EVALUATING THE IMPACT OF DISFLUENCIES ON THE PERCEPTION OF SPEAKER COMPETENCE USING NEURAL SPEECH SYNTHESIS

Ambika Kirkland[1], Marcin Włodarczak[2], Joakim Gustafson[1], Éva Székely[1]

[1]KTH Royal Institute of Technology, [2]Stockholm University
kirkland@kth.se, wlodarczak@ling.su.se, jkgu@kth.se, szekely@kth.se

ABSTRACT

Perceptions of speaker competence can be influenced by various factors, including the presence of disfluencies such as false starts and different types of repetitions. In this study, we generated utterances with false starts and each of three different types of repetitions using a text-to-speech system with implicit prosody control trained on spontaneous speech. A web-based listening task was conducted to evaluate the impact of these different types of disfluencies on perceptions of speaker competence. False starts were found to have the greatest negative impact, consistent with previous research showing a relationship between processing fluency and perceived competence. The results of this study have implications for public speaking training, as speakers can work on minimizing false starts and other disfluencies to improve perceived competence. Additionally, understanding the impacts of different types of disfluencies can help speakers choose strategies for minimizing their use and improving their overall fluency and effectiveness in communication.

Keywords: Speech perception, speech synthesis, public speaking, spontaneous speech, disfluencies

1. INTRODUCTION

Disfluencies, such as filled pauses, repetitions and speech repairs, are ubiquitous in spontaneous speech and play an important role in maintaining the smooth flow of conversation when delays occur in speech planning [1]. Disfluencies can inform listeners about upcoming delays and provide an insight into speech planning [2]. Nonetheless, there is evidence that disfluencies negatively impact a range of judgments about speakers, including their perceived competence [3, 4]. This is consistent with more general evidence that disrupting the fluent processing of information is one factor that can make the source of this information seem less competent [5]. However, the term “disfluency” encompasses a wide range of speech behaviors and not all of these affect the listener in the same way.

False starts, also called speech repairs, occur when a speaker begins a word or phrase and then stops and starts over to correct it. These follow the general pattern of *reparandum* (the original words), *editing term* (e.g. "I mean" or "uh"), and *alteration* (the correction) [6]. False starts have been found to negatively impact word monitoring latency when they occur mid-sentence, likely due to the disruption they cause in the flow of the speech [7]. In contrast, word repetitions seem to be less disruptive to comprehension. Operantly conditioning speakers to reduce their production of silent pauses increases the incidence of function word repetition [8], which may indicate that repeating function words plays a similar role to that of silent pauses in buying time while planning the next part of the utterance. On the other hand, fluency failures on content words do not seem to serve the same function and may occur when the speaker begins saying a word before the entire plan is ready [8].

Despite the link between speech disfluencies and perceived competence, and the evidence that repetitions and false starts reflect different cognitive processes and affect processing fluency to different degrees, some questions remain. Do specific types of repetitions have different effects on perceived competence? Do differing impacts on processing fluency actually translate to differing impacts on competence perception?

One challenge in trying to answer these questions is that disfluencies are peculiar to and characteristic of spontaneous speech [9]. This means that as an object of investigation, they are not very well suited to the common strategy of using lab-recorded speech to create stimuli for perception experiments. This is a more general dilemma in studying speech perception. Differences between read and spontaneous speech are apparent in turn-taking behaviors [10], stress position and boundaries between tone units [9], articulation rate [11], vowel reduction [11], and f0 range [12], and are even reflected in the neurophysiological correlates of speech processing [13]. Using spontaneous speech

would be ideal but this requires relinquishing control over the content of utterances. Many spontaneous speech corpora exist but it is highly unlikely that one could find specific utterances in these corpora which meet a given set of criteria in terms of structure and content and of which there are several instances from the same speaker varying only in terms of which disfluencies are present.

One potential alternative is to use neural text-to-speech (TTS) trained on spontaneous speech to create stimuli for speech perception experiments. Neural TTS has come to rival human speech in quality and naturalness in recent years [14], making spontaneous TTS an increasingly viable alternative to natural speech. Many stimuli can be produced without the time and expense of recording speech in a lab, and can be easily tailored to the specific demands of the experiment. Some spontaneous neural TTS systems also grant a measure of control over various prosodic features as well as the location and frequency of disfluencies such as filled pauses [15, 16]. Such systems have been leveraged recently to investigate the role of filled pauses and their interaction with prosodic features in the perception of uncertainty [17, 18].

In the present study, we used spontaneous TTS to investigate how false starts and different kinds of repetitions (function word repetitions, verb phrase repetitions, or the repetition of key content words) impact the perception of speaker competence. We hypothesized that false starts would have the most negative impact on perceived competence given their disruption of listening comprehension, while function word repetitions would have the smallest impact, with the other types of repetitions falling somewhere in between.

2. METHOD

2.1. Stimulus creation

2.1.1. Data and system

The training data for the TTS system were derived from a public-domain technology podcast which involves two male speakers of American English discussing technology news and reviewing products. Audio from the speaker with the most airtime was used in the training corpus. The audio was segmented into breath groups using a speaker-dependent breath detection method [19]. The Google Cloud Speech API [20] was used for the initial transcription. Filled pauses such as *uh* and *um* were identified using IBM Watson Speech to Text, combined with the output of the Gentle forced aligner [21]. These tokens were then inserted into the Google API transcription. Then the transcriptions were manually corrected, as especially the pronunciations of technology acronyms were not good enough. Finally, mean f0 and mean speech rate were automatically computed for each breath group in the training corpus using the Wavelet Prosody Analyzer toolkit [22].

The text-to-speech system was based on the sequence-to-sequence Tacotron 2 TTS engine, modified with a style-unit-level prosody control method similar to that described in [23] and [18] to enable implicit control of f0 and speech rate at inference. As opposed to setting specific values for speech rate and fundamental frequency, the system can be prompted to produce an f0 or speech rate based on a target percentile of the sampling distribution of these values in the training data. The neural vocoder HiFi-GAN [24] was used to decode the speech signal.

2.1.2. Stimuli

The sentences synthesized for the experiment all consisted of definitions of scientific terms with the general form: "The [description of the phenomenon] is called the **[eponym]**.", e.g., "The proportionality factor that relates temperature to kinetic energy in a gas is called the **Boltzmann constant**."

These sentences were created with a few factors in mind. First, we wanted to ensure that the repetition or false start would occur at a point in each sentence that was as similar as possible in terms of syntactic context, and that the actual repeated word or phrase would not vary. This allowed for some diversity in the content of the stimuli while constraining the immediate context of the disfluency. Secondly, we did not want participants to base their judgment on whether the sentences seemed accurate or plausible, so we chose factual statements about real-world phenomena.

Finally, we wanted to evoke a public speaking context where speaker competence would be of utility to the listener without actually testing comprehension by asking participants to place themselves in the shoes of someone trying to learn new information. Hence, we chose technical terms from domains such as physics, astronomy and medicine which would not be common knowledge to most people. These definitions were adapted from Wikipedia entries [25] to fit the stimulus format. We created four variations of 15 different sentences (60 items in total) by adding a repetition or false start: a function word repetition (*fw-rep*), verb

Condition	Example
fw-rep	"The proportionality factor that relates temperature to kinetic energy in a gas is called **the um the** Boltzmann constant."
vp-rep	"The proportionality factor that relates temperature to kinetic energy in a gas **is called um is called** the Boltzmann constant."
n-rep	"The proportionality factor that relates temperature to kinetic energy in a gas is called the **Boltzmann um Boltzmann** constant."
false-start	"The proportionality factor that relates temperature to kinetic energy in a gas **is known um is called** the Boltzmann constant."

Table 1: Examples of the four utterance types synthesized for the perception experiment (fw-rep = function word repetition, vp-rep = verb phrase repetition, n-rep = name repetition)

phrase repetition (*vp-rep*), name repetition (*n-rep*) and false start (*false-start*). A filled pause (*um*) was placed between the repeated word or phrase, or in the case of false starts, as an editing term between the reparandum and alteration. Examples of each utterance type are shown in Table 1. We initially considered including a fluent version of each sentence but ultimately used only non-fluent stimuli due to concerns that participants would attend only to the mere presence or absence of disfluencies. When synthesizing the stimuli we used prosody modification to ensure that all utterances had a falling intonation contour.

2.2. Perception experiment

A web-based subjective listening task was carried out to assess the impact of different kinds of repetitions on perceptions of a speaker's competence. We used a MUSHRA-like [26] design where participants viewed all four versions of each utterance side-by-side (in randomized order) and rated them each on a 5-point scale ranging from 1 ("not at all competent") to 5 ("extremely competent"). Participants were able to play each stimulus as many times as needed. To insure that participants were attending to the stimuli we implemented attention checks. Two extra sets of sentences were synthesized, with one audio clip in each set instructing participants to to give the

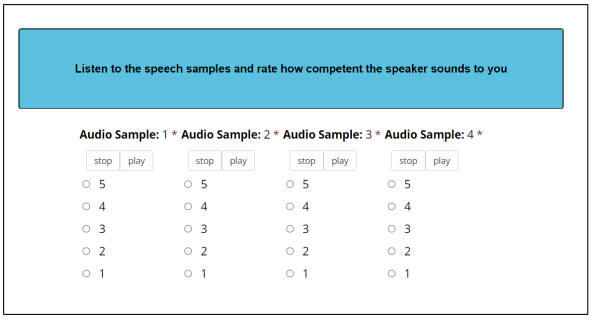

Figure 1: Interface of the perception experiment

stimulus a rating of "1". Responses to these stimuli were not included in the results.

Participants were native speakers of English recruited using the crowdsourcing platform Prolific. They were pre-screened to check that they met the inclusion criteria and affirmed that they were wearing headphones for the duration of the experiment and had no hearing impairments. We recruited new participants to replace any who failed one or both attention checks, encountered technical issues, or were unable to complete the experiment for other reasons, until we had collected data from a total of 40 participants. Of these, 43.6% were female and 56.4% were male.

3. RESULTS

A Friedman test with repeated measures on participants was carried out to determine whether responses differed between conditions. This test is a non-parametric equivalent to repeated-measures analysis of variance, suitable for non-normal ordinal-scale data. The test was significant, χ^2=14.03, $p < 0.005$, indicating a difference in participants' ranking of stimuli on the 5-point competence scale across conditions. Figure 2 shows the distribution of responses for each condition. Post-hoc Conover tests on mean ranks with a Holm-Bonferroni correction for multiple comparisons showed that utterances with false starts were rated as sounding significantly less competent than both utterances with function word repetitions (t=2.63, $p < 0.05$) and utterances with verb phrase repetitions

Condition	Mean	SD
fw-rep	3.11	0.77
vp-rep	3.17	0.80
n-rep	3.03	0.70
false-start	2.80	0.83

Table 2: Mean and standard deviation of competence ratings by condition

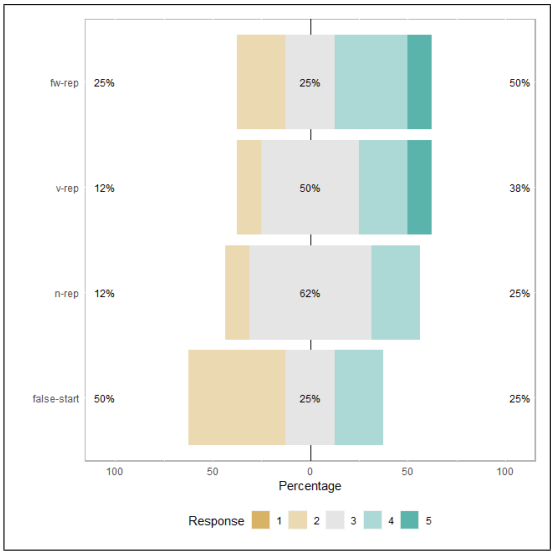

Figure 2: Distribution of responses by condition

(t=3.55, $p < 0.005$). No other pairwise comparisons were significant.

4. DISCUSSION

The results of the perception experiment partially confirmed our predictions. Utterances containing false starts were rated as sounding less competent than utterances with function word or verb phrase repetitions. However, there was no significant difference between false starts and name repetitions, or between any of the other repetition types. The relative predictability of where a disfluency might occur in an utterance may have attenuated some of their impact. The stimulus order was randomized in each trial, meaning participants would not know which type of repetition they were about to hear the first time they listened to a given audio file, but the amount of variation was nonetheless constrained by the stimulus design. If the effect of disfluencies is at least partly a function of how much they disrupt processing it makes sense that predictability would blunt this effect to some extent. Nonetheless, the disfluency that appears most disruptive to processing was associated with lower perceived competence. These findings are hence consistent with, and build upon, previous work showing a link between fluency and perceived competence [3, 4, 5].

It is also important to note that in addition to potentially indicating a speaker's competence, repetitions play a whole host of other roles in conversation. Bazzanella [27] provides a list of dozens of such roles, which include cognitive and meta-cognitive functions such as planning an utterance or marking or facilitating comprehension; interactional functions like marking surprise, agreement or disagreement; turn-taking behaviors such as holding or yielding a turn; and stylistic functions such as emphasis. The context, the speaker, and characteristics of the listener may all shape how disfluencies are interpreted and in turn how they influence perceptions of competence.

Spontaneous neural text-to-speech (TTS) has the potential to be a valuable tool for investigating various factors that impact speech perception. One advantage of this method is its ability to generate a wide range of stimuli, allowing researchers to systematically vary certain features while others are constrained. For example, sentences with the desired semantic content or syntactic structure could be synthesized with different speaker characteristics (such as age or gender, both of which have been shown to influence speaker competence), voice quality (e.g., creak/vocal fry), or modifications to pitch, intensity or speech rate. Since the use of spontaneous TTS as a tool for studying speech perception is relatively new, the methodology will need continued development as the capabilities of TTS evolve. More studies are needed to better understand the types of research questions this method is suited to investigate and to chart out pitfalls and best practices.

5. CONCLUSIONS

This study investigated the effects of false starts and different types of repetitions on a speaker's perceived competence, building upon previous research utilizing speech synthesis as a tool for studying speech perception. The results demonstrated that false starts had a negative impact on perceived competence, consistent with the idea that they may disrupt processing fluency for the listener. There were no significant differences observed between the various types of repetitions. These findings have practical implications for public speaker training, suggesting that minimizing false starts may improve perceived competence. While it has been generally assumed that all disfluencies negatively impact perceived competence, our results indicate that certain types of disfluencies may be more detrimental than others. The study demonstrates an example of how spontaneous neural TTS can be used to generate a range of stimuli that are systematically varied to investigate the various factors that influence speech perception.

6. ACKNOWLEDGEMENTS

This research is supported by the Riksbankens Jubileumsfond project CAPTivating – Comparative

Analysis of Public Speaking with Text-to-Speech (P20-0298).

7. REFERENCES

[1] H. H. Clark and T. Wasow, "Repeating words in spontaneous speech," *Cognitive Psychology*, vol. 37, no. 3, pp. 201–242, 1998.

[2] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[3] G. R. Miller and M. A. Hewgill, "The effect of variations in nonfluency on audience ratings of source credibility," *Quarterly Journal of Speech*, vol. 50, no. 1, pp. 36–44, 1964.

[4] J. K. Barge, D. W. Schlueter, and A. Pritchard, "The effects of nonverbal communication and gender on impression formation in opening statements," *Southern Communication Journal*, vol. 54, no. 4, pp. 330–349, 1989.

[5] D. M. Oppenheimer, "Consequences of erudite vernacular utilized irrespective of necessity: Problems with using long words needlessly," *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, vol. 20, no. 2, pp. 139–156, 2006.

[6] E. E. Shriberg, "Preliminaries to a theory of speech disfluencies," Ph.D. dissertation, University of California at Berkeley, Berkeley, CA, 1994.

[7] J. E. F. Tree, "The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech," *Journal of memory and language*, vol. 34, no. 6, pp. 709–738, 1995.

[8] P. Howell and S. Sackin, "Function word repetitions emerge when speakers are operantly conditioned to reduce frequency of silent pauses," *Journal of psycholinguistic research*, vol. 30, no. 5, pp. 457–474, 2001.

[9] P. Howell and K. Kadi-Hanifi, "Comparison of prosodic properties between read and spontaneous speech material," *Speech communication*, vol. 10, no. 2, pp. 163–169, 1991.

[10] C. Aruffo, "Reading scripted dialogue: Pretending to take turns," *Discourse Processes*, vol. 57, no. 3, pp. 242–258, 2020.

[11] G. P. Laan, "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style," *Speech Communication*, vol. 22, no. 1, pp. 43–65, 1997.

[12] P. Wagner and A. Windmann, "Re-enacted and spontaneous conversational prosodyâhow different?" *Proceedings of Speech Prosody 2016*, pp. 518–522, 2016.

[13] M. Drolet, R. I. Schubotz, and J. Fischer, "Authenticity affects the recognition of emotions in speech: behavioral and fmri evidence," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 12, no. 1, pp. 140–150, 2012.

[14] S. Shirali-Shahreza and G. Penn, "Mos naturalness and the quest for human-like speech," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 346–352.

[15] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis," in *The 10th ISCA Speech Synthesis Workshop*, 2019.

[16] S. Wang, J. Gustafson, and É. Székely, "Evaluating sampling-based filler insertion with spontaneous tts," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 1960–1969.

[17] E. Székely, J. Mendelson, and J. Gustafson, "Synthesising uncertainty: The interplay of vocal effort and hesitation disfluencies." in *INTERSPEECH*, 2017, pp. 804–808.

[18] A. Kirkland, H. Lameris, E. Székely, and J. Gustafson, "Where's the uh, hesitation? the interplay between filled pause location, speech rate and fundamental frequency in perception of confidence," in *Proceedings of Interspeech*, 2022, pp. 18–22.

[19] É. Székely, G. E. Henter, and J. Gustafson, "Casting to corpus: Segmenting and selecting spontaneous dialogue for tts with a cnn-lstm speaker-dependent breath detector," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6925–6929.

[20] G. LLC, "Google cloud speech api video model," https://cloud.google.com/speech-to-text.

[21] R. Ochshorn and M. Hawkins, "Gentle forced aligner[computer program]," 2017.

[22] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.

[23] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," *arXiv preprint arXiv:2009.06775*, 2020.

[24] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[25] Wikimedia Foundation, "Wikipedia, the free encyclopedia," http://en.wikipedia.org/, 2023, [Online; accessed 06-January-2023].

[26] International Telecommunication Union, Radiocommunication Sector, "Method for the subjective assessment of intermediate quality levels of coding systems," ITU Recommendation ITU-R BS.1534-3, 2015.

[27] C. Bazzanella, "Redundancy, repetition, and intensity in discourse," *Language sciences*, vol. 33, no. 2, pp. 243–254, 2011.