DISFLUENCIES IN CHANGE DETECTION IN NATURAL, VOCODED AND SYNTHETIC SPEECH

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

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

[2]School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, UK
r.dall@sms.ed.ac.uk, mwester@inf.ed.ac.uk, martin.corley@ed.ac.uk

ABSTRACT

In this paper we investigate the effect of filled pauses, a discourse marker and silent pauses in a change detection experiment in natural, vocoded and synthetic speech. In natural speech change detection has been found to increase in the presence of filled pauses, we extend this work by replicating earlier findings and explore the effect of a discourse marker, like, and silent pauses. Furthermore we report how the use of "unnatural" speech, namely synthetic and vocoded, affects change detection rates. It was found that the filled pauses, the discourse marker and silent pauses all increase change detection rates in natural speech, however in neither synthetic nor vocoded speech did this effect appear. Rather, change detection rates decreased in both types of "unnatural" speech compared to natural speech. The natural results provide support for the theory that it is the pause itself and not the type of pause which leads to increases in change detection rates. While the "unnatural" results suggest that it is not the full pipeline of synthetic speech that causes the degradation, but rather that something in the preprocessing, i.e. vocoding, of the speech database limits the resulting synthesis.

Keywords: change detection, filled pauses, speech synthesis

1. INTRODUCTION

Filled pauses (FPs) in naturally occurring spontaneous speech have received considerable attention and a variety of interesting phenomena have been found, such as faster reaction times [6, 7], faster word integration [3] and more accurate object identification [1].

This work explores the effect of filled pauses (‘uh’) in the context of "unnatural" speech, namely vocoded and synthetic speech, and compares it to the effects in natural speech. In other work we’ve explored the effects in various reaction time (RT) experiments [5, 11]. Here the same general tendency has been found. Vocoded speech generally mirrors natural speech effects, however no effects are found in synthetic speech except a generally slower RT in response to synthetic speech compared to the other types.

While the reaction time experiments provides evidence that FPs affect peoples’ on-line processing, FPs may have other, and longer term, effects. Change Detection [10] is a paradigm in which participants are asked to listen to short paragraphs of speech and are subsequently presented with the contents of the speech in writing. It is then the task of the participant to detect if a single change has occurred in the text as compared to the speech. This requires participants to not only process the speech as it is heard, but also to memorise it long enough to detect a change at a later point. Thus change detection, as opposed to reaction time, experiments provides a measure of the memorability of the speech in a slightly longer term context.

The basic effect reported by Collard [2] (Chapter 6 & 7), is that the presence of an FP prior to the changing word, as compared to fluent speech, increases the change detection rate by 10-15%. Collard [2] concludes that the acoustic quality of the FP is responsible for the effect (Chapter 7.6, pp. 128). His conclusion was based on manipulating silences around the FP but [ref snaford molle 2006] has shown that a simple silent pause can make the same effect appear, something also foudn in other related studies [4]. We therefore extend this work by including silent pauses and a discourse marker (‘like’) in natural speech to see if the effect is unique to FPs.

As we are interested in the effects of "unnatural" speech types on listeners, we also perform the experiment using vocoded and synthetic speech. Vocoding, in speech synthesis, is the step of parametrising the speech in a manner suitable for statistical machine learning. This parametrised version can be reformed directly by the vocoder, with some loss in quality, and this is what constitutes vocoded speech. Alternatively, a statistical model can be learned from which speech parameters can be generated, this is the method of synthesis applied in this paper.

The working hypothesis was that a similar pattern to the RT experiments would appear, in which the effect of disfluencies is present in natural and vocoded speech, but not in synthetic. This is motivated by the results of the prior experiments, but also by the assumption that current vocoding techniques do not degrade the quality of the speech in a way would prevent the effect from appearing. It is possible however, that a differing pattern will appear due to the differences between the two paradigms. In RT experiments we are testing people's online monitoring and recognition of speech, whereas in change detection people are required to memorise the heard speech in order to detect the change at a later point. This means that even though participants may understand the speech, they may not be able to efficiently memorise it.

2. CHANGE DETECTION EXPERIMENTS

To perform the change detection experiments 35 short paragraphs, 16 critical, 16 filler and 3 practice, said by the same speaker in a spontaneous conversation were prepared. In each paragraph a target word was chosen and four alternative paragraphs were created. One where the target was preceded by a FP ('uh'), a silent pause (SP), the discourse marker 'like' (DM) or by nothing (i.e. fluent speech). The original paragraph was of one of these four cases, and the alternatives were made by altering the original by splicing out the segment immediately preceding the target word and splicing in the relevant replacement. The change word was a near-synonym or semantically related to the target word (i.e. the close-change condition of [2]). For the filler sentences no change existed, however a dummy target word was still chosen in front of which either an FP, SP or DM was placed. The paragraphs potentially included other FPs, DMs and SPs than the critical one so participants could not learn to use those as cues for the change. Two practice sentences contained no change and one a change. Table 1 illustrates a few paragraphs.

The vocoded versions were created taking the natural paragraphs and vocoding them using STRAIGHT [9], no further modifications to the audio was made. The synthetic utterances were made using HTS [12] and a good-quality state-of-the-art HMM-based voice trained on approximately 8 hours of speech. The transcripts of the paragraphs were

Sample Paragraphs
Last week I was thinking actually about writing a book. The possibilities are endless, but I came down to it and thought, well, I'll start with a *cemetery/graveyard* setting, and, within a few minutes, the whole book started to write itself.

Table 1: Sample paragraphs presented to participants. The first of the italicised words is what is heard and the second what is seen by the participant.

used for the synthesis, and versions including a FP or DM was made by inserting these as words in the token stream, whereas the SP version was made in a similar way as in the RT experiments in [5], the length of the SP was thus similar to that of the FP.

2.1. Method

108 participants were recruited, 36 listened to natural speech only, 36 to vocoded and 36 synthetic speech. Each participant only heard samples with either an FP, SP or DM such that for each type of speech and each type of pause there were 12 participants. Each participant listened to the practice sentences and then to each of the 32 paragraphs in a random order, of the 16 critical half contained the appropriate form of pause, and the other half no pause (with 6 participants getting one set and other 6 the other set). In total this yielded 576 (36*16) critical evaluations per speech type and 192 (12*16) per condition (FP, SP or DM) within each speech style.

3. RESULTS

Due to an error in the experiment scripts 96 trials were invalid (5.5%) and were removed from the analysis. In 116 of the remaining trials (7.1%) participants correctly detected a change but incorrectly specified which change. In 16 of these the participant answered that the DM was the change which can arguably be considered correct. Therefore, two analysis were carried out - with (Exact) or without (Permissive) the exact specification of change. Notably however, the pattern of the results are identical. Please note that in the following analysis *disfluent* speech includes FPs, DMs and notably SPs, *fluent* speech is thus only speech with none of these present.

A two-way ANOVA over the by-subject mean scores per condition was run. There was no overall effect of Disfluency Type (FP, DM, SP) or Disfluency Condition (Fluent or Disfluent), however a significant effect of Speech Type (Permissive: $F(2, 99)=5.917$, $p<0.005$, Exact: F(2,

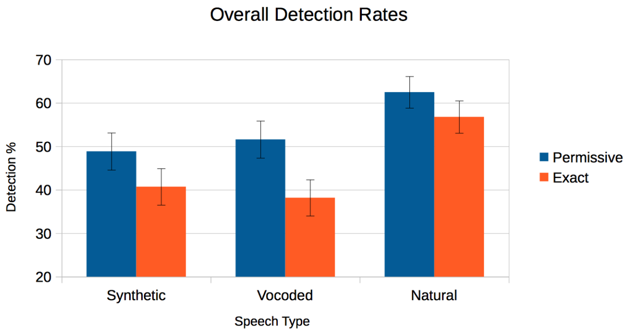

Figure 1: Detection rates per speech type. Permissive includes correct detection of change but incorrect identification. Exact does not.

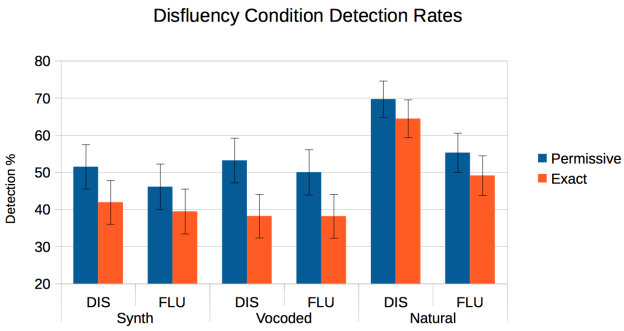

Figure 2: Detection rates divided by disfluency condition and speech type. DIS are disfluent conditions and FLU the fluent condition.

99)=10.377, p<0.0001) was found and an interaction between Speech Type and Disfluency Condition for the Exact analysis (F(2, 99)=5.180, p<0.01) which was only marginal in the Permissive (F(2, 99)=2.788, p=0.066). Using Bonferroni correction the effect of Speech Type is such that for the Natural Speech detection rates were significantly *higher* than Vocoded (Permissive: t(139)=2.692, p<0.05, Exact: t(140)=4.745, p<0.0001) and Synthetic (Permissive: t(142)=3.878, p<0.001, Exact: t(139)=4.699, p<0.0001), but no difference existed between Synthetic and Vocoded (Permissive: t(138)=0.870, p=1, Exact: t(133)=0.662, p=1), see Figure 1. That is, changes are generally detected better in natural speech than in synthetic (by 13.6% in the Permissive and 16.1% in the Exact case) and vocoded (by 10.9% in the Permissive and 18.6% in the Exact case).

The interaction effect (see Figure 2) was explored as it was significant in the Exact case and near significant in the Permissive. Using Bonferroni correction, there was no effect of disfluency condition in synthetic (Permissive: t(70)=1.374, p=0.521, Exact: t(70)=0.582, p=1) and vocoded speech (Permissive: t(70)=0.355, p=1, Exact: t(70)=0.075, p=1), however a significant effect was present in natural speech (Permissive: t(70)=3.326, p<0.005, Exact: t(70)=3.307, p<0.005). The presence of a disfluency did not have any effect on detection rates in synthetic and vocoded speech, however in natural they increased detection rates by 14.4% in the Permissive and 15.3% in the Exact case.

Notably in neither the Exact nor Permissive case was there any significant effect of the type of disruption (FP, SP, DM), see Figures 3, 4 and 5.

4. DISCUSSION

In natural speech all of ‘uh’ (FP), ‘like’ (DM) and a silent pause (SP) increase the change detection rate compared to fluent speech with no disruption.

The SP results are opposed to [2] who concluded that the acoustic quality of the FP was important. While Collard investigated varying the length of SPs surrounding the FP he did not evaluate SPs on their own as done here. Our SP results was also found by [molle 2006] in a very similar experimental setting. These results would therefore be in line with the temporal delay hypothesis of [4] that it is simply the disruption which causes the increase in change detection rates. It may, however, be that the lack of a difference between SP and FP results is a consequence of our many tests. This is as the tendency was for the SP to have lower detection rates than either FP or DM (by about 8%) and testing this in isolation may show an additional advantage of the FP and DM. That the effect appears with the DM ‘like’ can support both the hypothesis that it is the disruption which is the cause but also the idea that the use and purpose of DMs and FPs is similar [refs]. In order to find which is more likely to be true using a non-speech condition as in [4] could be considered in future studies.

Current Synthesis and Vocoding techniques do not produce speech for which the change detection results observed for natural speech are replicated. Where FPs, DMs and SPs increase the detection rate with 11-17% in natural speech there is no discernible pattern in synthetic and vocoded speech, rather, they tend to produce the same detection rates. Not only did the natural effect not appear, for both vocoded and synthetic speech the overall detection rate dropped as compared to natural speech by 11 to 18%. This is not just an effect of increased detections in the disfluency conditions of the natural speech, but rather an overall effect of the speech type. It is notable that this inability to replicate the effect occurs in *both* synthetic and vocoded, as the initial expectation was that current vocoding techniques were good enough to replicate

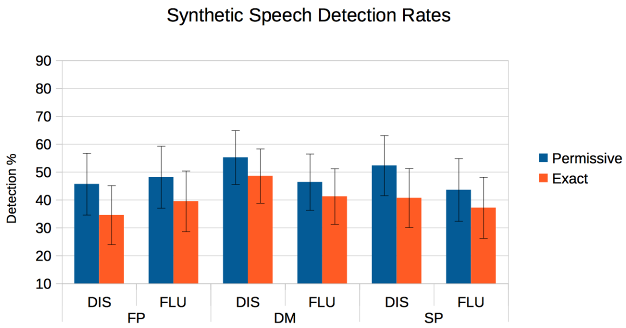

Figure 3: Detection rates per disfluency type (filled pause (FP), discourse marker (DM) and silent pause (SP) for synthetic speech. The FLU sentences are the corresponding results for the same paragraph as the DIS.

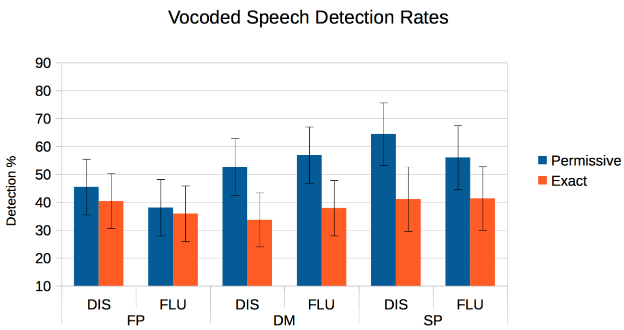

Figure 4: Detection rates per disfluency type (filled pause (FP), discourse marker (DM) and silent pause (SP) for vocoded speech.

the effect. However, they are not. This suggests that it is not simply a matter of the speech prosody and general naturalness being poor, but rather that there is something about the inherent speech quality of the vocoder which limits synthetic speech in this regard.

In reaction time experiments we have found that vocoded speech [5, 11] elicits the same patterns as natural speech, which is in contrast to current results. Vocoding is known to introduce a buzzy character to the speech, while we are aware of the effects on the speech' perceived naturalness of this [8], other possible psychological effects of this buzziness is unknown. It is possible that this demonstrates one of them. To detect a change the participant must necessarily be able to commit to (short term) memory what was being said in the paragraph in order to compare with the text later. Thus if the effect of vocoding decreases participants ability to memorize the salient elements of the paragraph, it should show an overall decrease in participant's ability to detect changes, something which is the case. This decrease is likely due to an additional strain on the participant's cognitive resources and can also explain the lack of disruption/temporal delay effect. The participant must use so many resources to simply process

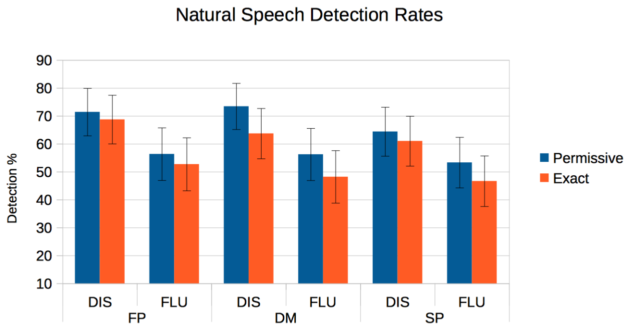

Figure 5: Detection rates per disfluency type (filled pause (FP), discourse marker (DM) and silent pause (SP) for natural speech.

the incoming speech stream that any potential benefit to be had from the disruption is lost. Following [2], the effect of disfluency found in natural speech is due to heightened attention to the target word, resulting in better recall and notice of changes. While durational and prosodic cues may still be present after vocoding, if the participant is already straining their cognitive resources to simply understand and commit the content to memory, it is likely that these does not result in any attentional shift. This is, however, speculative and further experimental evidence would be needed. Experiments consciously manipulating the cognitive strain on participants such as dual-attention [ref] could be used in combination with a change detection paradigm using natural speech, if this changes the natural results to look similar to those of the vocoded and synthetic it would provide evidence toward a cognitive strain hypothesis.

5. CONCLUSION

We have shown that FPs, DMs and SPs increase change detection rates in natural speech, but that this effect is not present in either vocoded or synthetic speech. The SP results are in contrast to [2], but support the temporal delay hypothesis of [4], and the vocoding results to [5, 11]. We have discussed why this may be and suggestions for further work which may resolve these tensions which include using a non-speech condition and a dual-attention paradigm.

6. ACKNOWLEDGEMENTS

The authors would like to thank Amelie Osterrieth and Anisa Jamal for help collecting the natural speech data. This work has been partially funded by the JSTCREST uDialogue project and EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

7. REFERENCES

[1] Brennan, S. Feb. 2001. How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language* 44(2), 274–296.

[2] Collard, P. 2009. *Disfluency and listeners' attention: An investigation of the immediate and lasting effects of hesitations in speech*. PhD thesis University of Edinburgh.

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

[4] Corley, M., Hartsuiker, R. J. Jan. 2011. Why um helps auditory word recognition: the temporal delay hypothesis. *PloS one* 6(5), e19792.

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

[6] Fox Tree, J. E. 1995. The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language* 34(6), 709–738.

[7] Fox Tree, J. E. 2001. Listeners' uses of um and uh in speech comprehension. *Memory and Cognition* 29(2), 320–326.

[8] Henter, G. E., Merritt, T., Shannon, M., Mayo, C., King, S. September 2014. Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech. *Proc. Interspeech* volume 15 1504–1508.

[9] Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A. 1999. 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* 27(3-4), 187–207.

[10] Sturt, P., Sanford, A. J., Stewart, A., Dawydiak, E. 2004. Linguistic focus and good-enough representations: An application of the change-detection paradigm. *Psychonomic Bulletin and Review* 11(5), 882–88.

[11] Wester, M., Corley, M., Dall, R. 2015. The temporal delay hypothesis: natural, vocoded and synthetic speech. *Submitted to DiSS* Edinburgh, UK.

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