The Effect of Filled Pauses and Speaking Rate on Speech Comprehension in Natural, Vocoded and Synthetic Speech

Rasmus Dall[1], *Mirjam Wester*[1], *Martin Corley*[2]

[1]The Centre for Speech Technology Research, University of Edinburgh, UK
[2]PPLS, University of Edinburgh, UK

r.dall@sms.ed.ac.uk, mwester@inf.ed.ac.uk, martin.corley@ed.ac.uk

10.21437/Interspeech.2014-12

Abstract

It has been shown that in natural speech filled pauses can be beneficial to a listener. In this paper, we attempt to discover whether listeners react in a similar way to filled pauses in synthetic and vocoded speech compared to natural speech. We present two experiments focusing on reaction time to a target word. In the first, we replicate earlier work in natural speech, namely that listeners respond faster to a target word following a filled pause than following a silent pause. This is replicated in vocoded but not in synthetic speech. Our second experiment investigates the effect of speaking rate on reaction times as this was potentially a confounding factor in the first experiment. Evidence suggests that slower speech rates lead to slower reaction times in synthetic *and* in natural speech. Moreover, in synthetic speech the response to a target word after a filled pause is slower than after a silent pause. This finding, combined with an overall slower reaction time, demonstrates a shortfall in current synthesis techniques. Remedying this could help make synthesis less demanding and more pleasant for the listener, and reaction time experiments could thus provide a measure of improvement in synthesis techniques.

Index Terms: HMM-synthesis, speech synthesis, reaction time, filled pause, disfluency, speaking rate, speech perception

1. Introduction

Filled pauses (FPs) are generally not considered in speech synthesis systems. This is likely due to the lack of FPs in text, however in speech they are very common [1] and have been shown to provide a variety of benefits to the listener [2, 3, 4], as such, they should be considered for synthesis systems attempting to replicate spontaneous human speech. To date, few attempts have been made at modelling and inserting FPs. Previous studies by Adell and colleagues [5, 6, 7] included FPs in concatenative speech synthesis using the underlying fluent sentence [7]. Another approach [8, 9], which uses Hidden Markov Model (HMM) synthesis, treats FPs as normal word tokens in the speech stream when building the models. Common to both approaches is that they match state-of-the-art systems in terms of perceived naturalness. Andersson et al. [9] also showed improvements in perceived conversationality and Adell et al. [7] showed users prefer a system which includes FPs. However, the evaluation of the effect of FPs in synthetic speech is, unfortunately, not entirely convincing. For example, the evaluation in [5] consisted of comparing pairs of sentences with/without FPs asking questions specifically regarding FPs (e.g. "Do you think that filled pauses make the voice (more/equal/less) suitable for a dialogue?"). The perceptual results supported what the authors were hoping to find, i.e., sentences with FPs were judged to be more natural, equally suitable for dialogue, and more human-like. We are concerned that the listeners were primed to prefer sentences containing FPs due to the experimental set-up [10].

This paper aims to investigate the evaluation issue of including FPs in speech synthesis by approaching it from an angle supported by research findings from the field of psycholinguistics, i.e., by measuring listeners' reaction times. Fox Tree showed, in a series of experiments, that reaction time (RT) to a target word following an FP decreases compared to a pause of equal length or complete omission. This was shown for repetitions [11], 'uh's [12] and 'oh's [4] in English, was also found for Dutch [12], and was borne out by other researchers in [2]. Additionally, other benefits have been shown, including better recall of target words [4, 13], identifying target objects more accurately [3] and easier integration of words in their contexts [13]. Corpus studies also suggest certain regularities in their use [14, 1, 15] such as their appearance around phrase boundaries and before multi-syllabic words. The focus in this paper is on 'uh', 'oh' and repetitions in natural, vocoded and synthetic speech. We investigate whether current state-of-the-art HMM synthesis systems are good enough to reproduce the effects found in response to natural speech, and if not what the cause of this may be. Natural speech was included to ensure our experiments could replicate earlier findings. Vocoded speech was included as a stage between natural and synthetic and can be seen as an upper bound to what is possible, in terms of speech quality, with current HMM-synthesis.

2. Experiment 1: Filled pauses in natural, vocoded and synthetic speech

We follow Fox Tree's method [11, 4, 12] with a few small adjustments. The general method involves visually presenting a target word to listeners and asking them to react as quickly as possible when they hear it by pushing a button. In Fox Tree's experiments, each stimulus was presented in three conditions: i) FP included, ii) FP replaced by a pause of equal length or iii) FP spliced out of the sentence. In our experiments, we included the first two of Fox Tree's conditions: the FP and pause of equal length (silent pause "SP") conditions. Our goal is to show how information may be understood faster with FPs than without and it has been argued in [16] that there may be issues with preceding and trailing silences in Fox-Tree's third condition. Participants hear only one version of each sentence, and these critical stimuli are interspersed with fillers. The critical measurement is the participants' RT to the target word. We define the target word as the first non-determiner after the FP as in [11], see [12, 4] for variants. Finally, we also produce vocoded and synthetic versions of each stimulus.

2.1. Experiment 1: Data

116 utterances containing an instance of either 'uh', 'oh' or a repetition were selected from the AMI Meeting corpus [17], a corpus of spontaneous speech. Care was taken that the FP in these utterances was followed by a word that did not appear earlier on in the sentence. There were 79 critical stimuli and 37 filler stimuli (the target word in the fillers never occurred directly following a FP). The natural stimuli were digitally edited such that the FP was replaced with a silent pause (SP). Note that silent does not mean silence, but rather that a pause (not containing speech) was taken from another point in the same recordings and copied in. A speech expert, in addition to the authors, checked the data for edits but was unable to identify them reliably. Next, the utterances were vocoded using STRAIGHT [18], a state-of-the-art vocoding technique. The HMM-synthesis voice was based on HTS 2 [19] in a system that was newer than, but broadly similar to, that in [20], which is representative of the state-of-the-art. During synthesis, the FPs were treated as regular word tokens in the input stream, as argued for in [21], which was also the case for silent pauses (SPs) and ensures both are treated the same by the system. When producing the SP versions of the sentences the FPs were not edited out, but rather replaced with a pause in the input specification (the full-context labels). Due to the nature of HMM-synthesis, the exact durations of filled and silent pauses can deviate by a few frames (frame = 5 ms). The deviations occur as the system goes through different sequences of phoneme/pause in the input specification. Although there are slight differences in timings, this was deemed the better solution as it avoids manual editing of the synthesis output and better modelling of, e.g., co-articulation.[1]

2.2. Experiment 1: Method

Thirty native English speakers (mainly students at the University of Edinburgh) participated in this experiment. The test took approximately 25 minutes and participants were paid. Each participant was seated in front of a computer screen in a sound-attenuated booth. A fixation point was presented visually on screen for 500 ms, this was followed by a blank screen for 500 ms, and then by the target word for 1000 ms. 500 ms after the target word disappeared, an utterance was played and participants were instructed to press a button as quickly as possible if they heard the target word. Participants were instructed to only press the button if they heard the word. The test was split into four parts. The first part was a trial run; this always consisted of the same four stimuli, one each of the natural, vocoded and synthetic filler stimuli and one critical stimuli in one of the three speech types. Participants were encouraged to ask clarifying questions after the trial run. The remainder of the stimuli - 34 filler and 78 critical - were randomly divided into three roughly equal sized parts. All participants were presented with each utterance only once in any of its forms (with an equal amount of each form). In total, the experiment consisted of six conditions, two versions (SP and FP) of each of the three types of speech (natural, vocoded and synthetic).

2.3. Experiment 1: Results

Due to experimenter error three synthetic sentences were wrongly synthesised and excluded from the analysis. Of the remaining 2305 critical responses 141 were null responses.

[1] A sample of each is available at the conference repository as natural[pau].wav, vocoded[pau].wav and synthetic[pau].wav respectively.

Condition	Mean RT (SD)	Diff	Adjusted p	N
Natural FP	532.4 (146.9)	-38.9	<0.05	312
Natural SP	571.2 (150.8)			312
Vocoded FP	541.5 (146.2)	-41.5	<0.005	322
Vocoded SP	583.0 (146.3)			303
Synthetic FP	554.5 (140.9)	14.6	= 1	342
Synthetic SP	539.9 (141.8)			335

Table 1: Mean reaction times (RT) with standard deviations (SD) in ms for filled pause (FP) and silent pause (SP) conditions for the three types of speech and the RT difference from FP to SP conditions. N is *after* outlier removal.

Outliers were determined using the median absolute deviation (MAD) [22] instead of standard deviation (SD), because SD is itself subject to outliers, MAD is not. We used the moderately conservative threshold of 2.5 times the MAD [22] to detect outliers over all critical stimuli (Median=546, MAD=166.8). This value included exactly all negative reaction times (RTs) - that is RTs where the button was pressed prior to target word presentation - as these are evidently wrong it provides support to remove high RT outliers as well. Furthermore, in some sentences the target word was repeated later in the sentence, if a participant missed the first instance they may have reacted to the second one, which would also be captured as an outlier by 2.5 times the MAD. In total, 238 outliers were pruned leaving a final dataset of 1926 responses. Table 1 gives the mean RTs, standard deviation, number of stimuli for the different conditions and the difference in RT between FP and SP conditions.

A two-way ANOVA over the by-subject mean scores per condition showed a significant effect of pause type ($F(1, 29)=12.73$, $p<0.005$), no effect of speech type ($F(2, 58)=0.805$, $p=0.452$) and an interaction effect of pause and speech types ($F(2, 58)=8.359$, $p<0.001$). Bonferroni correction showed RT to be significantly faster in both the natural and vocoded FP conditions than in the corresponding SP conditions (see Table 1), however no significant differences exist between the synthetic conditions. Exploring the interaction, we find the effect in the SP conditions with the RTs for synthetic speech significantly faster than for vocoded speech ($t(636)=3.778$, $p<0.005$) and marginally faster than for natural speech ($t(645)=2.729$, $p=0.059$). Thus, we have replicated the results of [11, 12, 4], and shown that current vocoding techniques are able to represent the acoustic cues that are used by listeners to react more quickly to a target word after an FP. However, this effect is not replicated when hearing synthetic speech. In fact, in both synthetic FP and SP conditions RTs are found that are similar to the RTs found in the other FP conditions, showing that there is currently no advantage to including FPs in synthetic speech.

2.4. Experiment 1: Discussion

So far, we have reported a replication of earlier findings which were based on natural speech, showing that participants are quicker to identify a target word when it follows a FP than when it follows a silent pause of equal length. This finding extends to vocoded speech in a straightforward manner, but not to synthetic speech. This raises the question: what is it about synthetic speech that makes the effect disappear. Our results show that vocoding is not the issue even though a slight decrease in general speech quality is to be expected [23]. One issue present in the creation of the synthetic speech was that of speaking rate, as we did not ensure that the durations of the synthetic and natural stimuli were matched. Synthesisers model duration based on the available training data, and the synthesis system which

Speech Type	Mean	SD	Diff	p	N
Natural	3.921	1.163	0.353	<0.05	77
Synthetic	3.569	0.581			75

Table 2: Mean speaking rate (in sylls per s) for natural and synthetic speech critical stimuli.

we used was trained on read prompts. However, it has been shown that spontaneous speech tends to have a faster speaking rate than read speech [8, 24]. It is therefore likely that the synthetic speech durations were in general longer than the natural speech durations. Furthermore, one of the roles of FPs is to signal upcoming new information [13] and it can be argued that at lower speaking rates this role of FPs is superfluous as new information will be integrated fast enough without the need for the additional marker of an FP.

To confirm this, we compared the speaking rate (SR) of the natural and synthetic utterances. SR was defined as syllables per second up to the target word. This definition ensures that spurious changes in SR after the critical point do not influence the overall SR and yields a measure of the time available to participants to react in. The SR of the natural speech was significantly higher than that of the synthetic, see Table 2. Furthermore, in a rough comparison of the data, we split the natural sentences into quantiles based on SR and found that the slowest quantile RTs were at least 30ms slower than their faster counterparts (Q1: 563ms, Q2: 512ms, Q3: 521ms, Q4: 532ms). Notably the SR of the synthetic speech falls within the lower quantile range. While there is not enough data to calculate reliable statistics on this, it suggests a trend in which it is possible that the FP advantage only appears at higher SRs not present in the synthetic speech. Accordingly, this slower SR in synthetic speech may have affected results and led us to further investigation.

3. Experiment 2: Speaking rate and filled pauses in natural and synthetic speech

The above difference in SR prompted us to carry out a second experiment in which the overall SR of the synthetic speech was controlled to match the natural speech. The goals of this second experiment were i) to find whether speech rate could explain the lack of effect of FPs on RT in synthetic speech and ii) to investigate the effect of FPs in natural speech at different SRs, which has not previously been investigated. For simplicity, and because of its similar effect to natural speech, vocoded speech was not included in this experiment.

3.1. Experiment 2: Data

Selecting from the same AMI corpus as before, 80 critical and 40 filler stimuli were chosen which included either a repetition or 'uh'. The critical stimuli were chosen to represent three speaking rates, Fast, Medium and Slow. We used the speech from the previous experiment as a guide, faster sentences were chosen to match faster natural speech (Fast), slower sentences to match slower synthetic speech (Slow) and a medium category to match the average natural speaking rate (Medium). The total duration, excluding initial and final silences, of each natural stimulus was measured and used to define the length of the synthetic stimuli. To avoid further editing of the synthetic speech, we decided not to simply stretch or compress each synthetic stimulus to match the natural, but rather to require the stimuli to be a certain total duration. Again, this allows the system to deviate slightly from the prescribed lengths, but it ensures the system produces as natural an utterance as it is capable of.

Speech Type	FP	SP	Diff	Adjusted p
Natural	511.6	533.2	-21.5	< 0.05
Synthetic	547.7	524.5	23.2	< 0.05

Table 3: Mean RT for filled pause (FP) and silent pause (SP) conditions for natural and synthetic speech (SR combined) and the RT difference from FP to SP conditions.

3.2. Experiment 2: Method

Thirty-two native English speakers (mainly students at the University of Edinburgh) participated in the experiment. None of them had participated in the first experiment. The experimental procedure was similar to that in the previous experiment.

3.3. Experiment 2: Results

Due to experimenter error three natural sentences were incorrect and removed from the analysis. Of the remaining 2496 critical observations, 406 were null responses and pruned. Using the 2.5 MAD threshold to detect outliers (Median=514, MAD=139.36), a further 205 responses were pruned. A two-way ANOVA over the by-subject mean scores per condition was run. For pause and speech type it showed no significant effect of pause type ($F(1, 30)=0.098$, $p=0.757$), a significant effect of speech type ($F(1, 30)=5.112$, $p<0.05$) and an interaction effect of pause and speech types ($F(1, 30)=22.19$, $p<0.001$). Investigating the speech type effect using Bonferroni Correction, we see that natural speech results in a mean faster RT of 13.7ms ($p<0.05$) compared to synthetic speech. Exploring the interaction effect (see Table 3) we find that FPs in natural speech, as in Experiment 1, result in *faster* RTs, however for synthetic speech they result in *slower* RTs. That is, we again have a benefit of FPs in natural speech, but we now see the opposite effect in synthetic speech with FPs resulting in slower RTs. Looking at the SR effect we find an overall effect of SR ($F(2, 61)=6.083$, $p<0.005$), an interaction of SR and speech type ($F(2, 61)=3.770$, $p<0.05$), no interaction with pause type ($F(2, 61)=0.016$, $p=0.984$) and no interaction between all conditions ($F(2, 61)=1.656$, $p=0.199$), see Table 4 for an overview. Using Bonferroni correction the overall effect of SR is that we find slower RTs in the slow speed condition compared to the medium ($p<0.05$) and fast ($p<0.01$) but no difference between medium and fast ($p=1$). So, slower speech results in slower RTs for participants. For the interaction effect, we find that RTs in the synthetic conditions are generally slower than in the natural, except for in the fast condition, this is due to the natural fast SP condition (Figure 1) which is the only condition that does not follow the general pattern of RTs becoming faster as the speech becomes faster.

3.4. Experiment 2: Discussion

Again we see that FPs in natural speech provide a benefit in terms of faster RT, however for synthetic speech we now find the opposite effect with FPs resulting in slower RT. The slower speaking rate of synthetic speech compared to natural speech was not the reason for the lack of an effect in synthetic speech in the first experiment. Rather we see that a slower speaking rate results in slower RTs in both natural and synthetic speech. This is unexpected, if RTs represent a measure of comprehension time we would expect slower speech to have at least as fast, if not faster, RTs than faster speech. It is possible that people adapt their processing speed to the rate of incoming information and thus the slower speech yields overall slower processing. While the results could be interpreted as showing faster speech to provide a benefit, we would caution against this de-

Condition	Mean SR	Mean RT	RT SD	N
Natural FP				
Slow	2.607	520.3	125.6	172
Medium	3.817	511.9	128.6	138
Fast	5.561	499.5	133.1	126
Natural SP				
Slow	2.650	537.4	108.7	173
Medium	3.846	523.2	124.7	130
Fast	5.551	537.9	123.4	119
Synthetic FP				
Slow	2.751	563.0	129.6	134
Medium	4.021	545.1	115.8	162
Fast	5.173	539.7	121.9	204
Synthetic SP				
Slow	2.702	550.3	122.2	150
Medium	4.007	522.4	121.1	191
Fast	5.098	505.8	122.6	186
All				
Slow	2.672	542.8	121.4	629
Medium	3.935	525.7	126.4	621
Fast	5.299	520.7	129.6	635

Table 4: Overview of speed divided conditions. SR = Speaking Rate in syllables per second.

spite the Fast condition generally resulting in faster RTs than the Medium condition. Rather, we think there may be a 'sweet' spot SR range, around normal conversational SRs, in which we see the lowest RTs, speaking much faster listeners are likely faced with intelligibility issues which would hamper RTs and result in more null responses.

While SR did not account for the difference, it is possible that the nature of the natural SP condition did. Where the synthesis system creates the sentence to a given specification, for the natural speech the FPs were digitally edited out and replaced with a pause. This editing may have influenced our findings. To test this, each of the 80 critical sentences from the second experiment were used in a spot-the-edit test. Two groups of 8 participants were presented with 10 critical and 10 filler sentences. In the first group, none of the stimuli had been edited, and in the second group the critical sentences were edited. The rate at which subjects believed an edit to be present did not differ (edits: 35%, no edits: 31%) and of guessed edits only 67% were correct, suggesting subjects were unable to correctly identify the edits. While it is possible that the edits may still have had a subconscious effect, it is beyond the scope of this paper to test this. Considering that similar testing of the methodology has been done [11, 12, 4], and splicing in FPs instead of SPs results in the same effect [12], this seems unlikely to be the reason.

4. Overall discussion and conclusions

FPs in synthetic speech do not behave in a similar manner to FPs in natural speech. Where natural FPs provide a benefit in terms of faster RT to a target word compared to a pause of equal length, synthetic FPs give rise to the opposite effect, namely a slower RT. This is not due to the effect of vocoding as vocoded speech follows the same pattern as natural speech. We tested whether the generally slower speaking rate of synthetic speech caused the effect to appear and found this not to be the case. In fact, we found that a slower speaking rate tended to produce slower RTs *also* in natural speech, this is a new effect which has not been reported in the literature before. Furthermore, we have shown it is unlikely to be due to the edited nature of the natural pause samples. Another potential reason for the difference results from the synthetic speech is that the synthesised FPs are of a much lower quality than the surrounding speech, e.g. some are very long and some very short.[2] We believe this is due to the nature of the training data for the synthesis system. Current synthesis systems rely on recordings of read aloud scripted prompts which do not contain any FPs at all. While HMM synthesis is well know for its greater robustness to missing training data over concatenative synthesis [25] a particular problem appears in the representation of FPs. This is particular pertinent to 'uh' which is represented in the synthesis dictionary as a schwa. Unfortunately, while schwa is the most common phoneme, it is also one of the least stable in its realisation [26] meaning that it is greatly affected by surrounding phonemes and often the result of a reduced vowel. FPs are much more stable in their realisation than a schwa, but without training samples the system will attempt to infer the synthesis parameters (duration, f0 and spectrum) from mid-word and mid-syllable schwas not like those found in an FP, similar acoustic differences have been noted for repetitions [27]. One could attempt to mitigate this by including FPs in the scripts for recording, however as spontaneous speech has been found to be preferred over read prompts [10] it seems likely that these FPs would not be well received. We therefore suggest to follow [9] in training HMM-voices from spontaneous speech which includes natural examples of FPs. Potentially this would not only provide the necessary data, but also naturally speed up the synthetic speech due to the generally higher SR in spontaneous speech, removing the need to enforce specific duration requirements on the synthesis system.

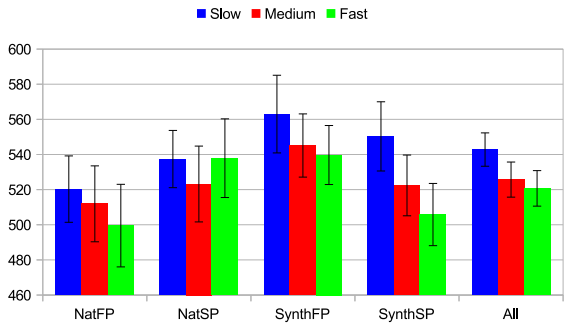

Figure 1: Mean RTs over each condition and speaking rate category in ms. Error bars show 95% confidence interval.

To conclude, FPs result in faster RTs in natural and vocoded speech, but slower RTs in synthetic speech. SR did not account for this difference, however we show a tendency for RTs to slow down in response to slower speech, the norm in synthetic speech. To enable speech synthesisers to show the same effect we therefore propose that the synthesis system includes FPs in the training data and its SR is increased. We also recommend that for the best results this is done in as natural a way as possible by training the synthesiser on spontaneous conversational speech.

5. Acknowledgements

We would like to thank Catherine Lai for evaluating the edited stimuli from our first experiment. This research was jointly funded by the JST Crest uDialogue Project and by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

[2] Samples available at the conference repository as fpshort.wav and fplong.wav

6. References

[1] E. E. Shriberg, "Disfluencies in SWITCHBOARD," in *Proceedings International Conference on Spoken Language Processing*, Philadelphia, PA, USA, 1996, pp. 11–14.

[2] M. Corley and R. J. Hartsuiker, "Hesitation in speech can. . . um. . . help a listener understand," in *Proceedings of the twenty-fifth meeting of the Cognitive Science Society*, Boston, USA, 2003.

[3] S. Brennan, "How Listeners Compensate for Disfluencies in Spontaneous Speech," *Journal of Memory and Language*, vol. 44, no. 2, pp. 274–296, Feb. 2001.

[4] J. E. Fox Tree and J. C. Schrock, "Discourse Markers in Spontaneous Speech: Oh What a Difference an Oh Makes," *Journal of Memory and Language*, vol. 40, no. 2, pp. 280–295, Feb. 1999.

[5] J. Adell, A. Bonafonte, and D. Escudero, "Filled pauses in speech synthesis: towards conversational speech," in *Proceedings 10th International Conference on Text, Speech and Dialogue*, vol. 1, 2007, pp. 358–365.

[6] J. Adell, A. Bonafonte, and D. Escudero-Mancebo, "Modelling Filled Pauses Prosody to Synthesise Disfluent Speech," in *Proceedings Speech Prosody*, Chicago, USA, 2010.

[7] J. Adell, D. Escudero, and A. Bonafonte, "Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence," *Speech Communication*, vol. 54, no. 3, pp. 459–476, Mar. 2012.

[8] S. Andersson, J. Yamagishi, and R. Clark, "Utilising Spontaneous Conversational Speech in HMM-Based Speech Synthesis," in *Proceedings SSW*, Kyoto, Japan, 2010.

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

[10] R. Dall, J. Yamagishi, and S. King, "Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation," in *Proceedings Speech Prosody*, Dublin, Ireland, 2014.

[11] J. E. Fox 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.

[12] ——, "Listeners' uses of um and uh in speech comprehension," *Memory and Cognition*, vol. 29, no. 2, pp. 320–326, 2001.

[13] 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. 658–68, Dec. 2007.

[14] E. E. Shriberg, "Preliminaries to a Theory of Speech Disfluencies," Ph.D. dissertation, University of California at Berkeley, 1994.

[15] E. R. Blackmer and J. L. Mitton, "Theories of monitoring and the timing of repairs in spontaneous speech." *Cognition*, vol. 39, no. 3, pp. 173–94, Jun. 1991.

[16] M. Corley and R. J. Hartsuiker, "Why um helps auditory word recognition: The temporal delay hypothesis." *PloS one*, vol. 6, no. 5, p. e19792, 2011.

[17] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus*," *Language Resources and Evaluation Journal*, vol. 41, no. 2, pp. 181–190, 2007.

[18] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[19] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based Speech Synthesis System Version 2.0," in *Proceedings SSW*, Bonn, Germany, 2007, pp. 294–299.

[20] J. Yamagishi and O. Watts, "The CSTR/EMIME HTS System for Blizzard Challenge 2010," in *Proceedings Blizzard Challenge Workshop*, 2010.

[21] H. H. Clark and J. E. Fox, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, pp. 73–111, 2002.

[22] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764–766, Jul. 2013.

[23] Q. Hu, K. Richmond, J. Yamagishi, and J. Latorre, "An experimental comparison of multiple vocoder types," in *Proceedings SSW*, Barcelona, Spain, 2013.

[24] E. Blaauw, "Phonetic differences between read and spontaneous speech," in *Proceedings ICSLP*, no. October, Banff, Canada, 1992, pp. 751–754.

[25] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed. Pearson, 2008.

[26] D. R. Van Bergem, "A model of coarticulatory effects on the schwa," *Speech Communication*, vol. 14, no. 2, pp. 143–162, 1994.

[27] J. Adell, A. Bonafonte, D. Escudero, and D. Informatics, "Disfluent Speech Analysis and Synthesis: A preliminary approach." in *Proceedings Speech Prosody*, Dresden, Germany, 2006.