DOI: 10.17469/O2109AISV000010

LOREDANA SCHETTINO, ANTONIO ORIGLIA, GIACOMO MATRONE

Modeling Hesitations. Speech Synthesis Application and Evaluation[1]

Studies have shown that elements like silent pauses, segmental lengthenings, and fillers are naturally involved in the economy of speech and, in specific patterns, may contribute to communication in both human-human and human-machine interactions. Therefore, research on speech synthesis aimed at developing more natural-sounding systems by inserting hesitation phenomena. However, audio issues were found to arise when synthesising filled pauses. Only recently, speech synthesisers based on Deep Neural Networks achieved better performances. In this study, we provide a first perceptual evaluation of a model of occurrence of hesitations (lengthenings, silent pauses as well as fillers) in Italian utterances using a state-of-the-art neural TTS system. A set of experimental stimuli were synthesized and subjected to listeners' evaluations in a discrimination test. Results show that synthetic utterances that include hesitations, according to the linguistic model, are judged as more natural sounding than utterances that do not include any.

Keywords: disfluency, pauses, speech synthesis, Deep Neural Network, perception.

1. *Introduction*

Spontaneous human speech is usually characterised by the occurrence of pauses, fillers, repetitions, corrections, change of planning, various phenomena that seem to alter its fluency and, hence, have been commonly referred to as speech "disfluencies". However, studies on spontaneous speech in different languages have highlighted that the occurrence of disfluency phenomena is not to be considered as exceptional with reference to "normal fluency". Indeed, they report a rate of around 6 to 10 phenomena per 100 words, which suggests that "fluency is the exception, rather than the rule" (Lickley, 2015: 451). Moreover, it has been observed that disfluencies may occur in regular patterns, as they actually serve as tools that the speakers may use to monitor and manage their own speech production by repairing something already uttered, abandoning already started utterance, taking extra-time needed for the planning and construction of the message that is about to be conveyed (Levelt, 1993; Shriberg, 1994).

In particular, speakers can temporarily delay the speech delivery by producing fillers, prolonging speech segments or just being silent. These pauses, prolongations

[1] This article is the result of the collaboration among the authors. However, for academic purposes only, Loredana Schettino is responsible for § 1, 2, 3, 4, 5 and 6 and all the authors are responsible for § 3.1 and 6.

and fillers are also commonly referred to as "hesitation phenomena" (Lickley, 2015). They contribute to communication and can be considered to be beneficial for both speakers and listeners by gaining extra-time for planning as well as for information processing. This claim is corroborated by the fact that these phenomena were also observed to consistently occur in informative speech, e.g., lecturers' or tourist guides' speech (Moniz, Batista, Mata & Trancoso, 2014; Schettino, Betz, Cutugno & Wagner, 2021a). Moreover, studies have shown that hesitation phenomena can bear procedural meaning and convey information about speech planning, structuring, and speakers' disposition (Chafe, 1980; Levelt, 1993; Schegloff, 2010; Tottie, 2016) as listeners learn to exploit the regular occurrence of such phenomena and use it for the interpretation of the ongoing discourse (Corley, Stewart, 2008; Finlayson, Corley, 2012).

These observations on the relevance of hesitation phenomena in communication sparked the interest in developing synthesis systems that were able to insert hesitations in synthesised utterances, in order to obtain more natural-sounding, likeable productions and, eventually, more effective human-machine interactions.

This study integrates the linguistic and computational perspectives while testing the hypothesis that utterances produced using a neural TTS synthesis system that is trained to synthesise disfluencies and where selected phenomena are inserted according to a previously proposed model based on corpus observation are perceived as more natural-sounding and more desirable.

The paper is structured as follows: § 2 provides an overview on previous studies concerning listeners' perception of disfluency phenomena occurring either in spontaneous stimuli or in synthesised ones. Then, in § 3, the approach adopted to evaluate how specific disfluency patterns may affect listeners' perception is described, including the linguistic model, the computational model (the speech synthesiser) and the experimental setting. Finally, the experimental results are presented and discussed in § 4 and § 5.

2. *Disfluency perception and speech synthesis*

Considering that linguistic perception does not necessarily correspond with what has been actually produced but is rather constantly influenced by the communicative context and listeners' selective attention (Levelt, 1993; Voghera, 2017), Lickley (2015) interprets fluency, and disfluency, as a multidimensional concept. The author distinguishes three dimensions that are related to the underlying processes of speech planning, production and perception identified by Levelt (1993) and highlights that speech may be perceived as *fluent* even when containing minor surface disfluencies only detectable on closer inspection of the speech signal, which means that *perception fluency*, does not necessarily imply *planning* and/or *surface fluency*.

In fact, it has been experimentally observed that in disfluency detection tasks, listeners tend to miss out various phenomena (see Collard, 2009 for an overview).

In particular, listeners' perception and awareness of fillers was found to depend on whether they attended to discourse content or style of delivery. Christenfeld (1995) shows that filled and silent pauses negatively affected the perception of the speaker only when listeners' attention was focused on style but not when it was drawn on content. In the latter case, filled pauses tended to be missed. Furthermore, the author observes that filled pauses are perceived as a more «relaxed-sounding» time-buying strategy than silent pauses.

Moreover, it was found that some hesitation phenomena may be perceived as more disruptive than others according to their phonetic features and positioning in sentences and discourse.

Investigating hesitation phenomena in European Portuguese in spontaneous and prepared non-scripted speech, Moniz et al. (2009, 2010) showed that the prosodic phrasing and contour shape exert an influence on participants' ratings of speakers' fluency, defined as *ease of expression.* Lengthenings, filled pauses, and repetitions were most likely rated as *felicitous* when occurring at prosodic breaks and with a flat or ascending pitch contour shape and *infelicitous* when occurring within intonation units or with descending contours.

More recently, Niebuhr and Fischer (2019) investigated the effect of filled pause occurrences on listeners' perception of a speaker's public-speaking and found that shorter and largely nasal filled pause realisations made listeners underestimate their actual number and improved their ratings of the speaker's performance, which was assumed to derive from the lower "saliency" linked to such realisations.

The acknowledgement of the role played by hesitation phenomena in speech challenges rhetoricians' warning against littering speech with them. In fact, it raised the attention of researchers interested in modelling human communicative behaviours to develop speech synthesis systems that would support human-machine interaction systems. So, different state-of-the-art synthesis methods and approaches were implemented to insert hesitations in speech synthesis, most of which focused on the synthesis of filled pauses.

Among the first attempts in this direction is the system developed by Adell, Escudero and Bonafonte (2012) within a rule-based framework. They built a model to generate filled pauses based on the modelling of human fillers prosodic features. A perception rating test conducted to evaluate the system showed that filled pauses introduced with this approach did not increase the degree of listening effort necessary to process the sentences nor decreased their naturalness.

On another account, Dall, Tomalin and Wester (2016) tested the synthesis of filled pauses using HMM-based Speech Synthesis System conducting various evaluation perception experiments. They first found that a voice trained on standard read speech was judged more natural than one trained on spontaneous speech, even when including filled pauses. Hence, the authors tested data-mixing techniques which consisted in combining a synthesis system based on read speech corpora, for the synthesis of general speech, and a system trained on spontaneous speech, for the synthesis of fillers. They observed that this approach together with obtaining a better phonetic representation of filled pauses improved the overall quality of the synthesis. However, the developed system did not apparently produce satisfying performances.

A HMM-based Speech Synthesis System was also proposed by Betz, Carlmeyer, Wagner and Wrede (2018) who developed and evaluated a model for hesitation insertion in Incremental Spoken Dialogue Systems. The original model included lengthenings, silences and filled pauses, but the perceptual experimental evaluation involved a reduced model without fillers because of the acoustic artefacts produced when synthesising fillers.

Only more recently, Székely, Henter, Beskow and Gustafson (2019) have developed a neural TTS system (Tacotron) trained on a large single-speaker corpus of spontaneous conversational speech. They evaluated the synthesis of filled pauses obtained using models trained on the basis of different types of filled pauses annotation by conducting a pairwise listening test with utterances that both contained filled pauses but were produced using different models: one where the annotation did not account for non-verbal elements so that the system would generate them automatically given a fluent text as input; one where the annotation included two different labels for «uh» and «uhm» instances which allows control on location and type of filled pauses; another one based on an annotation that associated all types of non-verbal vocalisations with one generic label, which only allows to control for their location and was found to provide more natural sounding utterances.

3. *Method*

The evaluation of the way the insertion of disfluency phenomena can affect listeners' perception is no easy task. Common approaches to prepare the experimental stimuli concern repetition tasks and signal manipulations (Fraundorf, Watson, 2011; Mühlack, Elmers, Drenhaus, Trouvain, van Os, Werner, Ryzhova & Möbius, 2021), whereby, however, different issues arise. Repetition tasks consist in asking subjects to listen to a disfluent recording of themselves in a natural setting and to repeat their utterances without disfluencies, which alters the recording setup and lacks in spontaneousness. The second approach consists in intervening on the signal by manually removing/inserting disfluencies, which only involves the disfluent portion of the speech signal thus leaving the immediate prosodic context unchanged and not considering the way it is influenced by the presence of disfluency phenomena. A possible solution to avoid these problems is synthesising experimental stimuli using a system that can be trained to generate utterances, also containing disfluencies, in a highly plausible way.

More specifically, Deep Neural Networks (DNNs) can learn a speaker's intonational patterns both from disfluent and non-disfluent portions of the training data. This allows to generate speech stimuli with and without disfluencies using a probabilistic, trained model generating the most *probable* speech signal that would correspond to an input text. By providing as input to the synthesiser two versions of the same text with and without annotated disfluencies, the obtained stimuli come from a coherent model of natural speech, unbiased by previous productions and contextually coherent. These can be used to investigate perceptual differences with a solid set of stimuli.

So, to test the effect of disfluency phenomena on listeners' perception, a discrimination test has been conducted whereby subjects were asked to rate stimuli produced using neural synthesis and where disfluency phenomena were inserted according to a linguistic model derived from previous corpus-based observations (Schettino et al. 2021a; Schettino, Betz & Wagner, 2021b). The next sections briefly describe how the synthesis system works (§ 3.1), the previously defined linguistic model (§ 3.2), and the setting of the perception experiment (§ 3.3).

3.1 Computational Model

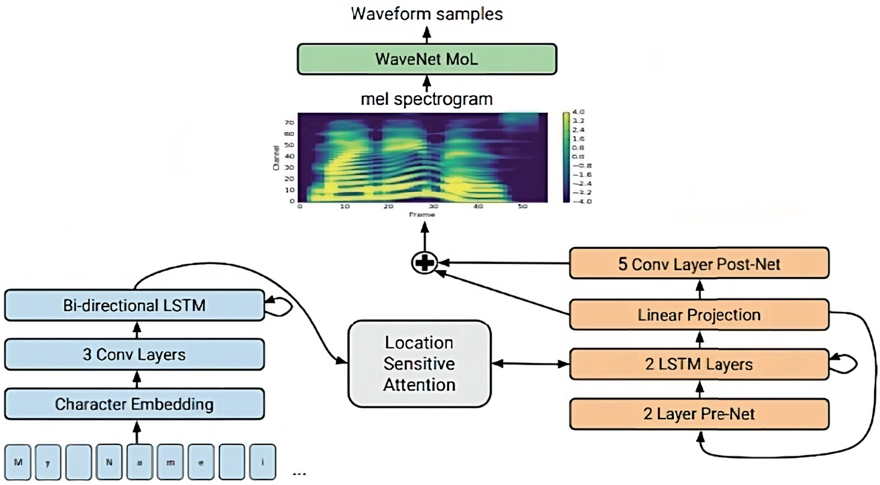
Neural speech synthesisers represent one of many applications of DNNs and can be trained to replicate the voice of a target speaker or even speaking styles. Previous works have concerned the building of models of the same speaker in different speaking styles (Wang, Stanton, Zhang, Ryan, Battenberg, Shor, Xiao, Jia, Ren & Saurous, 2018). However, it is also possible to target a specific style by training the model using only data representative of that style. In our case, the model has been trained on data that represent the *informative* speaking style.

More specifically, the model has been trained on two-and-a-half-hour single-speaker speech extracted from the CHROME corpus (Origlia, Savy, Poggi, Cutugno, Alfano, D'Errico, Vincze & Cataldo, 2018). It consists of Italian semi-spontaneous speech by a female expert guide leading visits at San Martino's Charterhouse and is supplied of orthographic transcriptions (Savy, 2005) and disfluency annotations (Schettino et al., 2021a). In particular, vowel lengthenings, filled pauses, and silent pauses were manually annotated by expert linguists and labelled using grapheme sequences that do not occur anywhere in the corpus other than in correspondence of the considered phenomena:

- Lengthenings (LEN), marked prolongation of segmental material (Betz, Wagner & Eklund, 2017), labelled with repetitions of vocalic sounds for prolongations, i.e., "vv";
- Filled Pauses (FP), non-verbal filler, vocalisations and/or nasalizations, i.e., *eeh, ehm,* annotated using a generic label for both the nasalized and non-nasalized versions of filled pauses, i.e., "ehm";
- Silent Pauses (SP), marked silences perceived as stalling pause in the context of occurrence (Lickley, 2015), labelled using the sequence "hh".

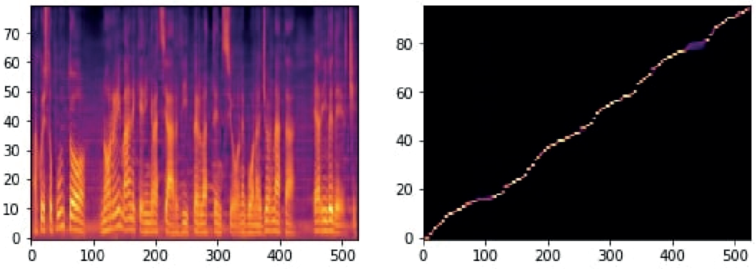
In this study, utterances were synthesised using a state-of-the-art system, namely Tacotron 2 (Shen, Pang, Weiss, Schuster, Jaitly, Yang, Chen, Zhang & Wang, 2018). A network pre-trained on English was fine-tuned on the CHROME transcribed audio material, including disfluency annotations, to generate Italian speech spectrograms. These were, then, fed to a Waveglow model (Prenger, Valle & Catanzaro, 2019) to produce the waveform and to apply denoising. At inference time, the model can be used to generate new speech sounds mimicking both voice and speaking style from the example speaker (see Fig. 1).

Figure 1 - *The Tacotron 2 architecture as described in (Shen et al., 2018, p. 4780). In this work, the WaveNet model is replaced by the more recent Waveglow model*



The synthetic spectrogram is accompanied by the corresponding *alignment graph* (Fig. 2) that represents to what extent the network decoder was able to select the correct states, among all the possible ones, making reference to frames in the training audio. The *hotter* the pixel, the higher the attention given to a certain state. Ideally, a diagonal line of *hot* pixels indicates that the decoder focused on the correct states from the encoder to generate the mel spectrum.

Figure 2 - *A synthetic spectrogram (left) together with its alignment graph (right). A good alignment between the encoder vectors (y axis) and the decoder steps (x axis) is represented by a line that tends to a diagonal*



Summarising, DNNs produce the most probable speech output given a character sequence. So, the same text can be submitted, with and without disfluencies, to the same network that is able to generate productions that are not influenced by what has been synthesised before, as would happen with humans. Also, the whole utterance prosodic representation is coherent with the presence or absence of the disfluency phenomena, as opposed to interventions with manual cuts.

3.2 Linguistic Model

The described machine learning solution allows the generation of stimuli where hesitation phenomena are produced after specification of their location in the text and, given their context of occurrence, the synthesis system computes their surface realisation.

This work concerns the perceptual evaluation of the following patterns of hesitations occurrence that emerged from previous corpus analyses:

- Lengthenings are placed: a) before semantically *heavy*, key constituents (*focusing function*, Schettino et al. 2021a); b) toward the end of the clause (Schettino et al. 2021b); c) on content or functional words (following the distribution found in the dataset, i.e., content words: 51%, function word: 49%);
- Silent or Filled Pauses are placed between two clauses (*structuring function*, Schettino et al. 2021a, b). Half of the stimuli also contained a Filled Pause and the other half a Silent Pause.

In the following two utterances of examples, hesitations phenomena are placed according to the just described patterns:

(1) *"Nella prima metà del diciottesimo secolo i lavori passarono a**aa** Nicola Tagliacozzi Canale **ehm** che farà rifare gli**ii** spazi del priore."*

"In the first half of the Eighteenth century the work was handed to[**LEN**] Nicola Tagliacozzi Canale [**FP**] who will redo the[**LEN**] prior's place."

(2) *"La certosa fu inaugurata e consacrata**aa** nel 1368 **hh** seppur i certosini avevano preso possesso del monastero**oo** dal 1337"*

"The Cartherhouse was inaugurated and consecrated[**LEN**] in 1368 [**SP**] although the Carthusians had taken possession of the monastery[**LEN**] since 1337."

3.3 Experimental Evaluation

3.3.1 Experimental Setting

The evaluation of synthesised utterances commonly involves judgements of their perceived *naturalness*. However, Wagner, Beskow, Betz, Edlund, Gustafson, Henter, Le Maguer, Malisz, Székely, Tånnander & Voße (2019) highlight that naturalness is not an inherent property of speech, but is specified with reference to the communicative context of application. Therefore, the evaluation of synthesis systems should refer to the principle of *contextual appropriateness* given a specific situation or application.

In this study, the context of application for the evaluation of the *disfluent* vs. *non-disfluent* synthetic utterances is to provide voice to a Virtual Avatar, i.e., Embodied Conversational Agent, serving visitors in museums. Given this context, the perception experiment consists in explicitly asking the listeners whether the system meets the estimated needs of *naturalness* and *appropriateness* with reference to the envisioned application.

More specifically, participants were subjected to a pairwise listening test where they were asked to listen to pairs of synthetic utterances and then select the one that sounded more natural to them (*naturalness*), and the one they would choose to give voice to a Virtual Avatar serving in museums like the San Martino Charterhouse in Naples (*appropriateness*).

The set of stimuli is composed by complex phrases, meaning a main clause and a dependent (mostly relative) clause, which describe point of interest of the Charterhouse and comprised 10 target pairs of utterances, one with hesitations (*Disf* condition) and one without any (*no_Disf* condition) and 10 filler couples paired as *Disf-Disf* and *no_Disf-no_Disf*. These stimuli were presented to participants in randomised order.

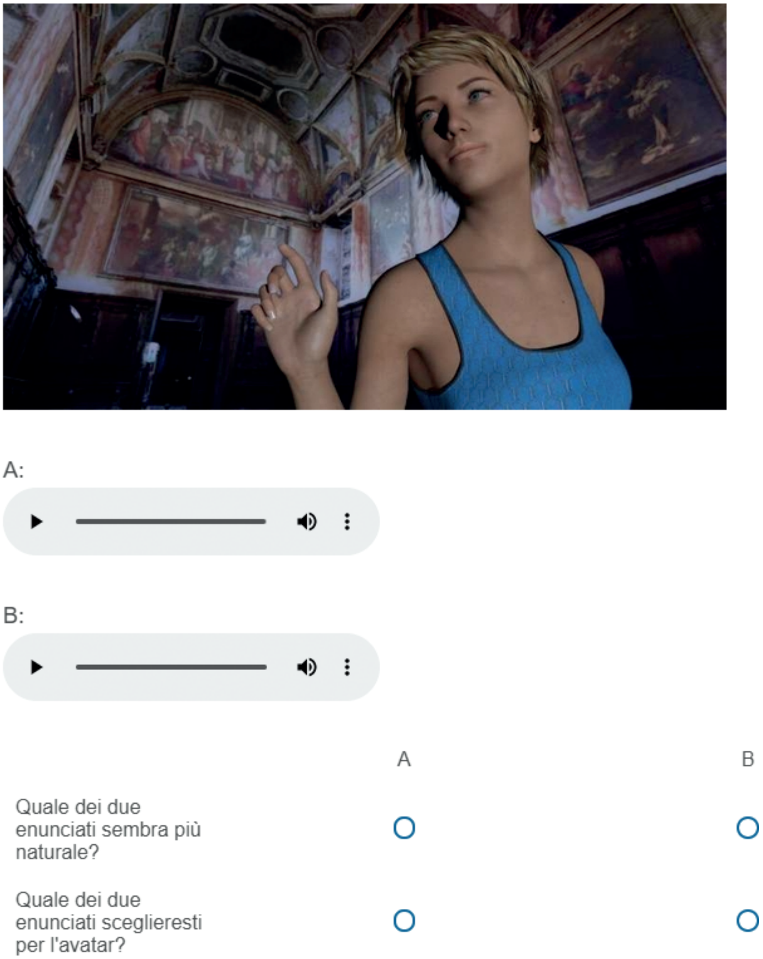
The test was set up and distributed on social media channels for university students using the QUALTRICS software for online surveys.[2]

Participants were asked to fill in a sociolinguistic questionnaire collecting information concerning the age, sex, country and city where they spent most of their life, whether they regularly listen to synthetic voices such as Siri or Cortana. Then, they were asked to make sure they were in a quiet closed area and wearing headphones throughout the experiment duration (approximately 15 minutes, including an initial training phase).

A picture of the CHROME avatar "Maya" (Origlia et al., 2018) was enclosed to each question to provide graphical support to the contextualization (Fig. 3).

[2] Version [2021] of Qualtrics – www.qualtrics.com

Figure 3 - *Example of a question of the Discrimination Task as visible by participants on Qualtrics*



A:

B:

	A	B
Quale dei due enunciati sembra più naturale?	○	○
Quale dei due enunciati sceglieresti per l'avatar?	○	○

3.3.2 Statistical Analysis

The statistical analysis is conducted using the R software (R Core Team 2021). A Generalised Linear Mixed Model ("lme4" package, Bates, Mächler, Bolker & Walker, 2015) is built including subjects' responses, i.e., the condition chosen between *Disf* and *no_Disf*, as dependent variable; the question, i.e., *naturalness* or *appropriateness*, and type of phenomena occurring in the disfluent stimuli, i.e., *LEN_FP* or *LEN_SP* as interacting independent variables and, to control for individual variability, participants as random effect. Sociolinguistic variables are also controlled considering sex, age, and familiarity with synthetic voices ("yes" or "no") as independent variables.

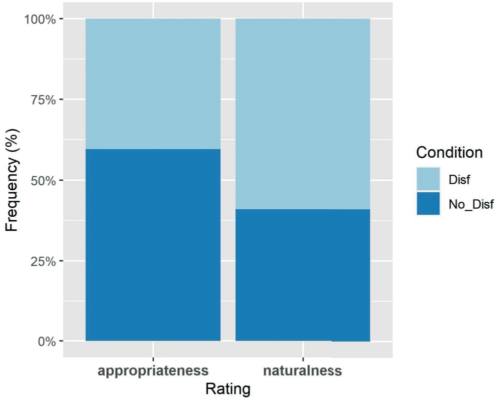
Post-hoc analyses are conducted to inspect the levels within the main effects and the interactions using pairwise comparisons (emmeans package, Lenth, Singmann, Love, Buerkner & Herve, 2018). P-values were calculated using Tukey's HSD adjustment.

4. *Discrimination Test Results*

The experiment was conducted with 22 participants (7 female, 15 male) with age ranging between 22 and 50 years (M = 30, StDev = 7). Among these participants, only 6 (27%) reported to regularly listen to synthetic voices such as Siri or Cortana.

As reported in Table 1, the statistical analysis yields significant results. Synthesised disfluent utterances are significantly more often judged as more natural. Conversely, non-disfluent utterances are significantly more frequently selected to give voice to a virtual agent, than disfluent ones (see Fig. 4).

Figure 4 - *Frequency (%) of appropriateness and naturalness ratings per condition*



Also, this effect is only significant when considering ratings of utterances that contain filled pauses and represents just a trend for judgements related to the utterances with silent pauses.

Table 1 - *Pairwise comparison among the levels of the question variable in interaction with the LEN_FP and LEN_SP stimuli groups*

Stimuli	*Contrast*	*Estimate*	*Std. Error*	*z value*	*p value*
LEN_FP, LEN_SP	*appropriateness – naturalness*	0.809	0.202	4.002	0.0001
LEN_FP	*appropriateness – naturalness*	1.270	0.292	4.353	<.0001
LEN_SP	*appropriateness – naturalness*	0.349	0.279	1.249	0.2115

5. *Discussion*

The participants' responses to the discrimination test attest an inverse direction for naturalness and appropriateness judgements. The insertion of hesitation phenomena according to the linguistic model significantly affects the listeners' perception of the synthesised utterances in that they are perceived as more natural sounding than non-disfluent utterances, but less luckily to be associated with a virtual avatar. These opposed tendencies may reflect the fact that people would not expect a *machine* to produce spontaneous physiologic (*natural*) elements such as disfluencies so that synthetic utterances containing disfluencies are not (yet) customarily associated with a virtual system, despite being rated as more natural sounding.

In fact, while testing suitable voices for robots, studies have found that a robotic voice is often preferred over human-like voices (Hönemann, Wagner, 2015; Wagner et al., 2019). According to Moore (2017) using human-like voices for artefacts might lead users to overestimate their abilities. Hence, intelligible but robotic voices could be considered more *appropriate* and better systems to manage the users' expectations of *conversational* artefacts, like *Google Now* or *Amazon Echo*. On the other hand, Rodero (2017) showed that human voices perform better in narrative tasks, such as telling an advertising story, as they are rated as more effective and enhance listeners' attention and recall. Based on this picture, synthesis systems that are able to generate human-like voices by reproducing plausible prosodic realisations including hesitation phenomena could provide effective and desirable voices for informative speech.

Looking more closely at the results, the observed effect of the insertion of hesitations on participants achieve significant values only for stimuli pairs where the disfluent utterance contains a filled pause (*LEN_FP*), whereas for those where the disfluent utterance contains a silent pause (*LEN_SP*), a weak trend emerges. This may suggest that filled pauses, being a voiced element that is perceptually independent from the other sounds in the speech chain, are more evident phenomena within the immediate prosodic surroundings and are more likely to be detected as speech planning devices, unlike lengthenings and silent pauses which may be considered as more subtle phenomena. Hence, disfluent *LEN_FP* utterances would stand out more clearly from non-disfluent ones with respect to disfluent *LEN_SP* utterances. The latter, instead, seem to be perceived as less markedly different than non-disfluent utterances, which would hamper the emergence of clear-cut preferences for selecting one type of utterances over the other (disfluent vs. non-disfluent).

6. *Conclusions*

The study described in this article has been conducted to provide a perceptual evaluation of previously observed patterns of occurrence of hesitation phenomena, i.e., silences, fillers, lengthening, in informative speech (Schettino et al., 2021a, b).

The preparation of the sets of experimental stimuli has been supported by a computational model of speech. More specifically, a neural synthesis system has been trained to generate utterances including hesitations in a contextually plausible way. Then, a discrimination test was designed to evaluate how lengthenings, filled pauses and silent pauses inserted according to the corpus-based model can affect the listeners' perception of the synthesized utterances.

The main results of the pairwise listening test highlight that disfluent utterances, especially when containing filled pauses, are perceived as more natural sounding, though less appropriate to the specific application in supporting virtual avatars serving in museums. However, Wagner and colleagues (2019) suggest that an accurate evaluation of synthesis, beside being based on participants' subjective ratings, should also consider behavioural assessment, which consists in the indirect evaluation of the users' comprehension and preferences while fulfilling a task. Therefore, a follow-up test has been designed to integrate the subjective evaluation with a behavioural one involving a task more closely related to the specific application, that is to give a voice for a virtual agent designed to serve visitors in cultural sites by showing relevant points of interest.

More generally, the study provides first evidence that modern technologies, such as neural synthesis systems, being able to produce highly plausible and, to a certain extent, controllable stimuli, may represent valuable tools for testing relevant hypothesis of linguistic, and phonetic, interest, especially when concerning speech phenomena that are difficult to elicit in natural settings such as disfluency phenomena (Malisz, Henter, Valentini-Botinhao, Watts, Beskow & Gustafson, 2019).

Acknowledgements

Funding: This work was supported by the Italian National Project PRIN Cultural Heritage Resources Orienting Multimodal Experiences (CHROME) [grant number B52F15000450001].

References

Adell, J., Escudero, D. & Bonafonte, A. (2012). Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. In *Speech Communication*, 54, 459-476. doi:https://doi.org/10.1016/j.specom.2011.10.010.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. In *Journal of Statistical Software*, 67, 1-48. doi:https://doi.org/10.18637/jss.v067.i01.

Betz, S., Carlmeyer, B., Wagner, P. & Wrede, B. (2018). Interactive hesitation synthesis: modelling and evaluation. In *Multimodal Technologies and Interaction*, 2 (1), 9. doi:https://doi.org/10.3390/mti2010009.

Betz, S., Eklund, R. & Wagner, P. (2017). Prolongation in German. In Rose, R., Eklund, R. (Eds.), *Proceedings of the 8th Workshop on Disfluency in Spontaneous Speech*, Stockholm, Sweden, 18–19 August 2017, 13-16.

Christenfeld, N. (1995). Does it hurt to say um? In *Journal of Nonverbal Behavior*, 19, 171-186. doi:https://doi.org/10.1007/BF02175503.

Collard, P. (2009). Disfluency and listeners' attention: An investigation of the immediate and lasting effects of hesitations in speech. Ph.D dissertation, The University of Edinburgh.

Corley, M., Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. In *Language and Linguistics Compass*, 2, 589-602. doi:https://doi.org/10.1111/j.1749-818X.2008.00068.x.

Dall, R., Tomalin, M. & Wester, M. (2016). Synthesising filled pauses: Representation and datamixing. In Bonafonte, A, Prahallad, K (Eds.), *Proceedings of 9th Speech Synthesis Workshop*, Sunnyvale, California, USA, 13–15 Septembre 2016, 7-13. doi:https://doi.org/10.21437/SSW.2016-2.

Finlayson, I. R., Corley, M. (2012). Disfluency in dialogue: An intentional signal from the speaker? In *Psychonomic bulletin & review*, 19, 921-928. doi:https://doi.org/10.3758/s13423-012-0279-x.

Fraundorf, S. H., Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. In *Journal of memory and language*, 65, 161-175. doi:https://doi.org/10.1016/j.jml.2011.03.004.

Hönemann, A., Wagner, P. (2015). Adaptive Speech Synthesis in a Cognitive Robotic Service Apartment: An Overview and First Steps Towards Voice Selection. In *Tagungsband Elektronische Sprachsignalverarbeitung ESSV 2015*, 135-142.

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. (1993). *Speaking: From intention to articulation*. Cambridge/London: MIT press. doi:https://doi.org/10.7551/mitpress/6393.001.0001.

Lickley, R. J. (2015). Fluency and disfluency. In Redford M. A. (Ed.), *The handbook of speech production*. Chichester: Wiley Online Library, 445-474. doi:https://doi.org/10.1002/9781118584156.ch20.

Malisz, Z., Henter, G. E., Valentini-Botinhao, C., Watts, O., Beskow, J. & Gustafson, J. (2019). Modern speech synthesis for phonetic sciences: A discussion and an evaluation. In Calhoun, S., Escudero, P., Tabain, M. & Warren, P. (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences*. Melbourne, Australia, 487-491. doi:https://doi.org/10.31234/osf.io/dxvhc.

Moniz, H., Batista, F., Mata, A. I. & Trancoso, I. (2014). Speaking style effects in the production of disfluencies. In *Speech Communication*, 65, 20-35. doi:https://doi.org/10.1016/j.specom.2014.05.004.

Moniz, H., Trancoso, I. & Mata, A. I. (2009). Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts. In *Proceedings of Interspeech 2009*. Brighton, United Kingdom, 6-10 September 2009. doi:https://doi.org/10.21437/Interspeech.2009-518.

Moniz, H., Trancoso, I. & Mata, A. I. (2010). Disfluencies and the perspective of prosodic fluency. In Esposito, A., Campbell, N., Vogel, C., Hussain, A. & Nijholt, A. (Eds.), *Development of multimodal interfaces: active listening and synchrony*. Berlin, Heidelberg: Springer, 382–396. doi:https://doi.org/10.1007/978-3-642-12397-9_33.

Moore, R. K. (2017). Appropriate voices for artefacts: some key insights. In Dassow, A., Marxer, R. & Moore, R., K. (Eds.), *Proceedings of the 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*. Skövde, Sweden, 25-26 August 2017, 7-11. doi:https://doi.org/10.3389/frobt.2016.00061.

Mühlack, B., Elmers, M., Drenhaus, H., Trouvain, J., van Os, M., Werner, R., Ryzhova, M., & Möbius, B. (2021). Revisiting recall effects of filler particles in German and English. In *Proceedings of Interspeech 2021.* Brno, Czechia, 30 August / 3 September 2021, 2021-1056. doi:https://doi.org/10.21437/Interspeech.

Niebuhr, O., & Fischer, K. (2019). Do not hesitate!-unless you do it shortly or nasally: How the phonetics of filled pauses determine their subjective frequency and perceived speaker performance. In *Proceedings of Interspeech 2019,* 15-19 September 2019 Graz, Austria, 2019-1194. doi:https://doi.org/10.21437/Interspeech.

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 AVI-CH Workshop on Advanced Visual Interfaces for Cultural Heritage (vol. 2091).* Grosseto, Italy, 1-4.

Prenger, R., Valle, R., & Catanzaro, B. (2019). Waveglow: A flowbased generative network for speech synthesis. In *Proceedings of the 44th International Conference on Acoustics, Speech and Signal Processing*. Brighton, United Kingdom, 12-17 May 2019, 3617-3621. doi:https://doi.org/10.1109/ICASSP.2019.8683143.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. URL: https://www.R-project.org/.

Rodero, E. (2017). Effectiveness, attention, and recall of human and artificial voices in an advertising story. Prosody influence and functions of voices. In *Computers in Human Behavior*, 77, 336–346. doi:https://doi.org/10.1016/j.chb.2017.08.044.

Savy, R. (2005). Specifiche per la trascrizione ortografica annotata dei testi. In Albano Leoni, F., Giordano, R. (Eds.), Italianoparlato. Analisi di un dialogo. Napoli: Liguori, 1-37.

Schettino, L., Betz, S., Cutugno, F., & Wagner, P. (2021a). Hesitations and individual variability in Italian tourist guides' speech. In Bernardasci, C., Dipino, D., Garassino, D., Negrinelli, S., Pellegrino, E., & Schmid, S. (Eds.), *Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications, STUDI AISV 8*. Milano: Officinaventuno, 243-262.

Schettino, L., Betz, S., & Wagner, P. (2021b). Hesitations distribution in Italian discourse. In *Proceedings of the 10th Workshop on Disfluency in Spontaneous Speech.* Paris, France, 25-27 August 2021, 29-34.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, A.R., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. In *Proceedings of the 43th International Conference on Acoustics, Speech and Signal Processing*. Calgary, Canada, 15-20 April 2018, 4779-4783. doi:https://doi.org/10.1109/ICASSP.2018.8461368.

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

Student (1908). The probable error of a mean. In *Biometrika*, 6, 1-25. doi:https://doi.org/10.2307/2331554.

Székely, É., Henter, G. E., Beskow, J., & Gustafson, J. (2019). How to train your fillers: uh and um in spontaneous speech synthesis. In *Proceedings of the 10th Speech*

Synthesis Workshop. Vienna, Austria, 20-22 September 2019, 245–250. doi:https://doi.org/10.21437/SSW.2019-44.

Voghera, M. (2017). *Dal parlato alla grammatica*. Roma: Carocci.

Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Eje Henter, G., Le Maguer, S., Malisz, Z., Székely, É., Tånnander, A., Vosse, J. (2019). Speech synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program. In *Proceedings of the 10th Speech Synthesis Workshop*. Vienna, Austria, 20-22 September 2019, 105-110. doi:https://doi.org/10.21437/SSW.2019-19.

Wang, Y., Stanton, D., Zhang, Y., Ryan, R.-S., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., & Saurous, R. A. (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In Dy, J., Krause, A. (Eds.), *International Conference on Machine Learning*. Stockholm, Sweden, 10-15 July 2018, 5180-5189.

Appendix 1

Target Stimuli

Disf_No Disf

LEN SP

1. La Certosa di San Martino costituisce in assoluto uno deiii maggiori complessi monumentali religiosi della città hh che è fra i più riusciti esempi diii architettura e arte barocca.
2. La certosa fu inaugurata e consacrataaa nel mille trecento sessantotto hh seppur i certosini avevano preso possesso del monasterooo dal mille trecento trentasette.
3. Il pavimento fu realizzato da Bonaventura Prestiii in preziosi marmi di diversi colori hh che produconooo un'apparente tridimensionalità.
4. La seconda cappella a sinistra della navata è quella diii San Bruno hh le cui decorazioni marmoree sonooo del Fanzago.
5. La seconda cappella di destra è quellaaa di San Giovanni Battista hh che fu decorata dal Fanzagooo nel mille seicento trentuno.

LEN FP

1. All'inizio del Seicento la direzione del cantiere passa aaa Giovan Giacomo di Conforto ehm che completaaa il progetto del Dosio.
2. In questa fase di ristrutturazione del complesso lavoraronooo pittori ehm che furono fra i più grandi artistiii del Seicento.
3. I lavori vennero affidati aaa Giovanni Antonio Dosio ehm che fu di fatto il primo responsabile delleee trasformazioni del complesso.
4. La facciata della chiesa trecentesca fu rimaneggiata sul finire del Cinquecentooo dal Dosio ehm a cui si deve il pronaooo a tre arcate.
5. Nella prima metà del diciottesimo secolo i lavori passarono aaa Nicola Tagliacozzi Canale ehm che farà rifare gliii spazi del priore.

No Disf_No Disf

1. Cronologicamente la Certosa di San Martino è la seconda certosa della Campania essendo nata diciannove anni dopo quella di San Lorenzo a Padula.
2. Le transenne di tutte le cappelle sonoo del Fanzago a cui si devono anche i festoni di frutta sui pilastri.
3. Nel registro inferiore della sala sono collocati alle pareti arredi mobiliari intarsiati i cui intagliatori furono Nunzio Ferraro e Giovan Battista Vigilante.
4. La chiesa delle donne era destinata ad uso esclusivo delle donne alle quali era proibito l'accesso alla certosa.
5. Le esecuzioni marmoree interne sono frutto dell'opera di Cosimo Fanzago che fu chiamato a ristrutturare la certosa dal mille seicento ventitrè al mille seicento cinquantasei.

Disf_Disf

1. Solo verso la seconda metà del Sedicesimo secolo il complesso fu dedicato aaa Martino di Tours hh probabilmente per la presenza nel luogo diii un'antica cappella preesistente a lui dedicata.
2. La terza cappella di sinistra è quellaaa dell'Assunta hh la quale presenta unaaa decorazione seicentesca.
3. Sul piazzale esterno al complesso certosino ehm è defilata sulla sinistra laaa chiesa delle Donne ehm che è opera diii Giovanni Antonio Dosio.
4. La facciata della chiesa trecentesca fu rimaneggiata successivamente daaa Cosimo Fanzago ehm che costruì nella prima metà del Seicento unaaa serliana.
5. La chiesa si compone di una navata unica e delleee cappelle laterali ehm che si succedono aiii lati della zona absidale.

Appendix 2

Test link:
https://phdmglunina.fra1.qualtrics.com/jfe/form/SV 6yAWNyk5xCDMHz0