1 has a median AoM rate of change significantly lower than Group 2 ($p<0.01$, ES= 0.4). Moreover, while the median AoM of Group 1 is significantly lower than 0 ($p<0.01$), the median AoM of Group 2 is not significantly different ($p>0.7$). This appears to indicate that, if the considered guide is moving, an FP/VFP occurrence is usually accompanied by a drop in AoM while, if the guide is not moving, AoM does not change across the FP/VFP occurrence.

5. *Discussion*

The results of the present study can be divided into three main parts. The first part that appears to be particularly noteworthy concerns the differences among the three tourist guides' individual speech. Data on pauses' occurrence and incidence presented in § 4.1 reflect the idiosyncratic linguistic behavior displayed by the tourist guides; in fact, they employ different speech strategies aimed at performing their non-linguistic task of leading a tourist tour.

The high incidence of pauses' occurrences registered in the first guide (G01) reveals that this speaker adopts what has been termed an "on the fly" strategy of formulation (Ferreira, Lau & Bailey, 2004): she manages to sound as spontaneous as possible to her audience and to produce an error-free speech using pauses as devices of wellformedness.

In contrast, both the second guide (G02) and the third guide (G03) make a more limited use of pauses, although not employing the same strategy.

G02 tends to avoid both silent and voiced pauses, which could be perceived as signals of hesitation by the listeners, filling the moments of hesitation with a strategy of "juxtaposition" of utterances, clauses, sentences. Such a resulting speech turns out to be highly error-full, mainly at the morphosyntactic level: some of the errors are corrected by means of retrospective disfluency phenomena (see § 3.2.1), while others are simply ignored or passed by.

On the other hand, G03 plans her speech very carefully and adopts a strategy of "rhetorical control". She resorts almost exclusively to SPs and avoids both FPs and VFPs, producing a more error-free and high-quality speech.

The second part of results provides a general comment about the employment of pauses in the tourist guides' speech. What systematically emerges is the recurring distinction between silent (SPs) and voiced (FPs and VFPs) pauses. In fact, from the functional perspective, SPs efficiently perform both the demarcative (DEM) and the strategic-rhetorical (STR-R) functions, showing a certain degree of awareness of the role of such pauses in emphasizing key words or concepts and attracting the audience's attention. The resulting overlapping between these two functions confirms the distinction provided by O' Shaughnessy (1992) between grammatical and intentional pauses on one side and ungrammatical and hesitation pauses on the

other, strengthening the link between grammaticality and willfulness. By contrast, both types of voiced pauses turned out to be rarely used for grammatical (DEM) and rhetorical (STR-R) goals but widely used to express speakers' indecisiveness while planning the discourse. Indeed, almost the totality of voiced pauses performs a hesitative (HES) function. What appears more interesting are the resulting overlapping and relationship between HES and PROG (programmative) functions. Firstly, both functions share the same characteristics of lack of grammaticality and intentionality (ungrammatical and unintended pauses, see O' Shaughnessy, 1992). Secondly, PROG is always included into the HES function; in other words, all pauses classified as PROG are at the same time classified as HES, but not vice versa. Hence, HES can be thought of as a "macro function" carried out by unintentional pauses; it assumes an added feature when it includes the PROG "micro function". Indeed, PROG pauses convey a particular subtype of hesitancy, which turned out to be strongly linked to the punctual search for lexical items, whose access is highly demanding for the speakers, termed Word Searching (WS). As concerns FPs, specific phonetic-acoustic features were found to correlate with the functional distinction between mere HES pauses and combination of PROG (WS) and HES pauses. WS FPs appear to be more disrupting: on the temporal level, they are twice longer than HES FPs; on the intonation level, they result embedded within their co-text (flat and valley F_0 patterns). On the contrary, realizations of HES FPs cause neither temporal nor melodic breaks within the utterance. In conclusion, duration and range level work as relevant phonetic-acoustic cues of an ongoing WS process.

Other results concern the relationship between pauses and gestures. The data on the gestures co-occurring with pauses definitely confirm the deep difference between Silent (SPs) and Voiced pauses (FPs and VFPs taken together). The gestures having a higher and lower communicative import, respectively, show a complementary distribution with the two types of Pauses.

As already hypothesized, VFPs, and even more FPs, may be seen as a cue to an underlying high cognitive effort: when a FP occurs, the speaker is struggling to plan her speech or committed to a word search activity; this is why she needs to freeze in an idle position or to self-assure herself, performing manipulators.

Such freezing of the body in voiced pauses is corroborated by the drop in Amount of Motion described in § 4.3.4.

One more reason for such a cognitive effort may be that, given the asymmetrical communication between the guide and her audience, she may deeply feel the need to be seen as trustworthy, reliable and confident; a FP or a VFP collides with such an image and conversely stands for an uncertain epistemic stance, that is what a guide needs to avoid the most.

On the contrary, SPs are cases in which the Speaker intentionally stops due to communicatively strategic reasons: she gives the time to the Addressee to process what she has just said, which is very clear to herself, and possibly she may perform semantically loaded gestures to make her message clearer or to reinforce the meaning she intends to convey. In doing so, her meanings flow out easily, no cognitive ef-

fort is present, and she makes deictic, coded, iconic or metaphoric gestures to clarify her semantic content, and beats to emphasize it by asking for attention.

To sum up, gestures may be seen as a cue of different mental states underlying voiced and silent pauses, respectively: they tend to be semantically empty in voiced pauses, but semantically loaded in SPs.

Overall, this account of the comparison between gestures and pauses might also shed light on the more general issue of the relationship between gesture and speech. Our view seems to support the idea of gesture and speech as an integrated system (McNeill, 1992; Kendon, 2004) in which they are but two different yet related routes for meanings to be expressed: if the mind is engaged in a high cognitive load, both speech and gesture suffer from this and they both freeze, stop, withdraw.

Future works might deepen subtler aspects of the temporal and semantic relations between gestures and pauses, e.g., by taking into account gesture's segmentation into phases and semantic relationship with speech, respectively.

6. *Conclusions*

This work has analyzed the phenomena of disfluency in a particular setting: the speech of tourist guides. In a corpus of guided tours, we have analyzed the silent and unlexicalized filled pauses (voiced pauses) performed by the guides, and the co-occurring gestures. The distribution of pauses and corresponding gestures, and the relations among types of pauses, their functions, and the types of gestures produced reveal that while in voiced pauses the speaker finds herself in a communicative *impasse*, and even gestures are less communicative, in silent pauses the speaker is in total control of her communication, and her gestures too are fully intentional and meaningful. Beside providing theoretical insights about speech disfluencies and concerning the relationship between speech and gesture, the proposed characterization of pauses and concurrent gestures and their functional role in spontaneous speech can be used to build a computational model predicting the occurrence of such elements, given a target text. Such a model can be used to control the synthesis process in text-to-speech systems: this will allow to investigate if the introduction of voiced pauses has an impact on the perceived naturalness of synthetic speech, in the cultural heritage presentation domain. In general, these data will support the development of interactive 3D avatars generating presentations of cultural heritage material on-the-fly.

Bibliography

Alibali, M.W., Kita, S. & Young, A.J. (2000). Gesture and the process of speech production: we think, therefore we gesture. In *Language and Cognitive Processes*, 15(6), 593-613.

Berruto, G., Cerruti, M.S. (2011). *La linguistica. Un corso introduttivo*. Novara: UTET De Agostini.

Betz, S., Carlmeyer, B., Wagner, P. & Wrede, B. (2018). Interactive Hesitation Synthesis: Modelling and Evaluation. In *Multimodal Technologies and Interaction*, 2(1), 9.

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.

Betz, S., Wagner, P. & Schlangen, D. (2015). Micro-structure of disfluencies: Basics for conversational speech synthesis. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, 6-10 September 2015, 2222-2226.

Betz, S., Wagner, P. & Vosse, J. (2016). Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data. In *Phonetik und Phonologie*, 12, 19-22.

Blankenship, J., Kay, C. (1964). Hesitation phenomena in English speech: A study in distribution. In *Word*, 20(3), 360-372.

Boersma, P., Weenink, D. (2018). Praat: Doing phonetics by computer (Version 6.0. 44) [Computer software]. Amsterdam: Institute of Phonetic Sciences.

Capirci, O., Volterra, V. (2008). Gesture and speech. The emergence and development of a strong and changing partnership. In *Gesture*, 8(1), 22-44.

Chafe, W. (1980). Some reasons for hesitating. In Dechert, H.W., Raupach, M. (Eds), *Temporal variables in speech: Studies in Honour of Frieda Goldman-Eisler*. The Hague: Mouton, 169-180.

Christenfeld, N., Schachter, S. & Bilous, F. (1991). Filled pauses and gestures: it's not coincidence. In *Journal of Psycholinguistic Research*, 20(1), 1-10.

Cibulka, P. (2016). On how to do things with holds: Manual movement phases as part of interactional practices in signed conversation. In *Sign Language Studies*, 16(4), 447-472.

Clark, H.H. (2002). Speaking in time. In *Speech Communication*, 36:1-2, 5-13.

Colletta, J.M., Kunene, R.N., Venouil, A., Kaufmann, V. & Simon, J.P. (2009). Multi-track annotation of child language and gestures. In Kipp, M., Martin, J., Paggio, P., & Heylen, D. (Eds.), *Multimodal corpora*. Berlin/Heidelberg: Springer, 54-72.

Crocco, C., Savy, R. (2003). Fenomeni di esitazione e dintorni: una rassegna bibliografica. In Crocco, C., Savy, R. & Cutugno, F. (Eds.), *API. Archivio di Parlato Italiano*, DVD.

d'Urso, V., Zammuner, V. (1990). The perception of pause in question-answer pairs. In *Bulletin of the Psychonomic Society*, 28, 41-43.

De Mauro, T. (1999). *Gradit*. Torino: UTET.

Duez, D. (1997). Acoustic markers of political power. In *Journal of Psycholinguistic Research*, 26(6), 641-654.

Duncan, S., Pedelty, L. (2007). Discourse focus, gesture, and disfluent aphasia. In Duncan, S., Cassell, J. & Levy, L.T. (Eds.), *Gesture and the dynamic dimension of language*. Amsterdam / Philadelphia: John Benjamins, 269-284.

Eklund, R. (1999). A comparative study of disfluencies in four Swedish travel dialogue corpora. In *Disfluency in Spontaneous Speech Workshop*, Berkeley, California, 1 July 1999, 3-6.

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.

Ekman, P., Friesen, W.V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. In *Semiotica*, 1,1, 49-98.

ELAN (Version 5.2) [Computer software]. (2018, April 04). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tla-tools/elan/.

Esposito, A., Marinaro, M. (2007). What pauses can tell us about speech and gesture partnership. In Esposito, A., Bratanić, M., Keller, E., & Marinaro, M. (Eds.), *Fundamentals of verbal and nonverbal communication and the biometric issue*. Amsterdam: IOS Press, NATO Publishing Series, 45-57.

Ferreira, F., Lau, E.F. & Bailey, K.G.D. (2004). Disfluencies, language comprehension, and tree adjoining grammars. In *Cognitive Science*, 28, 721- 749.

Gabrea, M., O'Shaughnessy, D. (2000). Detection of filled pauses in spontaneous conversation speech. In *Proceedings, 6th International Conference on Spoken Language Processing (ICSLP 2000*), Beijing, China, 16-20 October 2000, vol. 2, 678-681.

Giannini, A. (2003). Prolungamenti vocalici e vocalizzazioni. In Cosi, P., Magno Caldognetto, E. & Zamboni, A. (Eds), *Voce, canto, parlato. Studi in onore di Franco Ferrero*. Padova: Unipress, 163-172.

Ginzburg, J., Fernández, R. & Schlangen, D. (2014). Disfluencies as intra-utterance dialogue moves. In *Semantics and Pragmatics*, 7(9), 1-64.

Graziano, M., Gullberg, M. (2018). When speech stops, gesture stops: evidence from developmental and crosslinguistic comparisons. In *Frontiers in psychology*, 9, 879.

Hartsuiker, R.J., Notebaert, L. (2009). Lexical access problems lead to disfluencies in speech. In *Experimental psychology*, 57, 169-177.

Hieke, A.E. (1981). A content-processing view of hesitation phenomena. In *Language and Speech*, 24(2), 147-160.

Holler, J., Schubotz, L., Kelly, S., Hagoort, P., Schuetze, M. & Özyürek, A. (2014). Social eye gaze modulates processing of speech and co-speech gesture. In *Cognition*, 133, 692-697.

Kendon, A. (1980). Gesticulation and speech: two aspects of the process of utterance. In Key, M.R (Ed.), *Nonverbal communication and Language*. The Hague: Mouton, 207-227.

Kendon, A. (2004). *Gesture. Visible action as utterance.* Cambridge: Cambridge University Press.

Kita, S. (1990). The temporal relationship between gesture and speech: a study of Japanese-English bilinguals. Master's Thesis, Department of Psychology, University of Chicago.

Kita, S. (2000). How representational gestures help speaking. In McNeill, D. (Ed.), *Language and gesture*. Cambridge: Cambridge University Press, 162-185.

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. In *Language and cognitive processes*, 24(2), 145-167.

Kita, S., Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. In *Journal of Memory and Language*, 48(1), 16-32.

Kita, S., Alibali, M.W. & Chu, M. (2017). How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. In *Psychological Review*, 124(3), 245-266.

Kita, S., Davies, T.S. (2009). Competing conceptual representations trigger co-speech representational gestures. In *Language and Cognitive Processes*, 24(5), 795-804.

Kong, A.P., Law, S., Kwan, C.C., Lai, C. & Lam, V. (2015). A Coding System with Independent Annotations of Gesture Forms and Functions during Verbal Communication: Development of a Database of Speech and GEsture (DoSaGE). In *Journal of Nonverbal Behaviour*, 39(1), 93-111.

Krauss, R.K., Chen, Y. & Gottesman, R.F. (2000). Lexical gestures and lexical access: a process model. In McNeill, D. (Ed.), *Language and gesture*. Cambridge: Cambridge University Press, 261-283.

Krauss, R.M., Hadar, U. (1999). The role of speech-related arm/hand gestures in word retrieval. In Messing, L., Campbell, R. (Eds.), *Gesture, sign and speech*. New York: Oxford University Press, 93-116.

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.

Lickley, R.J. (2015). Fluency and Disfluency. In Redford, M.A. (Ed.), *The handbook of speech production*. John Wiley & Sons, 445-474.

Lickley, R.J. (1998). HCRC Disfluency Coding Manual. Edimburgh: Human Communication Research Centre Technical Report TR-100, University of Edimburgh.

Liszkowski, U. (2008). Before L1: a differentiated perspective on infant gestures. In *Gesture*, 8, 180-196.

Maclay, H., Osgood, C.E. (1959). Hesitation Phenomena in Spontaneous English Speech. In *WORD*, 15:1, 19-44.

Mahl, G.F. (1956). Disturbances and silences in the patient's speech in psychotherapy. In *The Journal of Abnormal and Social Psychology*, 53(1), 1-15.

Mayberry, R.I., Jaques, J. (2000). Gesture production during stuttered speech: insights into the nature of gesture-speech integration. In McNeill, D. (Ed.), *Language and gesture*. Cambridge: Cambridge University Press, 199-214.

McNeill, D. (1992). *Hand and mind*. Chicago: The University of Chicago Press.

McNeill, D. (2005). *Gesture and thought*. Chicago: The University of Chicago Press.

Mondada, L. (2007). Multimodal resources for turn-taking: Pointing and the emergence of possible next speakers. In *Discourse Studies*, 9(2), 195-226.

Morsella, E., Krauss, R.M. (2005). Muscular activity in the arm during lexical retrieval: Implications for gesture-speech theories. In *Journal of Psycholinguistics Research*, 34(4), 415-427.

Müller, C. (2004). Forms and uses of the Palm Up Open Hand: A case of a gesture family? In Müller, C., Posner, R. (Eds.), *The Semantics and Pragmatics of everyday Gestures*. Berlin: Weidler.

Nobe, S. (2000). Where to most spontaneous representational gestures actually occur with respect to speech? In McNeill, D. (Ed.), *Language and gesture*. Cambridge: Cambridge University Press, 186-198.

Nobili, C (2019). *I gesti degli italiani*. Roma: Carocci.

Origlia, A., Savy, R., Cataldo, V., Schettino, L., Ansani, A., Sessa, I., Chiera, A. & Poggi, I. (2019). Human, all too human. Towards a disfluent Virtual Tourist Guide. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization,* Larnaca, Cyprus, 9-12 June 2019, 393-399.

Origlia, A., Savy, R., Poggi, I., Cutugno, F., Alfano, I., D'Errico, F., Vincze, L. & Cataldo, V. (2018). An Audiovisual Corpus of Guided Tours in Cultural Sites: Data Collection Protocols in the CHROME Project. In *Proceedings of the 2018 AVI-CH Workshop on Advanced Visual Interfaces for Cultural Heritage*, Castiglione della Pescaia, Italy, 29 May 2018, 1-4.

O'Shaughnessy, D. (1992). Recognition of hesitations in spontaneous speech. In *Proceedings: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)*, San Francisco, USA, 23-26 March 1992, 521-524.

Pine, K.J., Bird, H. & Kirk, E. (2007). The effects of prohibiting gestures on children's lexical retrieval ability. In *Developmental Science*, 10(6), 747-754.

Poggi, I. (2007). *Mind, hands, face and body. A goal and belief view of multimodal communication.* Berlin: Weidler.

Ragsdale, J.D., Fry Silvia, C. (1982). Distribution of kinesic hesitation phenomena in spontaneous speech. In *Language and Speech*, 25(2), 185-190.

Schnadt, M.J., Corley, M. (2006). The influence of lexical, conceptual and planning based factors on disfluency production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, (Vol. 28, No. 28), Vancouver, Canada, 26-29 July 2006, 750-755.

Senigalliesi, S., Sparano, A., Schettino, L., Savy, R., Dell'Orletta, F., Lubello, S. & Basile, G. (2018). Frequenze lessicali e sintattiche nei testi scritti e orali della descrizione artistico/architettonica: analisi sul corpus delle Certose Campane. Poster presented at *Gruppo di Studio sulla Comunicazione Parlata* (GSCP 2018), Università degli Studi di Napoli "L'Orientale", 12-14 dicembre 2018.

Seyfeddinipur, M. (2006). Disfluency: Interrupting speech and gesture. PhD thesis, Radboud University Nijmegen, Nijmegen.

Shriberg, E. (1994). Preliminaries to a theory of speech disfluencies. PhD dissertation, University of California, Berkeley.

Shriberg, E. (2001). To "err" is human: Ecology and acoustics of speech disfluencies. In *Journal of the International Phonetic Association*, 31, 153-169.

Shriberg, E.E., Lickley, R.J. (1993). Intonation of clause-internal filled pauses. In *Phonetica*, 50.3, 172-179.

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.

Stam, G., Tellier, M. (2017). The sound of silence: the functions of gestures in pauses. In Breckinridge Church, R. Alibali, M.W. & Kelly, S.D. (Eds.), *Why gesture? How the hands function in speaking, thinking and communicating*. Amsterdam: Benjamins, 353-377.

Strangert, E. (2003). Emphasis by pausing. In *Proceedings of the 15th international congress of phonetic sciences*, Barcelona, Spain, 3-9 August 2003, 2477-2480.

Swerts, M. (1998). Filled pauses as markers of discourse structure. In *Journal of pragmatics*, 30(4), 485-496.

Viola, I., Madureira, S. (2008). The roles of pause in speech expression. In *Proceedings of the 4th Conference on Speech Prosody*, Campinas, Brazil, 6-9 May 2008, 721-724.

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

Yasinnik, Y., Shattuck-Hufnagel S. & Veilleux N. (2005). Gesture marking of disfluencies in spontaneous speech. In *The 4th Workshop on Disfluency in Spontaneous Speech*, 173-178.

Zellner, B. (1994). Pauses and the temporal structure of speech. In Keller, E. (Ed.), *Fundamentals of speech synthesis and speech recognition.* Chichester: John Wiley, 41-62.

Acknowledgments

Work funded by the Italian National Project PRIN "Cultural Heritage Resources Orienting Multimodal Experiences (CHROME)" (#B52F15000450001).