INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Pardon my disfluency: The impact of disfluency effects on the perception of speaker competence and confidence

Ambika Kirkland[1], *Joakim Gustafson*[1], *Éva Székely*[1]

[1]Division of Speech, Music & Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

kirkland@kth.se, jkgu@kth.se, szekely@kth.se

Abstract

Disfluencies are a hallmark of spontaneous speech and play an important role in conversation, yet have been shown to negatively impact judgments about speakers. We explored the role of disfluencies in the perception of competence, sincerity and confidence in public speaking contexts, using synthesized spontaneous speech. In one experiment, listeners rated 30-40-second clips which varied in terms of whether they contained filled pauses, as well as the number and types of repetition. Both the overall number of disfluencies and the repetition type had an impact on competence and confidence, and disfluent speech was also rated as less sincere. In the second experiment, the negative effects of repetition type on competence were attenuated when participants attributed disfluency to anxiety.

Index Terms: speech perception, speech synthesis, public speaking, spontaneous speech, disfluencies

1. Introduction

1.1. Disfluencies in spontaneous speech

Disfluency is a hallmark of spontaneous speech. Rather than delivering completely well-formed utterances, speakers frequently hesitate, repeat themselves, and perform self-repairs. Although some perspectives (e.g., [1]) might view disfluencies as errors or deficits in realizing fluent speech, they are more than mere noise in the speech signal. They index underlying cognitive processes such as lexical retrieval and speech planning [2, 3, 4], telegraph information about upcoming delays [5] and give listeners insight into what Brennan and Williams refer to as the *feeling of another's knowing* [6], allowing for the coordination of mental states during conversation.

Although disfluencies serve a functional role in speech and can convey friendliness [7] or spontaneity [8], most prior evidence suggests a negative impact of disfluencies on judgments about speakers. More disfluent speakers can appear more anxious [9] and less confident [10, 11], competent and dynamic [12, 7]. Disfluency can unfavorably affect a range of trait judgments related to social desirability [13, 14] as well as secondary judgments based on speaker's statements [7].

Moreover, the type and location of disfluencies, as well as the context in which they occur, can reflect different cognitive processes and speaker strategies for maintaining continuity during speech planning [2, 4] and impact processing fluency to different degrees. For example, filled pauses (FPs) make it easier for listeners to integrate words into their contexts [15], which is not the case for repetitions [16], and false starts increase word monitoring latency relative to simple repetitions [17].

It is still unclear, however, whether these differences translate to different effects on judgments about a speaker. In addition to a general link between the processing burden imposed by disfluency and its impact on evaluations [18, 19], recent research has shown that the type and location of disfluencies play a role in their effect on perceived competence [12] and confidence [10]. These findings are consistent with the notion that more severe processing disruptions might more negatively affect judgments, but relatively little is known about how variations within a disfluency type (for example, the syntactic context of disfluent repetitions) or interactions between commonly co-occurring disfluencies (such as repetitions and FPs) impact evaluations. Addressing these questions might shed light on how listeners perceive subtly "atypical" use of disfluencies, displayed for example by some second-language speakers or women and girls with autism spectrum disorders [20, 21, 22].

Another consideration is that listeners' theories about the cause of disfluencies may shape their judgments. Disfluency can trigger negative evaluations when listeners assume a disfluent speaker is not willing or able to communicate effectively [18] but it has been shown in non-speech contexts that providing an obvious explanation for disfluency can attenuate its effect [19], a phenomenon known as *discounting* [23]. If discounting effects apply to speech disfluencies as well, this could give speakers a concrete tool for potentially mitigating the effects of disfluency on how listeners perceive them.

1.2. Methods for studying disfluencies in speech

In order to measure the effect of disfluencies on listener judgments we need to create speech stimuli with varying levels of disfluency. This is no trivial task and previous studies have used different methods to address this challenge. One method (used for example by [12] and [7]) is to ask voice actors to produce fluent and disfluent versions of a script. A limitation of this approach is that read or acted speech and spontaneous speech vary on a number of dimensions and are perceived differently by listeners [24, 25, 26], so this type of stimuli not be well suited for studying characteristics of spontaneous speech.

Another approach is to begin with disfluent natural speech and then excise disfluencies, as used, e.g., in [13] and [9]. This approach is suitable for evaluating how accurately listeners can infer ground-truth speaker states or characteristics from speech but affords little control over the content of the utterances or the exact placement or types of disfluencies. Furthermore, FPs are often cliticized onto the previous word to form phonological words [5] (for example "we um" might be realized as "we.yum") which means that some FPs cannot be removed without cutting off part of the adjacent word.

Our approach is to create stimuli using a neural text-to-speech (TTS) system trained on spontaneous speech. The use of neural TTS for studying speech perception was suggested by [27] some years ago, and has become an increasingly viable

10.21437/Interspeech.2023-887

option as the quality and naturalness of synthesized speech improves. Neural TTS has been used recently to investigate the effect of filled pause location and prosodic features on perceptions of speaker confidence [28] and the impact of disfluencies on ratings of personality traits in the context of different speaking styles [8]. This method provides control over both the content and prosodic aspects of utterances and allows for the creation of stimuli with the characteristics of spontaneous speech.

1.3. Research questions and hypotheses

In the present work we sought to better understand how disfluencies affect listeners' evaluations on five different dimensions: general competence, task-specific competence, confidence, sincerity and friendliness. We included competence and confidence since the impact of disfluency on these dimensions has been investigated previously [12, 7] but the importance of specific characteristics of disfluencies is underexplored. Friendliness and sincerity were included to gain more insight into how speech disfluencies impact the perception of positive traits.

We used a TTS system trained on spontaneous speech to synthesize stimuli with different types and numbers of disfluencies. According to [4] speakers repeat certain words in certain contexts more often because those repetitions best facilitate smooth and timely communication. We therefore reasoned that more "typical" repetitions, i.e. words that are repeated more often relative to the overall frequency of that word in a given context, would have less of an impact on judgments because they are more expected and less disruptive. The repetitions used in the experiments are summarized in Table 1. We also explored whether discounting effects might mitigate negative impacts of disfluency on evaluations by carrying out an experiment in which participants had the opportunity to attribute disfluency to anxiety. We propose the following hypotheses:

H1: Perceived competence and confidence will decrease as the overall number of disfluencies increases.

H2: The effect of FPs and repetitions should not be directly additive because FPs often occur with repetitions and give listeners a "heads up" about delays [4].

H3: Repetitions that are less common in spontaneous speech will have a greater impact on competence and confidence.

H4: More disfluent speech will be rated as friendlier and more sincere.

H5: If listeners are able to attribute disfluency to anxiety, the effects of disfluency on competence will be reduced.

2. Data and Synthesis

For the synthesis of the samples we use a TTS model built on the ThinkComputers Corpus (TCC), described in [29]. The corpus is created from recordings of a podcast which is available in the public domain. [1] In the podcast, two male speakers of American English discuss technology news and review computer hardware and software. Their speaking style can be described as extemporaneous, as they speak freely around a prepared outline. As a result, the corpus naturally contains disfluencies. The corpus includes 9 hours of speech from one of the speakers. To improve audio quality, the utterances in the corpus were processed using the Adobe Podcast enhance function[2].

[1] https://archive.org/details/podcasts_miscellaneous Creator: ThinkComputers

[2] https://podcast.adobe.com/enhance

The TTS system was trained using a modification of a PyTorch implementation[3] of the sequence-to-sequence neural TTS engine Tacotron 2 [30]. To control the level of fluency specifically with regard to filled pauses (FPs), the corpus is divided into two parts: utterances that contain FPs, and utterances that do not. An 8-dimensional speaker style embedding is added to the Tacotron 2 (the implementation closely following [31]), and each utterance in the training data is given one of two speaker IDs. One is reserved for utterances without filled pauses (1436 out of the 4906 samples), which we refer to as **ID-noFP**, while training samples with at least one FP ('uh' or 'um') receive the other, which we call **ID-FP**. Both speaker embeddings include data containing repetitions. In parallel to the embedding, we introduce an utterance-level prosody control, similar to [32]. Mean f0 and speech rate at the utterance level are added to the training. Both speaker embedding and prosodic features are added to a model, transfer learned on a model trained on the same corpus on a base Tacotron 2 architecture for 92.5k iterations. To allow for the additional features the relevant input dimensions to the attention, LSTM, projection and gate layers are padded with zeros. These additions increase the number of parameters to 28.28M from 28.19M in the base model. The model with embedding and prosodic features is trained for an additional 45k iterations on 4 GPUs with 28 batch size.

3. Perception experiments

3.1. Experiment 1

The first experiment investigated whether the number and type of disfluencies affect ratings of a speaker's competence, confidence, sincerity and friendliness. We varied the number and type of repetitions as well as the presence or absence of FPs in 14 different public speaking scenarios to create audio clips of 30-40-seconds in length. In half of the scenarios (the "lecture" context) the speaker explained scientific concepts, such as how scientists measure the expansion of the universe. In the other half (the "instruction" context) the speaker gave instructions about outdoor skills, such as how to build a shelter.

The category and number of repetitions were varied by repeating words at locations that were consistent across stimuli, resulting in three different repetition versions. Version A contained four repetitions and versions B and C contained these same four repetitions plus four additional repetitions, for a total

[3] https://github.com/NVIDIA/tacotron2

Table 1: *Repetitions with frequencies per 1000 occurrences*

	freq.	**examples**
Repetitions used in all stimuli		
contraction	33.2	we'll
relative pronoun	37.7	what
conjunction	30.8	and
determiner	28.8	some
Less typical repetitions (Condition B)		
preposition	14.3	in, between
misc. function words	22.3	by, how
More typical repetitions (Condition C)		
"the", complex subject NP	65.0	**The** number that helps measure
"the", complex object NP	55.0	measure **the** echoes
"a", complex object NP	59.0	there's **a** direct relationship
"a", complex pred. nominative NP	55.0	stretch **a** long sturdy branch

of eight each. The difference between version B and C is that the unique repetitions in version B were less common repetitions, while those unique to version C were more common. The repetitions used in each version and their frequency per thousand occurrences as reported by [4] are shown in Table 1.

We also varied whether or not FPs were present. Stimuli with FPs were created by synthesizing utterances with the filler "um" between repeated words using the **ID-FP** embedding (see Section 2). Utterances without FPs were synthesized with the non-hesitant version of the TTS system (**ID-noFP**). This resulted in six disfluent versions of each utterance: three repetition versions with FPs and three repetition versions without FPs. Figure 2 shows an example of these variations. Audio samples are available on the demo page [4].

Fluent: Good morning everyone and welcome to the first lecture of Understanding the Early Universe. **In** this lecture **we'll** discuss how scientists can measure **the** echoes of the distant past. Let's start **by** learning the definitions of **some** important terms. **The** number that helps us estimate **how** fast the universe is expanding is called the Hubble constant. **What** this constant tells us is that there's **a** direct relationship **between** a galaxy's distance **and** how quickly it's moving away from us.
A: In this lecture **we'll [um] we'll** discuss how scientists can measure the echoes of the distant past.
B: **In [um] in** this lecture **we'll [um] we'll** discuss how scientists can measure the echoes of the distant past.
C: In this lecture **we'll [um] we'll** discuss how scientists can measure **the [um] the** echoes of the distant past.

Figure 2: *Examples of experimental stimuli. To create the disfluent versions, the highlighted words were repeated with or without FPs (blue in version A, red and blue in version B, green and blue in version C)*

Twenty self-reported native speakers of English were recruited via the crowdsourcing platform Prolific [5] and participated in a web-based experiment on cognition.run. Half of participants identified as male and half as female. They listened to and rated 14 audio clips in random order. Each participant heard stimuli from every condition, but heard only one version of each scenario. They were randomly assigned a unique combination of scenarios and conditions. The number of times a particular version of a scenario was presented was balanced across participants. Participants could listen to the stimuli as many times as needed, and rated the speaker from 1 to 7 on how competent, confident, sincere and friendly they sounded, and how much they would rely on the speaker to teach them something new.

[4] www.speech.kth.se/tts-demos/disfluency2023
[5] http://prolific.co

Table 2: *ANOVA results summary with F scores and p values. Significant effects are marked with an asterisk (*).*

	Rep. type		**Filled pause**		**Rep * FP**	
	F	*p*	*F*	*p*	*F*	*p*
Competence (general)	**15.99**	**< .001***	2.841	0.11	0.38	0.68
Competence (task-related)	**13.99**	**< .001***	**9.96**	**< .01***	**3.26**	**< .05***
Confidence	**34.11**	**< .001***	**5.86**	**< .05***	1.34	0.27
Sincerity	**3.71**	**< .05***	<0.01	0.95	**5.11**	**< .05***

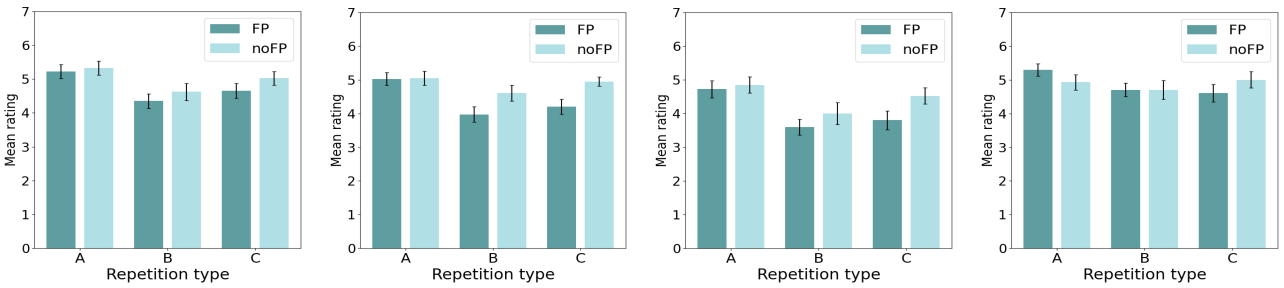
3.1.1. *Experiment 1 results*

We carried out a 3 (repetition type: A, B, C) x 2 (FP, no FP) x 2 (context: lecture, instructions) repeated measures ANOVA on each measure, with post-hoc Holm-Bonferroni tests. Significant effects are summarized in Table 2. As shown in Figure 1, participants rated speakers higher on both confidence and task-specific competence when the utterances contained few disfluencies (condition A), but between the two conditions with eight disfluencies, those with more uncommon disfluencies (B) were rated less competent and confident than those with more common disfluencies (C). There was no effect of FPs on general competence, but utterances with FPs were rated less confident and lower on task-related competence. Sincerity was affected by the number but not the type of repetitions. Condition C was rated as more sincere than both A and B, which were not significantly different from one another. Though there was no main effect of FPs on sincerity, the effects of repetition on sincerity were only significant for utterances with FPs. There were no effects of any of the manipulations on perceived friendliness, and the effect of context was not significant for any measure.

3.2. Experiment 2

The second experiment explored whether the effects of repetition type and frequency on perceived competence would be attenuated if participants attributed the disfluency to anxiety. Disfluency is associated with state anxiety [9] so this provides a realistic alternative explanation. Half of participants (the "control" condition) listened to an unmodified subset of the stimuli from Experiment 1. Four versions of each of four randomly chosen lecture scenarios without FPs (one fluent, one with each set of repetitions) were used in this condition. The other half of the participants (the "anxiety attribution" condition) received the same stimuli but with a small modification: Each clip included a short apology for being anxious, synthesized with the same voice. The utterances were otherwise identical to those used in the control version and in Experiment 1.

Thirty-two self-reported native speakers of English were re-



(a) *Competence (general)* (b) *Competence (task-specific)* (c) *Confidence* (d) *Sincerity*

Figure 1: *Mean ratings of competence (general and task-specific) confidence and sincerity in Experiment 1*

cruited via Prolific and randomly assigned to either the "control" condition or the "anxiety attribution" condition (16 per condition). The task followed a similar procedure to Experiment 1, but each participant rated only 5 stimulus items. Control participants rated the speaker's general and task-specific competence, as in Experiment 1, while participants in the other group also rated how anxious the speaker sounded.

3.2.1. Experiment 2 results

A 4 (repetition type) x 2 (condition) mixed factorial ANOVA showed a significant main effect of repetition type on both general competence, F (3) = 10.77, p $<$ .001 and task-related competence F (3) = 18.77, p $<$ .001 as well as a significant interaction between condition and repetition for both general F (3) = 9.74, p $<$ 0.001 and task-related competence F (3) = 9.57, p $<$ .001. The simple main effect of general competence was significant only in the control condition, and post-hoc tests show that all differences between levels of repetition (shown in Table 3) were significant in the control group. The simple main effect of task competence was significant in both conditions, but only the difference between fluent utterances and repetition type C was significant in the anxiety attribution condition, while in the control condition all means except repetition type A and C were significantly different. There was no main effect of condition. Because we purposely did not mention anxiety to the control group, we cannot compare anxiety ratings between groups. However, a one-sample t-test showed that the mean anxiety rating of 4.69 was significantly higher than the middle value of 4 on the scale, t = 3.92, p $<$ .001.

4. Discussion

Our results confirm hypothesis **H1**, that disfluencies negatively impact perceived competence and confidence. However, general competence was affected only by repetitions, while task-specific competence was also impacted by FPs. In **H2** we predicted that a combination of FPs and repetitions would not be strictly additive. This seems to be true for general competence, where no interaction was found between repetitions and FPs, but for task-related competence, the effect of repetitions was stronger when FPs were present. This may be due to how we operationalized task-related competence, by asking participants if they would rely on the speaker to teach them something new. Teaching ability depends on not only competence, but also factors such as how engaging a speaker is, so this question may have captured dimensions that we did not intend to measure.

We confirmed the hypothesis **H3** that less typical repetitions (those found by [4] to occur less often in spontaneous speech) would have a greater impact on evaluations. Between utterances with the same number of repetitions, those with more uncommon repetitions were rated less competent and confident. This pattern only held for general competence in Experiment 1, but affected both measures of competence in Experiment 2. It may be that the smaller number of stimuli in Experiment 2 made subtle differences more salient. Although we have shown that the type of repetition does matter, we cannot be certain why less common repetitions had a stronger impact on evaluations. These repetitions differ in a number of ways from more common repetitions: they tend to be lower-frequency words, are a more diverse category, and tend to disrupt larger syntactic constituents. One approach in future work could be to repeat the same lexical item in different syntactic contexts.

Table 3: *Means and standard deviations of competence ratings. Greater than or less than symbols indicate significant differences between means, $p < 0.05$*

	fluent		**Rep. A**		**Rep. B**		**Rep. C**
			General competence				
Control	5.69 (1.01)	$>$	4.69 (1.01)	$>$	3.38 (1.36)	$<$	4.13 (1.31)
Anxiety	5.13 (1.02)	=	5.19 (1.17)	=	4.81 (1.11)	=	4.63 (1.09)
			Task competence				
Control	5.5 (1.10)	$>$	4.56 (1.21)	$>$	3.19 (1.22)	$<$	4.00 (1.27)
Anxiety	5.25 (1.18)	=	5.00 (1.10)	=	4.69 (0.95)	=	4.63 (1.02)

Despite previous findings that disfluencies can make speakers seem more friendly [7], **H4** was not confirmed. There was no effect of disfluency on perceived friendliness, and more repetitions (combined with FPs) made speakers seem less sincere. This may be because listeners did not find judgments about friendliness or sincerity pertinent to delivering a lecture or instructions. Sincerity ratings may have been affected by the general negative impression of disfluencies, whereas participants may not have felt that they had enough information to judge friendliness. Future work could explore this further by including scenarios where friendliness or sincerity are more relevant.

The second experiment provided additional confirmation of **H3**, and also supports **H5**. When participants heard that the speaker was anxious about giving a lecture, the effect of repetitions on competence was eliminated (and it is also worth noting that the "anxious" voices were not rated as sounding less competent overall). Presumably, in line with [19], this represents a discounting effect: listeners may have concluded that the repetitions were unrelated to the speaker's competence because they were offered a better explanation. One of the participants commented that admitting to anxiety was a "classic public speaking mistake" but our results suggest otherwise. On the contrary, when speakers anticipate that they will be disfluent (whether because they really are nervous or because they are unfamiliar with the material they need to present) an excuse might lessen the impact of a halting delivery. Another implication of these findings is that listeners actually seem to attribute states, such as anxiety, to "speakers" they know are not real people.

5. Conclusions

Our results show that the characteristics of different types of disfluencies and the contexts in which they appear may play a role in how they impact evaluations. We found that both the number and type of speech disfluencies affect perceived competence, sincerity and confidence. The effects on competence are attenuated by discounting effects, so asking listeners to pardon our disfluencies may be an effective way to reduce their impact how we are perceived. This may be particularly relevant for speakers who use disfluencies atypically, such as some L2 speakers or women with ASD. Future studies could also examine how speaker characteristics such as gender or age may moderate the effects of disfluencies on evaluations.

6. Acknowledgements

This work is supported by the Swedish Research Council project Perception of speaker stance (VR-2020-02396), and the Riksbankens Jubileumsfond project CAPTivating – Comparative Analysis of Public speaking with Text-to-speech (P20-0298).

7. References

[1] N. Chomsky, *Aspects of the Theory of Syntax*. MIT press, 2014, vol. 11.

[2] 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.

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

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

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

[6] S. E. Brennan and M. Williams, "The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers," *Journal of Memory and Language*, vol. 34, no. 3, pp. 383–398, 1995.

[7] 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.

[8] J. Gustafson, J. Beskow, and É. Székely, "Personality in the mix – investigating the contribution of fillers and speaking style to the perception of spontaneous speech synthesis," *Proc. SSW*, pp. 48–53, 2021.

[9] J. A. Harrigan, I. Suarez, and J. S. Hartman, "Effect of speech errors on observers' judgments of anxious and defensive individuals," *Journal of Research in Personality*, vol. 28, no. 4, pp. 505–529, 1994.

[10] A. Kirkland, M. Włodarczak, J. Gustafson, and É. Székely, "Perception of smiling voice in spontaneous speech synthesis," in *Proc. SSW*, 2021, pp. 26–28.

[11] É. Székely, J. Mendelson, and J. Gustafson, "Synthesising uncertainty: The interplay of vocal effort and hesitation disfluencies." in *Proc. Interspeech*, 2017, pp. 804–808.

[12] 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.

[13] C. H. Lay and B. F. Burron, "Perception of the personality of the hesitant speaker," *Perceptual and Motor Skills*, vol. 26, no. 3, pp. 951–956, 1968.

[14] M. Wester, M. Aylett, M. Tomalin, and R. Dall, "Artificial personality and disfluency," in *Proc. Interspeech*, 2015.

[15] M. Corley, L. J. MacGregor, and D. I. Donaldson, "It's the way that you, er, say it: Hesitations in speech affect language comprehension," *Cognition*, vol. 105, no. 3, pp. 658–668, 2007.

[16] L. J. MacGregor, M. Corley, and D. I. Donaldson, "Not all disfluencies are are equal: The effects of disfluent repetitions on language comprehension," *Brain and Language*, vol. 111, no. 1, pp. 36–45, 2009.

[17] 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.

[18] M. Dragojevic and H. Giles, "I don't like you because you're hard to understand: The role of processing fluency in the language attitudes process," *Human Communication Research*, vol. 42, no. 3, pp. 396–420, 2016.

[19] 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.

[20] J. Cenoz, "Pauses and communication strategies in second language speech." ERIC Document ED 426630, 1998.

[21] J. Parish-Morris, M. Y. Liberman, C. Cieri, J. D. Herrington, B. E. Yerys, L. Bateman, J. Donaher, E. Ferguson, J. Pandey, and R. T. Schultz, "Linguistic camouflage in girls with autism spectrum disorder," *Molecular Autism*, vol. 8, no. 1, pp. 1–12, 2017.

[22] J. K. Lake, K. R. Humphreys, and S. Cardy, "Listener vs. speaker-oriented aspects of speech: Studying the disfluencies of individuals with autism spectrum disorders," *Psychonomic Bulletin & Review*, vol. 18, pp. 135–140, 2011.

[23] R. F. Bornstein and P. R. D'Agostino, "The attribution and discounting of perceptual fluency: Preliminary tests of a perceptual fluency/attributional model of the mere exposure effect," *Social Cognition*, vol. 12, no. 2, pp. 103–128, 1994.

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

[25] 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.

[26] P. Wagner and A. Windmann, "Re-enacted and spontaneous conversational prosody – How different?" *Proc. of Speech Prosody*, pp. 518–522, 2016.

[27] S. King, "A reading list of recent advances in speech synthesis," in *Proc. ICPHS*, 2015.

[28] A. Kirkland, H. Lameris, É. 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 *Proc. Interspeech*, 2022, pp. 18–22.

[29] 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 *Proc. ICASSP*, 2019, pp. 6925–6929.

[30] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[31] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *Proc. ICASSP*, 2020, pp. 6189–6193.

[32] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," *Proc. Interspeech*, pp. 4432–4436, 2020.