THE TEMPORAL DELAY HYPOTHESIS: NATURAL, VOCODED AND SYNTHETIC SPEECH

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

[1]The Centre for Speech Technology Research, The University of Edinburgh, UK
[2]PPLS, The University of Edinburgh, UK
m.wester@inf.ed.ac.uk, martin.corley@ed.ac.uk, r.dall@sms.ed.ac.uk

ABSTRACT

Including disfluencies in synthetic speech is being explored as a way of making synthetic speech sound more natural and conversational. How to measure whether the resulting speech is actually more natural, however, is not straightforward. Conventional approaches to synthetic speech evaluation fall short as a listener is either primed to prefer stimuli with filled pauses or, when they aren't primed they prefer more fluent speech. Psycholinguistic reaction time experiments may circumvent this issue. In this paper, we revisit one such reaction time experiment. For natural speech, delays in word onset were found to facilitate word recognition regardless of the type of delay; be they a filled pause (*um*), silence or a tone. We expand these experiments by examining the effect of using vocoded and synthetic speech. Our results partially replicate previous findings. For natural and vocoded speech, if the delay is a silent pause, significant increases in the speed of word recognition are found. If the delay comprises a filled pause there is a significant increase in reaction time for vocoded speech but not for natural speech. For synthetic speech, no clear effects of delay on word recognition are found. We hypothesise this is because it takes longer (requires more cognitive resources) to process synthetic speech than natural or vocoded speech.

Keywords: delay hypothesis, disfluency

1. INTRODUCTION

Various studies have shown that speech understanding can sometimes benefit from the presence of filled pauses (e.g., *um* and *uh*) and that words following a filled pause are recognised more quickly [8, 4, 3]. A study by Corley and Hartsuiker [5] showed that not just filled pauses but delays of any kind help auditory word processing. This study investigates whether synthetic speech understanding also benefits from delays in the form of either filled pauses or silence.

The end objective of this work is to produce synthetic conversational speech (interesting for e.g., artificial personalities, more natural speech synthesis) and including disfluencies in the synthetic speech is a possible way of achieving this. Evaluating a synthetic system which includes disfluencies however is not straightforward. The standard preference tests used in the synthetic speech field result in listeners either being primed to prefer sentences with fillers [1] or when they are not primed they prefer stimuli without fillers [6].

Experimental paradigms (e.g., reaction time and change detection experiments) borrowed from the field of psycholinguistics may be a way of circumventing this issue. First of all, in these paradigms listeners are not primed regarding the presence or absence of disfluencies. Secondly, listeners are not asked to judge the quality of the synthetic speech, they are asked to react to the speech they have processed. The idea is that if listeners respond to filled pauses in synthetic speech in the same way as they do to filled pauses in natural speech, we will have an indirect measure of the quality of synthetic speech and the validity of including disfluencies in synthetic speech will be strengthened.

A previous reaction time study [7], including filled pauses in synthetic and vocoded speech, showed that processes observed for natural speech were also observed for vocoded speech but not synthetic speech. The lack of effect for synthetic speech was hypothesised to be due to the poor quality of the synthetic filled pauses. In Corley and Hartsuiker [5] it was shown that delays in word onset facilitate word recognition regardless of the type of delay, whether they were filled with *um*, silence or even non-speech sounds (a tone). Corley and Hartsuiker refer to this as "the temporal delay hypothesis", i.e., it is the temporal delay that facilitates word recognition rather than that speech understanding benefits from the presence of filled pauses such as *um*, *uh*, or similar. If this temporal delay hypothesis applies to

synthetic speech then the quality of the filled pause should be of less importance. The Corley and Hartsuiker study, a reaction time experiment, was replicated here for natural, vocoded and synthetic speech.

2. METHOD

The experiment consists of participants viewing pairs of images on a computer screen and following instructions to press a button corresponding to one of the images as quickly as possible. Details of the materials, speech types and experimental procedure are given below.

2.1. Materials

The same experimental materials were used as in [5]. The materials consisted of both auditory and visual stimuli. The auditory stimuli were instructions to press a button corresponding to one of the pictures in a pair. In the delay conditions, listeners heard an instruction with a delay directly preceding the target word. In the control conditions, the delay was earlier on in the sentence. The instructions were either:

1. Now press the button for the <delay> <target>, please.
2. Now press the <delay> button for the <target>, please.

The delay was either a filled pause *um* or a silent pause of the same length. In addition to the delay there was also a task difficulty manipulation. In the difficult condition the words were low-frequency (LF) words and visually blurred. In the easy condition the target words were high-frequency (HF) words and visually intact. Two sets of 16 pictures were used (examples of LF words: kite, snail, vase, etc., HF examples: bed, foot, tree, etc.). For details of how the frequency category of the words was determined see [5]. Each LF picture was paired with four HF pictures (never in the same combination) resulting in 64 picture pairs. Each picture was shown twice on the left, twice on the right and was a target twice: once in an instruction with an early delay and once with a late delay. The delay was either a filled pause (*um*) or a silent pause. Three picture pairs with mid-frequency items (lamp-cake, clock-knife, wheel-cow) were used for practice trials at the start of the experiment. Figure 1 shows an example of the picture pair snail/tree.

2.2. Speech types

The above described experiment was run using natural, vocoded and synthetic speech. The natural speech recordings have been described in detail in [4, 5]. To summarise, a female native speaker of English was recorded reading the list of target words embedded in the above carrier sentence. Target words, together with the word *please*, were removed from their original contexts and spliced into one version of the carrier sentence that had not originally included any of the target items. The delay was created by asking the speaker to insert an *um* "as naturally as possible" when reading a list of low-frequency items in carrier sentences. A single *um* that was judged most natural was selected and spliced in before the target word (delay condition) and before the word *button* (control condition). All targets start at 2297 *ms*, the *um* or silent pause is 1078 *ms* long.

Figure 1: Example of picture pair snail/tree.

The vocoded and synthetic speech were generated as in [7]. The vocoded speech was created by taking the natural speech and vocoding the stimuli using STRAIGHT [10]. The durations output by STRAIGHT are not exactly the same as the natural speech durations, due to the way silence is dealt with. The target onset for vocoded speech with *um* matches the natural speech at 2297 *ms*. For the silence condition, the target onset time is at 2270 *ms* (the length of the pauses is 1078 *ms* in both conditions).

The HMM-based synthetic speech was generated using HTS 2 [18] in a system newer than but roughly similar to [17]. All the target words were synthesised in the carrier sentence. The *um* was generated by the system but some additional padding was added by hand to make the pause the same length as the pause in natural speech. The silence delay was spliced in by hand. Target onsets were measured by hand and vary from 2413 *ms* to 2507 *ms*.

2.3. Procedure

The experiment was run using OpenSesame [12]. The auditory stimuli were presented to native British English speakers with no hearing problems over Beyerdynamic DT770 headphones in individual sound-treated booths. In total 120 subjects took part, twenty per speech type. The participants were in-

Table 1: Reaction time results for the six experiments: Natural, Vocoded and Synthetic Speech including either *um* or silence delays. Participant mean correct reaction time (ms) relative to target onset. Standard error in brackets.

(a) Experiment 1– Natural Speech; *um*

Instructions	Target Type	
	clear HF	blurred LF
control (*um* early)	376 (10.4)	429 (10.9)
delay (*um* late)	369 (10.0)	417 (12.0)

(b) Experiment 2 – Natural Speech; silence

Instructions	Target Type	
	clear HF	blurred LF
control (silence early)	422 (12.4)	461 (12.5)
delay (silence late)	382 (11.4)	432 (12.0)

(c) Experiment 3– Vocoded Speech; *um*

	clear HF	blurred LF
control (*um* early)	398 (10.5)	444 (10.6)
delay (*um* late)	368 (11.0)	389 (11.2)

(d) Experiment 4 – Vocoded Speech; silence

	clear HF	blurred LF
control (silence early)	440 (10.1)	474 (10.2)
delay (silence late)	389 (9.0)	451 (11.1)

(e) Experiment 5 – Synthetic Speech; *um*

	clear HF	blurred LF
control (*um* early)	557 (9.2)	603 (10.0)
delay (*um* late)	564 (9.7)	592 (10.1)

(f) Experiment 6 – Synthetic Speech; silence

	clear HF	blurred LF
control (silence early)	553 (11.5)	582 (11.4)
delay (silence late)	561 (10.5)	586 (11.4)

formed that the study was about sentence comprehension and that the aim of the study was to follow instructions given in stressful situations. This minor deception was necessary to justify the disfluencies in the study. Ethical approval was obtained from the Ethics Committee of PPLS, University of Edinburgh. The participants were explicitly told to be as fast and accurate as they could. Prior to the experiment starting, the subjects were given the three practice trials to familiarise themselves with the procedure. Following this, the 64 items were presented in a random order. The experiment took just over 5 minutes to complete.

3. RESULTS

For each of the experiments, there are 1280 responses (20*64). Before the data was analysed some of the responses had to be removed: One participant in Experiment 3 did not complete the task and so was disregarded, reducing the number of responses for that experiment to 1216. Furthermore, all incorrect responses (e.g., a subject clicking left when it should have been right) were removed, as well as all responses with a reaction time (RT) smaller than 0 *ms* and all RTs larger than 1100 *ms*. RTs < 0 indicate a participant responded before the target started, RTs > 1100 correspond to button pushes well after the end of the utterance. The number of discarded responses and the total responses included in the analyses per experiment are given in Table 2. Analyses were carried out by fitting Generalized Linear Mixed-Effects models, as implemented in the lme4 library in R [14, 2].

Table 1 shows mean correct reaction times (RTs) relative to target onset with standard error between brackets. Experiments 1 and 2 here are the same as Experiments 1 and 2 in [5].

In Experiment 1, which included natural speech and *um* as the local delay, we found that the delay led to only very small decreases in RT (7 *ms*) for the clear HF words, and marginally larger decreases for LF blurred words (12 *ms*). This effect was not significant, in contrast to what was reported in [5]. On the other hand, the effect of task difficulty was found to be significant with participants taking 48 *ms* longer to react to blurred LF images ($p = .003$) which is more in line with the results reported in [5].

In Experiment 2, which again included natural speech but this time with a silent pause as the delay, a significant effect of delay was found with participants faster by 36 *ms* in the delay condition ($p = .03$). The effect of task difficulty was also significant with participants 50 *ms* slower to respond to blurred images ($p = .003$).

Experiments 3 and 4 show results for vocoded speech. In the *um* condition (Experiment 3), a significant decrease of 42 *ms* ($p = .008$) in RT was found due to the local delay. The effect of task diffi-

Table 2: Number of discarded trials per experiment, and the total number of included trials.

Experiment #	1	2	3	4	5	6
# Incorrect	38	26	39	17	22	42
#RT<0	20	0	3	2	1	0
#RT > 1100	19	59	14	21	23	92
Total responses	1203	1195	1160	1240	1234	1146

culty was significant ($p = .03$) with participants taking 33 *ms* longer to respond to blurred LF pictures. In the silence condition (Experiment 4), significant effects of both delay and task difficulty were found, participants were respectively 36 *ms* faster after a silence delay ($p = .003$) and 50 *ms* slower in the blurred LF condition ($p = .003$).

Experiments 5 and 6 show the results for synthetic speech. Overall the RTs are slower for synthetic speech than for natural and vocoded speech. There is no significant effect of delay for synthetic speech. In both *um* and silence conditions there is a main effect of task difficulty with blurred LF words processed less quickly than HF clear words 42 *ms* ($p = .0003$) and 33 *ms* ($p = .02$), respectively.

Cross-experiment comparisons in which we incorporate an additional "experiment" factor show significant effects of speech type ($p < 0.0001$) and of frequency ($p = 0.0002$). This frequency, or task difficulty, effect corresponds to the findings reported above per experiment. Regarding speech type, synthetic speech is 146 *ms* slower in the silence condition and 181 *ms* slower in the *um* condition than natural speech, and 131 *ms* and 179 *ms* slower than vocoded speech. There is no significant difference in RT between natural and vocoded speech (p=0.96).

4. CONCLUSIONS

The only robust result across all three types of speech (natural, vocoded and synthetic) is that it takes approximately $30 - 50$ *ms* longer to react to blurred images than to visually intact images.

Natural and vocoded speech show a similar picture to the findings in Corley & Hartsuiker's paper [5]. Experiments 2, 3 and 4 support their conclusion "... any delay in word onset can help word recognition". There was a main effect of delay after a silent pause in both natural and vocoded speech and there was a main effect of the *um* delay for vocoded speech. However, the results for Experiment 1 only show a slight increase in the speed of word recognition after *um*.

No effect of delays was observed for synthetic speech. Neither the *um* nor the silence conditions led to increases in RT. This is in line with our previous RT experiments [7], which followed Fox Tree's method [9, 8], and showed that filled pauses (*uh*) led to faster reaction times in natural and vocoded speech but slower reaction times in synthetic speech. At the time this was hypothesised to be due to the poor quality of the filled pauses and prosody. In the current study, not only did we consider filled pauses but also silent pauses.

We found that listeners are significantly slower in experiments 5 & 6 (synthetic speech) compared to the same experiments using natural and vocoded speech. Our findings give support to the theory that processing synthetic speech requires a listener to apply more cognitive resources than when processing natural speech. In [16], Pisoni and colleagues illustrate that the perception of synthetic speech requires more cognitive resources citing studies from the eighties using formant synthesis. For instance, listeners took more time to process synthetic stimuli than natural stimuli in a speeded lexical decision task [13] and they needed to hear more of synthetic speech before reliably identifying whole words [11]. Our RT studies suggest that this also holds for modern statistical parametric speech synthesis (SPSS). Future work revisiting some of the lexical decision, word recognition and sentence verification tasks comparing SPSS and natural speech should further inform how the acoustic-phonetic characteristics of SPSS influence speech perception and how speech synthesis has evolved psycholinguistically compared to earlier approaches.

For now, there is no evidence that including disfluencies is beneficial in synthetic speech as the disfluencies do not seem to be processed in the same way as in natural speech. We hypothesise that no clear effects of delay on word recognition are found because synthetic speech takes so much longer to process than natural or vocoded speech.

All research data associated with this paper can be found at Edinburgh DataShare [15] (http://hdl.handle.net/10283/806).

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

5. REFERENCES

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

[2] Bates, D., Maechler, M., Bolker, B. M., Walker, S. 2014. lme4: Linear mixed-effects models using Eigen and S4. Available: http://CRAN.R-project.org/package=lme4. R package version 1.1-7.

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

[4] Corley, M., Hartsuiker, R. J. 2003. Hesitation in speech can... um... help a listener understand. *Proceedings of the 25th Meeting of the Cognitive Science Society* 276–281.

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

[6] Dall, R., Tomalin, M., Wester, M., Byrne, W., King, S. 2014. Investigating automatic & human filled pause insertion for speech synthesis. *Proceedings Interspeech* Singapore.

[7] 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.

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

[9] Fox Tree, J. E., Schrock, J. C. Feb. 1999. Discourse Markers in Spontaneous Speech: Oh What a Difference an Oh Makes. *Journal of Memory and Language* 40(2), 280–295.

[10] 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.

[11] Manous, L., Pisoni, D. 1984. Effects of signal duration on the perception of natural and synthetic speech. *Research on Speech Perception Progress Report No* 10.

[12] Mathôt, S., Schreij, D., Theeuwes, J. 2012. Opensesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods* 44(2), 314–324.

[13] Pisoni, D. 1981. Speeded classification of natural and synthetic speech in a lexical decision task. *The Journal of the Acoustical Society of America* 70(S1), S98–S98.

[14] R Development Core Team, 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. ISBN 3-900051-07-0.

[15] Wester, M., Dall, R., Corley, M. 2015. Experiment materials for "The temporal delay hypothesis: Natural, vocoded and synthetic speech.", [dataset]. University of Edinburgh, School of Informatics, Centre for Speech Technology Research. http://dx.doi.org/10.7488/ds/272.

[16] Winters, S., Pisoni, D. 2005. Speech synthesis: Perception and comprehension. In: Brown, K., (ed), *Encyclopedia of Language and Linguistics* volume 12. 31–49.

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

[18] 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 of SSW6* Bonn, Germany. 294–299.