Investigating Automatic & Human Filled Pause Insertion for Speech Synthesis

Rasmus Dall[1], *Marcus Tomalin*[2], *Mirjam Wester*[1], *William Byrne*[2], *Simon King*[1]

[1]The Centre for Speech Technology Research, University of Edinburgh, UK
[2]Cambridge University Engineering Department, University of Cambridge, UK
r.dall@sms.ed.ac.uk, mt126@cam.ac.uk, mwester@inf.ed.ac.uk, wjb31.cam.ac.uk, Simon.King@ed.ac.uk

Abstract

Filled pauses are pervasive in conversational speech and have been shown to serve several psychological and structural purposes. Despite this, they are seldom modelled overtly by state-of-the-art speech synthesis systems. This paper seeks to motivate the incorporation of filled pauses into speech synthesis systems by exploring their use in conversational speech, and by comparing the performance of several automatic systems inserting filled pauses into fluent text. Two initial experiments are described which seek to determine whether people's predicted insertion points are consistent with actual practice and/or with each other. The experiments also investigate whether there are 'right' and 'wrong' places to insert filled pauses. The results show good consistency between people's predictions of usage and their actual practice, as well as a perceptual preference for the 'right' placement. The third experiment contrasts the performance of several automatic systems that insert filled pauses into fluent sentences. The best performance (determined by F-score) was achieved through the by-word interpolation of probabilities predicted by Recurrent Neural Network and 4gram Language Models. The results offer insights into the use and perception of filled pauses by humans, and how automatic systems can be used to predict their locations.

Index Terms: filled pause, HMM TTS, SVM, RNN

1. Introduction

Currently, filled pauses (FPs) are rarely modelled in speech synthesis systems. This is usually due to the absence of FPs in the kinds of (formal) texts such systems take as input. However, FPs are common in conversational speech [1], and in the psycholinguistic literature they have been shown to provide various benefits such as better word recall [2], faster reaction times [3, 4], faster word integration [5] and more accurate object identification [6]. They are often subclassified into distinct subtypes - some (e.g., UH) indicate a minor delay, while others (e.g., UM) indicate a major delay - and they are often associated with planning problems, discouraging of interruptions, and highlighting of discourse structure [7, 1, 8, 9, 10, 11]. Consequently, there has been extensive research into the identification of FPs in the context of speech recognition, sometimes with a view to removing them from the output text [12, 13, 14, 15]. Given their prevalence and psychological importance in conversational speech, it is desirable for them to be incorporated into any synthesis system that seeks to produce 'natural' spontaneous speech. We therefore focus on the task of automatically predicting when to insert FPs into sentences. To date, there have been only a few attempts at modelling and inserting FPs in speech synthesis systems. Adell and colleagues [16, 17, 18] included FPs in concatenative speech synthesis using the underlying fluent sentence [18]. Another approach [19, 20], which uses Hidden Markov Model (HMM) synthesis, treats FPs as normal word tokens in the speech stream when building models based on spontaneous speech. Both approaches achieve naturalness scores comparable to state-of-the-art non-disfluent systems and Andersson et al. [20] also showed improvements in perceived conversationality, while Adell et al. [18] showed that users prefer systems which include FPs. Predicting when to use FPs, Adell et al. [16] used a combination of ngrams and decision trees to predict FPs based on a 317,000 word corpus. They obtain a high F-score, however the possible insertion points (IPs) were limited to those occurring after the 20 words most commonly followed by an FP. These kinds of distribution patterns are well-modelled by ngram language models (LMs). Anderson et al. [21] combined ngrams and the Viterbi algorithm to find the best possible IPs of fillers and discourse markers using a limited training set of 2120 sentences, these were the transcriptions of the data used for training the actual synthesis system. The method is geared toward picking examples which exist in the limited training data, something which is especially important in concatenative synthesis, but which limits the domain.

In this paper, we present our initial attempts at utilising very large corpora of spontaneous speech (see Section 4) for FP prediction. We focus on the two most common FPs, UH and UM, and present three experiments. The first experiment determines whether such corpora can act as a gold standard of FPs and whether people's predictions about FPs are consistent with reality and with each other. Secondly, we test the claim that there are 'right' and 'wrong' places to insert FPs. We compared synthetic sentences with either an FP inserted at the most frequently used IP from the first experiment, an FP inserted at an unused position or no FP at all. This also allowed us to determine whether a state-of-the-art HMM-synthesis system could produce convincing FPs. Finally we present results for initial attempts at FP insertion prediction using ngram LMs, a recurrent neural network (RRN) LM, support vector machines (SVMs) and decision trees (DT).

2. Experiment 1: Filled pause insertion

While corpus studies suggest certain regularities in the use of FPs [22, 23, 1], such as their appearance around phrase boundaries and before multi-syllabic words, it is uncertain whether naturally occurring FPs can represent a gold standard for automatic methods to predict. By asking participants to insert FPs in sentences we set out to investigate (i) if humans agree on where to insert FPs, (ii) whether different types of data (i.e., sponta-

Source	Example Sentence
WSJ	It is important for actors to reinvent themselves.
AMI	I think it's a multiple chip design and it's maybe printed on to the circuit board.

Table 1: Example sentences from the WSJ and AMI isolated sentences, AMI sentences were presented without FPs.

Cond	Pos	Used	Ins	Top	Top 3
WSJ	12.77	80.68%	1.40	28.07%	59.14%
AMI	16.48	75.73%	1.67	24.86%	54.19%
Isolated	14.59	77.54%	1.51	26.78%	58.17%
Paragraph	14.60	78.31%	1.55	25.98%	54.91%
All	14.59	77.93%	1.53	26.35%	56.49%
Chance	14.59	97.23%	1.53	9.54%	28.61%

Table 2: Mean values over all sentences for Possible IPs (Pos), Used IPs (Used), Inserted FPs (Ins), most (Top) and three most (Top 3) used IP agreements.

neous speech and written news texts) yield different agreements and (iii) if people's judgements coincide with the IPs actually encountered in the data. If (iii) can be substantiated, then it is possible to use data extracted from transcribed corpora as a gold standard, and, as multiple IPs may be valid in any given sentence, if (i) holds, then this would allow us to gather gold standard data from informants which would provide multiple potential IPs for the automatic methods to predict.

2.1. Materials (Exp. 1)

Sentences were selected from two different corpora: the WSJ corpus [24] and the AMI corpus [25]. The WSJ data comprises news texts, the AMI data is spontaneous speech from meetings. The two data types were chosen to allow us to investigate whether subjects behave differently for spontaneous speech vs written news texts. For both AMI and WSJ sentences, a set of 15 isolated sentences and 15 paragraphs was selected, a sentence in the middle of each paragraph was identified as the target sentence for people to insert FPs into. This allows us to investigate if context affects people's choice of IP. The use of the AMI data allowed us to choose sentences already containing FPs and compare the predicted IPs to actual use. Sentences contained at least one and maximally five FPs (mean = 2.9) and were required to be well-formed sentences containing no other types of disfluencies. In total, we had 60 distinct sentences. Table 1 show a few examples.

2.2. Method (Exp. 1)

72 paid native English University of Edinburgh students were recruited. The 60 sentences were divided into two sets with equal amounts of each text and sentence type. Each participant rated one of the two sets with sentences presented in a random order. The participants were instructed to imagine they were saying the sentence in a conversation, and then determine where they would be most likely to insert an FP. They were told to insert at least one FP, but were free to insert several if they thought it was natural. The possible IPs were at any point between the words in the sentence, including the beginning or end.

2.3. Results and discussion (Exp. 1)

Due to experimenter error one AMI sentence contained an FP when presented to subjects and was excluded from the analysis. For comparison, a chance category was created. Using the overall statistics of potential IPs and mean number of inserted FPs, a simulation of the experiment was run 10,000 times to find the Chance values presented in Table 2, which gives statistics of the data and shows subjects' agreement results.

Subjects insert FPs in a similar way regardless of the type of data they are faced with: paragraph or isolated sentences, news or spontaneous text (hypothesis (ii)). The AMI and WSJ differences are the largest, but still insignificant, which is possibly due to the lower number of potential IPs in the WSJ sentences which result in slight agreement increases. Focusing then on the overall results; subjects are quite consistent with each other, as the top used IP represents 26.35% of all insertions and the top three IPs represent 56.49%, which is way above chance levels at 9.54% and 28.61%. We also see that 22.07% of IPs are never used compared to 2.77% by chance. Comparing the actual IPs from the AMI data to the manually chosen IPs, we find that for 40.3% of the sentences there was a match between an IP in the original data and the most frequently chosen IP in the test data, this is compared to a lower internal consistency in the manually chosen IPs of 24.86%. Almost all (96.6%) of the original AMI sentences had an FP in one of the three most likely chosen IPs, compared to a 54.19% internal consistency, and only four (4.82%) of the original IPs were not predicted at any time in the AMI test data. This demonstrates a very good consistency between subjects' predicted usage and their actual usage (iii), and a good consistency in subjects' predictions (i). We can therefore use these values as a guide to compare automatic methods against when using transcribed spontaneous speech as a gold standard.

3. Experiment 2: Perception experiment

While humans are consistent in where they use, and predict, FPs, this does not demonstrate that the IP makes a perceptual difference to listeners. In speech synthesis it is generally accepted that incorrect pausing is detrimental to the processing of speech synthesis (e.g., [26], p.142), however, there is, to our knowledge, only one paper that has investigated this. Scharpff and van Heuven [27] measured the effect pausing has on intelligibility of low quality speech synthesis. They conclude that the intelligibility of low quality speech improves when pauses are inserted at prosodic boundaries, but deteriorates when other locations are chosen. It is likely that an FP may behave similarly, that there are 'right' and 'wrong' places for FPs, however this has not been tested (e.g., [21] assumed this to be true). While it has been found [20] that a TTS system based on spontaneous speech does not decrease the naturalness of FP-containing sentences, such a system cannot be considered 'standard', and as such we wanted to see how well-placed FPs in a state-of-the-art TTS system based on read speech compare to no FP at all. The objective of this second experiment is therefore to analyse whether there are IPs where one should not insert an FP and if well-placed FPs are preferred over no FP insertion.

3.1. Materials (Exp. 2)

Twenty of the thirty AMI sentences from Experiment 1 were used in the perception experiment. FPs (in this case UH) were inserted either at the most likely place (according to the judgements from Experiment 1) or randomly in one of the unused IPs (i.e., an IP that wasn't chosen by any of the participants in Experiment 1). The sentences were synthesised using a female voice based on HTS 2 [28] and about 8 hours of read speech, in a system that was newer than, but broadly similar to, that in [29], which is representative of state-of-the-art HMM-synthesis. During synthesis, the filled pauses were treated as regular word

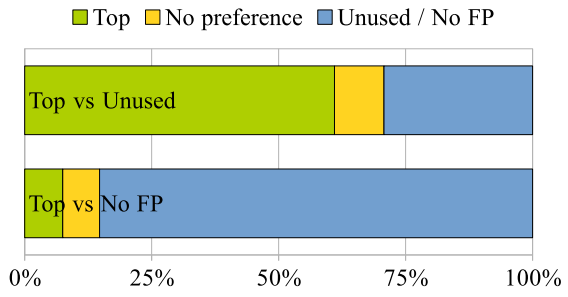

Figure 1: Results from the preference test comparing the most used position (Top) to unused positions or no FP.

tokens in the input stream, as argued for in [9, 30]. However, it has been shown [20] that producing spontaneous speech elements using a voice based on read speech results in less natural sounding synthesis than if it was based on spontaneous speech. To circumvent the worst quality problems with the synthetic UHs, we selected a good example of a synthetic UH and spliced it into the synthesised sentences at the appropriate locations.[1]

3.2. Method (Exp. 2)

The listening test was conducted through Amazon Mechanical Turk. Two conditions were created each consisting of 20 sentence pairs. Pairs were presented in a random order, and the comparisons within a condition were either: FP in top position versus FP in an unused position or FP in top position versus no FP. The task for the Turkers was to listen to the two sentences and choose which version they preferred, if any. The instructions were: "You will be listening to pairs of sentences. Please choose which one you think sounds most natural". The options were: "Sample 1", "Sample 2" or "No preference". We requested that only native speakers of English carried out the work. Quality control is an issue with AMT, to overcome this we offered the work only to Turkers with master worker status. Master status indicates a worker has completed work to a high level of satisfaction. This criterion was to ensure that we got workers who would carry out the task diligently as they would not want to risk their master worker status. In addition, we included three control questions in which the Turker was instructed, in the sound file (synthetic speech), to select a certain option. Turkers that failed to respond correctly were excluded. 44 Turkers completed the work, four were excluded as they failed the control questions. In total, responses from 20 workers per condition were considered.

3.3. Results and discussion (Exp. 2)

Figure 1 shows the results of the listening test. Listeners have a clear preference for FPs inserted in the top IP (61%) compared to FPs inserted at a random unused position (29%). However, when given the choice between a sentence containing a top FP versus a synthetic sentence without an FP listeners overwhelmingly choose the fluent sentence as the more natural. This is not surprising since FPs are often considered disfluencies, and these are generally judged to be undesirable if naturalness is equated with formal correctness [31]. However, as mentioned above, FPs are, in practice, very 'natural' since they are prevalent in spontaneous speech.

[1] Samples available at the conference repository as [top/unused/none].wav

Our results seem to be the opposite of Adell et al. [16] where listeners found sentences with FPs more natural than sentences without. A large difference between the current experiment and [16] is the way the question was framed. Listeners in [16] heard pairs of sentences with and without FPs and were asked whether the FP increased the naturalness of a voice for a dialogue system. Their focus was drawn to the FPs explicitly and the question was further framed by specifying the style of speech. In the current study, we purposefully did not specify that the sentences contained FPs and were from conversational speech, as we felt that might prime the participants towards choosing FPs. It has been shown that how listeners are asked to focus [31] and how questions are framed [32] strongly influences the resulting judgements.

Another possible reason for the difference is that Adell et al. [16] used a concatenative system in which they hand-picked samples of actual FP recordings based on their earlier work [33]. By contrast, our system used a voice trained on read speech containing no FPs. The FPs in [16] most likely sound more natural than ours, despite our splicing, and our results may partly reflect poor synthetic FP quality. As mentioned above, [20] found that a voice trained on spontaneous speech containing FPs did not degrade naturalness compared to a read voice. This is in line with recent findings that listeners prefer spontaneous speech over read speech when considering naturally produced utterances [32]. This leads us to conclude that while there are 'right' and 'wrong' places to insert an FP, and these places conform to human usage, it is not enough to simply insert FPs in the right places, they also have to sound right.

4. Experiment 3: Automatic FP Prediction

Experiment 1 indicates regularities we can predict, and experiment 2 that both quality and position of the FP is important. Focusing on the position (see Section 5 for a short quality discussion), we explored various techniques for automatic FP insertion. A training data set (1,164,938 sentences; 19,467,756 words) was defined using data from AMI [25], Fisher [34], Switchboard [35] and an unreleased corpus of British conversational telephone speech. The two most common kinds of FPs (UH and UM) were mapped to a single type, UH, since we were primarily concerned with finding the most likely IP irrespective of FP subtype. Sentences containing fewer than two words were removed as backchannels were not of interest. Development (dev) and test sets were defined using the same corpora. They each contained 2000 sentences, half with FPs and half without, consisting of 35,131 and 35,100 words respectively. The FP-containing sentences were designed to be similar to the sentences used in Experiment 1: word length was restricted in a similar way, and they contained exactly three FPs. We chose three FPs because this is similar to the average number of FPs in the real sentences used earlier (2.9), and because it allows us to make comparisons similar to the top three used IPs from Experiment 1. Using the training data, six automatic FP insertion systems were built:

1. *Random*: Randomly inserts a single UH into a sentence
2. *Ngram LM*: A standard 4gram LM was built using the SRILM toolkit [36] and the training data (68K wordlist, KN discounting).
3. *RNN LM*: A Recurrent Neural Network LM was built using the RNNLM Toolkit [37] and the training data. The RNN was 500 neurons wide and, for speed reasons, was trained using 250 classes.
4. *Interpolated RNN and ngram LM*: The Ngram and RNN

System	Precision	Recall	All F	UH F
Random				
dev	0.13	0.16	0.14	0.16
test	0.14	0.17	0.15	0.18
4-gram				
dev	0.49	0.15	0.23	0.26
test	0.48	0.16	0.24	0.27
RNN				
dev	0.31	**0.51**	0.39	0.51
test	0.32	**0.52**	0.40	0.53
RNN/4-gram				
dev	**0.53**	**0.51**	**0.52**	**0.57**
test	**0.50**	0.47	**0.48**	**0.54**
SVM All				
dev	0.24	0.17	0.20	0.20
test	0.26	0.16	0.20	0.19
SVM Best				
dev	0.27	0.22	0.24	0.27
test	0.29	0.23	0.25	0.27
DT All (Best)				
dev	0.25	0.16	0.19	0.19
test	0.25	0.17	0.20	0.21

Table 3: Overview of the 1-best output, the best system scores are bold-face. 'All F' is the F-score when considering the full dev/test set. 'UH F' is when only considering sentences containing FPs. 'All' refers to the system using all features. 'Best' refers to the best performing feature combination.

LMs were linearly interpolated on a by-word basis to re-rank the potential sentences.

5. *SVM-based*: A vector of features was extracted for each IP in each given sentence in the training data:
   (a) syllable count of word following IP
   (b) phrase boundary associated with IP
   (c) clause boundary associated with IP
   (d) 4g log prob for sentence with UH in IP
   (e) Part-of-Speech associated with word following IP

   (a) was obtained using tsylb;[2] (b), (c), and (e) were obtained using the Stanford Parser,[3] while (d) was obtained using the 4gram LM. All features were scaled and normalised so they could be expressed as floating point integers between log(0) and log(1). SVM models were built for all possible feature combinations using SVM-Perf.[4]
6. *Decision Tree-based*: A CART-style Decision Tree was built using R[5] and the same features as for the SVM above. The tree was pruned by selecting the complexity parameter associated with the smallest cross-validated error.

Outputs were produced for all systems, and they were scored using precision, recall and F-score. All systems predict maximally one FP, however, the FP containing sentences contains three FPs. A 'correct' prediction therefore occurs when the system predicts no FP when there is none or correctly predicts one of the three IPs in a given dev/test sentence. This is similar to the situation in Experiment 1 for the 3-best consistence, and we can thus make a (cautious) comparison.

4.1. Results and discussion (Exp. 3)

From Table 3 we can see that the best performing system is the RNN/4-gram interpolation. It is clear that the RNN and Ngram LMs complement each other. Although the Ngram is conservative in its prediction of FPs (only 359), it is much more exact (42% of predicted correct) than the other systems, which yields the high overall precision. By comparison, the RNN massively over predicts FPs (1877) and therefore is not very precise overall, but it has a much better recall. By interpolating the two, we predict a number of FPs much closer to the actual 1000, namely 1217, and obtain the best precision and second best recall, yielding the highest F-score. It is likely that the reason for this difference lies in the way the two LMs work, where the 4-gram will only output an FP when there is sufficient local evidence, the RNN is capable of considering longer-range dependencies. This is important since, e.g., sentence length has an impact on the likelihood of an FP being used. Both the SVM and DT perform disappointingly, with the simple 4gram LM performing as well as the best SVM and also being the SVM's most useful feature. Whereas the DT achieves the best results with all features, features (a)-(c) and (e) only confuse the SVM, presumably because they are not context-dependent. This is in stark contrast to the DT performance in [16] which was much better than ours. The difference is likely due to [16] limiting the number of IPs to those 20 words most often followed by an FP, which simplifies the task significantly for an ngram model.

The performance of the RNN/4gram system is encouraging as it shows we can quite reliably predict where to insert FPs in text, and the performance rivals that of the human top-3 performance (56.49%). If we allow the system to produce a 3-best list the precision (85%) and recall (81%) improves even further (F=0.83), demonstrating that reasonable IPs are being identified.

5. Conclusions

This paper has focused on the task of identifying IPs for FPs in different kinds of data (e.g., conversational speech and written news text). The results of experiments 1 and 2 demonstrate that the type of data does not affect results, that there is good consistency between human subjects' predictions of FP usage and their actual usage, and also a good consistency between predictions from different people. This confirms that FP insertion is not merely random, and therefore it can be modelled in speech synthesis systems. As an initial step towards this, the performance of various automatic systems was compared and contrasted, and it was shown that an interpolated Ngram and RNN LM produced the best output. The superior performance of this system suggests that the accurate modelling of local and long-range lexical and syntactic contexts is central to this task, and the systems described here could be improved by the inclusion of additional features or complementary modelling methods (e.g., dependency grammars). There remains, of course, the problem of handling inserted FPs convincingly in speech synthesis systems (e.g., imposing a plausible intonation contour and generating phonetically-plausible FPs), here using spontaneous conversational training data may be key. It is also possible to extend the methods summarised here to the task of inserting other kinds of spontaneous speech phenomena such as repetitions, restarts and discourse markers.

6. Acknowledgements

This research was jointly funded by the JST Crest uDialogue Project and by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

[2] ftp://jaguar.ncsl.nist.gov/pub/tsylb2-1.1.tar.Z
[3] http://nlp.stanford.edu/software/lex-parser.shtml
[4] http://www.cs.cornell.edu/people/tj/svm_light/svm_perf.html
[5] http://www.r-project.org

7. References

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

[2] 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.

[3] ——, "Basic Meanings of You Know and I Mean," *Journal of Pragmatics*, vol. 34, pp. 727–747, 2002.

[4] 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.

[5] 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.

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

[7] S. Schachter, N. Christenfeld, B. Ravina, and F. Bilous, "Speech disfluency and the structure of knowledge," in *Journal of Personality and Social Psychology 60*, 1991, pp. 362–367.

[8] M. Swerts, "Filled pauses as markers of discourse structure," in *Journal of Pragmatics 30*, 1998, pp. 485–496.

[9] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speech," in *Cognition 84*, 2002, pp. 73–111.

[10] S. H. Fraundorf and D. G. Watson, "The disfluent discourse: Effects of filled pauses on recall," *Journal of Memory and Language*, vol. 65, no. 2, pp. 161–175, Aug. 2011.

[11] S. Benus, R. Levitan, and J. Hirschberg, "Entrainment in spontaneous speech: the case of filled pauses in Supreme Court hearings," in *In Proceedings of the 3rd IEEE Conference on Cognitive Infocommunications*, 2012.

[12] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper, "Structural Metadata Research in the EARS Program," in *Proceedings ICASSP*, 2005.

[13] K. Audhkhasi, K. Kandhway, O. D. Deshmukh, and A. Verma, "Formant-based technique for automatic filled-pause detection in spontaneous spoken English," in *Proceedings ICASSP*, 2009.

[14] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, pp. 181–200, 2010.

[15] D. Brizan, G. An, and A. Rosenberg, "Detecting laughter and filled pauses using syllable-based features," in *Proceedings Interspeech*, 2013.

[16] 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. Springer, 2007, pp. 358–365.

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

[18] 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.

[19] S. Andersson, J. Yamagishi, and R. Clark, "Utilising Spontaneous Conversational Speech in HMM-Based Speech Synthesis," in *Proceedings SSW 7*, 2010.

[20] 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.

[21] S. Andersson, K. Georgila, D. Traum, M. Aylett, and R. A. J. Clark, "Prediction and Realisation of Conversational Characteristics by Utilising Spontaneous Speech for Unit Selection," in *Proceedings Speech Prosody*, 2010.

[22] 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.

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

[24] D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," in *Proceedings Workshop on Speech and Natural Language*, 1992, pp. 357–362.

[25] 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.

[26] P. Taylor, *Text-to-speech synthesis*. Cambridge University Press, 2009.

[27] P. J. Scharpff and V. J. van Heuven, "Effects of pause insertion on the intelligibility of low quality speech," in *Proceedings 7th FASE Symposium*, 1988.

[28] 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 *SSW6*, Bonn, Germany, 2007, pp. 294–299.

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

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

[31] N. Christenfeld, "Does it hurt to say um?" *Journal of Nonverbal Behavior*, vol. 19, no. 3, pp. 171–186, 1995.

[32] 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.

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

[34] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: A Resource for the Next Generations of Speech-to-Text Fisher," in *Proceedings LREC*, Lisbon, Portugal, 2004.

[35] J. J. Goodfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings ICASSP*, San Francisco, CA, USA, 1992, pp. 517–520.

[36] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "SRILM at Sixteen: Update and Outlook," in *Proceedings ASRU*, Hawaii, USA, 2011.

[37] T. Mikolov, S. Kombrink, A. Deoras, L. Burget, and J. Černocký, "RNNLM - Recurrent Neural Network Language Modeling Toolkit," in *Proceedings ASRU Demo Session*, Hawaii, USA, 2011.