Perception of Natural and Synthetic Speech After Time Compression

Abstract
Even though synthetic speech is generally perfectly intelligible when presented at a normal speech rate, the intelligibility of synthetic speech suffers more than that of natural speech when both types of speech are time-compressed. However, since some studies suggest that perception of fast speech is helped by segmental redundancy, the hyperarticulation often found in synthetic speech might turn into an advantage at a fast rate. If the segmental redundancy of hyperarticulated diphone speech, consisting only of hyperarticulated stressed building blocks, is helpful, then the processing advantage of natural over synthetic speech might decrease after artificial time compression. A second expectation was that processing time-compressed speech was expected to put a higher processing load on listeners than normal-rate speech. A phoneme detection experiment was set up to test processing speed of normal-rate and time-compressed natural and synthetic speech. The results showed that the processing advantage of natural over synthetic speech did not decrease, but rather tended to increase. Although the fact that all syllables in diphone speech are stressed and neatly articulated might help phonetic processing, the alternation of weak and strong syllables helps listeners to group syllables together and to start lexical access. Synthetic speech which, for the most part, lacks such speaking effort fluctuation becomes rather blurred at faster playback rates, which in turn hinders processing of, particularly, polysyllabic words. Phoneme detection times, which were assumed to be an indication of processing load, were faster in the time-compressed conditions, contrary to our expectation. This suggests that listeners adjust their response deadline to the input rate.

3.1 Introduction

In the previous chapter, segmental intelligibility, as measured in nonwords, was shown to be severely affected by time compression. The recognition scores of real words, on the other hand, are relatively high at these very fast rates. Lexical redundancy in real words helps listeners to fill in the ‘difficult’ segments. Thus, listeners have two sources of information when listening to real words: segmental intelligibility and lexical redundancy. In this chapter, segmental intelligibility is looked at from a different angle, namely by comparing the intelligibility of natural versus synthetic speech. Natural and synthetic speech differ in ease of processing at normal rates, and the central question in this chapter is how and whether the difference in processing speed between the two speech types changes under the influence of increased playback rate. The study described in this chapter focusses on low-level ease of processing of both types of speech, rather than on differences in perceived quality or naturalness between natural and synthetic speech.

Segmental intelligibility seems to be important for the perception of fast speech. In the study described in Chapter 4, speakers were asked to produce sentences at normal and very fast rates (cf. section 4.2). In a pilot study (presented in section 5.2.2), the intelligibility of this naturally produced very fast speech was compared with speech which was originally spoken at a normal rate, and then time-compressed to that same fast rate. The intelligibility of the time-compressed speech appeared to be much higher (90% correct word identification) than that of the fast articulated speech (64% correct identification). The slurring, coarticulation and assimilation processes that inevitably accompany a very fast speech rate obviously do not contribute to the intelligibility of speech, even though listeners might expect these processes to occur at such a fast rate.

Quené & Krull (1999) used a word detection task to investigate whether word recognition is speeded up by assimilation or is hampered by it. The type of assimilation was deletion of /t/ between consonants, as in Dutch /pos/t/ /brongen/ ‘mail deliver’. Various studies (Gaskell & Marslen-Wilson 1996, 1998; Marslen-Wilson, Nix & Gaskell 1995) had suggested that, at normal speech rates, there is no perceptual advantage for assimilated over unassimilated articulations of a word form, given the appropriate phonological context. Quené & Krull (1999) thought that this could have been due to the rate and style of the experimental material in those three studies. If rate and style were intermediate, it could have been truly optional whether assimilation occurred or not. In that situation, listeners may not be biased towards either the assimilated, or the unassimilated form. Given a faster speech rate, however, listeners should expect certain assimilation processes to occur. Therefore, Quené & Krull (1999) argued, if listeners
expect assimilation, as in fast speech, then assimilated forms should be recognised faster than unassimilated forms. The results of the word detection study were rather surprising: whereas people detected the assimilated form of the word *post* faster than the unassimilated form at normal speech rate, the reverse was found for fast speech rate. Even though the unassimilated form was rather unnatural given the fast speech rate, listeners were faster in recognising it.

Thus, assimilation and coarticulation seem to deteriorate the intelligibility of fast speech. Word recognition and intelligibility in fast or time-compressed speech are helped by segmental redundancy, even if that segmental redundancy is artificially high. How does this relate to the idea that speakers take into consideration the needs of the listeners? Normally, speakers would only speed up if they assume that the listener can handle this loss of information. In our laboratory situations, speakers are asked to speed up their global speech rate, regardless of the communicative situation. Hence, they may be forced to neglect the needs of the listener, for articulatory reasons.

In this chapter, the question is raised whether the segmental redundancy, or hyperarticulation, present in synthetic diphone speech could be turned into an advantage, when the perception of time-compressed natural speech is compared with that of time-compressed synthetic speech. One of the most widely used speech synthesis systems for Dutch is Fluent Dutch (a commercial product of Fluency), which is based on diphone concatenation. The diphones of the Fluent Dutch diphone database are all cut from neatly articulated, stressed nonsense syllables. Concatenated strings of diphones are therefore, in a sense, segmentally maximally redundant: all syllables are originally stressed and carefully articulated. Since unstressed syllables in synthetic speech are given a short duration, as in natural speech, the hyperarticulation is all the more overdone. This hyperarticulation may sound rather unnatural for speech presented at a normal rate in good listening conditions. Still, although unnatural, segmental redundancy may be helpful for perception in difficult listening situations. At faster playback rates, the unnaturalness of hyperarticulated speech may be outweighed by its increased segmental redundancy (Quené & Krull 1999). Under difficult listening situations, perception might be helped by this type of overspecification. Even though natural speech is expected to have a processing advantage over synthetic speech, this advantage is expected to decrease when the playback rate is increased. Thus, we are not so much interested in the absolute difference in processing time between natural and synthetic speech, but rather in how the difference between the two changes when playback rate is increased. This means that the difference in perception between the two speech conditions at a normal rate must be evaluated against the difference between the two after time compression. As a consequence, intelligibility is not suitable for measuring differences between the two conditions, as both speech conditions are...
perfectly intelligible at a normal rate. Phoneme detection speed has been shown to differ for natural and synthetic speech conditions when both conditions are perfectly intelligible. Phoneme detection thus provides a useful and sensitive measure of low-level acoustic/phonetic processing difficulty (Nix, Mehta, Dye & Cutler 1993; Pisoni 1997). Research on the intelligibility of time-compressed speech has also shown that speech time-compressed to about 1.5 times normal rate is still almost perfectly intelligible (cf. Chapter 1). Hence, phoneme detection can be used to evaluate the difference in processing speed between natural and synthetic speech both at a normal rate, and after moderate time compression.

The prediction is that after time compression, the processing advantage of natural over synthetic diphone speech is smaller than at a normal rate, because processing of time-compressed synthetic speech is helped by its greater segmental redundancy. Furthermore, the fact that all diphones have been cut from stressed syllables might mean that the advantage over natural speech comes out most clearly for polysyllabic words. Polysyllabic synthetic words are more redundant than polysyllabic natural words, as each syllable stems from an originally stressed syllable. Attempts have been made to improve the perceived naturalness of synthetic speech by recording both stressed and unstressed diphones (Drullman & Collier 1991), or by controlling articulation effort (Wouters & Macon 2002a, 2002b). Wouters & Macon’s acoustic analyses of natural speech showed that spectral rate of change of vowel transitions increases with linguistic prominence (Wouters & Macon 2002a). The spectral rate of change can be predicted on the basis of the prosodic structure of the utterance. They describe an approach to integrate this knowledge into a concatenative speech synthesis system. Their results show that controlling the degree of articulation improves the perceived naturalness of speech (Wouters & Macon 2002b). Conversely, the use of unstressed diphones for unstressed syllables did not systematically result in more natural-sounding speech than when temporally reduced - stressed diphones were used (Drullman & Collier 1991). The present study investigates whether the unnatural aspect of having equal stress (spectrally, but not temporally) on all segments might be turned into an advantage.

A second prediction is that increased playback rate makes perception more difficult. Even though speech is still perfectly intelligible when speech is accelerated about 1.5 times, time compression is expected to put a higher processing load on the listeners. This higher processing load may either be the result of the increased rate of information, or of the reduced segmental intelligibility of time-compressed speech (witnessed by the lower identification percentages in nonwords than in real words, cf. Chapter 1), or indeed of both. This higher processing load is expected to translate into slower detection times in the fast condition than in the normal rate condition.
The two hypotheses are repeated below:

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

To put these hypotheses to the test, a phoneme detection study was set up. This study is presented in section 3.3. Section 3.2 contains an experiment which was set up to compare the intelligibility of heavily time-compressed natural speech and synthetic speech. This intelligibility test was used as a first naïve attempt to find out whether synthetic speech has a higher intelligibility than natural speech after severe time compression. In section 3.4 the results of the pilot experiment and of the phoneme detection study will be discussed.

### 3.2 Experiment 1: Intelligibility of time-compressed natural and synthetic speech

A pilot test was set up to compare the intelligibility of heavily time-compressed natural speech and synthetic speech. This pilot test can only indicate the difference in intelligibility between the two speech types after rather severe time compression. The hypothesis is that the intelligibility of synthetic speech is higher than that of natural speech because the high segmental redundancy of synthetic speech turns into an advantage when this type of speech is time-compressed.

#### 3.2.1 Synthesis system

Speech synthesis must find a way to model phonetic transitions as well as the more stationary parts of speech. There are two major families of speech synthesis systems: rule-based systems and concatenation-based systems. In rule-based systems, the phonetic transitions and stationary parts of speech are modelled explicitly, in the form of rules that describe the influence of phonemes on one another. MITALK is an example of such a rule-based system (Allen, Hunnicutt & Klatt 1987). Rule-based

---

5 This intelligibility test was carried out by two undergraduate students in phonetics, Fiona Sely and Eva Sittig, as part of a practical course.
synthesisers are always formant synthesisers which describe speech in terms of up to 60 parameters, mostly related to formant and antiformant frequencies and bandwidths (Klatt 1980). Exact modelling of the parameters is very time-consuming and error-prone, and achieving a high degree of naturalness is problematic (Dutoit 1997). Whereas rule-based synthesis requires much explicit knowledge, this type of phonetic knowledge is implicitly embedded in the stored segments in concatenation synthesis. Concatenation-based synthesis uses pre-recorded tokens of phonetic transitions and coarticulations into a database. The MBROLA system (Multi-Band Resynthesis Overlap-Add) is based on the concatenation of diphones (cf. section 1.2.2 on PSOLA and Dutoit (1997)). 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 standard Dutch diphone database stores over 2300 of such transition. The position of the boundary 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 amplitude mismatches at concatenation, the energy levels at the beginnings and at the end of segments are modified during an equalisation stage. This equalisation stage entails setting the energy of all phones of a given phoneme to their average value before storage (Dutoit 1997). The spectral envelopes, pitch, and phase of concatenated segments must somehow be adapted to one another to avoid audible discontinuities. The solution to this consists of resynthesising the original speech segments in the database. This is performed by resynthesising the voiced parts of all segments with constant synthesis pitch and fixed initial phases for each period, as performed by the MultiBand Resynthesis-PSOLA (MBR-PSOLA) algorithm (Dutoit & Leich 1993).

Spectral smoothing, or the attenuation of spectral mismatch, is solved only 'at runtime', by linear interpolation in the time-domain. Thus, the naturally introduced coarticulation is still maintained (Dutoit 1997). The MBR-PSOLA technique was turned into the more efficient MBROLA technique. 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 text file, containing a list of phonemes in SAMPA transcription (a machine-readable phonetic alphabet), together with prosodic information (phoneme durations and a piecewise linear description of pitch).
3.2.2 Material

Forty sentence pairs were constructed for the present pilot experiment. The intelligibility test was set up such that the listeners had to fill in the missing word in a sentence. This word should therefore not be predictable from the sentence context. The sentences were selected randomly from a number of books, but were modified somewhat if necessary. Short sentence fragments were cut out of these sentences: these fragments were to be presented to the listeners. Two examples of sentence fragments are presented in (1) below (target words are in bold).

(1) Vorige week had de juwelier een grap gemaakt over dwergen7
Ik hoorde Robert zeggen dat hij nog een goede kaart zocht8

The target nouns contained one to three syllables and occurred in different positions in the sentence. The 40 sentences were read aloud by the same reference male speaker (Arthur Dirksen) whose diphones were used as the standard Dutch diphone database (NL2) for the MBROLA synthesiser (Dutoit 1997; Dutoit, Pagel, Pierret, Bataille & van der Vreeken 1996).

The natural speech material, as produced by the speaker, was recorded on DAT tape in a sound-treated booth with a Sennheiser ME30 microphone. The material was then fed as digital input to a computer, downsampled to 16 kHz, and then the sentence fragments were selected. These fragments were then segmented by hand. A close diphone copy was made based on the SAMPA transcription of the fragments, adjusted to the segment durations and pitch contours found in the natural sentences. All phoneme durations were exactly equal to those measured in the natural speech. The $F_0$ contour was a rough piece-wise close-copy version (time – log$F_0$ domain) of the $F_0$ contour found in the natural version. In this way the natural and synthetic conditions could be compared within a single speaker. Mean intensities of the synthetic and natural version of each sentence were made equal.

In Figure 3.1 below, a segmented waveform (plus SAMPA labels) of a natural sentence is shown, together with its close synthetic copy.

---

7 ‘Last week the jeweller had made a joke about dwarfs’
8 ‘I heard Robert say he was still looking for a good map’
3.2.3 Design and Procedure

There were two experimental conditions (natural vs. diphone). The 40 items were rotated over the 2 conditions, yielding 20 items per condition. The design was a within-subject design and the two conditions were balanced (Latin square) over two experimental lists. Items were balanced over the two test conditions, in such a way that condition was not confounded with the monosyllabic/polysyllabic distinction.

First, by way of pre-tests, the degree of time compression was determined: intelligibility should not be too high to avoid ceiling effects and it should not be too low to avoid floor effects. After compression to 50% of the original duration,
intelligibility was still almost 100%. Therefore, a compression rate to 30% of the original duration was chosen. Time compression was carried out by means of PSOLA.

During the test session, subjects were seated in sound-treated booths, wearing closed-ear headphones. They were first presented with the sentence on a computer screen from which the target word was missing. Then the whole time-compressed sentence was presented to them over the headphones, including the target word. They were then asked to fill in the missing word. There was no time-pressure: only after they had hit the Enter key, would the following sentence appear on the screen.

3.2.4 Subjects

To each list, 18 subjects were assigned. The 36 subjects were all students at Utrecht University.

3.2.5 Results

A higher intelligibility score was expected for the synthetic speech than for the natural speech because of the higher segmental redundancy in synthetic speech. The overall raw correct recognition percentages (at compression to 30%) are shown in Table 3.1.

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>Monosyllabic</th>
<th>Polysyllabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural version</td>
<td>48%</td>
<td>51%</td>
<td>46%</td>
</tr>
<tr>
<td>Synthetic version</td>
<td>32%</td>
<td>34%</td>
<td>31%</td>
</tr>
</tbody>
</table>

The recognition percentages were computed for each item and for each subject in both conditions and (after arcsine transformation) were fed into ANOVAs in which either subjects or items were treated as repeated measures. First, the results do not support the hypothesis: the intelligibility of synthetic speech is lower than that of natural speech. The effect of Speech Type was significant in both the item and the subject analysis ($F_1(1,35)=36.5$, $p<0.001$; $F_2(1,39)=15.2$, $p<0.001$). Secondly, this lower intelligibility of synthetic speech holds for both monosyllabic and polysyllabic words. Although there were more polysyllabic words ($N=28$) than monosyllabic words ($N=12$) in the material, the effect of Syllable Number and the interaction between Syllable Number and Speech Type were analysed statistically in separate Repeated Measures ANOVAs. The effect of Syllable Number (monosyllabic vs. polysyllabic items) was not significant ($F_1(1,35)=4.31$, $p=.045$; $F_2(1,38)<1$, n.s.), and neither was the interaction between Speech Type and Syllable Number ($F_1(1,35)<1$, n.s.; $F_2(1,38)<1$, n.s.). Thus, our data
do not show different patterns for monosyllabic and polysyllabic words with respect to the intelligibility difference between the two types of speech.

A third observation is that the intelligibility of the polysyllabic words is overall lower than that of the monosyllabic words. Although this effect is not significant, it is rather surprising, because longer polysyllabic words are more redundant than shorter words. Another experiment (not described here) excluded the possibility that the monosyllabic words were more predictable from the sentence context. The very fast speech rate may have caused unstressed syllables to become too short to be perceived. Some of the unstressed syllables may have been extra vulnerable because of their segmental content. The segmental intelligibility results presented in Chapter 2 showed how some segments suffer more from time compression than others, in particular when they are part of the unstressed syllable. Thus, at this rate of speech, the higher lexical redundancy of disyllabic words is outweighed by unstressed syllables becoming perceptually obliterated, even though they were originally hyperarticulated.

Now that synthetic speech turns out to have a lower intelligibility than natural speech after severe time compression, the processing difference between the two speech types at a normal rate should also be established. Only if the processing differences at both normal rate and fast rate are known, do we know whether time compression does make the difference between the two smaller. Intelligibility at a normal rate is too high to find any differences in intelligibility by means of an intelligibility test, so a different type of test must be used to evaluate the differences between natural and synthetic speech at a normal rate and at a fast rate. For this purpose the phoneme detection task was selected, which has been shown to be a useful tool to compare the perception of highly intelligible speech types. The next section reports on a phoneme detection experiment, set up to test whether and how the processing advantage of natural over synthetic speech is affected by an increase in playback rate.

### 3.3 Experiment 2: Processing speed

The results of experiment 1 showed that, after heavy time compression, the intelligibility of synthetic speech is lower than that of natural speech. This means that the hypothesis for the intelligibility pilot was rather naïve: diphone speech does not only differ from natural speech in a positive, i.e., hyperarticulation sense. Diphone speech may be rich in acoustic cues, but it is also rich in false or misleading acoustic cues. Diphone /pe/ for the word pen may have been cut from the syllable pet and thus
still contains some cues for a coda /t/. Furthermore, it lacks cues for the actual nasal coda /n/, which may be equally disruptive for speech perception. Diphones can only account partially for the coarticulatory effects in speech because these often affect a whole segment rather than only its first or second half independently (O'Shaugnessy 1990). Concatenation of diphones also yields spectral discontinuities at the diphone edges. Although spectral smoothing is applied to make these discontinuities less audible, the signal is not as smooth as natural speech. Therefore, the processing difference at a normal rate should also be established. These negative aspects of synthetic diphone speech were expected to be equally harmful at a normal and at a fast playback rate. The hypothesis is not that synthetic speech is more intelligible than natural speech after time compression, but that the processing advantage of natural speech over synthetic speech decreases after time compression.

In Pisoni (1987; 1997) and Nix et al. (1993), reaction time measures are described as a tool for comparing perception of natural and synthetic speech. Nix et al. (1993) found longer response times for synthetic speech, relative to natural 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. The difficulty of listening to synthetic speech has been argued to occur mainly at the lower phonetic level, and not at higher prosodic levels. Although improvements towards more appropriate and more natural prosody are certainly preferred by listeners (Terken & Lemeer 1988), the perceptual disadvantages at the lower phonetic level are assumed to demand the greater part of the extra processing capacity. Pisoni (1997) mentions several experiments with natural and synthetic speech. In an auditory lexical decision task subjects responded significantly faster to natural words and nonwords than to synthetic words and nonwords. The differences in response time between words and nonwords were equal for natural and synthetic speech. Thus, the “extra processing effort appears to be related to the initial analysis and perceptual encoding of the acoustic-phonetic information, and not to the process of accessing words from the lexicon” (Pisoni 1997: p.550). Similar results were obtained in a naming task using natural and synthetic words and nonwords. In this experiment subjects were asked to repeat the stimulus words which were presented to them auditorily. So, these two experiments suggest that early stages of perceptual encoding require more time for synthetic speech than for natural speech. Reaction time measures are assumed to give a better insight into processing difficulty than other measures taken after processing is complete (Levelt 1978).

Pisoni (1997) argues that if initial acoustic/phonetic analysis is slowed down, both pre-lexical processing, and consequently, lexical processing are slowed down as well. It is important to keep in mind that speech processing may work in this serial fashion, but phoneme detection responses can still be based on pre-lexical representations. Cutler &
Norris (1979) argued that phoneme detection can either be the result of a target detection procedure carried out on the pre-lexical representation or on the basis of phoneme information associated with a lexical representation. These two procedures run in parallel, and whichever is the fastest, wins the race. If the target is detected on the basis of pre-lexical information before lexical access is completed, the pre-lexical route wins. If lexical access is achieved before the target can be detected via the pre-lexical representation, the lexical route wins and, consequently, the response is based on the lexical representation. In the more recent Merge model (Norris, McQueen & Cutler 2000) phonemic decisions are argued to be based on the merging of pre-lexical and lexical information.

Note that the experiments of Pisoni (1997) and Nix et al. (1993) involved speech synthesis by rule. In the present study, diphone synthesis is used. This type of synthesis depends on the concatenation of naturally produced units. Nusbaum, Dedina & Pisoni (1984) argued that phonetic information is often redundantly specified in natural speech, but that in synthetic speech the cues are sparser, which implies harder work for the phonetic processor. Diphone speech is richer in phonetic cues than rule-based synthetic speech because of the inherent natural phoneme-to-phoneme transitions. In the present study the prosodic pattern (i.e., intonation pattern and durations) of the natural utterance is applied to its synthetic counterpart. This should help prosodic processing. In the Discussion section, the question at which level processing difficulties occur will be taken up again.

Phoneme detection can be used to evaluate the processing advantage of natural speech over synthetic diphone speech at a normal rate. The difference in response time can be investigated for highly intelligible time-compressed speech as well. By comparing the processing difference between the two types of speech at normal and fast rates, the main hypothesis can be tested that the processing advantage of natural over synthetic speech decreases with increasing rate. A second expectation is that time-compressed speech will elicit slower phoneme detection times than normal rate speech. Although speech 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 was expected to translate into slower phoneme detection times.

3.3.1 Method

Material

The material that was used in the two pilot tests did not contain enough suitable target phonemes. Those sentences formed only a small sub-set of the recording from which they had been taken. Therefore, a new sample of 100 sentences with suitable target
phonemes was selected from the recording of the same male speaker whose diphones
are used as the standard MBROLA diphone set for Dutch. These sentences had all
been taken from books, and had been modified slightly, if necessary. 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: there were many /b/s and only few /d/s in the material. The
target word could be either a noun, a verb or an adverb. Target words were
monosyllabic (32) or polysyllabic (68). If possible, the target phoneme did not occur
elsewhere in the sentence, either word-initially, word-medially or word-finally. Three
target words are presented in (2) below (target word in bold):

(2)   target /p/ De pater en de non ruimden gezamenlijk de tafel op het terras af9
      target /d/ Alleen één ding is in elke auto onzeker10
      target /b/ Het mooie servies van oma was weer bijna compleet11

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 the same Dutch diphone set was used as in the experiment
described above (NL2). Grapheme-to-phoneme conversion was done automatically,
but was corrected manually if necessary. Fluent Dutch also computes a natural pitch
contour and suitable durations. By default, the program assigns sentence accent to all
content words, including 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 target word
was made just as long as 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. This was done by means of
PSOLA time scaling as implemented in the speech editing program GIPOS. Note that,
unlike in the previous pilot tests, the natural sentences were not segmented manually
phoneme-by-phoneme. Only the durations of the target words, and the parts preceding
and following the target words were made equal in duration. 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 test sentences were
time-compressed linearly to 65% of their original duration by PSOLA time scaling. This

9 ‘The priest and the nun together cleared the table on the terrace’
10 ‘Only one thing is in each car uncertain’
11 ‘Grandmother’s beautiful crockery set was almost complete again’.
is about the fastest speech rate speakers can attain if they try very hard (cf. Chapter 4). Speech that is time-compressed to this rate is almost perfectly intelligible. The 100 test sentences were rotated over the 4 conditions (Natural-normal rate, Synthetic-normal rate, Natural-fast, Synthetic-fast), using a Latin square design. Because each subject could be presented with each item only once, there were 4 experimental lists. Apart from the 100 test sentences, there were 10 practice sentences, 10 warming-up sentences and 70 filler sentences. The filler sentences did not contain the phoneme the subjects were asked to detect, and were included in order to prevent subjects from pressing the button randomly. The fillers were rotated over the 4 test conditions and interspersed with the material.

Subjects
Ten subjects were assigned to each of the 4 experimental lists. The 40 subjects were all students at Utrecht University, and were paid a small amount of money for their participation. None of them reported any hearing or reading problems.

Procedure
Subjects were seated in a sound-treated booth and were tested individually. The speech material was presented to them over closed headphones. They first read instructions on the computer screen in front of them 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 spelling of all test words was regular: if subjects were asked to monitor for /k/, there were no target words which are spelled with ‘c’. The onset of the target plosives was marked in the speech waveform by means of a time marker. The program then computed the phoneme detection time by measuring from that marker point in time to the moment that the button press was registered.

After the practice session, subjects could still ask questions if anything was unclear. Before subjects started with the actual test, 10 warming-up sentences were played to them after which they proceeded seamlessly with the actual test. All test and filler items were presented in random order. The experiment lasted 20 minutes.

3.3.2 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 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 observations. Missing observations were due to subjects missing the phoneme, or responding too early (i.e., to another phoneme). The raw detection times, plus the miss rates in each condition, are shown in Table 3.3.

<table>
<thead>
<tr>
<th></th>
<th>Normal rate</th>
<th>Time-compressed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>s.e.</td>
</tr>
<tr>
<td>Natural speech</td>
<td>598</td>
<td>8</td>
</tr>
<tr>
<td>Synthetic speech</td>
<td>677</td>
<td>11</td>
</tr>
<tr>
<td>Difference</td>
<td>-79</td>
<td>-71</td>
</tr>
</tbody>
</table>

The difference in detection time between synthetic and natural speech is quite large: 79 ms at normal rate. This difference is somewhat smaller after time compression. As a first (quick and dirty) analysis, all missing observations were replaced by the grand mean of 627 ms\(^{12}\). These results were entered into analyses of variance on items and on subjects (Repeated Measures) to test the effects of Speech Type and Rate. The effect of Speech Type was highly significant ($F_1(1,39)=77.0$, $p<0.001$; $F_2(1,99)=8.92$, $p=0.004$). This means that natural speech is easier to process than synthetic speech. A second hypothesis is that phoneme detection times are slowed down by time compression because of the higher processing load. The effect of Rate approached significance in the subject analysis, but was insignificant by items ($F_1(1,39)=4.02$, $p=0.052$; $F_2(1,99)<1$, n.s.). Thus, these data do not provide evidence for the idea that time compression makes speech more difficult to process. The main hypothesis, however, was that the processing advantage of natural speech over synthetic speech would decrease after both speech types had been time-compressed. The interaction between Speech Type and Rate was far from significant ($F_1(1,39)<1$, n.s.; $F_2(1,99)=1.1$, n.s.).

The miss rates (cf. Table 3.3) were also analysed to establish the effects of Speech Type and Rate on the number of missing observations. For both speech types, the miss rate increases as a result of time compression. The miss rates (after arcsine transformation) in all four conditions were analysed in ANOVAs treating either items or subjects as repeated measures. These analyses showed significant effects of Speech

---

\(^{12}\) Repeated Measures ANOVAs in SPSS cannot cope with missing values. There are more sophisticated ways to replace missing values (Girden 1992). Particularly when missing values are not distributed equally over the conditions, replacing them by the grand mean is not very elegant.
Type ($F_1(1,39)=43.7$, $p<0.001$; $F_2(1,99)=35.1$, $p<0.001$), and of Rate ($F_1(1,39)=16.0$, $p<0.001$; $F_2(1,99)=10.2$, $p=0.002$). The interaction between Speech Type and Rate was not significant in either analysis ($F_1(1,39)=1.41$, n.s; $F_2(1,99)=2.27$, n.s.).

In a second analysis, several items were left out of the analysis. Although subjects thought that intelligibility of the speech material was generally very high, those items that elicited many missing values may have been lower in intelligibility than the rest. There should be 10 observations for each item in each condition (10 subjects on each list). If the number of valid observations (excluding missing data) was lower than 7 out of 10, the item was left out of the analysis. This was the case for 15 out of 100 items. In order to obtain equal numbers of observations for all conditions on each list, 9 more items had to be left out of the analysis. These 9 items were chosen at random. Thus, in total, 24 items were left out of the analysis. Overall, the percentage of missing observations in the remaining subset of 76 items was 3%. The mean raw detection times for this subset are shown in Table 3.4.

<table>
<thead>
<tr>
<th></th>
<th>Normal rate</th>
<th>Time-compressed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>s.e.</td>
</tr>
<tr>
<td>Natural speech</td>
<td>598</td>
<td>10</td>
</tr>
<tr>
<td>Synthetic speech</td>
<td>654</td>
<td>11</td>
</tr>
<tr>
<td>Difference natural-synthetic</td>
<td>-56</td>
<td>-90</td>
</tr>
</tbody>
</table>

Because the missing observations are not distributed equally across the different conditions, the missing observations were replaced by the subject's mean for that condition in the subject analysis, and by the item mean in that condition in the item analysis (Girden 1992). Furthermore, after that, the reaction time data were transformed for the following reason. Reaction time data are not distributed in a normal (or Gaussian) way. Analyses of variances actually assume normally distributed data. In order to make the distributions more normal, the reaction time data were transformed to inverse reaction times (1/RT). A Kolmogorov-Smirnov test showed that even these transformed data were not entirely normally distributed ($z=5.63$, $p<0.001$). The statistical analyses were run again on the transformed data. As in the previous analysis, there was a significant main effect of Speech Type ($F_1(1,39)=60.9$, $p<0.001$; $F_2(1,75)=23.0$, $p<0.001$), but now, the effect of Rate was also significant ($F_1(1,39)=8.7$, $p=0.005$; $F_2(1,75)=6.2$, $p=0.015$). Note that this Rate effect is in the opposite direction from what was expected: generally, detection times turn out to be faster in the time-compressed condition than in the normal rate condition. The
interaction between Speech Type and Rate was significant as well ($F_1(1,39)=5.0$, $p=0.032$; $F_2(1,75)=4.8$, $p=0.031$).

Thus, it is clear that there is, at least, a tendency towards an increase of the processing advantage of natural speech over synthetic speech when both types of speech are time-compressed.

The miss rates (after arcsine transformation) in all four conditions in the selected subset (cf. Table 3.4) were analysed in ANOVAs treating either items or subjects as repeated measures. The analyses showed significant effects of Speech Type ($F_1(1,39)=38.2$, $p<0.001$; $F_2(1,75)=21.8$, $p<0.001$) and Rate ($F_1(1,39)=11.8$, $p=0.001$; $F_2(1,75)=7.06$, $p=0.01$). The interaction between Speech Type and Rate was not significant ($F_1(1,39)=2.77$, $p=0.104$; $F_2(1,75)=1.68$, $p=0.199$). So, also in this subset of items, miss rate increases after time compression.

Of the 76 remaining items, 25 items were monosyllabic and 51 were polysyllabic. Separate analyses were carried out for the monosyllabic and polysyllabic subsets. Note that each of the two subsets of data cannot be completely balanced over the 4 experimental conditions. The behaviour of the monosyllabic items differs considerably from that of the polysyllabic items. This is illustrated in Figure 3.2 below.

For the monosyllabic items, the processing advantage of natural over synthetic speech is substantial at both rates (122 ms at normal rate (529 natural vs. 651 synthetic); and 94 ms at fast rate (515 natural vs. 609 synthetic)). Univariate ANOVAs (which allows unequal numbers of observations over cells) of the inverse reaction time data ($1/RT$) showed a significant main effect of Speech Type ($F_1(1,39)=25.4$, $p<0.001$; $F_2(1,24)=10.6$, $p=0.003$). The effect of Rate was not significant ($F_1(1,39)<1$, n.s.; $F_2(1,24)<1$, n.s.), and neither was the interaction between Speech Type and Rate ($F_1(1,39)<1$, n.s.; $F_2(1,24)<1$, n.s.). For the polysyllabic items, the difference between the two speech conditions was 20 ms in the normal rate condition (633 natural vs. 653 synthetic), and 110 ms after time compression (575 natural vs. 685 ms. synthetic). For these items, the interaction between Speech Type and Rate was significant ($F_1(1,39)=6.24$, $p=0.017$; $F_2(1,50)=5.36$, $p=0.025$), and so were the main effects of Speech Type ($F_1(1,39)=36.7$, $p<0.001$; $F_2(1,50)=11.8$, $p=0.001$) and of Rate ($F_1(1,39)=13.4$, $p=0.001$; $F_2(1,50)=9.12$, $p=0.004$).
Three main observations can be made from these data. First, the detection times are not slowed down by increased rate, which is against our prediction. For both speech types, detection times are even somewhat faster in the time-compressed than in the normal rate condition. This effect is significant only when the least intelligible items are left out of consideration. In other words, items are processed faster when speech is time-compressed, but, at the same time, miss rates in the fast conditions are significantly higher than in the normal rate conditions. The faster detection times in the time-compressed conditions are mainly due to the natural (polysyllabic) items. The second observation is that there is (at least a trend towards) an interaction between Speech Type and Rate (significant in the ANOVAs on the subset). Processing is sped up by increased playback rate for the natural items, but remains relatively unaffected for the synthetic items. Contrary to the first hypothesis, the processing advantage of natural over synthetic speech does not decrease when both speech types are time-compressed, but rather tends to increase. The third observation is that this mainly goes for the polysyllabic items. Time compression has a different effect on polysyllabic natural items than on monosyllabic natural items.
3.3.3 Further analysis of phoneme detection data

The phoneme detection literature reports several effects which may have blurred the expected differences between monosyllabic and polysyllabic items in the data. Because the sentences had not been designed for the purpose of a phoneme detection experiment, these effects were not systematically controlled. If, accidentally, all monosyllabic items score differently on one of these variables than the polysyllabic items, this might have influenced the data. Five such variables, the same as investigated in Nix et al. (1993), will be discussed below in relation to the present data set.

1. Transition probability. Phoneme targets in contextually predictable words are detected faster than targets in unpredictable words (Dell & Newman 1980; Morton & Long 1976). By means of a paper-and-pen cloze test which was presented to 20 subjects, the predictability of the target items in our 100 sentences was established (following the procedure described in Nix et al. (1993)). For both the monosyllabic words and the polysyllabic words, we checked whether targets in predictable items were detected faster than targets in unpredictable items. This appeared not to be the case: in all four conditions, detection times were slower for the predictable items.

2. Preceding word length. If phoneme targets are preceded by long words, they are detected faster than when preceded by short words (Mehler, Segui & Carey 1978). The isolation or recognition point of longer words is often before the end of the word, whereas the isolation point of short words can be one or two syllables after the offset of the word (Grosjean 1985). However, in our material, targets which were preceded by monosyllabic words or by no word at all were detected faster than targets which were preceded by longer words.

3. Position of target bearing word in the sentence. The later the target bearing item occurs in the sentence, the faster the RTs are (Cutler & Fodor 1979; Foss 1969). This is caused by the context being more restrained towards the end of the sentence. As in Nix et al. (1993) detection times were compared for targets in syllable positions 1 to 5 (early) with detection times in later positions. There was some evidence for this factor in our data: collapsed over all four conditions, targets were detected 22 ms faster in late positions than in early positions. There was no indication of an interaction between this factor Position and Speech Type or Rate.

4. Sentence accent. Targets in words with sentence accent are detected faster than targets in unaccented words (Cutler 1976). There was no consistent effect of sentence accent on the detection times in the present study.

5. Lexical stress. Taft (1984) showed that initial phonemes were detected faster when they were part of the stressed syllable than when part of an unstressed syllable. There is evidence that this effect plays a role only in spontaneous and not in read speech (Mehta...
& Cutler 1988). For our polysyllabic items, we checked whether this was also the case in our data. Only in the natural-speech-normal-rate condition, were target items in initially stressed words detected faster (22 ms) than targets in non-initially stressed words.

Thus, of the five variables mentioned above, only the effect of position in the sentence appeared to play a role in our data. It is not clear how these five variables can shed any light on the results presented in Figure 3.2. What needs to be explained is not why there are differences between groups of items (i.e., monosyllabic vs. polysyllabic words), but rather how these groups are affected by an increase in rate. The key question remains why the difference between the natural and synthetic conditions remains stable for monosyllabic words, whereas the difference between the two speech conditions increases markedly for polysyllabic words. The variable ‘Position in the Sentence’ does not explain why this is the case. 13

3.4 Discussion

Despite the high intelligibility and quality of synthetic diphone speech, listeners still find natural speech easier to process than synthetic speech. Now that natural speech was found to be more intelligible than synthetic speech after equal rates of time compression (in experiment 1), it was investigated whether the processing advantage of natural speech is affected by an increase in playback rate. The hypothesis was that the hyperarticulation that is present in synthetic diphone speech, consisting only of neatly articulated stressed syllables, might become advantageous in difficult listening conditions.

A second hypothesis was that processing would be slowed down for time-compressed speech, relative to normal rate speech because of the higher processing load of fast speech. A phoneme detection study (experiment 2) was set up to investigate these two hypotheses.

Contrary to what we expected, phonemes were detected faster in time-compressed speech than in the normal rate condition. There are at least two possible explanations:

13 One of the remaining factors that might have influenced the detection times could be the intelligibility or processing difficulty of the preceding word. Although it is assumed that all sentences were perfectly intelligible, at least at sentence-level, it is possible that subjects experienced difficulty in processing certain pre-target words in certain conditions. The outlier data condition in Figure 3.2, the polysyllabic synthetic condition, may have suffered from ‘low-quality’ words preceding the target words. This can only be checked by way of another phoneme detection experiment.
for this. Listeners may adapt to the higher rate of information, and phonetic and lexical processing may not become more difficult when speech is time-compressed to this moderately fast rate. The shorter detection times in the fast condition could then be caused by syllable and word durations being shorter. A target phoneme can only be detected after the word or the syllable has been processed. When speech is time-compressed, the isolation point at which the word can be recognised is reached earlier. Thus, the contribution of the target’s duration to the total phoneme detection time is smaller after time compression. If this is the case, one would expect to find a significant correlation between detection time and length of the (monosyllabic) items in normal and fast condition, more specifically, between the difference in item duration and the difference in mean reaction time in the normal and fast condition. For all monosyllabic items, the difference in mean reaction time was computed between the normal and fast conditions, together with the difference in mean duration between the normal and fast condition. This correlation was not significant: neither when the results were collapsed for both speech types (Pearson correlation coefficient \( r=0.06, p=0.76 \)), nor when only the natural item results were analysed (\( r=0.26, p=0.22 \)). So, there is no linear correlation between the shorter duration of the monosyllabic words and the faster detection times.

An alternative explanation might be that listeners adjust their response deadline to the input rate. As a reaction to the fast rate of presentation, they adapt their processing rate. They succeed in reacting faster than normally, and thus in keeping up with the higher rate, but only at the expense of making more errors. This is witnessed by the higher miss rates in the time-compressed conditions, for both speech types.

Contrary to the hypothesis, the processing advantage of natural over synthetic speech did not become smaller after both speech types had been time-compressed. The phoneme detection study showed that the processing advantage of natural over synthetic speech was relatively large: phoneme detection time was 79 ms faster for natural speech than for synthetic speech when both speech types were presented at a normal rate. This supports the results of Nix et al. (1993) that phoneme detection can be used to test differences in speech communication quality. If the least intelligible items are left out, the processing advantage still amounts to 54 ms at the normal rate of presentation. This difference in quality must, at least partly, be attributed to phonetic processing difficulties. Pisoni (1997) and Nix et al. (1993) relate the slower phoneme detection times in synthetic speech to the extra processing effort in the initial analysis and perceptual encoding of the acoustic-phonetic information. If this initial pre-lexical analysis of the speech signal requires more processing time, word recognition will consequently be delayed as well, relative to natural speech. Importantly, the processing advantage of natural over synthetic speech did not decrease after time compression.
There was even a tendency towards the opposite: in particular for the polysyllabic items, the advantage of natural speech even increased at fast playback.

A first explanation concerns the initial assumption that the negative aspects of synthetic speech would be independent of compression rate. It is conceivable, however, that these negative aspects (i.e., misleading coarticulatory cues) do become more harmful to speech processing when speech is time-compressed. Misleading spectral cues make initial pre-lexical processing more difficult, so both the pre-lexical and lexical route are slowed down. Even if segmental redundancy has a positive effect in difficult listening conditions, this may have been outweighed by these negative aspects becoming more problematic.

A second possible explanation bears on the different patterns found for monosyllabic and polysyllabic items. On the basis of the hyperarticulation-is-helpful-in-adverse-listening-conditions hypothesis, one would have expected the processing advantage of natural speech to decrease mainly for polysyllabic items because polysyllabic items consist only of stressed-and-hyperarticulated syllables. Hyperarticulation of unstressed syllables should make initial low-level analysis easier, and, consequently, phoneme detection via the pre-lexical route should benefit from the higher segmental intelligibility. For the polysyllabic items presented at a normal rate, natural speech has no robust processing advantage over synthetic speech. For weak-strong words (non-initial stress), targets in synthetic speech items were even detected 16 ms faster than in natural items in the normal rate condition. This was not the case at the fast rate of presentation: in the fast condition, targets in natural versions are detected faster (84 ms) than in synthetic versions. For strong-weak polysyllabic items, targets in natural versions were detected faster than those in synthetic conditions, both at the normal rate (natural speech advantage is 49 ms), and at the fast rate (advantage is 85 ms).

With respect to the fast rate of presentation, note that some subjects complained that the time-compressed synthetic speech sounded blurred to them, and that they found it difficult to detect word boundaries. When segmental intelligibility is affected because of time compression, it is conceivable that the prosodic template of a word, consisting of the speaking effort and duration pattern, 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 speaking effort fluctuation due to different levels of stress is largely missing in synthetic speech. Speaking effort translates into loudness, but also into articularatory precision. The 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 these prosodic cues make lexical access easier, not initial pre-lexical analysis. If we assume that the lexical route contributes most to the ultimate phoneme decision response, the lack of proper stress information should slow down the lexical route.

There is some further evidence that the lexical route contributes more to the phoneme detection results than the pre-lexical route. Phoneme detection times are slower overall for polysyllabic words than for monosyllabic words. This holds for natural speech at both rates, but for the synthetic speech only at the fast rate. At a normal rate, detection times of synthetic monosyllabic words are about equal to those of synthetic polysyllabic words. It is also important to note that, for the natural speech, the difference between the average detection time of monosyllabic and that of polysyllabic items decreases at faster playback rate: the difference between monosyllabic and polysyllabic is 114 ms at the normal rate, and 60 ms at the fast rate. This agrees with the fact that the difference in duration between monosyllabic and polysyllabic words after time compression is only 65% of what it was at the normal rate. Note that these results are in conflict with the results of Dupoux & Mehler (1990) who found that even after time compression, phoneme detection does not necessarily depend on the lexical code. However, Dupoux & Mehler’s phoneme detection study (1990) was based on the presentation of single target words, which resulted in much faster detection times (about 430 ms) than we found in our present study (about 580 ms for the time-compressed natural speech). The design of their experiment may have caused subjects to rely more on the pre-lexical code than on the lexical code. When the targets are embedded in meaningful sentences, as was the case in the present experiment, the information from the lexical, rather than the pre-lexical route weighs more heavily (Cutler & Norris 1979).

The speaking-effort account for the increasing processing advantage of natural over synthetic polysyllabic words agrees with the metrical segmentation study of Cutler & Norris (1988), and the subsequent study by Young, Altmann, Cutler & Norris (1993). In Young et al. (1993) the question is raised whether speech is easier to recognise under difficult listening conditions when all strong syllables are word-initial. This question was based on Cutler & Norris (1988), who 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. (1993) 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 attractive. In the sense that speech recognition is pattern recognition, the speaking effort contour may be an important suprasegmental characteristic of polysyllabic words. Contrary to the hypothesis, hyperarticulation of diphone speech turns out to be harmful in difficult listening conditions.

Others have shown that there are ways of increasing intelligibility over normal intelligibility. Intelligibility of nonwords in noise can be improved by cue-enhancement of certain consonantal regions of rapid spectral change (Hazan & Simpson 1998). The tentative conclusion, however, is that the type of hyperarticulation present in diphone speech, i.e., having equal stress on all syllables, turns out to be harmful in adverse listening situations. For the recognition of polysyllabic words, a natural speaking effort contour is at least equally important.

3.5 Conclusion

Three conclusions can be drawn from our data. First, time-compressed speech is more difficult to process than speech presented at a normal rate, but this does not translate into slower detection times. Subjects adapt their response time in order to keep up with the higher input rate, but at the expense of making more errors.

Secondly, synthetic speech is more difficult to process than natural speech, both at a normal and at a fast rate. This is witnessed by lower scores in the intelligibility test, a higher miss rate in the phoneme detection experiment, and longer phoneme detection times. Misleading coarticulatory cues, consequent spectral discontinuities, and possibly, the lack of a natural prosodic pattern all contribute to this processing difficulty.

Thirdly, the data did not support the expectation that hyperarticulation that is found in synthetic speech is helpful at a faster playback rate. From a segmental intelligibility viewpoint, equal stress on all syllables might enhance intelligibility. However, the lack of speaking effort fluctuation, as an important suprasegmental characteristic of polysyllabic words, becomes more problematic for word recognition at a fast rate.

The results of this chapter show that segmental and suprasegmental factors both influence lexical processing. The question of how an increase in speech rate affects segmental and prosodic characteristics in natural speech plays an important role in the next chapter. The key question is again how segmental and prosodic factors contribute to speech intelligibility of artificially time-compressed speech.