Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®
Speech Communication 42 (2004) 155–173
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

# Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech

Esther Janse

*Utrecht institute of Linguistics OTS, Utrecht University, Trans 10, 3512 JK Utrecht, The Netherlands*

Received 25 July 2002; received in revised form 1 March 2003; accepted 29 July 2003

**Abstract**

Natural fast speech differs from normal-rate speech with respect to its temporal pattern. Previous results showed that word intelligibility of heavily artificially time-compressed speech could not be improved by making its temporal pattern more similar to that of natural fast speech. This might have been due to the extrapolation of timing rules for natural fast speech to rates that are much faster than can be attained by human speakers. The present study investigates whether, at a speech rate that human speakers can attain, artificially time-compressed speech is easier to process if its timing pattern is similar to that of naturally produced fast speech. Our first experiment suggests, however, that word processing speed was slowed down, relative to linear compression. In a second experiment, word processing of artificially time-compressed speech was compared with processing of naturally produced fast speech. Even when naturally produced fast speech is perfectly intelligible, its less careful articulation, combined with the changed timing pattern, slows down processing, relative to linearly time-compressed speech. Furthermore, listeners preferred artificially time-compressed speech over naturally produced fast speech. These results suggest that linearly time-compressed speech has both a temporal and a segmental advantage over natural fast speech.


*Keywords:* Timing; Prosody; Perception; Word recognition; Fast speech; Time compression; Segmental reduction

## 1. Introduction

Increases in speech rate are accompanied by changes in relative timing of speech units at syllable, word and sentence level. When people speak faster, consonant durations are reduced less, relatively, than vowel durations (Gay, 1978; Max and Caruso, 1997). When speech rate is increased, durations of unstressed syllables in polysyllabic words are reduced more than those of stressed syllables, relatively, which makes the prosodic pattern more prominent (Janse et al., 2003). Furthermore, at sentence level, durations of sentence-stressed syllables are reduced less, relatively speaking, than durations of unstressed syllables (Peterson and Lehiste, 1960; Port, 1981). The question is whether this enhanced prosodic pattern at word and sentence level in faster speech is the result of a strategic and communicative principle, namely that speakers tend to preserve the parts of information in the speech stream that are most informative. This reflects the predictions of the Hyper and Hypospeech theory, which states that much of the variability of speech stems from the ways speakers adapt their speech to what the

*E-mail address:* esther.janse@let.uu.nl (E. Janse).

0167-6393/$ - see front matter 
doi:10.1016/j.specom.2003.07.001

speaker thinks is needed for the listener to comprehend the message (Lindblom, 1990).

Alternatively, the changes in word- and sentence-level timing observed in naturally produced fast speech may be due to restrictions on articulation, rather than reflecting a strategic and communicative principle. Speakers may not be able to speed up in such a way that it approaches linear time compression. Lexical stress is specified in the mental lexicon, and as a result of this specification stressed syllables are produced with more articulatory precision. de Jong (1995) argues that linguistic stress can be seen as localised hyperarticulation. Stressed vowels are closer to their citation form (Lehiste, 1970; van Bergem, 1993). In the mental representation, the target values for stressed segments may be more strictly specified than for unstressed segments. Fowler (1981) found for English that lexically stressed vowels show less contextual variation than lexically unstressed ones: in other words, stressed vowels have a greater coarticulatory resistance than unstressed ones. Cho (2001) found that the same holds for sentence-stress in English: accented vowels show a greater coarticulatory resistance than unaccented vowels. This means that the speaker is forced to spend more energy on approximating the specified targets for stressed syllables than for the more loosely defined unstressed syllables. A faster articulation rate (i.e., beyond a 20% increase in rate) is almost inevitably accompanied by undershoot of the pre-defined targets because of the inertia of the speech organs (Lindblom, 1963; Moon and Lindblom, 1994). If more precision is required for the stressed syllables than for the unstressed syllables, the speaker simply cannot speed up that much during the production of stressed syllables. As a result, speakers compress unstressed syllables more than stressed syllables. However, this nonlinear type of speed-up need not be beneficial for the listener.

Janse et al. (2003) started from the assumption that the enhanced prosodic pattern found in naturally produced fast speech might improve the intelligibility of artificially time-compressed speech: if speakers time-compress speech in a selective way for the sake of the listener, listeners should benefit from such nonlinear time compression. The duration study showed that, when speakers speed up their speech rate, unstressed syllables in disyllabic words were affected more in duration, relatively, than stressed syllables. The changes in timing were extrapolated from the moderately fast speech rate reached by our speakers to a much faster speech rate: normal-rate speech was artificially time-compressed, either linearly or nonlinearly, to about three times faster than normal rate. This is much faster than can be attained by human speakers. Apart from the practical reason of avoiding ceiling effects in intelligibility, we chose to use exceptionally fast speech because we expected that a degraded speech signal would cause listeners to rely more on prosodic cues than when speech quality is high. Prosody has been shown to be an important source of information in lexical processing (Cutler and Clifton, 1984; Cutler and van Donselaar, 2001; Cutler and Koster, 2000; van Heuven, 1985; Slowiaczek, 1990), and listeners rely even more on prosodic information in case of adverse listening conditions (van Donselaar and Lentz, 1994; Wingfield, 1975; Wingfield et al., 1984). However, for this heavily time-compressed speech, making the timing pattern more similar to that of natural fast speech turned out to have a negative effect on intelligibility, relative to linear time compression (Janse et al., 2003).

Three objections may be raised against the experimental set-up in (Janse et al., 2003) that might explain why imitating the rules of natural fast speech timing showed no improvement over linear time compression.

The first objection is that the results might have been due to an unwarranted extrapolation of the timing rules of natural fast speech to speech rates that are much higher than what human speakers can achieve. Selective compression, based on natural fast speech timing, might in fact have improved intelligibility at the speech rate at which this timing pattern was observed (about 1.5 times normal rate), but not at the very fast rate employed in (Janse et al., 2003) perception study (about three times normal rate). If, on the other hand, the changed timing in natural fast speech is indeed due to articulatory restrictions, and not to a communicative strategy, then word perception of artificially time-compressed speech is not helped

by making its temporal pattern closer to that of natural fast speech, not even at a rate of speech that speakers can attain. This first objection against the results of Janse et al. (2003) will be referred to as the 'extrapolation problem' and will be addressed in Experiment 1.

A second objection against the validity of the results obtained in (Janse et al., 2003) is that the selective compression method applied in that study and in Experiment 1 here, may not have been typical of natural fast speech. The duration measurements which were used in order to establish the 'selective' or nonlinear type of compression (in Janse et al., 2003) were based on very fast and rather slurred speech. According to Lindblom's Hyper and Hypospeech theory, the listener's needs are always in the mind of the speaker. Horton and Keysar (1996) found evidence that this may not be true when speakers are under time or task pressure. In their experiment, speakers carried out a referential communication task in which speakers had to describe objects. Horton and Keysar's data showed that common ground, or shared knowledge, was used in the descriptions without time pressure, but that common ground was not used when speakers were under time pressure. This may have implications for the present study on natural fast speech. It is conceivable that the time pressure that was imposed on the speakers of the duration study reported in (Janse et al., 2003) may have made them lose sight of the listeners. We therefore cannot exclude the possibility that the speakers in this experiment just chose the easiest, and not necessarily the only possible way to speed up. The question remains then how imitating the temporal pattern of natural fast, *but intelligible*, speech affects processing of artificially time-compressed speech.

A third possible explanation for the failure to improve word perception over strictly linear compression might be that our selective time compression was not a faithful representation of what speakers do when they speed up. The duration study in (Janse et al., 2003) showed that speakers reduced stressed syllables in disyllabic words less than unstressed syllables. On the basis of this finding, the entire sentence was divided into stressed and unstressed syllables. This may have been far too coarse: it is questionable whether entire sentences can be treated as polysyllabic content words, consisting of stressed vs. unstressed syllables. The speed-up 'strategy' of the speaker may be far more complex than modelled in the previous paper (Janse et al., 2003).

These latter two issues, which will be referred to as the 'communicative-intention' problem and the 'unfair imitation' problem, will be addressed in Experiment 2. In this second experiment, the nonlinear type of time compression is based on natural fast speech that is perfectly intelligible. This nonlinear compression condition will be an exact copy of the timing pattern of the naturally produced fast speech, at least at syllable level. This condition will be compared with strictly linear compression. Furthermore, word processing of these two types of artificially time-compressed speech will be compared with processing of the natural fast speech itself. In this way, the natural fast speech, which is inevitably somewhat less carefully pronounced, can be compared with normal-rate careful speech that is time-compressed (linearly or nonlinearly) afterwards.

The segmentally reduced articulation of natural fast speech is expected to hinder perception, even though listeners might expect these processes to occur in fast speech. Coarticulation and assimilation may be helpful for listeners when speech is articulated at a normal rate (Whalen, 1991), but perception of speech presented at fast rates seems to be helped by a more redundant speech signal. In (Quené and Krull, 1999) a word detection task was used to investigate whether word recognition is sped up by assimilation or is hampered by it. The type of assimilation was deletion of /t/ between consonants, as in Dutch *pos/t/ brengen* '*mail deliver*'. The results of this word detection study were rather surprising: whereas listeners 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. The unassimilated form may have been rather unnatural given the fast speech rate, yet listeners were faster in recognising the unassimilated form than the assimilated form. Thus, increased assimilation and coarticulation are assumed to slow down the processing speed of fast speech. Even though speakers may produce fast but intelligible speech,

perception is predicted to be more difficult in the natural-fast condition than in artificially time-compressed conditions, in spite of its naturalness, due to its increased coarticulation and assimilation. As said, the 'communicative-intention' objection and the 'unfair imitation' objections against our previous results (Janse et al., 2003) will be addressed in Experiment 2.

Lastly, in view of possible technological applications in which artificially time-compressed speech is used, it is interesting to know which type of fast speech listeners actually prefer. Therefore, a preference study (Experiment 3) will be set up to investigate which type of speech listeners find more agreeable to listen to: natural fast speech (that is perfectly intelligible), or speech that is time-compressed artificially to the same fast rate, either linearly or nonlinearly. If listeners turn out to process linearly time-compressed speech faster than naturally produced fast speech, it would be reasonable to assume that they also find this type of speech the most agreeable one to listen to. A competing hypothesis would be that listeners' preference is based on naturalness: listeners may find the most natural version the most agreeable one to listen to, given that all three types of speech are perfectly intelligible.

In sum, the following three questions will be addressed in this paper:

1. Can word processing in moderately fast artificially time-compressed speech be improved by making its timing pattern more similar to that of natural fast speech?
2. Which is easier to process: naturally produced fast speech or artificially time-compressed speech? How do a changed timing pattern and segmental effects contribute to this difference?
3. When listening to fast speech, do listeners prefer either artificially time-compressed speech or naturally produced fast speech?

## 2. Experiment 1: linear vs. nonlinear time compression

In this experiment, the 'extrapolation problem' is addressed. The question is whether the failure to improve perception of artificially time-compressed speech by imitating the timing rules of natural fast speech might have been due to the unwarranted extrapolation of the rules of natural fast speech to much faster rates.

In the previous study (Janse et al., 2003), word intelligibility was measured in three different time compression conditions. In the Linear Compression condition (LC), all syllables were time-compressed to the same degree. Selective Compression (SC) was based on the nonlinearities found in a duration study of natural fast speech. Speakers in this duration study were pressed to speak very fast, and the resulting speech was relatively slurred. All stressed syllables were time-compressed to 65% of their original duration; and the unstressed syllables were compressed further, namely to 45% of their original duration. Conversely, in the Unnatural Compression condition (UC), all stressed syllables were compressed more (i.e., to 45%) than the unstressed syllables (to 65%). This extra condition was added to test the hypothesis that perception might be helped most by segmental intelligibility of both stressed and unstressed syllables. By comparing these three types of compression, it can be investigated which type of nonlinear compression (Selective or Unnatural) actually improves perception over linear compression.

If speech is time-compressed to the moderately fast rates of speech that speakers can attain, an intelligibility test is not viable because artificially time-compressed speech at this rate is still perfectly intelligible. In order to avoid these ceiling effects in intelligibility, Janse (2001) investigated whether selective compression would improve intelligibility at a moderately fast rate by means of a speech-interference test (Nakatani and Dukes, 1973). Unfortunately, no difference in speech communication quality was found, possibly due to the insensitivity of this particular test. Therefore, another technique is needed to address this question. On-line techniques may be more sensitive. Pisoni (1987, 1997) used reaction time measures to compare and evaluate perfectly intelligible synthetic speech systems. Reaction time measures, such as lexical decision time or phoneme detection time, are assumed to reflect the speed with which different types of speech can be processed. Phoneme

detection was used by Nix et al. (1993) to compare processing of natural and synthetic speech, and has also been used to compare spontaneous and read speech: two types of speech which differ both segmentally and with respect to timing (Mehta and Cutler, 1988). According to Cutler and Norris' dual route model, phonemes can be detected on the basis of either pre-lexical information or lexical information (Cutler and Norris, 1979), depending on which of the two routes is fastest in yielding a response. In the more recent Merge model (Norris et al., 2000), phonemic decisions in speeded phoneme detection tasks are argued to be based on the merging of pre-lexical and lexical information. For the present questions about word processing speed, the experimental task should allow us to draw conclusions about lexical, and not about pre-lexical processing. It is therefore important that the phoneme detection responses can be taken to result from lexical processing. Stimulus monotony (presenting lists of isolated words) can make subjects shift their attention to pre-lexical processing. By presenting entire meaningful sentences, this might be avoided (Cutler and Norris, 1979). The point of lexical vs. pre-lexical processing will be taken up again in the results section.

### 2.1. Method

#### 2.1.1. Material

The test material consisted of 81 meaningful sentences or sentence fragments taken from Dutch newspaper articles. These sentences or fragments were a subset of the material used in the previous study (Janse et al., 2003) because only a selection of the original material contained suitable target words for phoneme detection. The target words should have a word-initial plosive consonant, which does not occur anywhere else in the sentence. The material did not contain enough nouns which met these criteria. Thus, other word classes were used as well: the target items were nouns, verbs and adverbs.

The sentence fragments ranged from 5 to 16 words and formed complete clauses. Each sentence fragment contained a disyllabic monomorphemic target word. The position of the target word varied through the sentence fragment. The target words were semantically not predictable in their sentence contexts. An example sentence is provided in (1) below (target word underlined):

(1) Hij had de partij moeten vernietigen ('He should the batch have destroyed')

About half of the target words carried sentence accent; the other half was unaccented. Apart from the 81 test sentences, 60 catch trials (which did not contain the assigned phoneme) were interspersed with the material to keep subjects from pressing the button randomly.

Overall, the speech material in this experiment was time-compressed to 65% of its original duration (i.e., 1.5 times normal rate): this is an approximation of the compression factor that speakers can attain when they are asked to speak very fast. As in the previous study (Janse et al., 2003), Selective Compression (SC) was based on the nonlinearities at syllable level found in a duration study of naturally produced fast speech. There are also nonlinearities in the speaker's way of speeding up below the syllable level. Some phonemes, or even some parts of phonemes (such as the steady-state portion), may be affected more by an increase in rate than others. However, there are two reasons not to take these nonlinearities into account here. The first is that there are no data available, at least to our knowledge, on the precise small-scale effects of increased speech rate on all segments. Hence, segment durations cannot be manipulated validly in artificially time-compressed speech. A more fundamental reason for imitating only nonlinearities at the syllable level is that the major changes in temporal pattern are expected to occur mainly at higher levels: the amount of time compression in natural speech was found to depend strongly on the level of stress or on the status of the word (Peterson and Lehiste, 1960; Port, 1981; Janse et al., 2003). These effects are necessarily larger than phonemic or subphonemic effects.

All syllables in the sentence fragments were first given a [+stress] or [−stress] mark. Articles and auxiliaries, and unstressed syllables in all types of words got a [−stress] mark. All other syllables (such as prepositions, main verbs and all types of

stressed syllables within polysyllabic words) got a [+stress] mark; cf. the example sentence in (2).

(2) Hij+ had− de− **par−tij+** moe+ten− ver− nie+ti−gen− ('He should the batch have destroyed')

Then, all [+stress] syllables were time-compressed to 65% of their original duration; and the [−stress] syllables were compressed to 45% of their original duration. The Pitch-Synchronous Overlap Add (PSOLA) time-scaling technique, as implemented in the speech editing program GIPOS (version 2.3; http://www.ipo.tue.nl/ipo/gipos) was used to time-compress the fragments. In GIPOS, selected parts of the waveform can be time-compressed, while the remainder of the waveform remains unaffected. In this way, each syllable can be time-compressed according to its appropriate factor. Conversely, in the Unnatural Compression condition (UC), all syllables with a [+stress] mark were compressed to 45%, and all [−stress] syllables were compressed to 65% of their original duration. Thus, stressed syllables were reduced more, relatively, than unstressed syllables. Then, the nonlinearly compressed speech in conditions SC and UC was expanded again to an overall compression rate of 65% of the original duration. This was also done separately for the target words, such that the target word duration would be equal in all three compression conditions. The difference between the three conditions lies in the duration of the stressed and unstressed syllables. The repeated PSOLA manipulations did not create any audible artefacts.

For the linear time-compression condition, all syllables were compressed to the same degree (i.e., to 65% of their normal-rate duration).

Importantly, the PSOLA manipulations only affected the time scale of the fragments; the intonation contour remained unchanged.

### 2.1.2. Design and procedure

The three compression conditions, viz. Linear Compression, Selective Compression, and Unnatural Compression were rotated over the 81 items, yielding 27 items per condition. Items and conditions were combined orthogonally on three experimental lists. Subjects were seated individually in sound-treated booths, wearing closed-ear headphones. First, they were given a written instruction on their task during the experiment. Subjects were told that they would see a letter-sound on the computer screen in front of them. This was the sound they were supposed to detect during the upcoming sentence, which was to be presented to them over headphones. They were told to press either button of the button box in front of them as fast as possible whenever they detected a word-initial occurrence of the assigned phoneme. They were also told that there would be catch trials which did not contain the assigned plosive. Regardless of whether the subject had pressed the button, the next sentence was presented 2 s after the previous sentence offset. There were 10 practice items, after which additional instruction was given, if necessary. After the subjects had resumed the test, 10 warming-up filler items were presented before the actual test began. All test and filler items were presented in random order.

A time marker had been placed in each audiofile at the start of the silent interval of the target plosive (or at the start of the voice bar for voiced plosives). During the experiment, reaction times were computed by subtracting this marker time from the time until the button press was registered.

### 2.1.3. Subjects

Ten subjects were assigned to each list, so that 30 subjects, all students at Utrecht University, participated in this experiment. The subjects were paid for their participation.

## 2.2. Results

All subjects agreed that the time-compressed speech was perfectly intelligible. There were some negative reaction times, which indicates that subjects had responded to another initial plosive. These were considered missing observations, together with those instances in which subjects had not pressed the button at all. The raw detection times in the three time-compression conditions are shown in Table 1, together with standard errors of the mean and miss rates.

In the previous intelligibility results at a very fast speech rate (Janse et al., 2003), the results were quite different for the two stress positions. The effect of Selective Compression was mainly harmful (relative to Linear Compression) for the intelligibility of the finally stressed items, whereas Unnatural Compression was harmful only for the initially stressed items. In Table 2 the phoneme detection results are therefore broken down by Stress position to investigate whether the same interaction is found here. There were 35 initially stressed items, and 46 items with noninitial stress. For comparison, the previous intelligibility results are shown in parentheses.

For the statistical analysis of the results, all missing observations were replaced by the grand mean over the test conditions (554 ms). Although this is not a very sophisticated way to deal with missing values, the reported effects are large enough, however, to warrant this approach (i.e., $p$ values are small enough). Generally, reaction time data do not show normal (Gaussian) distributions. Since analyses of variance assume normally distributed data, reaction time data may pose a problem for ANOVA analysis. When reaction times are transformed to inverse reaction times (1/RT), the distributions are usually less skewed. The inverse detection time results were entered into analyses of variance (in which either items or subjects were treated as repeated measures) to establish the effects of Compression Type and Stress Position. The main effect of Compression Type was significant ($F_1(2,28) = 7.4$, $p = 0.001$; $F_2(2,78) = 5.9$, $p = 0.004$), and so was the interaction between Stress position and Compression Type ($F_1(2,28) = 4.4$, $p = 0.017$; $F_2(2,78) = 4.0$, $p = 0.022$). There was no main effect of Stress position ($F_1(1,29) < 1$, n.s.; $F_2(1,79) < 0$, n.s.).

The results in Table 2 are remarkably similar to the previous intelligibility results (in parentheses): for the initially stressed words Linear Compression wins; for the finally stressed items, there is hardly any difference between Linear Compression and Unnatural Compression. Importantly, for the finally stressed items, Unnatural Compression is less harmful than Selective Compression.

Univariate ANOVAs provide the possibility of doing post hoc analyses to find out which conditions differed significantly from each other. These analyses can also cope with missing observations, so that these do not have to be replaced by the grand mean.

In these ANOVAs with 1/RT as the dependent variable, Compression Type as fixed factor, and either subject or item as random factors, the effect of Compression Type was significant as well ($F_1(2,28) = 8.2$, $p = 0.001$; $F_2(2,79) = 5.0$, $p = 0.008$). The results of the post hoc tests (Scheffé) are shown in Table 3.

The post hoc analyses results in Table 3 show that Linear Compression and Selective Compression differed significantly from each other. The conditions LC and UC and SC and UC did not differ significantly from each other. Separate post hoc analyses on the inverse reaction time data were

Table 1
Mean raw detection time (in ms), plus standard error of mean and miss rate in three compression conditions

| | Mean (ms) | Standard error | Miss rate (%) |
|---|---|---|---|
| Linear Compression | 537 | 9 | 6 |
| Selective Compression | 576 | 11 | 8 |
| Unnatural Compression | 557 | 10 | 8 |

Table 2
Mean detection times in all three compression conditions, broken down by stress position

| | Initial stress ($N = 35$) | Final stress ($N = 46$) |
|---|---|---|
| Linear Compression | 532 (62) | 542 (68) |
| Selective Compression | 544 (54) | 597 (55) |
| Unnatural Compression | 586 (50) | 535 (68) |

For comparison, the previous intelligibility percentages (obtained in Janse et al., 2003) are shown in parentheses.

Table 3
Results of Scheffé post hoc test (significance values)

| | Subject analysis | Item analysis |
|---|---|---|
| LC–SC | $p = 0.008$ | $p = 0.011$ |
| LC–UC | $p > 0.1$ | $p > 0.1$ |
| SC–UC | $p > 0.1$ | $p > 0.1$ |

carried out for items with initial stress and for items with noninitial stress. For items with initial stress, the difference in phoneme detection time between the LC condition (532 ms) and the SC condition (544 ms) was not significant (Scheffé analysis on subjects $p > 0.1$; on items $p > 0.1$). The relatively large difference between the LC (532 ms) and UC condition (586 ms) was not significant either (subjects $p = 0.18$; items $p = 0.20$). Conversely, for the finally stressed items, the difference between LC (542 ms) and SC (597 ms) is significant (Scheffé analysis on subjects $p = 0.005$; on items $p = 0.006$). The difference between SC (597 ms) and UC (535 ms) is also significant (subjects $p = 0.006$; on items $p = 0.007$), but the difference between LC and UC is far from significant (both analyses $p > 0.1$).

Making the prosodic pattern less pronounced (i.e., Unnatural Compression) may slow down processing of words with initial stress, but it does not have a negative effect on the processing speed of words with final stress. This pattern of results for finally stressed words can be explained by two effects, working in opposite directions. The first effect is that of temporal alignment. The second effect concerns the duration of the stressed and most informative syllable. Regarding alignment, word recognition has a left-to-right aspect to it because speech unfolds over time. In Dutch some experimental results suggest that misstressing of finally stressed words is more disrupting than misstressing of initially stressed words (van Leyden and van Heuven, 1996). Misstressing a finally stressed word (i.e., making it initially stressed) might activate the wrong set of word candidates to start with. With respect to the present results, if the unstressed syllable of a finally stressed word is relatively long, it may be easier to start the correct alignment with possible word candidates at the start of the unstressed syllable, because the unstressed syllable is relatively salient, compared to the LC condition. On the other hand, the duration of the stressed syllable is shorter in the UC condition than in the other two conditions. This shorter duration of the stressed syllable probably makes the stressed syllable less intelligible, and may even lead to the activation of words with the wrong stress pattern. Hence, the positive effect of the UC condition on the initial alignment of finally stressed words against word candidates is counterbalanced by the short duration of the stressed syllable. The same two effects also explain the pattern observed for initially stressed words. Initial alignment with possible word candidates tends to be more difficult in the UC condition because the first syllable is shorter. As the first syllable is in this case also the stressed syllable, processing is only slowed down. Problems with correct initial alignment may also be the reason why the Selective Compression condition is mainly harmful for words with final stress (because of the very short initial unstressed syllables), and not for words with initial stress.

In the Merge model (Norris et al., 2000), information from the lexical and the pre-lexical route is always combined to yield a phoneme detection response. Pre-lexical processing feeds information to the lexical level to allow activation of lexical candidates. At the same time, this pre-lexical information is available for explicit phoneme decision making. The decision stage also continuously accepts input from the lexical level and merges the two sources of information. Therefore, phoneme detection responses are always a result of both processes. Still, the model allows the possibility to shift attention between the two outlets in order to explain why the experimental set-up can play such an important role (Cutler et al., 1987). Whereas stimulus monotony (e.g., lists of CVC words) may induce subjects to focus on the *pre-lexical* route, responses are more likely to depend mainly on the *lexical* route when the targets are embedded in meaningful (newspaper) sentences (Cutler and Norris, 1979). In the present experiment, subjects had to monitor different phoneme targets, and the stimulus materials were meaningful sentences. Furthermore, the decreased quality of the speech as a result of the moderate time compression may have caused the pre-lexical route to be rather inefficient. Additional evidence that speeded detection tasks tap into lexical processing comes from a study by Seidenberg and Tanenhaus (1979) who found effects of spelling on rhyme monitoring data. The striking similarity between the previous word-intelligibility results and the current detection time results (cf. Table 2) provides the strongest indication that the subjects focussed

mainly on lexical-level processing. As in (Janse et al., 2003), linear compression has a significant advantage over the selective compression condition, mainly for items with noninitial stress. Therefore, although the possibility of pre-lexical processing cannot be fully excluded, it seems more plausible to assume that the detection results reflect speed of word processing. Consequently, these results suggest that the attempt to preserve the intelligibility of the unstressed syllable, at the expense of a natural temporal pattern, again turns out to be less harmful than enhancing the natural prosodic pattern. Making the temporal pattern of time-compressed speech more like that of natural fast speech does not improve, but rather slows down word processing, particularly for items with final stress.

As argued in Section 1, the nonlinearities found in the duration study reported in (Janse et al., 2003) might be typical only of very fast and slurred speech. Although the nonlinearities may also result from the fact that the speaker simply cannot speed up in an approximately linear fashion, it cannot be excluded that the nonlinear speed-up is typical only of *slurred* fast speech. Before we can really conclude that the natural way of changing relative timing when speeding up is harmful, we have to be sure that the fast speakers were speaking with the intention of being understood. Furthermore, another explanation for the failure to improve word perception over strictly linear time compression might be that selective time compression, as applied in the previous study and in Experiment 1 here, was not a faithful representation of what speakers do when they speed up. Based on the observation that speakers reduced stressed syllables in disyllabic words less (to 65%) than unstressed syllables (to 45%), the entire sentence was divided into stressed and unstressed syllables, which were then reduced to 65% or 45%, respectively. This may have been far too coarse: entire sentences cannot be treated as polysyllabic content words, consisting of stressed vs. unstressed syllables. The selective compression applied in Experiment 1 could be an unfaithful simplification of what speakers actually do: the speed-up 'strategy' of the speaker may be far more complicated than modelled in Experiment 1.

In the next experiment word perception of normal-rate speech that is either linearly or nonlinearly time-compressed is compared with natural moderately fast speech. Importantly, the speaker is instructed to remain intelligible. The question is how a changed timing pattern and segmental effects contribute to a possible processing difference between this natural fast speech and artificially time-compressed speech (linear or nonlinear). In order to meet the second objection against the previous results, the nonlinear type of time compression will now be an exact copy (at syllable level) of the natural fast speech timing, and not a global extrapolated pattern of stressed vs. unstressed syllables.

## 3. Experiment 2: natural fast vs. time-compressed speech

In this experiment the 'communicative-intention' and the 'unfair imitation' objections against the previous results are addressed. The question is whether word-level timing is also different from that of normal rate for fast yet intelligible speech. If this is the case, how does it influence word perception, relative to linear compression? Secondly, when speakers are asked to speak fast and yet intelligibly, is naturally produced fast speech easier to process than artificially time-compressed speech? If a change in timing is found, it is expected to result from articulatory restrictions, rather than from a listener-oriented strategy. Hence, this change in timing should then slow down speech processing, in line with the previous results. Secondly, the increased coarticulation and assimilation that almost inevitably accompany a faster speaking rate are expected to make word perception more difficult relative to artificial time compression. Thus, the nonlinear way of speeding up, combined with the extra amount of segmental overlap, is expected to make word perception more difficult, because they are both due to the speaker not being able to speed up otherwise.

To put this to the test, word perception is compared in three experimental conditions:

- linear time compression
- copy-fast-speech-timing (all syllable durations of the normal condition are time-compressed

artificially to the syllable durations of the natural fast condition)
- natural fast speech

In this way, the effect of nonlinear time compression and of the combination of nonlinear time compression and increased segmental overlap can be investigated, relative to linear compression. The prediction is tested that the more similar fast speech is to linearly artificially time-compressed speech, the shorter the processing times. Removing only the segmental characteristics (as in the copy-fast-speech-timing condition) will make processing easier, relative to natural fast speech. Removing both the segmental and the temporal characteristics of natural fast speech (as in linearly time-compressed speech), will make processing even easier.

### 3.1. Method

As in Experiment 1, the phoneme detection task was used to evaluate speech processing difficulty.

#### 3.1.1. Material

Dutch news bulletin texts were collected and sentences were selected that contained nouns starting with a plosive. Eighty-four sentences or sentence fragments were chosen with a mean length of 23.4 syllables (s.d. 7.6). The target words were always nouns, and had two to four syllables. Half of these target words had initial stress; in the other half, stress was on the second or later syllable. The nouns were never compounds, but some were morphologically complex (e.g., as in *tuinder* 'market gardener'). Target words started with a plosive (18 with /t/, 10 with /d/, 20 with /p/, 21 with /b/, and 15 with /k/). The assigned word-initial plosives did not occur elsewhere in the sentence, neither word-initially, word-medially nor word-finally. Half of the target items carried sentence accent; the other half did not. An example sentence fragment is given in (3) below (target word underlined):

(3) Verf en tapijt brengen giftige stoffen in omloop. ('Paint and carpet spread toxic substances.')

One male speaker of Dutch, known for his clear speaking style, read the sentence material at a normal and at a fast rate. It was stressed that the fast rate should be fast, but still intelligible. The sentence fragments were cut out from the longer bulletin texts. The normal articulation rate in the sentence fragments was 6.1 syllables/s, and this rate was increased to 8.5 syllables/s in the fast condition. Thus, the overall fast-to-normal ratio was 0.72 (i.e., speed-up factor 1.4). Pairwise comparisons of the articulation rates of the normal and fast sentence fragments showed that the mean fast rate is significantly faster than the normal rate ($t(83) = 41.6$, $p < 0.001$). It is important to note that the fast rate in the present study is moderately fast: speakers in the duration study of Janse et al. (2003) increased their articulation rate to 10.5 syllables/s.

The 84 test sentences or sentence fragments that were articulated at a normal rate were somewhat louder than those articulated at a fast rate. For each sentence, the mean intensity of the fast version was therefore amplified to equal that of the normal-rate version.

The test sentences were labelled manually: time markers were placed in the waveform at each syllable boundary, both in the normal and fast rate versions of each sentence. For the copy-fast-speech-timing condition, all fast-to-normal ratios were computed by dividing the duration of each syllable in the fast condition by its corresponding duration in the normal-rate condition. Then, the durations of all syllables in the normal-rate condition were time-compressed one by one, using these specific fast-to-normal ratios. In this way, the timing structure of the copy-fast-speech-timing condition was an exact copy of that of the natural fast version, at least at the syllable and sentence level.

After the sentence had been time-compressed syllable-by-syllable, the end result was somewhat shorter than the natural-fast condition. This is due to a PSOLA artefact: repetitive time compression of successive small windows of speech in the end yields a slightly faster version than specified. In order to make the copy-condition and the natural-fast condition exactly equally long, the natural-fast condition was time-compressed slightly (linear

compression to approximately 98–99% of duration). In this way, the two conditions were of identical total duration. Furthermore, potential reaction time differences between the three conditions would not be due merely to the fact that two of them were PSOLA resynthesised, whereas one of them was not.

For the linear compression condition, the overall fast-to-normal ratio was computed for each sentence. This overall fast-to-normal ratio was then applied by linear time compression to the normal-rate version of each sentence. Lastly, the target word's duration was made equal to that in the natural fast condition so that target word offset would not be reached earlier in any of the three conditions.

To prevent subjects from pressing the button randomly, 80 catch trials, also taken from news bulletin items, were interspersed with the test material. These catch trials did not contain the assigned plosive.

An informal pilot study showed that the intelligibility of the natural-fast and time-compressed conditions was about 100%.

#### 3.1.2. Design and procedure

The 84 test sentences, in three experimental conditions, were distributed over three lists according to a Latin square design. There were 10 practice items, after which subjects could ask questions if anything was unclear. All test and catch trial items were presented in random order. The instructions and experimental setting for the subjects were identical to those in Experiment 1.

#### 3.1.3. Subjects

To each of the three lists, 10 subjects were assigned. The 30 subjects were all students at Utrecht University and they were paid for their participation.

### 3.2. Results

Before the perception results are discussed, the question whether word-level timing is also different from that of normal rate for fast yet intelligible speech is addressed first. As said, the fast speech rate in the present study is slower (8.5 syllables/s) than the fast rate reported in (Janse et al., 2003) (10.5 syllables/s). The duration measurements presented in that paper illustrated the nonuniform way of speeding up at word level: on average, stressed syllables had a mean fast-to-normal ratio of 0.65 (i.e., they were reduced to 65% of their normal-rate duration), whereas unstressed syllables had a mean fast-to-normal ratio of 0.45. In the present material, stressed syllables had a mean fast-to-normal ratio of 0.77; and the unstressed syllables had a mean fast-to-normal ratio of 0.71. This suggests that the change in timing is not due to sloppy articulation alone: even intelligible fast articulation is accompanied by a shift in the temporal pattern. Major shifts in word-level and sentence-level timing, however, apparently take place only when the speech rate is sufficiently high.

We now turn to the perception results of the phoneme detection study. Time markers had been placed in the audiofiles at the start of the silent interval of the target plosive (or at the start of the voice bar for voiced plosives). Reaction times were computed by subtracting this marker time from the time that the button press was registered. The raw mean detection times are presented in Table 4.

The results are relatively similar for items with initial stress vs. items with noninitial stress, and the results are therefore not broken down by stress position. The results are as predicted: detection times are fastest in the linear compression condition (572 ms), followed by the copy-fast-speech-timing (600 ms), and the natural fast speech condition (624 ms). The miss rates are low in all conditions, which provides further evidence that the speech in all conditions was highly intelligible. The missing observations were replaced by the subject's mean in that condition in the subject analysis, and by the item's mean in that condition

Table 4
Mean raw detection time (in ms), standard error of the mean, and miss rate for the three fast conditions

| | Mean (ms) | SE | Miss rate (%) |
|---|---|---|---|
| Linear time compression | 572 | 9 | 3 |
| Copy-fast-speech-timing | 600 | 13 | 3 |
| Natural fast speech | 624 | 14 | 3 |

in the item analysis. As in Experiment 1, statistical analyses were run on inverse reaction time data (1/RT). Two ANOVAs on the inverse detection times were carried out in which either items or subjects were treated as repeated measures. Condition and Stress position were analysed as fixed factors (with items nested under the Stress factor in the item analysis). The effect of Condition was significant in both analyses ($F_1(2,28) = 7.4$, $p = 0.002$; $F_2(2,81) = 4.3$, $p = 0.039$). The effect of Stress position was far from significant ($F_1(1,29) = 1.0$, n.s.; $F_2(1,82) < 1$, n.s.); and so was the interaction between Condition and Stress position ($F_1(2,28) < 1$, n.s.; $F_2(2,81) < 1$, n.s.).

Univariate analyses of variance allow the possibility of doing post hoc tests. Condition was taken as fixed factor, and either subjects or items as random factors. Missing observations were replaced by either the subject or the item mean in that condition, depending on the analysis. The effect of Condition on the inverse detection times was significant in both analyses ($F_1(2,28) = 5.5$, $p = 0.007$; $F_2(2,82) = 3.1$, $p = 0.048$). Separate post hoc tests (Scheffé) showed which fast speech conditions differed significantly from one another. The significance values of the post hoc analyses are shown in Table 5.

Only the linear time-compression condition and the natural fast condition differed significantly from each other: there was no significant difference between linear compression and the copy-fast-speech-timing condition, and the difference between the copy-fast-speech-timing condition and natural fast speech was not significant either.

Thus, the differences between the three experimental conditions are rather small. Only the 52 ms advantage of linear compression over natural fast speech is significant. The absence of a significant difference between the two time-compressed conditions may be attributed to the relatively small difference in speech rate between the normal and the fast speech conditions, which, in turn, induced only small changes in word- and sentence-level timing (cf. the fast-to-normal ratios of 0.77 for stressed vs. 0.71 for unstressed syllables). The phoneme detection task may be not sensitive enough to pick up any perceptual consequences of this timing shift. To some extent, then, the experimental set-up failed to answer the question, because the effects of changed timing and segmental reduction cannot be quantified separately. Although only the combined effect significantly slows down processing, the data in Table 4 suggest that both separate effects slow down detection times. Consequently, the results provide some evidence that the less *natural* the fast speech is (whether temporally or segmentally), the easier the process of word recognition.

Table 5
Significance values of post hoc analyses (Scheffé)

| | Subject analysis | Item analysis |
|---|---|---|
| Linear and Copy-fast | $p > 0.1$ | $p > 0.1$ |
| Linear and Natural-fast | $p = 0.03$ | $p = 0.009$ |
| Copy-fast and Natural-fast | $p > 0.1$ | $p > 0.1$ |

Analogous to the results of Experiment 1, the results are assumed to reflect lexical, rather than pre-lexical processing. As argued before, the decreased quality of the time-compressed speech is assumed to make pre-lexical processing rather inefficient. Furthermore, the use of meaningful sentences may also induce listeners to focus mainly on lexical processing.

This would then indicate that speakers cannot speed up their speech rate without making processing more difficult for the listener. In other words, the less words deviate from their normal-rate or 'canonical' form, the easier it is for the listener to map the incoming information onto the mental lexicon.

Given that the most natural condition is obviously not the easiest one to process, it would be interesting to find out whether listeners have a clear preference for either of the fast conditions. In view of technological applications, listeners' preference is an important factor. Therefore, subjective listeners' preference is tested in a third experiment. The careful pronunciation and 'optimal' prosodic pattern cause (linearly) time-compressed speech to be processed faster than naturally produced fast speech. It is conceivable that listeners will then also find linearly time-compressed speech the most agreeable type of fast speech to listen to. However, a competing prediction would be that listeners

might prefer to listen to a more natural type of fast speech than the unnatural artificially time-compressed speech. Given that all three types of speech are perfectly intelligible, listeners might prefer natural fast speech over the 'hyperarticulated' artificially time-compressed speech. Listeners' preference was tested in Experiment 3.

## 4. Experiment 3: listeners' preference

The three fast conditions of Experiment 2 were evaluated perceptually by way of a subjective preference test. The question was whether listeners could indicate whether one version (i.e., condition) of the same sentence sounded more 'agreeable' than another. This dimension was chosen to evaluate listening effort or overall perceived quality of the three conditions (cf. van Bezooijen and van Heuven (1997) for a comprehensive chapter on evaluation of text-to-speech systems). Van Bezooijen and van Heuven distinguish between functional and judgment testing. Judgment (or opinion) testing is a procedure whereby a group of listeners is asked to judge the performance of a speech output system. Functional testing evaluates how well a speech system actually performs (e.g., in terms of intelligibility or successful task completion in an information-retrieval system). There is evidence that the results of judgment and functional evaluations converge (Pavlovic et al., 1990). High correlations were found between paired comparison results and reaction time data (word monitoring) by Delogu et al. (1992), who evaluated the overall speech quality of synthesiser and vocoder systems and one human speaker. More importantly, however, Delogu et al. found that the best discrimination between the conditions was obtained with paired comparisons, whereas reaction time data showed the least discriminatory power of the four test methods that were used in their study. So, whereas no difference could be found between the two artificial time-compression (linear vs. nonlinear) conditions in Experiment 2, the subjective preference test may bring out differences between the two.

The three fast conditions of Experiment 2 were evaluated perceptually using the Comparative Mean Opinion Score (CMOS) test (ITU-P.800, 1996). Listeners' subjective preference was tested by presenting pairs of utterances, and asking them whether version B sounded more agreeable than version A. By doing this, listeners focus on the differences between the two versions.

### 4.1. Method

#### 4.1.1. Material

A selection of the speech material of the latter experiment (Experiment 2) was used in the present experiment. In some of the 84 original sentences, there were minor differences between the normal and fast rate versions of the sentence with respect to their intonation patterns. Therefore, 45 test sentences (and five practice sentences) were selected in which the difference between the normal and natural-fast rate utterance was minimal. [1] For each of the 45 test sentences, three pairs of fast conditions were evaluated (Linear vs. Copy-fast, Linear vs. Natural-fast, and Copy-fast vs. Natural-fast).

#### 4.1.2. Procedure

A complementary (Latin square) design was set up in which these comparison pairs were rotated over the sentence pairs and over three different lists. This was done to avoid training effects (van Santen, 1993). Each subject evaluated all comparison pairs equally often, but he evaluated only

[1] The minor changes in intonation did not involve any important differences in word durations. The placing of accents was also similar in the three conditions, but there were some minor differences with respect to which pitch configurations were used. The subset of the material that was used in Experiment 3 was also analysed to investigate whether the effects reported in Experiment 2 were also found in this subset. Univariate analyses on the inverse detection times showed that Condition had an effect on the detection times ($F_1(2,28) = 3.3$, $p = 0.043$; $F_2(2,43) = 2.6$, $p = 0.084$), although the effect was not significant in the item analysis, due to the decreased statistical power. Post hoc Scheffé analyses showed that only the linear compression and natural-fast condition differed significantly from each other (subject analysis $p = 0.02$; item analysis: $p = 0.024$), as was the case for the complete set.

one pair per sentence. Different listener groups heard different pairs for each sentence.

Subjects were seated in front of a computer screen on which there were two buttons (one labelled 'version A' and one 'version B'). Subjects listened to both members of the pair by first clicking on the version A button and then on the version B button (or in the reverse order). After listening to both members of each utterance pair, the subject indicated his preference by clicking on a seven-point scale, ranging from 'B is much more agreeable than A' (+3) to 'B is much less agreeable than A' (−3). In between are 'B is more agreeable than A' (+2), 'B is a little bit more agreeable than A' (+1), 'B and A are equally (un)agreeable' (0), and the reverse scale options (−1, −2).

It is conceivable that listeners always first listen to sound A and then to sound B. Consequently, listeners might have had a bias towards perceiving sound B as more agreeable because they were by then familiar with the contents of the sentence. To avoid this effect, or any other unwanted bias effects, the assignment of a specific member of each comparison pair to the A or B button was varied. Each condition within a comparison pair appeared about equally often as version A or B. Since the order of the items was randomised for each subject, the conditions varied randomly between buttons A and B.

Subjects listened to the material over headphones while seated in a sound-treated booth. They were told that they could listen to the two members of the pair as often as they liked before giving their preference value. After they had indicated their preference judgment by clicking on one of the seven buttons on the scale, they could click a button labelled 'Next' in order to hear the next sentence pair. The test lasted about 12 min. Before subjects started with the actual experiment, they were presented with five practice sentences, after which additional feedback or instruction was given, if necessary.

#### 4.1.3. Subjects

To each of the three experimental lists, six subjects were assigned. They were all students at Utrecht University, and were paid €5 for their participation.

### 4.2. Results

Each of the 18 listeners evaluated one comparison per sentence, yielding 45 judgments per listener. The mean perceptual scores for the three comparison pairs, with their respective standard errors, are given in Table 6.

A negative CMOS value indicates that the second member of the pair (as indicated in Table 6) was judged as less agreeable than the first. Statistical analyses of these CMOS values take the form of one-sample $t$-tests (both on subject and on item means) to test the hypothesis ($H_1$) that the mean CMOS value per pair differs significantly from zero. The $t$-tests for the first comparison pair shows that, although the CMOS value is really small, the Copy-fast (nonlinear) time-compressed condition is judged as significantly less agreeable than the linearly time-compressed condition ($t$-test on subject means: $t_1(17) = -3.4$, $p = 0.003$; on item means: $t_2(44) = -4.6$, $p < 0.001$). The difference between Linear and Natural fast is also significant in both analyses ($t_1(17) = -3.7$, $p = 0.002$; $t_2(44) = -4.4$, $p < 0.001$). The difference between Copy-fast and Natural-fast is not significant in either analysis ($t_1(17) < 1$, n.s.; $t_2(44) < 1$, n.s.).

The results confirm the prediction that listeners find the natural-fast condition less agreeable to listen to than the linearly time-compressed condition. Interestingly, a significant difference between the two artificial time-compression conditions was found (in favour of linear compression), whereas this was not found in Experiment 2. The direction of this preference was as expected: linear compression is preferred over the nonlinear (copy-fast) condition. Lastly, listeners did not prefer the copy-fast (nonlinear) time-compression condition over natural-fast speech. In sum, even at a rate at which

Table 6
Mean perceptual scores of CMOS test, on a scale from +3 to −3, plus standard errors

| | Mean CMOS | SE |
|---|---|---|
| Linear and Copy-fast | −0.27 | 0.05 |
| Linear and Natural-fast | −0.50 | 0.09 |
| Copy-fast and Natural-fast | −0.02 | 0.09 |

all three fast conditions are still perfectly intelligible, listeners have a slight preference for the condition which also proved easiest to process in Experiment 2.

These results agree with the study by Delogu et al. (1992) who found that paired comparison results highly correlate with reaction time data. Furthermore, our results are consistent with those of Delogu et al. in that differences between the conditions were most clearly observed in the paired comparison data.

## 5. Discussion

The results reported in the previous study (Janse et al., 2003) and the results of the present experiments point in the same direction. Even when speakers succeed in producing relatively fast but still perfectly intelligible speech, the resulting speech is more difficult to process than speech which is articulated at a normal rate and which is time-compressed linearly afterwards. This can be attributed partly to increased segmental overlap and to a changed temporal pattern. Although the effect of changed timing was less clear for the moderately fast rate in Experiment 2, the results of Experiment 1 support the idea that making the temporal pattern of artificially time-compressed speech more similar to that of natural fast speech slows down word processing, mainly for stress-final words. This means that the effect of a changed timing only becomes problematic at very fast articulation rates. Despite this small and insignificant processing difference between the two time-compressed conditions in Experiment 2, the results of Experiment 3 showed that listeners find the nonlinear type of artificial time compression, i.e., copy-fast-speech-timing, less agreeable to listen to than linearly time-compressed speech.

Given the results of Experiments 1 and 2, it seems unlikely that the results reported in (Janse et al., 2003) were due to an unwarranted extrapolation of the rules of fast speech timing to even faster rates. Natural fast speech timing rules do not improve intelligibility nor ease of processing, not even at the rate of speech at which they were observed (Experiments 1 and 2). Secondly, listeners prefer artificially linearly time-compressed speech over naturally produced fast speech, and over copy-fast-speech-timing time-compressed speech, even when the natural fast speech is still perfectly intelligible (Experiment 3). This strengthens our belief that the timing changes that accompany natural fast speech are due to articulatory restrictions, and do not serve a communicative purpose. In the mental representation, the articulatory or acoustic target values for stressed segments are more strictly specified than for unstressed segments. Consequently, the speaker is forced to spend more energy on approximating the specified targets for stressed syllables than for the more loosely defined unstressed syllables. Although a moderate increase in speech rate is not necessarily accompanied by target undershoot (van Son and Pols, 1990, 1992), a considerable increase in articulation rate (of, say, more than 20%) is almost inevitably accompanied by undershoot of the pre-defined targets because of the inertia of the speech organs (Lindblom, 1963; Moon and Lindblom, 1994). The increase in rate in the present experiment (1.4 times normal rate in Experiment 2) was clearly beyond 20%, so the fast speech must have been segmentally 'reduced'. Articulatory structures such as the jaw are relatively slow (cf. Perkell, 1997). Hence, if more precision is required for the stressed syllables than for the unstressed syllables, the speaker simply cannot speed up that much during the production of stressed syllables.

Secondly, the advantage of artificially time-compressed speech over naturally produced fast speech can also be attributed to overall 'reduced' articulation: the increased segmental slurring seems to hinder perception because it blurs segmental distinctions. Various studies by Marslen-Wilson et al. (1995), Gaskell and Marslen-Wilson (1996), and Gaskell and Marslen-Wilson (1998) suggested that, at a normal rate of speech, there is no perceptual advantage for assimilated versions over unassimilated articulations of a word form. Kohler (1990) describes assimilation as perceptually tolerated articulatory simplification. In that view, assimilation takes place in order to make the speaker's job easier, but only if the communicative situation allows it. This reflects the ideas laid down

in the H&H theory (Lindblom, 1990). At faster rates, however, unassimilated pronunciations were preferred (Quené and Krull, 1999). Bard et al. (2000a,b) also show how natural variation in pronunciation may affect lexical access: in a cross-modal identity priming task, reduced tokens taken from running speech primed less than clear list-read items, although both the clear and reduced forms showed robust priming. The more distorted word forms are, that is, the more they deviate from their normal-rate or 'canonical' form, the more difficult it is for the listener to map the incoming information onto the mental lexicon. Similar effects were found in studies in which one segment of a word was acoustically altered to produce a poorer phonetic exemplar (e.g., acoustic manipulation of the VOT value of initial plosive consonant, as in the word *cat*). Reduced activation was found for these acoustically altered word forms, compared to their respective intact versions (Andruski et al., 1994; Aydelott Utman et al., 2001). The more carefully articulated items add redundancy to the speech signal which is beneficial for perception in difficult listening situations. At a normal rate, this redundancy may not be really necessary, but at fast rates, listeners are obviously helped by the extra segmental information. The less words deviate from their normal-rate or 'canonical' form, the easier it is for the listener to map the incoming information onto the mental lexicon.

Several studies have provided evidence that speakers are not as listener-oriented as some have thought them to be. Sotillo and Bard (1998) examined pronunciations of landmark names to investigate whether reductions in pronunciation are less where lexical competition is greater. They did not even find a trend towards less reduction for words with greater competitor sets. Stronger evidence against the H&H claims comes from (Bard et al., 2000a,b), who found that listener's knowledge was irrelevant to the reductive effect of Givenness on duration and intelligibility of words in semi-spontaneous dialogues. Conversely, the Givenness effect on pronunciation was shown to depend only on what the speaker knew. This had also been found by Hawkins and Warren (1994): local phonetic variables (such as sentence accent and phonological and phonetic properties of individual segments) exert a greater influence on intelligibility than whether or not the word had been used before in the conversation.

These studies demonstrate that speakers are not always as cooperative as the H&H theory claims them to be. When speakers are under time or task pressure, this will only become worse. Horton and Keysar (1996) observed that time pressure made speakers indifferent to what listeners knew. However, this indifference may be caused by restrictions on speech production. Under time or task pressure, speakers may not have the time to compute the listeners' needs. Furthermore, the present results suggest that the way in which speakers speed up is the only possible way. Speakers may be aware of the fact that the way in which they speed up a message is not beneficial for listeners, but they have no other way to do it. Natural prosodic patterns do not contribute to speech intelligibility of fast speech. The explanation for the prosodic characteristics of natural fast speech is not to be found in the assumption that speakers always try to help their listeners. They rather result from the fact that speakers just cannot speed up in any other way. Speakers will therefore only choose to speed up when the communicative situation allows it.

As mentioned before, the way in which the timing of natural fast speech was implemented on normal-rate speech is only a global imitation of what speakers do. In Experiment 2, copying fast speech timing per syllable is a relatively fair approximation of what happens to the natural temporal pattern, but there are all kinds of non-linearities below syllable level. Consonants are reduced less than vowels and the steady-state parts of vowels may be reduced more than the transitions. In averaging the compression ratio over the syllable, all these small-scale effects are ignored. Further research is necessary to get a better insight into these lower-level effects. It must be kept in mind, however, that the increased articulatory overlap in natural fast speech can never be imitated by nonlinear compression, because segments become more and more coarticulated. Even though the way in which the timing of natural speech has been imitated in this study is still not entirely fair, it seems reasonable to assume that imitating in a more precise way what the speaker

does will not improve perception either. Listeners' preference for artificially linearly time-compressed speech, and the fact that making artificially time-compressed speech *less* natural (Unnatural Compression) is generally less harmful for perception than making it *more* natural, strongly suggest that what the speaker does in speeding up, is not necessarily what the listeners needs or prefers.

The results reported in this study provide some evidence that speeding up speech rate, globally or locally, results in a heavier processing load for the listeners. But in normal everyday communication, a locally increased rate can still be functional, in that slower speech rate generally signals new and important information, and faster speech rate signals given or redundant information (Lindblom, 1990). When speakers speak faster during more redundant words, it seems rather unlikely that this should *in the end* be problematic for listeners. Bard and colleagues argue that natural variation in word pronunciation is not noise, but useful information (Bard et al., 2000a,b; Bard et al., 2001). Duration, prominence, and segmental reduction provide cues as to whether words are presented in isolation or in context, where the phrase boundaries are, whether the word is predictable or redundant, etc. This information is mostly related to higher-level factors. In most theories of auditory word recognition, successful lexical access is dissociated from the recovery of the information contained in the variability in pronunciation. Bard et al. argue for a theory in which lower-level lexical processes may suffer from variability in pronunciation, and may even fail to resolve lexical competition. This leaves room for higher-level information to aid the process of lexical competition. Even though lexical access may suffer from a faster speech rate at some points during the sentence, higher-level knowledge comes in later to resolve ambiguities and to make overall comprehension of the message faster. Local variation in speech rate then is "not noise, but useful information". It seems plausible that increased difficulty in word processing in itself may be informative for higher-level processing of the message: the increased difficulty signals the givenness or the redundancy of the word in question, and thus provides information on its role in the message as a whole. However, a *global* increase in speech rate, instead of *local* rate variation, inevitably leads to an overall higher processing load for the listener, at probably all levels of processing.

## 6. Conclusion

Natural fast speech seems to be more difficult to process than linearly time-compressed speech. The results of the present study have shown that this holds even when the naturally produced fast speech is perfectly intelligible. The processing disadvantage of naturally produced fast speech is partly due to its changed timing, but also to its increased segmental overlap. Natural prosodic patterns do not contribute to speech intelligibility of fast speech: they rather result from the fact that speakers just cannot speed up otherwise. Secondly, increased segmental overlap seems to hinder speech processing at a fast rate. Assimilation and increased coarticulation reduce the redundancy of the speech signal, and they blur segmental distinctions. Segmental redundancy is beneficial for perception of fast speech.

Practically, this suggests that all attempts to make artificially time-compressed speech more intelligible or easier to process by making it more similar to natural fast speech may be in vain. The only aspect of naturally produced fast speech that should be imitated in order to make time-compressed speech more intelligible may be compression of pause duration. The results obtained with the Mach1 algorithm (Covell et al., 1998) and with other pause-reduction algorithms (He and Gupta, 2001) have shown that by compressing pauses more than the remaining speech, intelligibility can be improved over linear time compression at fast playback rates.

In every-day communication, faster articulation of words or phrases (i.e., a *local* increase in speech rate) is not just noise, but useful information. Even though lexical access may be hindered or delayed, the reduced pronunciation provides information on the word's role in the entire message. The care of pronunciation is often linked to the givenness or redundancy of a word. Thus, in the listener's head, higher-level information may interact with

low-level lexical access processes in order to perceive and understand the message as a whole. However, speakers can only afford to increase speech rate *globally* if they know that the listeners are willing and able to put extra effort into the listening process.

## Acknowledgements

Thanks are due to Sieb Nooteboom, Anne Cutler, Vincent van Heuven, Louis Pols, and Jacques Terken for their constructive comments on this work. I particularly thank Mirjam Ernestus and Hugo Quené for their suggestions on how to improve the readability of this manuscript.

## References

Andruski, J., Blumstein, S.E., Burton, M., 1994. The effect of subphonetic differences on lexical access. Cognition 52, 163–187.

Aydelott Utman, J., Blumstein, S.E., Sullivan, K., 2001. Mapping from sound to meaning: reduced lexical activation in Broca's aphasics. Brain Lang. 79, 444–472.

Bard, E., Anderson, A.H., Sotillo, C., Aylett, M.P., Doherty-Sneddon, G., Newlands, A., 2000a. Controlling the intelligibility of referring expressions in dialogue. J. Memory Lang. 42, 1–22.

Bard, E., Sotillo, C., Aylett, M.P., 2000b. Taking the hit: Why lexical and phonological processing should not make lexical access too easy. In: Proc. Workshop on Spoken Word Access Process., Nijmegen, The Netherlands. pp. 3–6.

Bard, E.G., Sotillo, C., Kelly, M.L., Aylett, M.P., 2001. Taking the hit: leaving some lexical competition to be resolved post-lexically. Lang. Cognit. Process. 16 (5/6), 731–737.

Cho, T., 2001. Effects of prosody on articulation in English. Doctoral Dissertation, University of California, Los Angeles.

Covell, M., Withgott, M., Slaney, M., 1998. Mach1: nonuniform time-scale modification of speech. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Process., Seattle.

Cutler, A., Clifton, C.E., 1984. The use of prosodic information in word recognition. In: Bouma, H., Bouwhuis, D.G. (Eds.), Attention and Performance X: Control of Language Processes. Erlbaum, Hillsdale, NJ, pp. 183–196.

Cutler, A., Koster, M., 2000. Stress and lexical activation in Dutch. In: Proc. 6th Internat. Conf. on Spoken Lang. Process., Beijing, Vol. 1. pp. 593–596.

Cutler, A., Norris, D., 1979. Monitoring sentence comprehension. In: Cooper, W.E., Walker, E.C.T. (Eds.), Sentence Processing: Psycholinguic Studies Presented to Merrill Garrett. Erlbaum, Hillsdale, NJ, pp. 113–134.

Cutler, A., van Donselaar, W., 2001. Voornaam is not (really) a homophone: lexical prosody and lexical access in Dutch. Lang. Speech 44 (2), 171–195.

Cutler, A., Mehler, J., Norris, D., Segui, J., 1987. Phoneme identification and the lexicon. Cognit. Psychol. 19, 141–177.

de Jong, K.J., 1995. The supraglottal articulation of prominence in English: linguistic stress as localized hyperarticulation. J. Acoust. Soc. Amer. 97 (1), 491–504.

Delogu, C., Paoloni, A., Sementina, C., 1992. Comprehension of natural and synthetic speech: preliminary studies. ESPRIT Project 2589 (SAM) Multilingual speech input/output assessment, methodology and standardisation; SAM Internal Report II.c. University College London, London.

Fowler, C.A., 1981. Production and perception of coarticulation among stressed and unstressed vowels. J. Speech Hear. Res. 46, 127–139.

Gaskell, M.G., Marslen-Wilson, W.D., 1996. Phonological variation and inference in lexical access. J. Exp. Psychol.: Human Percept. Perform. 22, 144–158.

Gaskell, M.G., Marslen-Wilson, W.D., 1998. Mechanisms of phonological inference in speech perception. J. Exp. Psychol.: Human Percept. Perform. 24, 380–396.

Gay, T., 1978. Effect of speaking rate on vowel formant movements. J. Acoust. Soc. Amer. 63 (1), 223–230.

Hawkins, S., Warren, P., 1994. Phonetic influences on the intelligibility of conversational speech. J. Phonet. 22, 493–511.

He, L., Gupta, A., 2001. Exploring benefits of non-linear time-compression. In: Proc. Conf. on Multimedia, Ottawa, pp. 382–391.

Horton, W.S., Keysar, B., 1996. When do speakers take into account common ground? Cognition 59, 91–117.

ITU-P.800. 1996. Methods for subjective determination of transmission quality. International Telecommunication Union (ITU): Recommendation P.800.

Janse, E., 2001. Comparing word-level intelligibility after linear vs. non-linear time-compression. In: Proc. VIIth European Conf. on Speech Comm. Technol. Eurospeech, Aalborg, Denmark, Vol. II. pp. 1407–1410.

Janse, E., Nooteboom, S., Quené, H., 2003. Word-level intelligibility of time-compressed speech: prosodic and segmental factors. Speech Comm. 41, 287–301.

Kohler, K., 1990. Segmental reduction in connected speech in German: phonological facts and phonetic explanations. In: Hardcastle, W.J., Marchal, A. (Eds.), Speech Production and Speech Modeling. Kluwer Academic Publishers, Dordrecht, pp. 69–92.

Lehiste, I., 1970. Suprasegmentals. MIT Press, Cambridge, MA.

Lindblom, B., 1963. Spectrographic study of vowel reduction. J. Acoust. Soc. Amer. 35, 1773–1781.

Lindblom, B., 1990. Explaining phonetic variation: a sketch of the H&H theory. In: Hardcastle, W.J., Marchal, A. (Eds.), Speech Production and Speech Modelling. Kluwer Academic Publishers, Dordrecht.

Marslen-Wilson, W.D., Nix, A.J., Gaskell, M.G., 1995. Phonological variation in lexical access: abstractness, inference and English place assimilation. Lang. Cognit. Process. 10, 285–308.

Max, L., Caruso, A.J., 1997. Acoustic measures of temporal intervals across speaking rates: variability of syllable- and phrase-level relative timing. J. Speech, Lang. Hear. Res. 40, 1097–1110.

Mehta, G., Cutler, A., 1988. Detection of target phonemes in spontaneous and read speech. Lang. Speech 31 (2), 135–156.

Moon, S.-J., Lindblom, B., 1994. Interaction between duration, context, and speaking style in English stressed vowels. J. Acoust. Soc. Amer. 96 (1), 40–55.

Nakatani, L.H., Dukes, K.D., 1973. A sensitive test of speech communication quality. J. Acoust. Soc. Amer. 53 (4), 1083–1092.

Nix, A.J., Mehta, G., Dye, J., Cutler, A., 1993. Phoneme detection as a tool for comparing perception of natural and synthetic speech. Comput. Speech Lang. 7, 211–228.

Norris, D., McQueen, J.M., Cutler, A., 2000. Merging information in speech recognition: feedback is never necessary. Behavioral Brain Sci. 23 (3), 299–370.

Pavlovic, C., Rossi, M., Espesser, R., 1990. Use of the magnitude estimation technique for assessing the performance of text-to-speech synthesis system. J. Acoust. Soc. Amer. 87, 373–381.

Perkell, J.S., 1997. Articulatory processes. In: Hardcastle, W.J., Laver, J. (Eds.), The Handbook of Phonetic Sciences. Blackwell, Oxford, pp. 333–370.

Peterson, G.E., Lehiste, I., 1960. Duration of syllable nuclei in English. J. Acoust. Soc. Amer. 32 (6), 693–703.

Pisoni, D.B., 1987. Some measures of intelligibility and comprehension. In: Allen, J., Hunnicutt, S., Klatt, D.H. (Eds.), From Text to Speech: the MITALK System. Cambridge University Press, Cambridge.

Pisoni, D.B., 1997. Perception of synthetic speech. In: van Santen, J.P.H., Sproat, R.W., Olive, J.P., Hirschberg, J. (Eds.), Progress in Speech Synthesis. Springer-Verlag, New York.

Port, R.F., 1981. Linguistic timing factors in combination. J. Acoust. Soc. Amer. 69 (1), 262–274.

Quené, H., Krull, J., 1999. Recognition of assimilated words in normal and fast speech. In: Proc. 14th Internat. Congress Phonet. Sci., San Francisco. pp. 1831–1834.

Seidenberg, M.S., Tanenhaus, M.K., 1979. Orthographic effects on rhyme monitoring. J. Exp. Psychol.: Human Learn. Memory 5, 546–554.

Slowiaczek, L.M., 1990. Effects of lexical stress in auditory word recognition. Lang. Speech 33 (1), 47–68.

Sotillo, C., Bard, E.G., 1998. Is hypo-articulation lexically constrained? In: Proc. SPoSS, Aix-en-Provence. pp. 109–112.

van Bergem, D.R., 1993. Acoustic vowel reduction as a function of sentence accent, word stress, and word class. Speech Comm. 12, 1–23.

van Bezooijen, R., van Heuven, V.J., 1997. Assessment of synthesis systems. In: Gibbon, D., Moore, R., Winski, R. (Eds.), Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin, pp. 481–563.

van Donselaar, W., Lentz, J., 1994. The function of sentence accents and given/new information in speech processing: different strategies for normal-hearing and hearing-impaired listeners? Lang. Speech 37 (4), 375–391.

van Heuven, V.J., 1985. Perception of stress pattern and word recognition: recognition of Dutch words with incorrect stress position. J. Acoust. Soc. Amer. 78, s21.

van Leyden, K., van Heuven, V.J., 1996. Lexical stress and spoken word recognition: Dutch vs. English. In: Cremers, C., den Dikken, M. (Eds.), Linguistics in the Netherlands. John Benjamins, Amsterdam.

van Santen, J.P.H., 1993. Perceptual experiments for diagnostic testing of text-to-speech systems. Comput. Speech Lang. 7 (1), 49–100.

van Son, R.J.J.H., Pols, L.C.W., 1990. Formant frequencies of Dutch vowels in a text, read at normal and fast rate. J. Acoust. Soc. Amer. 88 (4), 1683–1693.

van Son, R.J.J.H., Pols, L.C.W., 1992. Formant movements of Dutch vowels in a text, read at normal and fast rate. J. Acoust. Soc. Amer. 92 (1), 121–127.

Whalen, D.H., 1991. Subcategorical phonetic mismatches and lexical access. Percept. Psychophys. 50, 351–360.

Wingfield, A., 1975. The intonation-syntax interaction: prosodic features in perceptual processing of sentences. In: Cohen, A., Nooteboom, S.G. (Eds.), Structure and Process in Speech Perception. Springer-Verlag, Berlin.

Wingfield, A., Lombardi, L., Sokol, S., 1984. Prosodic features and the intelligibility of accelerated speech: syntactic versus periodic segmentation. J. Speech Hear. Res. 27, 128–134.