TIME-COMPRESSING NATURAL AND SYNTHETIC SPEECH

Esther Janse

Utrecht institute for Linguistics OTS
University of Utrecht, the Netherlands
Esther.Janse@let.uu.nl

ABSTRACT

Phoneme detection is a useful tool to compare the perception of perfectly intelligible speech types. As previous research suggests that perception of fast speech is helped by segmental redundancy, we expected the hyperarticulation of synthetic speech to turn into an advantage at a fast rate. Consequently, the processing advantage of natural over synthetic speech was expected to decrease after time-compression. Secondly, detection times were expected to be slower after moderate time-compression because of the higher processing difficulty of fast speech. However, detection times tended to become shorter in the time-compressed condition. This was attributed to shorter durations of syllables and words. Furthermore, the processing advantage of natural over synthetic speech did not decrease, but rather tended to increase. This may be explained by the lack of a speaking effort pattern in synthetic diphone speech, which makes it rather blurred at faster playback rates.

1. INTRODUCTION

Segmental intelligibility is severely affected when speech is time-compressed to very fast rates of speech. The fact that speech remains intelligible at playback rates that are twice or almost three times as fast as normal rate is to be attributed to segmental intelligibility, lexical redundancy, and context information. Lexical redundancy and context information help listeners to fill in the ‘difficult’ segments of which segmental intelligibility is low.

In this study, perception of natural speech is compared with synthetic speech. Although synthetic speech may be perfectly intelligible, natural speech may still be easier to process for listeners. Differences in processing speed between natural and synthetic speech have been found, e.g., by using phoneme detection [1].

The central question in this study is whether and how the difference in processing speed between the two speech types changes under the influence of increased playback rate.

Synthetic diphone speech generally has a very high intelligibility and good quality. Yet, diphone speech can be said to be hyperarticulated: all building blocks are cut from neatly articulated stressed nonsense syllables. Thus, the diphone building blocks are, in a way, maximally redundant. The fact that synthetic diphone speech still sounds rather unnatural is, at least partly, a result of all segments being equally stressed and hyperarticulated. Several attempts have been carried out to improve the naturalness of synthetic diphone speech by introducing unstressed diphones [2] or by spectral reduction of certain transitions [3]. In this study, the question is raised whether the segmental redundancy, or hyperarticulation, present in synthetic diphone speech could turn into an advantage, when the processing speed of time-compressed natural speech is compared with that of time-compressed synthetic speech.

Previous research suggests that listeners are helped by segmental redundancy when speech is presented to them at very fast rates. At equal rates of speech, time-compressed speech is more intelligible than speech which is articulated rapidly. Even at a rate of speech at which intelligibility of naturally produced fast speech is still almost 100%, word perception is easier in time-compressed speech than in natural fast speech [4]. Secondly, there is some evidence that assimilation speeds perception at normal rate, but that it hinders perception when the speech is presented at a fast rate [5]. This is against the expectation that assimilation speeds up perception in a context in which listeners expect these assimilation processes to occur. Even though the unassimilated form was rather unnatural given the fast speech rate, listeners were faster in recognising the unassimilated form.

Could it be the case that the unnatural hyperarticulation of synthetic speech turns into an advantage when the speech is presented at a Faster rate? In other words, does the difference in processing speed between natural and synthetic speech decrease when both types of speech are time-compressed? In difficult listening conditions, the unnaturalness of hyperarticulated speech may be outweighed by the increased intelligibility due to hyperarticulation.

A second question, related to the first question, addresses the increased processing difficulty after time-compression. Time-compressed speech is expected to elicit slower phoneme detection times than normal rate speech. Although speech that is time-compressed to 65% of the original duration is still highly intelligible, it is expected to put a higher processing load on the listeners. This higher processing load is expected to translate into slower phoneme detection times.

The following two hypotheses were tested:

1. The difference in processing speed between natural and synthetic speech is smaller when both types of speech are time-compressed than at normal rate.
2. Time-compressed speech elicits slower detection times than speech which is presented at normal rate.

2. PERCEPTION OF NATURAL AND SYNTHETIC SPEECH

In [1] phoneme detection is described as a tool for comparing perception of natural and synthetic speech. Phoneme detection time is a good measure of the ease of processing, and thus of
the speech communication quality of highly intelligible synthetic speech types. Nix et al. [1] found longer response times for synthetic speech, relative to natural speech, which indicates that there is a processing advantage of natural speech over synthetic speech. The difficulty of listening to synthetic speech has been argued to occur mainly at the lower phonetic level, and not at higher prosodic levels [6]. Phonetic information is often redundantly specified in natural speech, but the cues are weaker in synthetic speech, which implies harder work for the phonetic processor. Although improvements towards more appropriate and more natural prosody are certainly preferred by listeners, the perceptual disadvantages at the lower phonetic level are assumed to demand the most extra processing capacity. Note that the experiments [1,6] both involved speech synthesis by rule. In the present study, diphone synthesis is used. This type of synthesis depends on the concatenation of naturally produced building blocks. Although diphone speech is richer in phonetic cues than rule-based synthetic speech because of the inherent natural phoneme-to-phoneme transitions, the difficulty of processing synthetic diphone speech is also assumed to be found mainly in phonetic processing. In the present study the prosodic pattern (i.e., duration and intonation) of the natural utterance is applied to its synthetic counterpart. Consequently, there is no reason to assume that any possible processing difficulties are to be found at higher prosodic levels.

Phonemes can be detected via pre-lexical processing, or via lexical processing, depending on which route is faster [7]. When low-level processing is more difficult in one condition than in the other, this will affect the speed with which syllables can be detected, and consequently, the speed of word recognition. Hence, for the present experiment, it may not really be relevant to know whether phonemes are detected on the basis of lexical or pre-lexical information.

2.1. Method

Synthetic diphone speech

The MBROLA system (Multi-Band Resynthesis Overlap-Add) is based on the concatenation of pre-recorded diphones. A diphone is a unit that begins in the middle of a phone and ends in the middle of the following one. Thus, diphones contain transitions between two speech sounds, recorded from natural speech. The position of the border between the two phones is also stored, so that the duration of one half-phone can be modified without affecting the duration of the other half. To avoid audible discontinuities, the spectral envelopes, pitch, and phase of the voiced parts of all segments are resynthesised with constant synthesis pitch and fixed initial phases for each period (MultiBand Resynthesis [8]). Spectral smoothing is solved ‘at run-time’, by linear interpolation in the time-domain. The overlap-add technique can be applied during concatenation in order to provide the correct pitch and duration to the speech segments. The input to the MBROLA synthesiser is a list of phonemes, together with prosodic information (duration of phonemes and a piecewise linear description of pitch).

Diphone speech may be rich in acoustic cues, but it is also rich in false or misleading acoustic cues. Diphone /p/ for the word pen may have been cut from the syllable pet and thus still contains some cues for a coda /t/. Although spectral smoothing is applied at diphone edges, the signal is not as smooth as natural speech. Furthermore, diphones can only account partially for the coarticulatory effects in speech because these often affect a whole segment rather than its first or second half independently [9]. These properties of diphone speech all contribute to a quality disadvantage relative to natural speech.

Material

100 Sentences with suitable target phonemes were selected from a recording of the male speaker, whose diphones are used as the standard MBROLA diphone set for Dutch. The read-aloud sentences had all been taken from books. The target phonemes were all word-initial plosives: /p,t,k,d,b/. Because the speech material had not been designed for the purpose of a phoneme detection experiment, all possible target words were chosen. Consequently, the 100 sentences were not balanced over these five phonemes. The target words (32 monosyllabic and 68 polysyllabic) were either nouns, verbs or adverbs. If possible, the target phoneme did not occur in the sentence, neither word-initially, word-medially nor word-finally, before the target word.

A synthetic copy of the 100 sentences was made with the help of the Dutch text-to-speech conversion program Fluent Dutch (version 1.6). Fluent Dutch is based on the MBROLA synthesiser, and on the standard MBROLA diphone set for Dutch. Grapheme-to-phoneme conversion was done automatically, and a natural pitch contour and suitable durations were computed. By default, the program gives sentence accent to all content words and main verbs. If certain words are to be accented or deaccented, this can be indicated in the orthographic input. The pitch contour was adapted manually to make it similar to that of the natural utterance. Secondly, the duration of the target word was equal to that in the natural version by means of linear time-scaling. Furthermore, the durations of the parts of the sentence preceding and following the target word were made equal to the natural version. All phoneme durations within the words were computed by the speech synthesis program Fluent Dutch.

For the fast condition, the natural and synthetic versions of the sentences were time-compressed linearly to 65% of their original duration by means of PSOLA. This is about the fastest speech rate speakers can attain if they try very hard. Speech time-compressed to this rate is almost perfectly intelligible [4].

In addition to the 100 test sentences, there were 70 filler sentences which did not contain the assigned phoneme. These filler items were interspersed with the material to prevent that subjects would press the button randomly. Filler items were rotated over the 4 experimental conditions.

Design and procedure

The 100 test items were rotated over the 4 conditions (Natural-normal rate, Synthetic-normal rate, Natural-fast, Synthetic-fast) (using a Latin square design) on 4 experimental lists.

Subjects were seated in sound-treated booths. The sentences were played to them over closed headphones. They first read an experiment instruction on the computer screen before they started the experiment. They were told to look at the screen because a letter would appear on the screen before the sentence was played to them. Once the sentence was playing, they were told to press the button as soon as they heard this sound in word-initial position. The onset of the target plosives was marked in the waveform by means of a label. The experiment program then computed the phoneme detection
time by measuring from that label point in time to the moment that the button press was registered.

There was a practice session of 10 items. Before subjects started with the actual test, 10 warming-up sentences were played to them after which they proceeded seamlessly with the actual test. Natural and synthetic items, both at normal and at fast rate, were presented in random order.

To each experimental list, 10 subjects were assigned. The 40 subjects were all students at Utrecht University and received a small payment for their participation.

2.2. Results

After subjects had finished the test, they were asked whether they thought that all speech conditions had been intelligible. Most subjects thought that all speech conditions were of very good intelligibility. However, some subjects thought that the time-compressed synthetic speech sounded a bit blurred and that it was difficult to detect word boundaries.

The raw mean phoneme detection times were computed, along with the percentage of missing values. Missing values were due to subjects missing the phoneme, or responding to an earlier non-initial occurrence of the phoneme. The detection times and percentages of missing values in each condition are shown in Table 1.

As a first analysis of the data, all missing values were replaced by the grand mean 627 ms. This is a conservative way of dealing with missing values because possible differences between conditions are made smaller. The data were entered into ANOVAs (either subjects or items as Repeated Measures) to test the effects of Speech Type and Rate. The effect of Speech Type was highly significant (F1(1,39)=77.1, p<0.001; F2(1,75)=21.1, p<0.001). The direction of the Rate effect is opposite to what was expected; this effect was significant (F1(1,39)=9.37, p<0.001; F2(1,75)=6.34, p=0.014). The interaction between Speech Type and Rate, however, failed to reach significance (F1(1,39)=3.26, p=0.079; F2(1,75)=2.15, p=0.15). One major problem with the analysis of reaction time data is that the data are not normally distributed. Since analyses of variance actually assume normal distributions, the data were transformed to reaction frequencies (1/RT) in order to make the distributions more normal. The Repeated Measures analyses were run again on this smaller dataset. There was a significant main effect of Speech Type (F1(1,39)=77.1, p<0.001; F2(1,75)=21.1, p<0.001). The interaction of the Rate effect is opposite to what was expected; this effect was significant (F1(1,39)=9.37, p<0.001; F2(1,75)=6.34, p=0.014). The interaction between Speech Type and Rate was significant (F1(1,39)=6.24, p=0.052; F2(1,75)=2.15, p=0.15). For the monosyllabic items, the difference between the two speech conditions was 20 ms at normal rate (612 ms). The remaining subset of 76 items was largely comprised of polysyllabic items, the difference between the two speech conditions was 575 ms (652 ms). The remaining missing values were replaced by this grand mean and the analyses (Repeated Measures) were run again on this smaller dataset. There was a significant main effect of Speech Type (F1(1,39)=77.1, p<0.001; F2(1,75)=21.1, p<0.001). The direction of the Rate effect is opposite to what was expected; this effect was significant (F1(1,39)=9.37, p<0.001; F2(1,75)=6.34, p=0.014). The interaction between Speech Type and Rate, however, failed to reach significance (F1(1,39)=3.26, p=0.079; F2(1,75)=2.15, p=0.15). One major problem with the analysis of reaction time data is that the data are not normally distributed. Since analyses of variance actually assume normal distributions, the data were transformed to reaction frequencies (1/RT) in order to make the interactions more normal. The Repeated Measures analyses were run again on this smaller dataset. There was a significant main effect of Speech Type (F1(1,39)=77.1, p<0.001; F2(1,75)=21.1, p<0.001). The interaction of the Rate effect is opposite to what was expected; this effect was significant (F1(1,39)=9.37, p<0.001; F2(1,75)=6.34, p=0.014). The interaction between Speech Type and Rate was significant (F1(1,39)=6.24, p=0.017; F2(1,50)=3.84, p=0.058). The interaction between Speech Type and Rate now only just missed significance (F1(1,39)=3.26, p=0.079; F2(1,75)=2.15, p=0.15). One major problem with the analysis of reaction time data is that the data are not normally distributed. Since analyses of variance actually assume normal distributions, the data were transformed to reaction frequencies (1/RT) in order to make the distributions more normal. The Repeated Measures analyses were run again on this smaller dataset. There was a significant main effect of Speech Type (F1(1,39)=77.1, p<0.001; F2(1,75)=21.1, p<0.001). The interaction of the Rate effect is opposite to what was expected; this effect was significant (F1(1,39)=9.37, p<0.001; F2(1,75)=6.34, p=0.014). The interaction between Speech Type and Rate was significant (F1(1,39)=6.24, p=0.017; F2(1,50)=3.84, p=0.058). The interaction between Speech Type and Rate now only just missed significance (F1(1,39)=3.26, p=0.079; F2(1,75)=2.15, p=0.15).

The grand mean detection time was computed for the subset (612 ms). The remaining missing values were replaced by this grand mean and the analyses (Repeated Measures) were run again on this smaller dataset. There was a significant main effect of Speech Type (F1(1,39)=77.1, p<0.001; F2(1,75)=21.1, p<0.001). The interaction of the Rate effect is opposite to what was expected; this effect was significant (F1(1,39)=9.37, p<0.001; F2(1,75)=6.34, p=0.014). The interaction between Speech Type and Rate was significant (F1(1,39)=6.24, p=0.017; F2(1,50)=3.84, p=0.058). The interaction between Speech Type and Rate now only just missed significance (F1(1,39)=3.26, p=0.079; F2(1,75)=2.15, p=0.15). One major problem with the analysis of reaction time data is that the data are not normally distributed. Since analyses of variance actually assume normal distributions, the data were transformed to reaction frequencies (1/RT) in order to make the distributions more normal. The Repeated Measures analyses were run again on this smaller dataset. There was a significant main effect of Speech Type (F1(1,39)=77.1, p<0.001; F2(1,75)=21.1, p<0.001). The interaction of the Rate effect is opposite to what was expected; this effect was significant (F1(1,39)=9.37, p<0.001; F2(1,75)=6.34, p=0.014). The interaction between Speech Type and Rate was significant (F1(1,39)=6.24, p=0.017; F2(1,50)=3.84, p=0.058). The interaction between Speech Type and Rate now only just missed significance (F1(1,39)=3.26, p=0.079; F2(1,75)=2.15, p=0.15).

3. DISCUSSION

Despite the high intelligibility and quality of diphone speech, listeners still find natural speech easier to process than synthetic speech. This difference in ease of processing must, at least partly, be attributed to phonetic processing difficulties. The processing advantage of natural speech was expected to
decrease at the faster than normal playback rate, but the results even showed a tendency in the opposite direction.

Secondly, contrary to our expectation, phonemes were detected somewhat faster in time-compressed speech than at normal rate, although this was only significant for the subset of items. This suggests that the extra processing load due to the faster playback rate is not that substantial. Theoretically speaking, because of the shorter syllable and word durations in time-compressed speech, the syllable or word recognition points are reached within a shorter time-interval from the onset of the plosive than at normal rate. The higher speech rate may also make listeners feel rushed: they perform their task faster, but they also make more mistakes (cf. the increase in miss rate for both speech types).

The first hypothesis was based on the assumption that the negative aspects of synthetic speech (misleading coarticulatory cues etc.) would not become more harmful with increased playback rate. Yet, it is conceivable that these aspects indeed do become more problematic to speech processing when speech is time-compressed.

An alternative explanation may be found in the fact that the processing advantage of natural over synthetic speech even increases after time-compression for the polysyllabic words. On the basis of the hyperarticulation-is-helpful-in-difficult-listening-situations-hypothesis, one would have expected the processing advantage of natural speech to decrease mainly for polysyllabic items because these items consist of all-stressed-and-hyperarticulated syllables. Note, however, that some subjects complained that the time-compressed synthetic speech sounded blurred to them, and that they found it difficult to detect coarticulatory cues etc. because of time-compression, it is conceivable that the prosodic template of a word becomes more important for the recognition of the word. Stress information is spread over longer chunks of the speech signal and is thus more robust against time-compression than segmental information. Although stressed syllables and unstressed syllables differ in duration in the synthetic condition, as in natural speech, the natural variation in speaking effort due to different levels of stress is missing in synthetic diphone speech. Speaking effort translates into loudness, but also into articulatory precision. Variation in speaking effort (i.e. a speaking effort contour) may be an important suprasegmental characteristic of speech, which helps listeners to group weak and strong syllables together. This grouping together is essential for the recognition of polysyllabic words and for the ease of processing of syntactic chunks. So, although we had expected hyperarticulation to work out positively, the absence of variation in speaking effort turns out to be harmful to the holistic processing of e.g. polysyllabic words. Note that this explanation only holds if we assume that the phoneme detection responses reflect lexical rather than pre-lexical processing. When the targets are embedded in meaningful sentences, as was the case in the present experiment, responses are in fact more likely to depend on the lexical route [7]. Therefore, we tentatively assume that the present results can be taken to indicate speed of word processing.

Cutler & Norris [10] demonstrated that speakers of English segment speech input at the onset of strong syllables in the absence of explicitly marked cues to word boundaries. Young et al. [11] tested whether time-compressed sentences in which all content words began with strong syllables proved easier to recognise than time-compressed sentences in which all content words began with weak syllables. No difference was found between the two metrical conditions, but the idea that listeners are highly sensitive to metrical stress under difficult listening conditions is quite plausible.

4. CONCLUSION

Phoneme detection speed tends to be affected by playback rate: the shorter durations of words and syllables lead to somewhat faster detection times. This suggests that the increase in processing difficulty at this moderately fast playback rate is not substantial. For polysyllabic words, the processing advantage of natural speech over synthetic speech increased when speech was played back faster than normal. These results do not prove our hypothesis that the hyperarticulation of diphone speech turns into an advantage, but rather indicate that synthetic speech has certain characteristics which may become all the more harmful at faster playback rates. From a segmental viewpoint, equal stress on all syllables might enhance intelligibility. However, the lack of speaking effort fluctuation, as an important suprasegmental characteristic of words and phrases, may become more problematic at a fast playback rate.

5. REFERENCES