Where's the uh, hesitation? The interplay between filled pause location, speech rate and fundamental frequency in perception of confidence

Ambika Kirkland, Harm Lameris, Éva Székely, Joakim Gustafson

KTH Royal Institute of Technology

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

Abstract

Much of the research investigating the perception of speaker certainty has relied on either attempting to elicit prosodic features in read speech, or artificial manipulation of recorded audio. Our novel method of controlling prosody in synthesized spontaneous speech provides a powerful tool for studying speech perception and can provide better insight into the interacting effects of prosodic features on perception while also paving the way for conversational systems which are more effectively able to engage in and respond to social behaviors. Here we have used this method to examine the combined impact of filled pause location, speech rate and f0 on the perception of speaker confidence. We found an additive effect of all three features. The most confident-sounding utterances had no filler, low f0 and high speech rate, while the least confident-sounding utterances had a medial filled pause, high f0 and low speech rate. Insertion of filled pauses had the strongest influence, but pitch and speaking rate could be used to more finely control the uncertainty cues in spontaneous speech synthesis.

Index Terms: speech synthesis, speech perception, expressive speech synthesis, paralinguistics

1. Introduction

Reducing uncertainty is an important goal in communication, allowing for smooth and coordinated interactions [1] and in order to achieve this goal in a conversation it is necessary to convey one's level of certainty to others. There are a number of ways of doing this in human communication. While the lexical content of utterances plays a large role [2], various prosodic characteristics also make important contributions to judgments of how certain or confident a speaker sounds. More specifically, loudness [3], falling intonation (or absence of a rising intonation) [3] [4], lower f0 [5] [6] and faster speech rate [7] [6] [3] have been shown to correspond to perceived certainty. The presence of disfluencies such as filled pauses (e.g., *um* or *uh*) can also convey uncertainty [4] [8] [6], and their influence seems to interact with that of prosodic correlates of uncertainty [9] [10].

There is some evidence that the position of filled pauses within a sentence may play a role as well. Dinkar et al. [8] have found that the effect of filled pauses on perceived speaker confidence seems to be somewhat stronger when they occur in the middle of an utterance rather than the beginning. However, this evidence comes in the form of small differences in the strength of correlations between ratings of speaker confidence as measured by a handful of annotators and how often fillers occur in a given position. Furthermore, the annotators rated videos of the speakers and hence had access to visual information as well when making their judgments. More research is needed to establish the role filler position plays, as well as its interaction with prosodic features.

2. Related Work

The majority of previous work investigating the role of prosodic features in expressing confidence or certainty has focused on natural, mostly read speech. However, it is essential to also study spontaneous speech. Perceptual analysis of spontaneous characteristics of speech is traditionally done through one of three methods: a) corpus-based studies using stimuli extracted from ecologically valid, real-life speech recordings b) controlled experiments using lab-recorded speech stimuli, prompting participants to mimic or re-enact certain desired characteristics c) controlled experiments with acoustically manipulated speech stimuli. Presenting listeners with previously recorded spontaneous speech implies that researchers surrender a level of control when designing their experiments, which makes rigorous hypothesis testing difficult. The problem with the approach of specifically recorded stimuli is that many of the studied speech phenomena are normally produced semi-subconsciously, and when subjects are prompted to reproduce speech featuring for example filled pauses in specific places and with specified lengths, the recordings often sound acted or forced and are not representative of real spontaneous speech [11]. There is even neurophysiological evidence that acted emotional speech is processed differently from authentic emotional speech [12]. Laan [13] and Wagner and Windmann [14] investigated the effect of using scripted dialogues in studies where their speakers re-enacted earlier spontaneous interactions in order to investigate how changes in intonation, duration, and spectral features are perceived. Finally, speech manipulation procedures produce stimuli that often sound so unnatural that they influence listeners' perception in ways that bias the conclusions of the experiment. Moreover, many characteristics of spontaneous speech are too complex and difficult to approximate by modifying and manipulating recorded speech segments [15]. Even if the manipulation is integrated in the synthesis system, its application at the word level risks breaking prosodic constructs which typically extend beyond the word level [16].

There have been several studies on making unit selection TTS trained on read speech more spontaneous and expressive by inserting fillers in its text input [17], [18]. When filled pauses were inserted into utterances selected from spontaneous speech corpora, no significant decrease in naturalness was observed. Filled pauses have also been added to a read speech unit selection synthesizer in order to alter the perceived personality of the voice [19]. Lasarcyk et al used an articulatory speech synthesizer to synthesize utterances with varying certainty [20]. Their stimuli consisted of one-word German utterances, with a rising or falling pitch, that were preceded with or without an initial filled pause, and with a long or short silence. They conclude that these cues are additive, so that more uncertainty cues lead to higher perceived level of uncertainty. The same additive effects have been found in the perception of turn taking cues [21].

We propose using spontaneous speech synthesis as a research tool for speech perception. This approach allows for controlled variation of both the linguistic content and the acoustic realisation. Our neural TTS [22] is built on ecologically valid spontaneous speech data and enhanced with capabilities to vary characteristics like breathing [23] and filled pauses [24] independently, as well as producing laughter and smiling voice [25]. In this paper we introduce implicit control of mean speaking rate and mean pitch, on breath group or word level, to our spontaneous speech synthesizer. This allows us to investigate how varying these features influences the perception of paralinguistic information and speaker characteristics. In the current study we will make use of our spontaneous TTS in order to investigate the interplay between filled pause location, speech rate and fundamental frequency in the perception of certainty.

Hypothesis 1 : The contributions of filled pause location, speech rate, and pitch to perceived certainty will be as follows: Utterances with no filler will be rated as more confident, followed by the initial and then medial filler position; faster speaking rate will be perceived as more confident and slower speaking rate as less confident; and lower f0 will be rated as more confident while higher f0 is rated as less confident.

Hypothesis 2 : There will be an additive effect of filled pause location, speaking rate and mean pitch on perceived speaker confidence.

3. Data and synthesis

3.1. Data

The corpus used for voice building originates from the audio recordings from the Trinity Speech-Gesture Dataset (TSGD) [26], comprised of 25 impromptu monologues performed over multiple recording sessions by a male speaker of Hiberno-English. The monologues are on average 10.6 minutes long, spontaneously and without interruption, on topics such as hobbies, daily activities, and interests. During the monologues, he addresses a person seated behind the cameras who is giving visual, but no verbal, feedback. The monologues are separated into breath groups to create the voice training data using the approach described in [23], whereby consecutive breath groups are combined to form overlapping utterances no longer than 11 seconds.

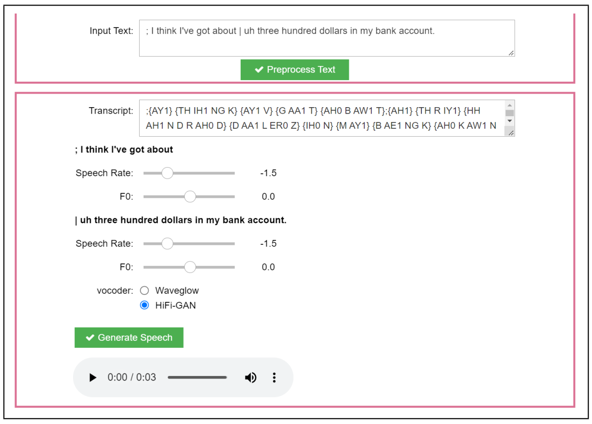
A breath or silent pause is the most probable location for a change in style for a speaker, although a speaker might also change the speaking style within a breath group or inter-pausal unit. To improve the consistency of the prosodic features, we identified these breaks in style through listening tests and where necessary split a breath group in multiple style units. The prosodic features are measured and summarised at the style unit level to provide mid-level control that is not distorted by audible changes in style within a breath group. In total, 284 additional style units were identified in a total of 3725 breath groups used in the corpus. Breaths, silent pauses and other style breaks are further identified in the corpus with a separate lexical token, allowing the system to include the prosodic features to the right parts of the utterances in both training and inference through identification of breaths (';'), silent pauses (','), and style tokens ('|').

3.2. System

The TTS system was trained using a modification of a PyTorch implementation[1] of the sequence-to-sequence neural TTS engine Tacotron 2 [27]. The modification implements a style-unit-level prosody control method, similar in approach to [28], to be able to direct f0 and speech rate at inference. As the inputs, speech rate (syllables/second) and mean f0 (measured over each style unit and excluding breaths and silent segments) were normalized. Normalization is performed by aligning the 1st and the 99th percentile points of the input data to the values of -1 and 1 respectively, while allowing outliers to go outside of that range. Normalized values for both features are appended to each utterance's encoded text for the tokens belonging to that style unit. The enhanced encoder output is then passed to the attention and decoder blocks from a pre-trained model. Transfer learning based on a model trained on a large read-speech corpus has been shown to improve the quality of spontaneous speech synthesis [22], and similarly it benefits the training of a modified TTS allowing prosody control to use a pretrained spontaneous TTS as a basis.

For this study, a voice was first trained on the TSGD corpus for 72.500 iterations using a pretrained model on the LJ Speech corpus [29] for transfer learning. In order to fit the additional features (speech rate and mean f0) to the model, the input dimension to the attention, LSTM, projection and gate layers in the decoder are expanded in the relevant dimension by two. The additional weights added to the model are initialized with zero values. As such, at the start of the training the model evaluates as the pre-trained model. This padded model was then used in the modified system to train on the TSGD corpus with f0 and speech rate features measured at the style unit level, for another 50.000 iterations. This method allows for modifying mean f0 and speech rate on utterance level based on the natural distribution of these features in the corpus, as opposed to direct manipulation. The speech signal is decoded from the output using the neural vocoder HiFi-GAN [30].

For inference, an interface was developed, allowing for easy placement of filled pauses, laughter, breath tokens, and style unit breaks in the input text and manipulation of the prosodic features of individual breath groups or style units (Figure 1).



Input Text: ; I think I've got about | uh three hundred dollars in my bank account.

Transcript: ;{AY1} {TH IH1 NG K} {AY1 V} {G AA1 T} {AH0 B AW1 T};{AH1} {TH R IY1} {HH AH1 N D R AH0 D} {D AA1 L ER0 Z} {IH0 N} {M AY1} {B AE1 NG K} {AH0 K AW1 N

; I think I've got about

Speech Rate: -1.5

F0: 0.0

| uh three hundred dollars in my bank account.

Speech Rate: -1.5

F0: 0.0

vocoder: ○ Waveglow
● HiFi-GAN

Figure 1: *The web-based TTS interface*

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

4. Evaluation

4.1. Stimuli

Stimuli were generated with the TTS system from 8 utterances of doxastic semantic modality: beginning with the phrase “I think”, and containing either an initial filled pause (*uh*), a medial filled pause, or no filled pause. The locations of the medial pauses were chosen based on a native speaker’s judgment of which position within each sentence made most sense semantically. Each version was synthesized with combinations of high, medium and low f0 and speech rate for a total of 216 different stimuli[2]. During training, values of -1 and 1 corresponded to the 1st and 99th percentile of the two normalized prosodic features in the training data. In order to achieve more perceptually clear low, medium, and high values, we set these prosodic features to -1.5, 0 and 1.5. The low and high settings produce stimuli that are moderately higher and lower in pitch and speaking rate, but without the artificial effects that accompany direct manipulation of the waveform, since these are based on the natural range of the speaker’s actual realizations.

Table 1: *Example of a test utterance, filled pauses are in bold.*

Filled Pause	Utterance
None	*I think that’s the more accurate version.*
Initial	***Uh**, I think that’s the more accurate version*
Medial	*I think that’s the more **uh**, accurate version*

4.2. Acoustic features

In order to ensure that high, medium and low speech rate and fundamental frequency were realized in the stimuli as intended, these features were measured using Praat [31]. Mean f0 was indeed highest in the high f0 condition (146.88) followed by the medium f0 condition (122.17) and lowest in the low f0 condition (106.82). Analysis of variance showed that the effect of f0 category on measured f0 was significant, and pairwise comparisons using the Šidák correction showed a significant difference between each of the three categories, $p < .001$. Speech rate in syllables per second was 3.24 in in the low speech rate condition, 3.57 in the medium speech rate condition and 3.84 in the high speech rate condition. Since speech rate measurements can be skewed by the actual words used and the presence of fillers, we compared utterances with matching content in our analysis. A linear mixed effects model with individual sentences treated as a subject variable showed that measured speech rate was significantly different across categories for each filler position, and pairwise comparisons corrected for familywise error with the Šidák correction confirmed that all three speech rate categories differed significantly from one another in measured speech rate, $p < .01$.

4.3. Perception test

The stimuli were evaluated using a web-based listening task. Thirty-five participants recruited via Prolific were asked to listen to and rate each item on a sliding scale anchored with “very hesitant” at the far left, “neutral” in the center and “very confident” to the far right. Participants responded by moving the slider, which began in the central position. The rating scale ranged from 0 to 100 with increments of 1, but participants did not see numbers when rating the stimuli.

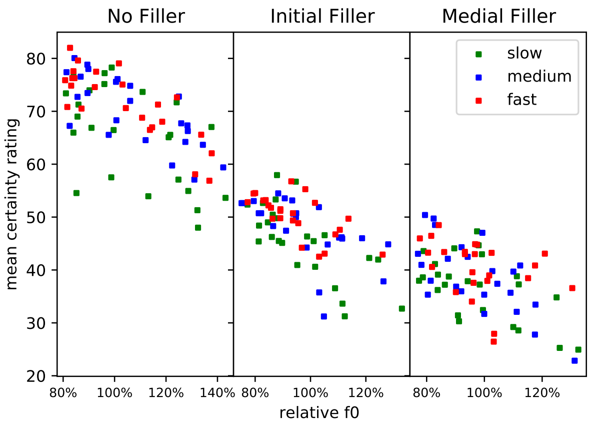


Figure 2: *Mean certainty ratings by stimulus. Relative f0 scores are obtained by dividing the measured f0 of an individual sample by the mean over all realisations of an utterance (with varying prosodic features and fillers).*

5. Results

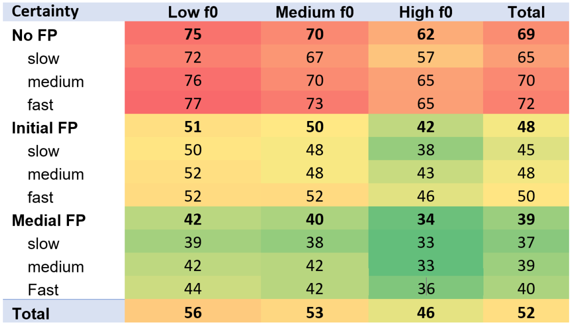
One participant was excluded from the final analysis because of several restarts of the experiment. A within-subjects factorial analysis of variance was conducted to evaluate the effects of fundamental frequency, speech rate and filler position on stimulus rating. F and p values are shown in Table 2. All three main effects were significant, and pairwise comparisons using the Šidák correction showed significant differences between all levels of each independent variable.

As shown in Figure 2, stimuli were rated as sounding most confident when no filler was present, and most hesitant with a medial filler. A faster speech rate and lower f0 were associated with greater confidence. In addition, there was a significant interaction between f0 and filler position. Tests of simple main effects of f0 at each level of filler position showed that the difference in ratings between low and medium f0 was significant only in the no filler condition. There was also a significant three-way interaction between filler position, speech rate and f0. The relationship between speech rate and f0 varied as a function of filler position.

Table 2: *Summary of effects*

effect	F	sig.
filler position	**598.35**	**< .001**
f0	**99.23**	**< .001**
speech rate	**52.18**	**< .001**
position * f0	**4.50**	**< .001**
position * sr	2.23	.18
sr * f0	1.89	.11
position * sr * f0	**5.01**	**< .001**

[2] Audio samples: www.speech.kth.se/tts-demos/interspeech2022



Certainty	Low f0	Medium f0	High f0	Total
No FP	**75**	**70**	**62**	**69**
slow	72	67	57	65
medium	76	70	65	70
fast	77	73	65	72
Initial FP	**51**	**50**	**42**	**48**
slow	50	48	38	45
medium	52	48	43	48
fast	52	52	46	50
Medial FP	**42**	**40**	**34**	**39**
slow	39	38	33	37
medium	42	42	33	39
Fast	44	42	36	40
Total	**56**	**53**	**46**	**52**

Figure 3: *Mean certainty ratings for the input combinations*

6. Discussion

All three of the features investigated (f0, speech rate and filled pause location) appear to have affected the perception of confidence in synthesized speech in accordance with Hypothesis 1. Synthesized speech sounded more confident when the fundamental frequency was lower and speech rate was higher, and the addition of filled pauses lowered perceived confidence, especially when the filled pause occurred in the middle of the utterance instead of at the beginning. Adjusting the presence and location of filled pauses produced the greatest changes in ratings, suggesting that this may have been the most impactful of the three cues, however the general trend seems to confirm Hypothesis 2 regarding the additive nature of these effects. The highest confidence ratings were for a combination of fast, lower-pitched speech with no filler, while the lowest ratings of speaker confidence were seen with a combination of slow speech rate, high f0 and a medial filler, see Figure 3.

The effect of fundamental frequency also interacted with filled pause location, and was more pronounced when filled pauses were absent, perhaps because more fine-grained variations were more apparent in the absence of the very salient cue provided by filler position. This did not seem to be the case for speech rate, however, as its effect seems to have remained constant even when a more salient cue was available. When the interaction between f0, speech rate and filler position is taken into account, however, the picture concerning speech rate is somewhat different. Ratings of low-f0 utterances were more strongly affected by speech rate (with slower utterances rated as less certain and vice versa) when fillers were absent. When fillers were present, some combinations of speech rate and f0 also corresponded to smaller differences in ratings. So it appears that while filled pauses do not override the effect of speech rate on perceived speaker confidence, and in fact these three features generally seem to have an additive effect, some specific configurations of acoustic features associated with confidence may be perceived more readily in utterances without a filled pause.

One prosodic cue for confidence that we did not investigate directly is utterance-final change in f0. Because of the way our system synthesizes different levels of pitch and speech rate based on the distribution of these features in the training data, each utterance has somewhat random variations in intonation. Rather than controlling pitch and speaking rate explicitly, we control it implicitly by allowing the system to meet utterance-level constraints on pitch and speaking rate in a way that is consistent with the speaker's behavior in the training data. This means that some of our stimuli contained a rise in pitch at the end while others did not. In fact, one way in which the system may have achieved an overall rise in f0 over the utterance is with a rise in pitch at the end of the utterance. While investigating the contribution of intonation to perceived certainty was not within the scope of this study, this should be considered in future work.

Another consideration is that the medial filled pause position used in our study actually encompasses a range of positions relative to the beginning and end of the sentence. Our findings could be expanded upon by looking in more detail at how more fine-grained changes in the position of a medial pause and its relation to syntactic structures affect perception.

7. Conclusions

Using synthesized spontaneous speech as a research tool, we were able to transcend some of the limitations of previous research on the role of filled pauses and prosodic features in influencing perceptions of speaker certainty. While much research investigating prosody in speech perception relies on lab-recorded speech, very short synthesized utterances or acoustically manipulated recordings, our system generated stimuli with the characteristics of spontaneous speech yet also allowed for more sophisticated control of prosody than would be possible by altering these features in recorded samples. To our knowledge, this is the first time this novel method of controlling speech rate and fundamental frequency has been utilized for studying speech perception. We found that filled pause location, mean f0 and speech rate contributed additively to listeners' perception of speaker confidence in synthesized utterances. The different contributions of these, as visualized in Figure 2, indicate that developers of future conversational systems could use filled pauses to get large variation in perceived uncertainty, and then change fundamental frequency and speaking rate for more fine-grained control. Future studies should look more closely at how other prosodic features such as intonation or voice quality interact with the features investigated here, as well as the role of semantic content and perhaps the different functions of filled pauses (e.g., as a reflection of cognitive load and searching for a word versus expressing a hesitant attitude).

Speech synthesis will likely become an increasingly promising tool for studying speech perception as it increases in sophistication and flexibility, allowing for even more fine-tuned control of prosody. This will allow us to learn even more about the ways in which prosody contributes to judgements about a speaker's attitudes, emotional state or personality traits in conjunction with other aspects such as content and speaker characteristics (gender, age, etc.). And even as speech synthesis can inform our understanding of speech perception, learning more about the acoustic correlates of paralinguistic information could help us build conversational agents which are more adept at performing and responding to social behaviors.

8. Acknowledgements

This research is supported by the Swedish Research Council projects: Perception of speaker stance – using spontaneous speech synthesis to explore the contribution of prosody, context and speaker (VR-2020-02396), Connected: context-aware speech synthesis for conversational AI (VR-2019-05003), and the Riksbankens Jubileumsfond project CAPTivating – Comparative Analysis of Public speaking with Text-to-speech (P20-0298).

9. References

[1] D. J. Goldsmith, "A normative approach to the study of uncertainty and communication," *Journal of communication*, vol. 51, no. 3, pp. 514–533, 2001.

[2] N. D. Goodman and D. Lassiter, "Probabilistic semantics and pragmatics: Uncertainty in language and thought," *The handbook of contemporary semantic theory. Wiley-Blackwell*, 2015.

[3] L. Goupil, E. Ponsot, D. Richardson, G. Reyes, and J.-J. Aucouturier, "Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature," *Nature communications*, vol. 12, no. 1, pp. 1–17, 2021.

[4] 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.

[5] J. J. Guyer, L. R. Fabrigar, and T. I. Vaughan-Johnston, "Speech rate, intonation, and pitch: Investigating the bias and cue effects of vocal confidence on persuasion," *Personality and Social Psychology Bulletin*, vol. 45, no. 3, pp. 389–405, 2019.

[6] X. Jiang and M. D. Pell, "The sound of confidence and doubt," *Speech Communication*, vol. 88, pp. 106–126, 2017.

[7] H. Pon-Barry, "Prosodic manifestations of confidence and uncertainty in spoken language," in *Proceedings of Interspeech*, 2008.

[8] T. Dinkar, I. Vasilescu, C. Pelachaud, and C. Clavel, "How confident are you? exploring the role of fillers in the automatic prediction of a speaker's confidence," in *Proceedings of ICASSP 2020*. IEEE, 2020, pp. 8104–8108.

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

[10] E. Lasarcyk and C. Wollermann, "Do prosodic cues influence uncertainty perception in articulatory speech synthesis?" in *Seventh ISCA Workshop on Speech Synthesis*, 2010.

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

[12] 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.

[13] 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, pp. 43–65, 1997.

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

[15] R. L. Rose, "The structural signaling effect of silent and filled pauses," in *The 9th Workshop on Disfluency in Spontaneous Speech*, 2019, p. 19.

[16] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 6264–6268.

[17] J. Adell, A. Bonafonte, and D. Escudero-Mancebo, "Modelling filled pauses prosody to synthesise disfluent speech," in *Speech Prosody 2010-Fifth International Conference*, 2010.

[18] S. Andersson, K. Georgila, D. Traum, M. Aylett, and R. A. Clark, "Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection," 2010.

[19] M. Wester, M. Aylett, M. Tomalin, and R. Dall, "Artificial personality and disfluency," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] E. Lasarcyk, C. Wollermann, B. Schröder, and U. Schade, "On the modelling of prosodic cues in synthetic speech–what are the effects on perceived uncertainty and naturalness?" *Proc. of NLPCS*, 2013.

[21] A. Hjalmarsson, "The additive effect of turn-taking cues in human and synthetic voice," *Speech Communication*, vol. 53, no. 1, pp. 23–35, 2011.

[22] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Spontaneous conversational speech synthesis from found data," in *Interspeech 2019, Graz*. ISCA, 2019, pp. 4435–4439.

[23] É.. Székely, G. Henter, J. Beskow, and J. Gustafson, "Breathing and speech planning in spontaneous speech synthesis," in *Proceedings of ICASSP 2020*. IEEE, 2020, pp. 7649–7653.

[24] É. Szekely, 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.

[25] A. Kirkland, M. Włodarczak, J. Gustafson, and E. Székely, "Perception of smiling voice in spontaneous speech synthesis," in *Speech Synthesis Workshop (SSW11), Budapest, Hungary August 26-28, 2021*, 2021.

[26] Y. Ferstl and R. McDonnell, "Investigating the use of recurrent motion modelling for speech gesture generation," in *Proc. IVA*, 2018, pp. 93–98. [Online]. Available: https://trinityspeechgesture.scss.tcd.ie

[27] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

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

[29] K. Ito, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset, 2017.

[30] 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.

[31] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.13)," 2009. [Online]. Available: http://www.praat.org