INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Synthesising uncertainty: the interplay of vocal effort and hesitation disfluencies

Éva Székely, Joseph Mendelson, Joakim Gustafson

KTH Royal Institute of Technology, Sweden
szekely@kth.se, josephme@kth.se, jocke@speech.kth.se

Abstract

As synthetic voices become more flexible, and conversational systems gain more potential to adapt to the environmental and social situation, the question needs to be examined, how different modifications to the synthetic speech interact with each other and how their specific combinations influence perception. This work investigates how the vocal effort of the synthetic speech together with added disfluencies affect listeners' perception of the degree of uncertainty in an utterance. We introduce a DNN voice built entirely from spontaneous conversational speech data and capable of producing a continuum of vocal efforts, prolongations and filled pauses with a corpus-based method. Results of a listener evaluation indicate that decreased vocal effort, filled pauses and prolongation of function words increase the degree of perceived uncertainty of conversational utterances expressing the speaker's beliefs. We demonstrate that the effect of these three cues are not merely additive, but that interaction effects, in particular between the two types of disfluencies and between vocal effort and prolongations need to be considered when aiming to communicate a specific level of uncertainty. The implications of these findings are relevant for adaptive and incremental conversational systems using expressive speech synthesis and aspiring to communicate the attitude of uncertainty.

Index Terms: speech synthesis, uncertainty, disfluencies, vocal effort, conversational systems

1. Introduction

Conversational systems are becoming more intelligent and adaptive, and are performing features of social behaviour, such as entraining to the conversation partner [1], adapting to the changing acoustic environment [2] and expressing emotions and attitudes [3]. Expressing and communicating internal uncertainty can contribute to a successful human-robot interaction. It has been demonstrated that people generally estimate artificial agents to be more knowledgeable in certain tasks than human conversational partners [4]. The agent's ability to express uncertainty could nuance this perception when required. [5] argues that when a robot is uncertain, it should be able to find a fluid way to ground the degree of commitment to its goal with the user. Moreover, signalling uncertainty about the correctness of one's answers can build trust, while expressing uncertainty about what the conversation partner is saying can communicate that the question under discussion is not yet resolved [6] [7]. Uncertain sounding synthetic speech can also aid in reinforcing character personality in conversational agents [8].

Along with rising intonation and increased pause duration, the insertion of filled pauses and prolongations have been linked to expressing the attitude of uncertainty [9] in speech. As the modelling of conversational characteristics in speech synthesis has become a focus of research interest, the synthesis of disfluencies such as prolongations and filled pauses has received a significant amount of recent attention [10], [11], [12].

Synthesis of vocal effort has been most widely studied in the context of intelligibility of speech in a noisy environment (Lombard speech) [13], [14]. Higher vocal effort has however also been shown to communicate emotion [15], dominance [16], speaker confidence and the speaker's degree of "feeling of knowing" [9], [17]. Decreased vocal effort, also characterised by a more breathy phonation, is more dependent on the social context than the environment or speaker distance [18], and has been shown to communicate a variety of attitude-related paralinguistic information, such as politeness, gentleness and tenderness [19].

There have been a handful of previous research attempts to express the attitude of uncertainty with synthetic speech, and to assess the perception of it. [7] applied copy synthesis on specific cue words (such as "yeah", "right", "really") for communicating (un)certainty with varying f0 contours and found that rising f0 contours indicate that the question under discussion is unresolved, but that the semantics of the cue words are more impactful on a credibility scale. [20] also uses intonation, as well as delays and filled pauses in synthesised single word utterances to demonstrate that in this context these cues are additive in increasing perceived uncertainty. [21] used a human production-based method of modelling acoustic-prosodic parameters in speech synthesis to allow differentiation between four attitudes within a sentence, including the attitude of uncertainty.

Because uncertainty is to a large extent determined by the lexical content, and thus the semantic meaning of an utterance [22], it is difficult to conduct corpus-based studies on spontaneous speech data to examine the interaction of two or more acoustic features influencing the perceived uncertainty. Using synthesised spontaneous speech in this study allows for control of lexical and semantic content. At the same time, it provides insight into what extent spoken dialogue systems using a synthetic voice would be capable of modelling these characteristics in conversation.

In this study, we introduce a synthetic voice that is built from recordings of spontaneous speech in dialogue, and is capable of both incorporating disfluencies and performing vocal effort changes with a corpus-based method. Using this voice, two listening tests were conducted: to ascertain that we are able to synthesise vocal effort independent of intensity, and to examine how disfluencies and vocal effort changes together influence the perception of uncertainty in a conversational sentence.

2. Method

2.1. Corpus and annotation

The corpus was recorded during an experiment originally conducted to study subjects' entrainment to their interlocutor's vo-

http://dx.doi.org/10.21437/Interspeech.2017-1507

cal effort level in different noise conditions, as described in [23]. The subjects were playing a card matching game with a confederate interlocutor who was asked to speak with a *soft*, *modal* and *loud* voice in three experimental conditions. In addition, a control condition was recorded, in which the interlocutor was asked not to be specifically conscious of his voice and focus on the game instead. The voice levels realised by the interlocutor can be described along the vocal effort continuum, where *soft* voice is decreased vocal effort with voicing (not whispering) and *loud* voice is increased vocal effort level that can be placed between modal voice level and shouting.

The participants' task consisted of describing the images on their six cards (originating from the Dixit^{TM} game), and finding a match with one of the cards of their conversation partner. The matches were not exact but rather based on topical similarity between the images. The images contained a lot of imaginative, unusual semantic combinations between figures and objects, and were therefore sometimes challenging to describe. This resulted in a varying degree of hesitation disfluencies in the participants' speech which makes the corpus particularly interesting for the purpose of synthesising conversational speech.

The interlocutor's recordings were manually transcribed, and filled pauses such as 'uh' and 'uhm' and feedback tokens like 'okay' and 'yeah' were annotated. Overlapping speech and noisy segments were excluded from the corpus. On average the corpus contains 4.8 filled pauses per 100 tokens, which is notably higher than what was identified for human-human interactions by [24] in the Switchboard (1.7) and the AMEX (2.8) corpora for informal and task-oriented telephone conversations, respectively. Each conversation was automatically segmented into turns, and the interlocutor's speech was further manually segmented into inter-pausal units (IPUs) to form utterances as input for the synthesiser. It was necessary to do this manually because the corpus contains a high degree of variability, both in speech rate across the experimental conditions and in pause length and frequency across turns. Therefore, we used variable pause lengths to separate lexical-prosodic phrases rather than a set minimum pause length to mark the boundary between utterances [25]. This resulted in a relatively short average utterance length of 7.1 words. After segmentation, the corpus consisted of 2 hours and 20 minutes of speech, including feedback tokens and filled pauses. The IPUs from the 4 conditions were then ordered on their average loudness value (normalised intensity raised to the power of 0.3 to simulate human sensitivity to loudness [26]) and divided into three subcorpora, in order to maximise the difference between the groups and to make use of the speech material from the 'control' condition.

2.2. Voice building

A DNN voice was built from all three subcorpora using MERLIN [27]. The system was set up to include 4 feed-forward (TANH) layers each containing 1024 hidden units, followed by a long short-term memory (LSTM) layer with 512 units. 4075 utterances from the three combined subcorpora were used for training, with a disjoint set of 450 for validation and a further 59 held out for testing. For the training of the duration model, only utterances between 5 and 15 tokens long were included. This resulted in 1905 utterances in the training and 170 in the validation set, with another 29 held out in a test set.

To synthesise varying levels of vocal effort, binary and normalised numerical linguistic features such as quinphone identity, part-of-speech and positional features were complemented by a linguistic feature corresponding to the subcorpus of that utterance. This feature represents *soft, modal* and *loud* with consecutive numerical values. Entering the discrete categorisation of the three subcorpora into a continuous input variable also allows for interpolation between the categories in duration modelling and synthesis. Varying the loudness feature throughout a sentence is also possible at the syllable level.

To synthesise lengthened filled pauses ('uh' and 'uhm') with a corpus based method another input feature was created reflecting the length of the specific filled pause on a continuous normalised scale (1 to 3). Values were based on the output from the forced alignment (excluding outliers), while all other words received a zero value for this feature. Because of this it was not necessary to introduce a new phone for filled pauses. The addition of these two normalised input features resulted in a 303-dimensional input feature set.

The synthesis of lengthened syllables is made possible by modification of the generated duration of the nucleus prior to synthesis to that of a specified (high) percentile of the corresponding phone in the corpus. The motivation behind this corpus-based approach was that research has shown that prolongations influence listeners' perception of naturalness to a different degree, depending on vowel identity [28]. This method also allows for leveraging the naturally occurring variability of prolongations in the corpus, without the need for the lengthened syllables to be specifically annotated.

Based on the output acoustic features (MCCs, BAPs and log F0s and their deltas, as well as a voiced/unvoiced binary value) the speech waveforms were reconstructed using the WORLD vocoder [29]. The feature-based synthesis of vocal effort levels is evaluated in Section 3.1. The methods described above for the synthesis of filled pauses and prolongations are used to create the stimuli in Section 3.2, but their separate evaluation is beyond the scope of this paper.

3. Experiments

3.1. Vocal effort difference in normalised samples

Two perceptual experiments were carried out. First, we needed to see whether listeners perceive a significant difference between the different vocal effort levels. While it is possible, as noted earlier, to interpolate between the vocal effort levels of the subcorpora, in the evaluation, we selected three levels directly estimated from the subcorpora, which we will here also refer to as *soft*, *modal* and *loud*. Higher vocal effort inherently means a relatively easily perceptible increase in sound pressure level (SPL) along with the more subtle changes in prosodic characteristics and voice quality due to increased subglottal pressure and vocal fold tension [13]. Because we wanted to ensure that the vocal effort difference is indeed perceptible and that listeners do not simply orient themselves by the amplitude differences, we normalised the amplitude of our stimuli to that of the modal voice for each utterance.

The alternative hypothesis is that the voices after normalisation of intensity can be ranked in increasing vocal effort. The listening test consisted of an AB pairwise comparison in the format of a similarity judgment test. Listeners were asked to choose which of the two stimuli sounded like the speaker was using more vocal effort, with a third option to indicate if they were not able to detect an audible difference. Ten conversational utterances for the stimuli were selected from movie subtitles from the OpenSubtitles2016 corpus [30], segmented into dialogue turns by [31]. Each utterance was synthesised in the 3 vocal efforts which yielded 30 pairwise comparisons in total.

3.2. Uncertainty rating of sentences with disfluencies

The aim of the second experiment was to assess how the speaker's vocal effort influences listeners' perception of uncertainty expressed through filled pauses and prolongations in an utterance. For this evaluation, a further 8 conversational utterances were selected from the OpenSubtitles2016 corpus. To make sure that the stimuli were comparable, only sentences starting with the phrase 'I think' were considered in the selection. [32] identify the verb 'think' as a lexical cue of the doxastic modality: hypothetical uncertainty reflecting the speaker's beliefs, which at the discourse level can reflect varying levels of uncertainty, ranging from low to high speaker confidence [33], [17]. Further overt expressions of certainty and uncertainty such as hedging strategies, adverbial markers of probability such as 'probably', 'maybe', 'possibly', and modal verbs like 'could', 'would' and 'may' were omitted from the selection. This was necessary to ensure that the perceived uncertainty of the synthesised stimuli were indeed reflecting the influence of vocal effort and the presence and absence of disfluencies and were not biased by a substantial difference in the baseline semantic uncertainty of the sentences. Sentiment analysis was performed with the Stanford Deep Learning for Sentiment Analysis parser [34] to select neutral sentences in order to avoid further bias resulting from highly positive or negative emotional content. All sentences underwent an additional manual check to ensure that the overall semantic meaning was appropriate. Four versions of each of the 8 sentences were synthesised with *soft* and *loud* vocal effort which resulted in 64 stimuli altogether. We chose to include only two different vocal effort levels to avoid fatigue in participants having to listen to 12 versions of each sentence.

The filled pause 'uh', with length feature set to 2.8 (see Section 2.2), was added to the original sentence (version A) to create version B. Version C contained a prolongation which was synthesised by modifying the duration of the nucleus of the stressed syllable of a function word [35] immediately preceding the filled pause in the sentence to the 99th percentile of the length of the corresponding vowel in the corpus. Finally, version D contained both types of disfluencies together. Table 1 displays the different versions of one of the test sentences.

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

(A)	*I think I hear the sound of running water.*
(B)	*I think I hear the sound of **uh** running water.*
(C)	*I think I hear the sound of running water.*
(D)	*I think I hear the sound of **uh** running water.*

Listeners were asked to rate each stimulus on a 0-100 scale, with increments of 5, how certain they perceived the speaker to be during that utterance, with 100 meaning most certain.

4. Results

4.1. Evaluation 1

The listening test was completed by 28 participants, of which 18 indicated that they have experience in speech research. All participants were wearing headphones, and the completion of the test took 10 minutes on average. Participants were given a short definition of vocal effort before starting the evaluation. In a similarity judgement test, answers in line with the alternative hypothesis are scored 1, those with the opposite view -1, while the answer "no difference perceived" is scored with 0 [36]. Table 2 summarises the listeners' responses for each comparison.

Table 2: *Listener responses in Evaluation 1*

	Loud vs. Soft			**Loud vs. Modal**			**Modal vs. Soft**		
	1	-1	0	1	-1	0	1	-1	0
SE	93%	5%	2%	84%	4%	12%	66%	2%	32%
NSE	88%	9%	3%	78%	10%	12%	40%	9%	51%

Figure 1 displays the calculated scores for each comparison. The null hypothesis, that the vocal effort in the (intensity-normalised) samples is not discernibly higher (mean score $\geq$ 0.5) could be rejected in the case of the *loud* vs. *modal* ($p < 0.001$) and *loud* vs. *soft* comparison ($p < 0.001$), but for the *modal* vs. *soft* comparison, only the speech expert group could identify a higher vocal effort correctly (p = 0.02).

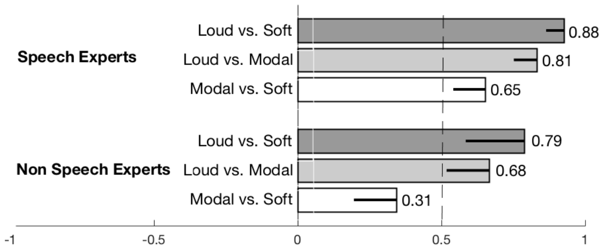

Figure 1: *Evaluation 1 results by participant and comparison type. The black bars represent the confidence interval (one-tailed t-test).*

An alternative way to interpret the results would be to proceed in the format of an AB preference test and equally distribute the undecided answers over the A and the B answers [37]. This would raise results of the *modal* vs. *soft* comparison above the 75% threshold for a two-choice forced choice task (77.5%). Despite the very similar set-up, this evaluation is assessing the discernibility of a specific aspect, rather than looking for a personal preference. Hence we decided to stay with the stricter interpretation of a similarity judgment test.

4.2. Evaluation 2

Evaluation 2 was conducted through the online crowdsourcing site Prolific Academic, and was completed by 80 participants, all native speakers of English from the United Kingdom and Ireland, aged between 22 and 66 years. On average, the listening test took 13 minutes to complete. The analysis included automatic cheat detection, proposed for crowdsourced listening tests by [38]. This involved a minimum sample correlation coefficient between worker and global mean opinion score estimates, with a conservative threshold. Six participants were rejected on this basis, and a further 8 were excluded based on self-reported incompetence in completing the test, such as indicating in the comment section that they lost interest or did not wear headphones. No statistically significant difference could be found across the ratings of the 8 sentences, which shows that all test sentences had a similar level of baseline semantic certainty.

A three-way ANOVA between the three factors, decreasing vocal effort, presence of a filled pause and presence of a prolongation, shows that the main effect of all three are significant ($p < 0.001$) in decreasing perceived certainty. Figure 2 shows the marginal means and their confidence intervals.

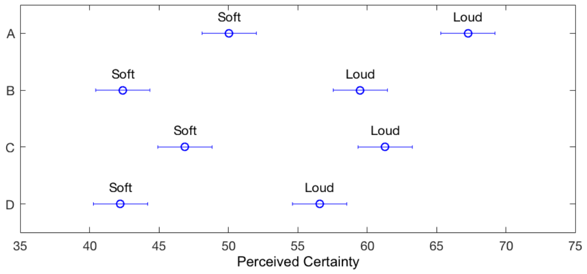

Figure 2: *Evaluation 2 results by disfluency type and vocal effort. The whiskers represent the confidence interval for the population marginal means.*

There is a significant interaction effect (p = 0.04) between vocal effort and the presence of a prolongation (see Figure 3). An added prolongation with a *loud* voice has a larger impact in decreasing perceived certainty than with the *soft* voice. This effect cannot be observed between vocal effort and filled pauses. Filled pauses however do have a significant interaction effect with prolongations (p = 0.03), whereby the decrease in perceived certainty of one of the cues lessens if the other is already present.

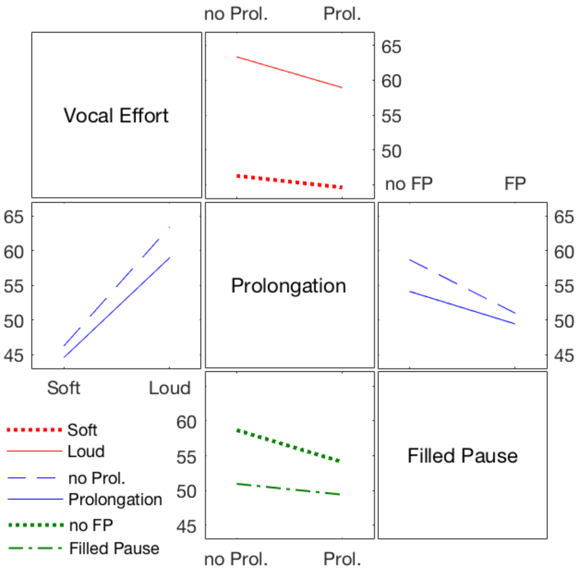

Figure 3: *Interaction graphs for all significant interactions in Evaluation 2. Each graph shows how the response to the factor noted besides the graph is impacted by the interaction noted above or below. For example, the top graph demonstrates how for a loud voice, a prolongation results in a greater decrease in perceived certainty than for a soft voice.*

5. Discussion

It was expected that the intensity normalisation will make listeners' task more difficult and thereby produce worse results than the original stimuli would have. However, these results reflect to what extent the synthetic voice is capable of producing a range of vocal effort that is closer to natural speech than the mere amplitude adjustment on the synthesised stimuli would be. The fact that listeners with experience in speech science were notably better at identifying the modal voice as higher vocal effort than the soft voice, can possibly be explained by vocal effort being a difficult concept to understand for naive listeners in the absence of an amplitude difference.

As mentioned in Section 3.2, only the two most extreme vocal effort levels were included in the experiment, which may have led to vocal effort being the strongest factor in impacting the degree of perceived certainty in the samples. But because the range of vocal effort levels that the synthetic voice is capable of producing can be found in between these two extremes on a continuous scale, we can estimate that intermediate levels of vocal effort would alleviate this effect. Within the current evaluation setup, we find support for using a range of uncertainty inducing cues for fine-tuning perceived uncertainty from the fact that the effect of the other cues is consistently measurable among these extreme vocal effort levels.

Knowledge on how the different influencing factors of a particular speaker attitude, such as uncertainty interact with each other, enables us to not only synthesise the attitude to the desired perceived degree, but to do so in the given situated interaction, where beside the attitude, other factors have to be taken into account in the voice style and prompt of a conversational system. Entraining to the conversation partner, background noise and speaker distance may require increased vocal effort, while delayed processing time may ask for the addition of prolongations and filled pauses. For example, in light of these findings, we can imply that if an incremental dialogue system inserts filled pauses to buy time, it can slightly increase its vocal effort to counteract potentially sounding too uncertain. Similarly, a system that uses Lombard speech as an adaptation strategy to increase intelligibility in background noise, but at the same time aspires to communicate an attitude of uncertainty, can likely achieve that by inserting prolongations, as the certainty-alleviating effect of these increases with vocal effort.

6. Conclusions and future work

We have built a DNN speech synthesiser that allows for a continuous control of vocal effort (from soft to loud), lengthening of individual syllables and insertion of filled pauses. Due to the design of the corpus recordings, where both variation in vocal effort and hesitation disfluencies have been elicited, the synthesiser is able to generate these with a corpus-based method, creating an approximation of how this variability is represented in natural speech. Evidence from the evaluation suggests that vocal effort, filled pauses and prolongations all contribute to the degree of perceived uncertainty of an utterance with doxastic semantic modality. The relationship between these three cues is not purely additive: the presence of one type of disfluency is a mitigating factor in the uncertainty-influencing effect of the other. At the same time, high vocal effort appears to increase the impact of prolongations in inducing listeners' perception of uncertainty. More research is needed to draw direct conclusions for specific conversational systems, but these results provide a starting point to modelling uncertainty in situated interaction. Future work involves investigating other markers of uncertainty: rising intonation and increased pause duration, as well as analysis and synthesis of the wide variety of filled pauses in the corpus.

7. Acknowledgements

The authors would like to acknowledge the support from the Swedish Research Council Project Incremental Text-To-Speech Conversion VR(2013-4935).

8. References

[1] R. Levitan, S. Benus, R. H. Galvez, A. Gravano, F. Savoretti, M. Trnka, A. Weise, and J. Hirschberg, "Implementing acoustic-prosodic entrainment in a conversational avatar," in *Proc. Interspeech*, 2016, pp. 1166–1170.

[2] S. Rottschäfer, H. Buschmeier, H. Van Welbergen, and S. Kopp, "Online Lombard-adaptation in incremental speech synthesis," in *Proc. Interspeech*, 2015, pp. 80–84.

[3] M. Schroder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. Ter Maat, G. McKeown, S. Pammi, M. Pantic *et al.*, "Building autonomous sensitive artificial listeners," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 165–183, 2012.

[4] B. R. Cowan, H. Branigan, H. Begum, L. McKenna, and É. Székely, "They know as much as we do: Knowledge estimation and partner modelling of artificial partners," in *Proc. of the 39th Annual Conference of the Cognitive Science Society*, 2017.

[5] J. Hough and D. Schlangen, "It's not what you do, it's how you do it: Grounding uncertainty for a simple robot," in *Proc. of the 2017 ACM/IEEE International Conf. on Human-Robot Interaction*, 2017, pp. 274–282.

[6] E. Marsi and F. V. Rooden, "Expressing Uncertainty with a Talking Head in a Multimodal Question-Answering System," *Communication and Cognition*, 2007.

[7] C. Lai, "What do you mean, you're uncertain?: The interpretation of cue words and rising intonation in dialogue," in *Proc. Interspeech*, 2010.

[8] J. Gustafson and K. Sjölander, "Voice creation for conversational fairy-tale characters," in *Proc. of the Fifth ISCA Workshop on Speech Synthesis*, 2004, pp. 145–150.

[9] V. L. Smith and H. H. Clark, "On the course of answering questions," *Journal of Memory and Language*, vol. 32, no. 1, 1993.

[10] R. Dall, M. Tomalin, and M. Wester, "Synthesising Filled Pauses: Representation and Datamixing," in *Proc. of the 9th ISCA Speech Synthesis Workshop (SSW9)*, 2016, pp. 2–8.

[11] S. Betz, P. Wagner, and D. Schlangen, "Modular Synthesis of Disfluencies for Conversational Speech Systems," in *Proc. Elektronische Sprachsignalverarbeitung (ESSV)*, 2015, pp. 128–134.

[12] S. Andersson, J. Yamagishi, and R. A. J. Clark, "Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis," *Speech Communication*, vol. 54, no. 2, pp. 175–188, 2012.

[13] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 543–571, 2014.

[14] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise," *Computer Speech & Language*, vol. 28, no. 2, pp. 648–664, 2014.

[15] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[16] M. Charfuelan and M. Schröder, "The vocal effort of dominance in scenario meetings." in *Proc. Interspeech*, 2011, pp. 2953–2956.

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

[18] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3438–3451, 2000.

[19] C. T. Ishi, H. Ishiguro, and N. Hagita, "Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, pp. 1–12, 2010.

[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?" in *Proc. of NLPCS*, 2013.

[21] A. Hönemann and P. Wagner, "Synthesizing Attitudes in German," in *Proc. of The Australasian International Conference on Speech Science and Techonology*, 2016.

[22] N. D. Goodman and D. Lassiter, "Probabilistic Semantics and Pragmatics: Uncertainty in Language and Thought," *Handbook of Contemporary Semantics, 2nd Edition*, no. June, pp. 1–45, 2014.

[23] É. Székely, M. T. Keane, and J. Carson-Berndsen, "The effect of soft, modal and loud voice levels on entrainment in noisy conditions," in *Proc. Interspeech*, 2015, pp. 150–154.

[24] E. E. Shriberg, "Disfluencies in Switchboard," in *Proc. of International Conf. on Spoken Language Processing*, 1996, pp. 11–14.

[25] S. Andersson, "Synthesis and Evaluation of Conversational Characteristics in Speech Synthesis," *Ph.D. dissertation, University of Edinburgh*, 2013.

[26] F. Eyben, F. Weninger, F. Groß, B. Schuller, F. Gross, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. of ACM International Conference on Multimedia*, 2013, pp. 835–838.

[27] Z. Wu and O. Watts, "Merlin : An Open Source Neural Network Speech Synthesis System," in *Proc. Interspeech*, 2016, pp. 218–223.

[28] N. Schaeffer and N. Eichorn, "The effects of differential vowel prolongations on perceptions of speech naturalness," *Journal of Fluency Disorders*, vol. 26, no. 4, pp. 335–348, 2001.

[29] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[30] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proc. of LREC*, 2016.

[31] P. Lison and R. Meena, "Automatic Turn Segmentation for Movie & TV Subtitles," *Proc. of the 2016 IEEE Workshop on Spoken Language Technology*, 2016.

[32] G. Szarvas, V. Vincze, R. Farkas, G. Móra, and I. Gurevych, "Cross-genre and cross-domain detection of semantic uncertainty," *Computational Linguistics*, vol. 38, no. 2, pp. 335–367, 2012.

[33] C. Caffi and R. W. Janney, "Toward a pragmatics of emotive communication," *Journal of Pragmatics*, vol. 22, no. 3, pp. 325–373, 1994.

[34] R. Socher, A. Perelygin, and J. Wu, "Recursive deep models for semantic compositionality over a sentiment treebank," *Proc. of the 2013 Conference on Empirical Methods in Natural Language Procesing*, pp. 1631–1642, 2013.

[35] S. Betz, P. Wagner, and J. Voße, "Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data," *Phonetik und Phonologie 12*, 2016.

[36] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Computer Graphics Forum*, vol. 31, no. 8, pp. 2478–2491, 2012.

[37] S. Buchholz, J. Latorre, and K. Yanagisawa, "Crowdsourced assessment of speech synthesis," *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*, pp. 173–216, 2013.

[38] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer, "Crowdmos: An approach for crowdsourcing mean opinion score studies," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2416–2419.