Synthesising Filled Pauses: Representation and Datamixing

Rasmus Dall[1], *Marcus Tomalin*[2], *Mirjam Wester*[1]

[1]The Centre for Speech Technology Research, The 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

Abstract

Filled pauses occur frequently in spontaneous human speech, yet modern text-to-speech synthesis systems rarely model these disfluencies overtly, and consequently they do not output convincing synthetic filled pauses. This paper presents a text-to-speech system that is specifically designed to model these particular disfluencies more efffectively. A preparatory investigation shows that a synthetic voice trained exclusively on spontaneous speech is perceived to be inferior in quality to a voice trained entirely on read speech, even though the latter does not handle filled pauses well. This motivates an investigation into the phonetic representation of filled pauses which show that, in a preference test, the use of a distinct phone for filled pauses is preferred over the standard /V/ phone and the alternative /@/ phone. In addition, we present a variety of data-mixing techniques to combine the strengths of standard synthesis systems trained on read speech corpora with the supplementary advantages offered by systems trained on spontaneous speech. In a MUSHRA-style test, it is found that the best overall quality is obtained by combining the two types of corpora using a source marking technique. Specifically, general speech is synthesised with a standard mark, while filled pauses are synthesised with a spontaneous mark, which has the added benefit of also producing filled pauses that are comparatively well synthesised.

Index Terms: TTS, Filled Pauses, HMM, Phonetic Representation, Speech Synthesis

1. Introduction

In most modern text-to-speech (TTS) systems, disfluencies are not normally modelled overtly. Filled pauses (FPs) such as UH and UM are normally categorised as disfluencies, yet a large body of psycholinguistic research has shown that they fulfil a divers set of roles in spontaneous human discourse [1]. FPs can improve reaction times to a target word [2, 3, 4], increase change detection rates [5], help word integration [6], and they can be used as a delay strategy to improve interaction perception [7, 8] amongst other things. These, often subconscious, benefits to the listener motivate the exploration of FPs in the context of TTS. In essence, TTS output that contains convincing FPs can produce more human-like speech that clarifies the discourse structure for the listeners, thereby reducing the cognitive load they experience while processing the synthetic speech. These desirable properties are of particular relevance given the recent burgeoning of hi-tech personal assistants, life-like robots, embodied agents, and the like. In these kinds of systems a much more 'natural' expression is desirable, and this could be achieved if they deploy FPs correctly.

In earlier work, we have shown that current TTS techniques cannot replicate the reaction time [9] nor change-detection effects [10] found for natural speech. In the first case, synthesis is the problem, while in the latter it is the vocoding. We have also shown that we can replicate human use of FPs using language modelling techniques [11], and we have presented a method for FP and discourse marker (DM) insertion through a controllable 'disfluency' parameter [12]. By contrast, in this paper, we focus on the question of how to realise FPs convincingly in TTS output. First, Section 2 presents an analysis of the FPs in a spontaneous speech corpus to illustrate how FPs are acoustically different to other phones. One clear problem with current synthetic voices is that few FPs are found in the standard training data corpora. Therefore, in Section 3, we compare two corpora of speech, one a standard TTS corpus, and one created from recordings of spontaneously produced speech. Voices based on both corpora are compared, and it is shown that, despite natural spontaneous speech (with its relatively high FP count) being more natural than standard corpus recordings [13], standard voices produce synthetic speech of a higher overall quality. Also, some of the common acoustic properties of FPs are not realised in the synthesis based on spontaneous speech, masking any potential benefits of the FPs. Consequently, in Section 4, an investigation into the phonetic representation of the FPs in the linguistic feature set is presented, in which we compare the phones /V/, /@/ and two non-standard representations based on these. Following that, in Section 5, we present a number of data-mixing techniques which are designed to retain the overall quality of the standard corpus while facilitating the synthesis of convincing FPs by drawing on the spontaneous corpus. Section 6 provides an overall discussion and conclusion of the findings of the current study.

2. Data Analysis

Previous studies have shown that the fundamental frequency (F0) contours and the duration of filled pauses are different to other phones in fluent contexts [14, 15, 16]. Another characteristic of FPs is the presence of silence before and/or after the filled pause [17, 16]. Since it is well-known that the particular characteristics of FPs are language-specific, we wanted to explore these claims in relation to a corpus of English. Therefore we examined the patterns of prosodic features associated with non-synthetic disfluencies in the test-set created by the Linguistic Data Consortium for the 2004 DARPA Effective Affordable Reusable Speech-to-Text (EARS) project meta-data evaluation (henceforth the CS corpus). This dataset consists of a total of 3 hours of high-quality conversational speech obtained from 72 speakers. The transcriptions of the speech were accurately produced by human annotators, and the disfluencies were overtly classified and labelled. Prosodic features were extracted for all

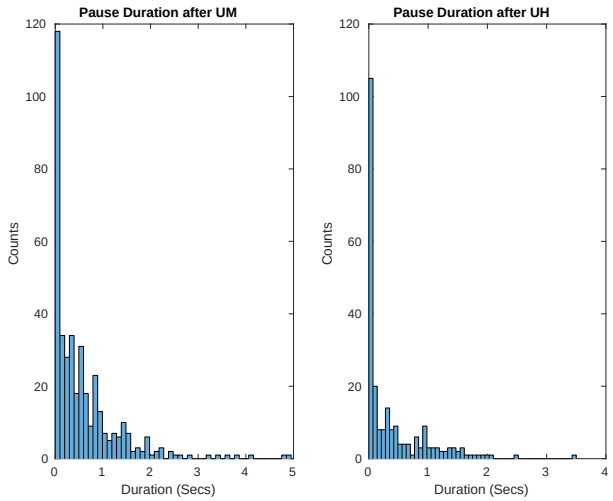

Figure 1: *Pause duration after UM and UH*

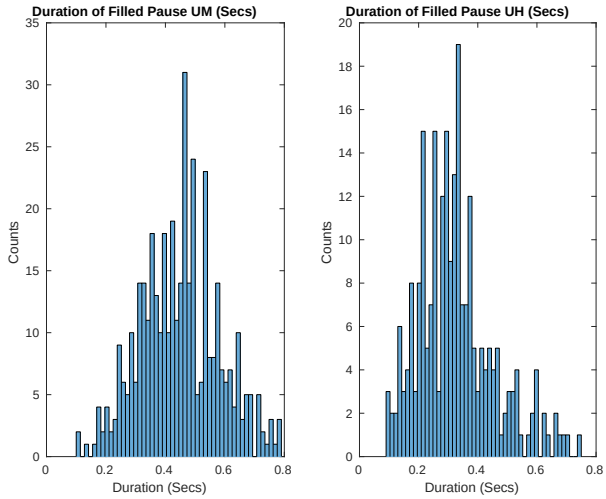

Figure 2: *FP duration UM and UH*

occurrences of the FPs UH and UM in the CS corpus. These features were extracted either from the waveform data, or from corresponding encoded data files, and some of the features were extracted from 0.2 sec windows either at the start (left window) or the end (right window) of each disfluency.

Adell et al. [16] found, in their material, that a preceding silence often occurs (60%) but after the filler silence only occurs in 24% of cases, though they noted that this finding was anomalous. Other papers, which explore these phenomena in different languages and different kinds of corpora, report varying patterns of silence after FPs. In the CS corpus, a silence occurs before 83% of the FPs UH and UM (specifically, 79% for UM, and 91% for UH). All the UH and UM tokens in the CS corpus are followed by a pause (though the duration of the trailing pause is, on average, less than half the duration of the preceding pause).

The duration of CS FPs are shown in Figure 2. The distribution of the FP durations peaks around 0.50 secs for UM and around 0.39 secs for UH. When similar vowels (e.g., schwa) occur in fluent context in the CS corpus, their duration is, on

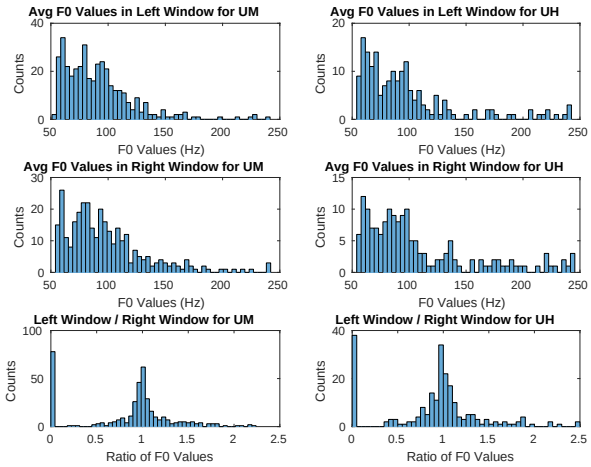

Figure 3: *F0 start and end of FP*

average, half the length of the vowels that occur in the FPs UM and UH. Figure 3 shows that there are a larger number of F0 values greater than 50 Hz in the left window (i.e., before the FP) than there are in the right window (i.e., after the FP). This quantifies the extent to which the F0 contour lowers when FPs occur. O'Shaughnessy has observed that the F0 values associated with FPs frequently end in the bottom 15% of the speaker's F0 range [14].

The above analysis of the FPs in the CS corpus provides a frame-of-reference for the analysis of the acoustic properties that characterise the synthetic FPs produced by the systems described in sections 4 and 5.

3. Read and Spontaneous Speech Based Voices

Most standard TTS corpora are based on phonetically balanced sets of prompts – the Arctic prompts [18] are the best known – read aloud by a voice talent in a studio under highly controlled conditions. In [13] it was shown that studio recordings of spontaneously produced speech are considered more natural than these standard "read" prompts, particularly when acoustic variation is considered. However, in [19] the overall naturalness of TTS output produced by a system trained on a spontaneous speech corpus never matched that of the voice trained on read speech. The same result persisted even after a pronunciation variant forced alignment method was applied to compensate for the additional reductions and deletions present in the spontaneous speech. However, the test sentences used in [19] did not contain FPs, and other researchers have found that including these in voices based on spontaneous speech can improve their perceived naturalness to be on par with, or even better than, voices trained on read speech [20, 16]. A test was thus performed to confirm whether this is also the case for the corpora used for the experiments reported in this paper.

3.1. Methodology

The read and spontaneous speech corpora used in these experiments were the same as those described in [19]. The read corpus consisted of studio recordings of a female British English speaker (recorded at 96khz, 32 bit, downsampled to 48khz, 16 bit), the prompts used were the Arctic sentences [18] and the

corpus contained a total of 1125 sentences (66 mins: 20 mins silence, 46 mins speech). The spontaneous corpus consisted of recordings of the same voice talent in the same studio but it consisted of unscripted spontaneous conversation. The voice talent and interviewer could hear each other through a headset connection and could see each other through a webcam to maintain as natural an interaction as possible. The resulting recordings were orthographically transcribed and segmented into utterances. The FPs were treated as a word token in the speech stream (similar to previous work). In total the corpus contained 1096 sentences (58 mins: 9 mins silence, 49 mins speech).

Synthetic voices for both speech types were trained based on HTS-2.3beta [21]. 60 sentences containing FPs were extracted from the text corpus of [11, 12] and a 5-point Mean Opinion Score (MOS) test and a preference test to rate naturalness were conducted. The sentences were split into two groups of 30 sentences for the preference pairs, and for the MOS test the 60 sentences were randomly divided into six groups, each containing 10 read and 10 spontaneous sentences. 30 native English speakers were recruited and they performed the test in a sound-proof booth in front of a screen wearing high-quality headphones. Each listener rated all 30 preference pairs, presented in a random order, and one of the 6 MOS groups for a total of 900 preference comparisons and 300 MOS ratings of each sentence.

3.2. Results and Discussion

Due to one participant misunderstanding the instructions for the experiment, this participant was removed from the analysis. For the preference test, participants preferred the read speech 59.5% of the time and this difference was significant using the exact binomial test ($p < 0.0001$). For the MOS test, the read voice (mean $= 2.54$) was rated significantly higher ($t(288) = 2.32, p = 0.021$) than the spontaneous (mean $= 2.26$). This means that the read voice was considered more 'natural' than the spontaneous. This is in contrast to the findings of [22, 20] and [16]. These earlier experiments both found that a spontaneous synthetic voice trained on data that contained FPs was rated at least as natural as a voice trained on read speech. There are, however, some differences in the experiments performed. In Andersson's work [22, 20], a data-mixing technique was applied to overcome data sparsity problems, in which one voice was trained using both types of speech, but a linguistic feature was added marking each sentence with the source speech type. At synthesis time, sentences could then be synthesised with either tag, and this technique enabled better FP-containing sentences to be synthesised by means of the spontaneous tag. In Adell et al. [23], synthesis of the FPs was based on a specific FP model, which was subsequently improved in [24]. FPs were modelled separately from other speech by applying modified search rules in a unit selection system. In both approaches, however, no results are given of the quality of synthesis based only on the data containing FPs. This suggests that their methods must be responsible for closing the gap that we measure in naturalness between synthetic speech based on read speech versus that based on spontaneous speech.

4. Phonetic Representation of Filled Pauses

As the improvements found by Adell et al. [24] came from a specific FP model, we propose something similar here, though in an HMM-based framework. The proposed model of [24] relies on an analysis of the acoustic features of FPs, such as increased duration and lowered F0, and was required to be explicit in order to guide the unit selection directly. However, in an HMM-based framework we are already building models of each context-dependent phone, and therefore we do not need an explicit FP model. Instead, we need the linguistic context features of the phones in an FP to distinguish them from other phones of the same type, which would allow the decision tree context-clustering to group the FP phones together. If a high-quality part-of-speech tager is used, FPs should be tagged as such, and this could serve as the distinguishing feature. However, FPs are not usually well modelled by sentence structure POS-taggers, and, in standard front-ends such as Festival and Flite, the tag set is reduced to one that does not contain the FP tag. In Festival 2.4 [25], when using the Combilex dictionary [26], an UH is phonetised as /V/ – a unrounded, open-mid, back vowel – and an UM as /V m/. However, for UM there are two additional alternatives in the dictionary: /@ m/ – schwa followed by a bilabial nasal - and /m!/ – a short bilabial nasal. While these are never used in standard transcriptions, they provide a convenient alternative representation. We here ignore the reduced form of /m!/, partly because it could arguably be considered to be the backchannel 'mhm' and not an FP, and partly because we are initially interested in fully pronounced FPs and not heavily reduced versions. Using /@/ for both UH and UM does not, however, uniquely identify FPs as the /@/ is the most common phone. It may however be a better representation of the sound of an FP, and so should be considered. In order to provide a distinguishing feature we suggest that a separate phone identity could be used for the FPs which could then borrow the features of the phone from either /@/ or /V/. In this way, the phone identity uniquely identifies the FP vowel, and consonants in the immediate context such as the /m/ in UM, while still sharing characteristics of its parent vowel.

4.1. Methodology

A preference test was performed to determine whether this alternative representation results in better FP realisation. SiRe [27] was used as the front-end and was modified to convert all vowels in FPs into each of four phones – /V/, /@/, /UHV/ and /UH@/. /UHV/ used the phone features of /V/, and /UH@/ the features of /@/. Four voices, each using one of the four FP phone representations, were trained using HTS-2.3beta [21] and a combined corpus of the read and spontaneous speech, this combination was done to ensure a higher overall quality of speech from the read speech while still retaining samples of the FP phone from the spontaneous corpus. Data mixing is discussed in more detail in Section 5. 20 sentences containing FPs were selected from a corpus of 'found data' derived from BBC's Desert Island Discs (DID) programme and made available as part of the EPSRC-funded *Natural Speech Technology* project. Specifically, the sentences were selected from the utterances of the presenter, Kirsty Young. The upper bound on the length was 25 tokens, the lower 5, and each utterance contained at least one FP. These sentences were synthesised using each voice.

30 paid native English speakers were recruited to take part, and each participant rated all 20 sentences for each preference pair. As there are four different voices this results in six pairs of 20 sentences for a total of 120 pairs rated by each participant and a total of 600 ratings of each pair. As we were particularly interested in the quality of the FPs, and not just the overall quality of the speech, participants were instructed to 'judge which of the two sentences you think sounds the most natural paying particular attention to the realisation of UH and UM'. As found

/V/	/@/	/UHV/	/UH@/	p
55.8%	44.2%	-	-	<0.05
48.2%	-	51.8%	-	=0.39
47.6%	-	-	52.4%	=0.25
-	44.1%	**55.9%**	-	<0.005
-	43.0%	-	**57.0%**	<0.001
-	-	49.8%	50.2%	=0.97

Table 1: Preference test results. P is calculated using the exact binomial test, the preferred phone in a pair is indicated using bold face.

	/V/	/@/	/UHV/	/UH@/
UH dur (s)	0.152	0.153	0.246	0.250
UM dur	0.313	0.303	0.431	0.404
vowel dur	0.074	0.074	0.079	0.081
UH F0 (Hz)	176	175	160	160
UM F0	174	177	171	173
vowel F0	169	170	168	169

Table 2: Mean duration and mean F0 for UH, UM, vowels. In all cases there were 30 UH, 9 UM and 717 vowels. The synthesis system used was the straight combination from Section 5.

in [13], naturalness ratings can be influenced by the instructions, and so the above wording was crafted to ensure participants focused primarily on the FP realisation. The options were 'Sample 1', 'Sample 2' or 'No Preference'.

4.2. Results and Discussion

Table 1 summarises the results of the preference test. 'No Preference' judgements were split evenly over the two systems. /@/ was significantly dispreferred compared to all other representations. There were no statistically significant differences between all other representations, although there was a tendency for the FP-specific phones to be slightly preferred over the /V/, with virtually no difference between the two FP-specific phones.

Although no perceptual preference was found for the FP-specific phones over the standard /V/, an analysis of the predicted acoustics show that both match the acoustic characteristics for FPs better. Table 2 shows the mean duration and mean F0 for UH, UM and vowels for each FP phone representation. These data show that both /UHV/ and /UHV/ have longer durations for FPs (about 100 ms longer) than /V/ and /@/ and that the duration for vowels is roughly equal across all FP models. Furthermore, for both /UHV/ and /UH@/, a lower mean F0 is found. These longer durations and lowered F0 values from /UHV/ and /UH@/ more closely match those found in Section 2 and thus show that the FP specific representations better capture the general acoustic characteristics of FPs than /V/ and /@/. The choice between /UHV/ and /UH@/ was made based on the finding that the /V/ phone was significantly preferred over /@/, thus favouring features borrowed from /V/ as it would seem /@/ is not a suitable underlying phone. Therefore, /UHV/ was used in the following investigation.

5. Data Mixing for FP Synthesis

The improved FP synthesis obtained by [22] was achieved by a data-mixing technique previously used for producing various emotions and speaking styles [28, 29]. The technique involves training a single model of speech using both read and spontaneous data simultaneously, but distinguishing the two speech types through an added linguistic feature which denotes the speech type the data came from. This affects the decision tree context-clustering, enabling each speech type to be clustered separately during training. At synthesis time each sentence is then marked with either the read or spontaneous tag, so that all speech is steered toward that particular style. The benefit of the method comes from the fact that not all context clusters will be split on the speech-type feature, and therefore some sharing of data is possible. [22] concludes that it is this which enables the spontaneous speech voice to match the read speech voice by overcoming data sparsity issues. It was not, however, reported how this method compares to a system that combines the two types of speech in training to produce a voice without marking the speech type. Consequently, we here present that system alongside the other methods.

It is possible that the TTS system trained on read speech in [22] faltered primarily because it was unable to utilise the FPs present in the spontaneous speech effectively, particularly considering the finding in Section 3 that a standard read speech voice is considered more natural than a spontaneous speech based voice, even when including FPs. This could happen due to the use of the /V/ phone, which would have samples very different from the FPs in the read speech. There are two possible ways to alleviate this. First, we can use an FP-specific phone, and the results in Section 4 suggest that /UHV/ seems most promising. This would distinguish the phone from those present in the read data, and therefore, during synthesis, there would be no samples of this particular phone with the read tag. This would force the system to use the spontaneous speech based /UHV/ model. It may, however, also simply result in the decision tree relying on the features of the phone to utilise the read speech samples of /V/, and so another method of synthesising from the data mixed voice was also applied. In the previous method, the sentence to be synthesised was tagged as either entirely read or else entirely spontaneous – but it is also possible to tag only parts of the sentence as coming from either type of speech. Specifically, we here propose to tag all of the sentence as read *except* the FPs themselves, and these we tag as spontaneous. This should allow us to retain the generally higher overall quality of the read speech, while still synthesising FPs from the more appropriate spontaneous speech model. The main potential problem with this method is that there is no data available of sentences in which this switch happens. However, it seems likely that the trajectory modelling applied should smooth the transitions effectively.

There is also an alternative way of mixing the data, namely, by using speaker adaptation. In this approach, an initial model from several speakers is usually trained, before being adapted to a target speaker using adaptation techniques such as the constrained structural maximum a posteriori linear regression (CSMAPLR) technique of [30]. We can apply this technique to the switch between read and spontaneous data by first training a voice on one type of speech and subsequently adapting it to the other. Adapting from read to spontaneous could solve data sparsity issues in a similar manner to the data marking technique, whereas adapting from spontaneous to read could retain read speech quality while still providing data for the FPs in a similar manner to the proposed switch in speech mark when using the marking technique.

5.1. Methodology

Four different voices were trained. One was a standard HMM voice in which both the read and spontaneous speech were pooled and used as training data. This provides the baseline approach (and therefore it is the system used in the phone experiment above). Another voice was trained using the data marking technique, and three methods of synthesis were applied: everything marked as read, everything marked as spontaneous, or everything marked as read except the FPs (which were marked as spontaneous). The final two voices were speaker-adaptive voices. One was adapted from a base read model to the spontaneous speech, and the other from spontaneous to read. In total six different synthesis methods were evaluated. The /UHV/ phone representation from the phone experiment was used in all cases, since it showed the most promise and could also potentially help the marked read synthesis in realising convincing FPs (as discussed above).

The same 20 sentences from the experiment in Section 4 were used, but this time a MUSHRA-style naturalness test was run. This was done in part due to the many preference pairs which would have been necessary, but also in part because we were interested in the overall quality of the resulting speech and not merely the synthesis of the FPs. An additional sentence from the DID corpus was synthesised and used as a training sample. The test was run without a natural reference as the DID data consisted of Kirsty Young's speech, not the voice talent's whose speech was used to base the synthetic systems on. The participants were instructed to rate how natural each sample sounded. However, it was explicitly mentioned that conversational phenomena such as FPs would occur, and that this was part of the test. This was done to ensure that participants paid attention to the FPs specifically, while also focusing on the overall naturalness of each voice. Besides that, the experiment was identical to a standard MUSHRA test with one slide per sentence where participants would listen to and rate all samples of that sentence for each system side by side. This provided both a measure of naturalness and preference between the synthetic voices. 30 paid native English speakers were recruited. Each participant sat in a sound-proof booth in front of a computer wearing high quality headphones and rated all 20 sentences for a total of 600 evaluations of each voice. The test took approximately 30 minutes to complete for each person.

5.2. Results and Discussion

The results are given in Figure 4. All the system pairs were compared using a Wilcoxon signed-rank test, after Holm-Bonferroni correction to avoid false positives. All the systems are significantly different ($p < 0.001$) from each other except for Mark_R and Mark_Sw. This finding is somewhat surprising as although we expected Mark_Sw to improve the naturalness of the speech, we did not expect Mark_R to do equally well. We hypothesise this is due to the fact that despite everything being marked read in the Mark_R system, spontaneous speech is necessarily used because there are no occurrences of /UHV/ in the read data.

An extra ad-hoc preference test was run to ascertain whether it is the specific phone /UHV/ for FPs that is benefiting Mark_R. 10 listeners took part in this listening test comparing a system with everything marked read using the /V/ phone for FPs versus a system using the /UHV/ phone for FPs. Each subject rated 45 sentences, an extended set of DID materials. Instructions and listening conditions were as detailed in Section 4. Listeners preferred the combination system with /UHV/ phone 63% of the time, significant using the exact binomial test ($p < 0.0001$). This suggests that the use of the /UHV/ phone did indeed allow the read marked speech to utilise the spontaneous FP occurrences to inform its model.

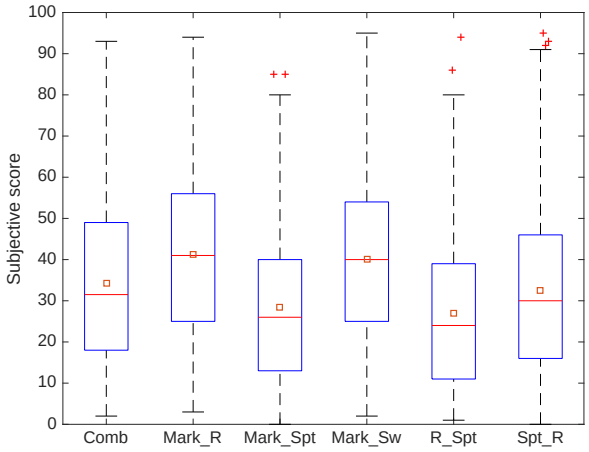

Figure 4: Results of the MUSHRA-style test. R = Read, Spt = Spontaneous, Sw = Switch. R_Spt = read adapted to spontaneous. Spt_R = Spontaneous adapted to read. Red line is the median, square the mean.

	Read /V/	Read /UHV/	Spont /UHV/	Switch /UHV/
UH dur (s)	0.059	0.143	0.241	0.242
UM dur	0.134	0.303	0.416	0.324
vowel dur	0.078	0.080	0.075	0.086
UH F0 (Hz)	182	183	171	166
UM F0	183	183	166	167
vowel F0	185	186	166	184

Table 3: Mean duration and mean F0 for UH UM and vowels. In all cases there were 30 UH, 9 UM and 717 vowels. Synthesis systems used: Mark_R (V), Mark_R (UHV), Mark_Spt, Mark_Sw.

Furthermore, if we compare the FP durations and F0 for the marked system using /UHV/ when either marking all as read, spontaneous, or switching with the read marked system using /V/ (Table 3) we can see that the read marked /UHV/ system produces durations and F0 closer to those expected for spontaneous speech (cf. Section 2), whereas the read marked system using /V/ does not capture this at all. However using the mark switching technique gave even better results. This suggests that although no perceptual preference difference was found between the read marked and switching system using /UHV/, the switching system still better captures the acoustic realisation of FPs.

Interestingly, neither adaptation system performed very well, suggesting that despite the two types of speech being from the same speaker, adaptation still introduces many serious artefacts which degrade the overall speech quality.

6. Overall Discussion and Conclusions

This paper has focused on the topic of modelling FPs overtly in a state-of-the-art TTS system. There are many reasons

why it is desirable for such systems to model these phenomena. Such as the wide range of functions in conversational interactions which FPs performs, where they can (amongst other things) indicate psychological states, structure spoken exchanges, facilitate word recall, and improve object recognition [3, 31, 17, 32, 33, 34, 13, 35]. Given their well-attested importance in spontaneous human speech, it is desirable to model these phenomena overtly in automatic TTS systems, to produce output that is more natural and human-like. In many respects, the broad motivations underlying research into disfluent synthesis are closely related to those that motivate the development of emotional or expressive TTS systems [36, 37, 38, 39, 40]. These closely-connected endeavours seek to create synthetic speech that is able to convey a wider range of emotional or psychological states, thereby producing synthetic voices that can simulate certain character and personality types more convincingly.

This paper has approached the problem of modelling FPs in a TTS system by developing an approach that exploits the most effective capabilities of synthetic voices trained on spontaneous and read speech. Specifically, it has been shown that a voice based on speech only from spontaneous unscripted conversation containing FPs is not considered as natural as a voice trained on standard read speech. The FPs in the synthetic speech did not exhibit the acoustic characteristics that have been shown in the literature [16, 15, 14] and in our own investigation of the CS corpus – notably longer durations and lower F0 compared to fluent speech. This contrasted with earlier findings that such voices could match voices trained on standard corpora [16, 22] when synthesising sentences containing FPs. However, in both cases the systems used were modified forms of a standard TTS system.

Consequently, we then looked into a number of different phonetic representations for modelling FPs to ascertain whether having a distinct phone for FPs would capture the acoustic properties of FPs more successfully (similar to the modelling of [16]). On the basis of a preference test and acoustic analyses, the FP specific phone /UHV/ was deemed the best for FP modelling. It was significantly preferred over /@/ and the longer durations and lower F0 more closely match the desired acoustic characteristics than /V/.

In addition to a specific phone for FPs, various data-mixing approaches to using both the read and spontaneous speech were explored - straight combination, data source marking and speaker adaptation. It was found that a data-marking technique similar to [20] performed the best. However, in contrast to [20], this technique did not improve the spontaneous speech-based voice to match that of a read-speech based voice. Our results were obtained using a specific FP phone representation, and a preference test showed that this representation improved the perceived synthesis quality. This suggests that the voice based on read speech, as in [20], suffered degrading quality issues due to the bad FP representation of /V/, as it tended to use read data in which no FPs occurred. By using a specific FP phone /UHV/ we found that the read voice could produce more convincing FPs such that perceptual quality did not degrade compared to a voice in which the FPs were synthesised using the spontaneous mark. However, using the spontaneous mark produced FPs even closer to the expected acoustic properties, and thus the switching of the mark, from read in general, to spontaneous when synthesising FPs, produced perceptually high quality speech while also retaining the defining characteristics of FPs. System samples and experimental materials currently available at www.dall.dk/rasmus/Samples.zip to be moved to to the NST data collection upon acceptance.

7. Acknowledgements

This work was supported by the EPSRC under Programme Grant EP/I031022/1 (Natural Speech Technology) and the JST Crest uDialogue programme. The full NST research data collection may be accessed at `http://hdl.handle.net/10283/786`.

8. References

[1] H. Nicholson, K. M. Eberhard, and M. Scheutz, "um... i don't see any: the function of filled pauses and repairs." in *DiSS-LPSS*, Tokyo, Japan, 2010, pp. 89–92.

[2] 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.

[3] ——, "Listeners' uses of um and uh in speech comprehension," *Memory and Cognition*, vol. 29, no. 2, pp. 320–326, 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] P. Collard, "Disfluency and listeners' attention: An investigation of the immediate and lasting effects of hesitations in speech," Ph.D. dissertation, University of Edinburgh, 2009.

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

[7] T. Shiwa, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "How quickly should communication robots respond?" in *Proceedings of the 3rd international conference on Human robot interaction - HRI '08*, New York, USA, 2008, p. 153.

[8] T. Baumann, "Incremental Spoken Dialogue Processing: Architecture and Lower-level Components," Ph.D. dissertation, Universität Bielefeld, Germany, 2013.

[9] R. Dall, M. Wester, and M. Corley, "The Effect of Filled Pauses and Speaking Rate on Speech Comprehension in Natural, Vocoded and Synthetic Speech," in *Proc. Interspeech*, Singapore, Singapore, 2014.

[10] ——, "Disfluencies in Change Detection in Natural, Vocoded and Synthetic Speech," in *Proc. DiSS The 7th Workshop on Disfluency in Spontaneous Speech*, Edinburgh, Scotland, UK, 2015.

[11] R. Dall, M. Tomalin, M. Wester, W. Byrne, and S. King, "Investigating Automatic & Human Filled Pause Insertion for Speech Synthesis," in *Proc. Interspeech*, Singapore, Singapore, 2014.

[12] M. Tomalin, M. Wester, R. Dall, B. Byrne, and S. King, "A Lattice-Based Approach to Automatic Filled Pause Insertion," in *Proc. DiSS The 7th Workshop on Disfluency in Spontaneous Speech*, Edinburgh, Scotland, UK, 2015.

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

[14] D. O'Shaughnessy, "Recognition of hesitations in spontaneous speech," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 521–524.

[15] E. Shriberg, "To 'errrr' is human: ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 153–169, 2001.

[16] 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.

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

[18] J. Kominek and A. W. Black, "CMU ARCTIC databases for speech synthesis," Carnegie Mellon University, Tech. Rep., 2003.

[19] R. Dall, S. Brognaux, K. Richmond, C. Valentini-Botinhao, G. E. Henter, J. Hirschberg, J. Yamagishi, and S. King, "Testing the Consistency Assumption: Pronounciation Variant Forced Alignment in Read and Spontaneous Speech Synthesis," in *Proc. ICASSP*, Shanghai, China, 2016.

[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] 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.

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

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

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

[25] A. W. Black, P. Taylor, and R. Caley, "Festival 2.4 Documentation," 2014. [Online]. Available: http://www.festvox.org/docs/manual-2.4.0/festival{_}toc.html

[26] K. Richmond, R. a. J. Clark, and S. Fitt, "Robust LTS rules with the Combilex speech technology lexicon," in *Proc. Interspeech*, Brighton, UK, 2009, pp. 1295–1298.

[27] R. Dall, "Experiment materials for "Testing the consistency assumption: pronunciation variant forced alignment in read and spontaneous speech synthesis"," 2016.

[28] J. Yamagishi, K. Onishi, and T. Masuko, "Acoustic Modeling of Speaking Styles and Emotional Expressions in HMM-Based Speech Synthesis," *IEICE Transactions on Information and Systems*, vol. E88-D, no. 3, pp. 502–509, 2005.

[29] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of Various Speaking Styles and Emotions for HMM-Based Speech Synthesis," in *Eurospeech*, Geneva, Switzerland, 2003, pp. 2461–2464.

[30] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[31] J. E. Fox Tree and J. C. Schrock, "Basic Meanings of You Know and I Mean," *Journal of Pragmatics*, vol. 34, pp. 727–747, 2002.

[32] M. Corley and R. J. Hartsuiker, "Hesitation in speech can. . . um. . . help a listener understand," in *Proceedings 25th meeting of the Cognitive Science Society*, Boston, USA, 2003.

[33] P. Collard, M. Corley, L. J. MacGregor, and D. I. Donaldson, "Attention orienting effects of hesitations in speech: evidence from ERPs." *Journal of experimental psychology. Learning, memory, and cognition*, vol. 34, no. 3, pp. 696–702, may 2008.

[34] M. Corley, P. H. Brocklehurst, and H. S. Moat, "Error biases in inner and overt speech: evidence from tongue twisters." *Journal of experimental psychology. Learning, memory, and cognition*, vol. 37, no. 1, pp. 162–75, jan 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21244112

[35] C. M. Laserna, Y.-T. Seih, and J. W. Pennebaker, "Um... Who Like Says You Know: Filler Word Use as a Function of Age, Gender, and Personality," *Journal of Language and Social Psychology*, vol. 33, no. 3, pp. 328–338, mar 2014.

[36] M. Schroder, "Emotional Speech Synthesis : A Review," in *Eurospeech*, Aalborg, Denmark, 2001, pp. 561–564.

[37] C. Nass and S. Brave, "Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship." The MIT Press, 2005.

[38] L. He, H. Hyang, and M. Lech, "Emotional Speech Synthesis Based on Prosodic Feature Modification," in *In Proceedings of the 8th International Conference on Bioinformatics and Biomedical Engineering*, 2013.

[39] M. Aylett, B. Potard, and C. Pidcock, "Expressive Speech Synthesis: Synthesising Ambiguity," in *In Proceedings of 8th ISCA Speech Synthesis Workshop*, 2013.

[40] F. Burkhardt and N. Campbell, "Emotional Speech Synthesis," in *The Oxford Handbook of Affective Computing*, 2014.