ELSEVIER

Speech Communication 41 (2003) 287–301

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Word-level intelligibility of time-compressed speech: prosodic and segmental factors

Esther Janse *, Sieb Nooteboom, Hugo Quené

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

Received 21 August 2001; received in revised form 9 June 2002; accepted 10 June 2002

Abstract

In this study we investigate whether speakers, in line with the predictions of the Hyper- and Hypospeech theory, speed up most during the least informative parts and less during the more informative parts, when they are asked to speak faster. We expected listeners to benefit from these changes in timing, and our main goal was to find out whether making the temporal organisation of artificially time-compressed speech more like that of natural fast speech would improve intelligibility over linear time compression. Our production study showed that speakers reduce unstressed syllables more than stressed syllables, thereby making the prosodic pattern more pronounced. We extrapolated fast speech timing to even faster rates because we expected that the more salient prosodic pattern could be exploited in difficult listening situations. However, at very fast speech rates, applying fast speech timing worsens intelligibility. We argue that the non-uniform way of speeding up may not be due to an underlying communicative principle, but may result from speakers' inability to speed up otherwise. As both prosodic and segmental information contribute to word recognition, we conclude that extrapolating fast speech timing to extremely fast rates distorts this balance between prosodic and segmental information.

Keywords: Perception; Time-compression; Prosody; Intelligibility; Timing; Fast speech

1. Introduction

Artificial time compression of speech, e.g. for the purpose of fast playback of long recordings, e-mails or voicemail messages, is normally performed in a linear way. This means that all segments are reduced by the same proportion, so that the relative timing pattern of speech played back at a fast rate is that of the original rate at which this speech was produced. It is a matter of debate whether speakers, when they are forced to speak faster than normal, also apply this approximately linear time compression. Kozhevnikov and Chistovich (1965) supported this notion of invariant timing patterns in speech movements. According to them, it is unrealistic to assume that there are separate motor programs for each rate at which an utterance can be produced. They therefore suggested that the rate of production may not be specified in the motor program but presents the "speed of realisation of the program" (Kozhevnikov and Chistovich, 1965). Kozhevnikov and Chistovich found temporal invariance for the relative duration of words in a phrase: regardless of speech rate, the duration

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

0167-6393/02/$ - see front matter
doi:10.1016/S0167-6393(02)00130-9

of each word was a constant proportion of the duration of the entire sentence. This temporal invariance was also found for the relative duration of the syllables in a word across different speech rates. However, they also found that the relative duration of the sounds within a syllable does vary as a function of speech rate.

Later studies found that rate-dependent changes in timing are not confined to the within-syllable level, but also occur between syllables and between words. When people speak faster, consonant durations are reduced less, relatively, than vowel durations (Lehiste, 1970; Gay, 1978; Max and Caruso, 1997). Furthermore, durations of sentence-stressed syllables are reduced less, relatively speaking, than durations of unstressed syllables (Peterson and Lehiste, 1960; Port, 1981). As a result, the relative difference in duration between stressed and unstressed syllables (i.e. the stressed/unstressed ratio) increases in faster speech, thereby making the prosodic pattern more prominent. This non-linear way of increasing speech rate might be 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 is then our particular interpretation of the Hyper- and Hypospeech theory (H&H theory). The H&H theory states that much of the variability of speech stems from the ways speakers adapt their speech to what they think that is needed by the listener to comprehend the message (Lindblom, 1990). On the one hand, the speaker wants to be understood, and this output-oriented goal forces him to use hyperspeech. On the other hand, speakers do not want to spend too much energy on redundant parts of speech, and this system-oriented, low-cost goal drives speakers to use hypospeech. In this way, the speaker continuously estimates how much care of articulation is minimally needed or is permitted by the audience.

If speakers, for communicative reasons, do indeed speed up most during the least informative parts of the sentence, and thereby assign a more prominent role to word-prosody, lexically stressed syllables might also be shortened less than unstressed syllables. In English the stressed syllable is the most informative syllable (Altmann and Carter, 1989), and it is likely that the same goes for Dutch (cf. van Heuven and Hagman, 1988). Furthermore, if the H&H principle of preserving the most informative parts holds, unaccented words would be affected more by an increase in speech rate than accented words. As unaccented words often refer to already given information, one might expect the speaker to choose a higher speech rate during unaccented words than during new and highly informative accented words.

We assume that, according to our particular interpretation of the H&H theory, speakers make prosodic patterns more pronounced in order to help the listeners. The non-linear way of speeding up is taken to be a strategic communicative principle, and thus, we also expect the listeners to benefit from these changes in timing. Consequently, the intelligibility of artificially time-compressed speech is expected to be improved if its temporal organisation is closer to that of natural fast speech. In other words, if speakers, or experimenters, i.e. by manipulation, deliberately assign a more prominent role to word-level prosody, this should be helpful to listeners.

The importance of prosody, and thus of temporal organisation, in word recognition and sentence processing in normal listening conditions has been shown in several studies. Cutler and Clifton (1984); van Heuven (1985) and Slowiaczek (1990) showed that word recognition is delayed when words in stress languages such as English and Dutch are deliberately mis-stressed. Cutler and Koster (2000) showed that stress information plays an important role in lexical activation in Dutch, and Cutler and van Donselaar (2001) also showed that listeners effectively use suprasegmental cues in Dutch. Under difficult listening conditions, prosodic factors seem to play an even more important role than they normally do (Wingfield, 1975; Wingfield et al., 1984; van Donselaar and Lentz, 1994). When the speech signal is degraded, prosodic information is usually preserved better than segmental information because it is spread over larger chunks of the speech signal. Furthermore, prosodic information is relatively well preserved in degraded speech also because the information (such as silence, pitch) is spread out over the entire spectrum. A degraded speech signal may therefore cause listeners to rely more on prosodic cues than

when speech quality is high. Secondly, correct sentence-level phrasing is helpful in the understanding of artificially time-compressed speech (Wingfield et al., 1984). Wingfield (1975) showed that intelligibility of sentences with anomalous intonation declined steeply as time compression increased, whereas the decline was much more gradual for sentences with normal intonation. Wingfield explains this in terms of the correct intonation pattern adding redundancy to the speech signal: this redundancy can be exploited in difficult listening situations. Furthermore, van Donselaar and Lentz (1994) investigated the use of the interdependence between information and accent structure and how this is affected by speech intelligibility. Hearing-impaired subjects interpreted the accented words as being new, regardless of their information value. Only when speech quality was degraded, did the normal-hearing subjects switch to the strategy of interpreting accented words as being new: they also made use of the interaction between information and accentuation.

There is at least one study that seems to show that imitating natural fast speech timing leads to significant improvement over linear time compression at very heavy rates of time compression. The time-compression algorithm Mach1 (Covell et al., 1998) is based on the compression strategies found in natural fast speech timing, such as compressing pauses most and compressing stressed (i.e. sentence-accented) vowels least. Covell et al. (1998) compared comprehension and preference for Mach1-compressed and linearly time-compressed speech. Mach1 did not only offer significant improvement in comprehension over linear compression, but was also preferred by the listeners. However, it is not entirely clear what the contributions are of each of the non-linear compression strategies at the different levels. By compressing pauses most, the remaining sentence or paragraph of text can be compressed less, in order to be equal in total duration, than in case of linear time compression. The lower articulation rate in the Mach1 compressed sentence is likely to make it more intelligible than the linearly time-compressed sentence. In other words, it is not clear what the separate contributions are of the word-level, sentence-level and paragraph-level changes in timing to the improvement in comprehension. In this paper we will only focus on the word-level changes in timing between normal and fast speech rate. The main question is whether taking these into account can improve the intelligibility of artificially time-compressed speech.

For the present study, we assume that speakers behave in line with the H&H theory, and assign extra importance to the most informative parts when they are forced to speak fast. This leads us to the following hypotheses:

1. When speaking at increased rate, speakers will reduce lexically unstressed syllables more, relatively, than stressed syllables.
2. The durational correlate of pitch accent will become more prominent at faster speech rates because unaccented words (referring to 'given' information) are reduced more, relatively, than accented words (containing 'new' information).
3. Word-level intelligibility of artificially time-compressed speech can be improved by taking into account the changes in temporal organisation going from normal to fast speech.

To investigate our hypotheses 1 and 2, we established how word-level timing is affected by an increase in speech rate in Dutch. This production study is described in Section 2. In Section 3 a perception experiment, set up to test the third hypothesis, is described.

2. Fast speech timing

2.1. Introduction

When asked to speak fast, speakers may have different ways to achieve this goal. It is assumed here that whatever speakers do when speaking fast, they will always choose a communicative strategy in accordance with the H&H principle. Hence, speakers will speed up most during the least informative parts of their speech. For practical reasons, this study focused on the shortening behaviour of stressed and unstressed vowels (and not syllables). This avoids the problem of resyllabification at syllable boundaries. Another practical

advantage is that vowels are relatively easy to segment in the wave form. The main reason for measuring vowels is that increasing speech rate has its strongest effect on the duration of the vowels, as these are the most elastic segments. As the vowel's duration thus mainly determines the length of the syllable, we compared the difference between the shortening behaviour of stressed and unstressed syllables by looking at vowel durations.

In English, most unstressed vowels are reduced to schwa: lexical stress has a strong effect on the colour of the vowel. In Dutch, on the other hand, vowel quality is less dependent on the stress level of the syllable: the unstressed syllable may well contain a full vowel (Kager, 1989; van Bergem, 1993). In order to investigate whether there is a difference between the shortening behaviour of unstressed full vowels or schwa, we measured the durations of Dutch disyllabic words containing schwa and of words containing two 'full' vowels. This was done to ascertain that the syllables are reduced according to their stress level, and not because of their segmental quality.

2.2. Method

2.2.1. Material

Thirty-two disyllabic nouns were selected in order to measure the durations of stressed and unstressed vowels: half of them with schwa, and half of them with full unstressed vowels. Stress position and vowel length were balanced wherever possible. The 32 target nouns are listed as Appendix A. The target words were embedded in long meaningful sentences because pilot work had shown that it is easier to attain a high speech rate in longer sentences. The target words appeared at the beginning of the sentence to avoid final lengthening. The sentences were recorded in two conditions: one in which the disyllabic target word had a pitch accent, and one in which the word was unaccented. A context sentence preceded the test sentence to indicate which words were to receive pitch accent in the following sentence. In the [−pitch accent] condition, the first adverb of the test sentence received pitch accent instead of the target word. The sentence structure was always the same, and so was the position of the target word.

2.2.2. Speakers

Four female native speakers of Dutch were asked to read the test material aloud at normal and very fast speech rates. They were paid for their participation.

2.2.3. Procedure

The recording session lasted about an hour and a half. First, the speakers were asked to read the material at a normal rate. If accentuation was not correct or if the sentence was not read out fluently (as judged by the experimenters), the speaker was asked to repeat the sentence. After all material had been recorded at the normal rate, the speaker was asked to aim for a fast speech rate without abnormal slurring. Speakers were encouraged to use a stopwatch, so they could get an impression of how fast they could speak, and they could try to outdo themselves in their speech rate. In order to increase the speech rate, they were asked to read out each sentence four or five times, and to keep an eye on the articulation time for each attempt. Again, the experimenters judged the speaker's performance. The material was recorded onto digital audiotape in a sound-proof cabin with a Sennheiser ME 10 microphone. The speech was then copied onto a computer disk and downsampled to 16 kHz.

2.2.4. Duration measurements

One version of each test sentence was selected on the basis of the correctness of the accent pattern. As there were about four versions of each sentence pair in the fast rate condition, the fastest trial with a correct accent pattern was selected for analysis. The durations of the stressed and unstressed vowels were measured. Based on the waveform and spectrogram displays, the criteria formulated below were used for the segmentation. In all cases, the segmentation was verified by auditory feedback. The vowel onset corresponded with the first positive zero crossing of the first periodic waveform at which an increased amplitude and a clearly visible change in the wave form due to a change in the harmonic structure occurred. The offset of the

vowel corresponded with the positive zero crossing of the last periodic waveform before the following plosive or fricative started. In case the target word ended in a vowel, the next word always started with a plosive or a fricative. Some unstressed schwas were followed by a coda /r/ consonant (e.g. the target word *beker* 'beaker'). The durations of these vowels were very difficult to measure because of the short duration of the syllables containing schwa and because of the vowel-like articulation of /r/ in coda position. Segmentation was rather difficult in the fast speech because of heavy coarticulation. In case no periodic vowel signal could be traced in the waveform and spectrogram, a minimum duration of 5 ms was postulated. This was established as a minimum duration because it corresponded to about one period (as the speakers were female with an average F0 of about 200 Hz). Furthermore, this minimum duration of 5 ms enabled us to compute fast/normal ratios, which would have been impossible if we had assumed a duration of 0 ms.

All measurements were carried out by two undergraduate students in phonetics who checked each other's measurements. These measurements were then checked by the first author. In most cases, the difference between the boundary locations was <10 ms. In case the measurements differed more than 10 ms (e.g. before /r/ or in the fast condition), the three judges decided on a 'compromise' duration. This procedure ensures a relatively high reliability of the vowel duration measurements.

Articulation rates were computed for the normal and fast speech rates by measuring the duration of the first part of the test sentence (containing the target word) and dividing it by the number of syllables (as counted in the canonical written version). The average articulation rate was 6.7 syllables/s at the normal speech rate, and it was increased to 10.5 syllables/s at the fast rate.

2.3. Results

Our first hypothesis was that as speech rate increases, the relative shortening of unstressed syllables is greater than that of stressed syllables. The mean durations of stressed and unstressed vowels at normal and fast speech rate are shown in Table 1 and in Fig. 1, together with the fast/normal ratios (vowel duration at fast rate/duration at normal rate).

Table 1
Mean durations (in ms) of stressed and unstressed vowels at normal and fast speech rate. Mean fast/normal ratios are also given

	Normal rate	Fast rate	Fast/normal ratio
Stressed vowel	114	75	0.67
Unstressed vowel	55	21	0.42

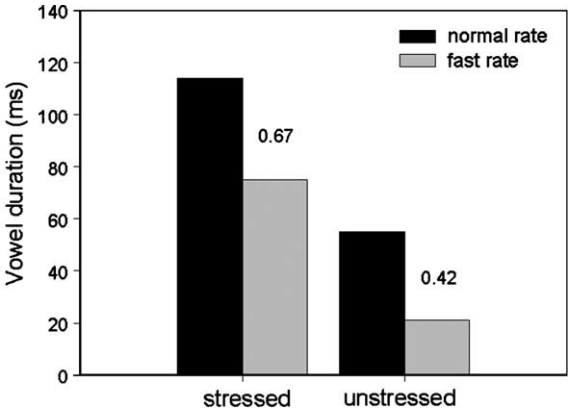

Fig. 1. Mean vowel durations (in ms) of stressed and unstressed vowels, broken down by speech rate. Fast/normal ratios are given above the paired bars.

At fast speech rate, unstressed vowels were reduced to 42% of their normal rate duration, and stressed vowels were reduced to 67% of their normal rate duration.

It was checked whether increasing speech rate had a similar non-linear effect on the level of the entire syllable, and not only on the vowel durations. The syllable durations of the disyllabic target words in the [+pitch accent] conditions were measured for one speaker. In the fast rate condition, the stressed syllable was reduced to 64% of its normal rate duration, and the unstressed syllable was reduced to 45%. These data suggest that the entire syllable is reduced according to its stress level.

The fast/normal ratios of the stressed and unstressed vowels (within each item, per vowel) were analysed in two Repeated Measures ANOVAs,

one on the 32 items and one on the four speakers, with stress (stressed vs. unstressed) and accent (accented vs. unaccented) as fixed factors. The analyses show a significant effect of stress on the fast/normal ratios ($F_1(1,3) = 158.6$, $p = 0.001$; $F_2(1,31) = 64.0$, $p < 0.001$).

Half of the disyllabic words contained two full vowels, and these were balanced for vowel length. For this subset of items, the fast/normal ratios of the stressed and unstressed syllables show the same difference, with ratios of 0.66 and 0.42 for the stressed and unstressed vowels, respectively. Thus, the first hypothesis is confirmed: regardless of vowel length, unstressed vowels in fast speech are compressed more, relatively, than stressed vowels.

To make sure that syllables are reduced according to their stress level, and not because schwa may be more compressible than other full unstressed vowels, the compression behaviour of the two types of unstressed vowels was checked. In Table 2 the duration results are shown for the two types of unstressed vowels.

Table 2 shows that the fast/normal ratio of unstressed 'full' vowels equals that of unstressed schwa vowels. The fast/normal ratios of the unstressed vowels were entered into repeated measures ANOVAs on items and on speakers, with vowel type and accent as fixed factors. In the item analysis, the items were nested under vowel type (schwa vs. full vowel). The effect of vowel type on the fast/normal ratios is not significant ($F_1(1,3) < 1$, n.s.; $F_2(1,30) < 1$, n.s.). Although the absolute duration of schwa was shorter on average than that of the full unstressed vowels at both rates, schwa is not more compressed than full unstressed vowels.

The second hypothesis was that vowels in words bearing a pitch accent on the stressed syllable reduce relatively less, with increasing speech rate, than vowels in words without a pitch accent. In Fig. 2 mean vowel durations of the stressed and unstressed vowels are shown, at both speech rates, and in [+pitch accent] and [−pitch accent] conditions.

Table 2
Mean vowel duration (in ms) of unstressed vowel at normal and fast speech rate (plus fast/normal ratio)

	Normal rate	Fast rate	Fast/normal ratio
'Full' unstressed vowel	66	24	0.42
Schwa	44	17	0.41

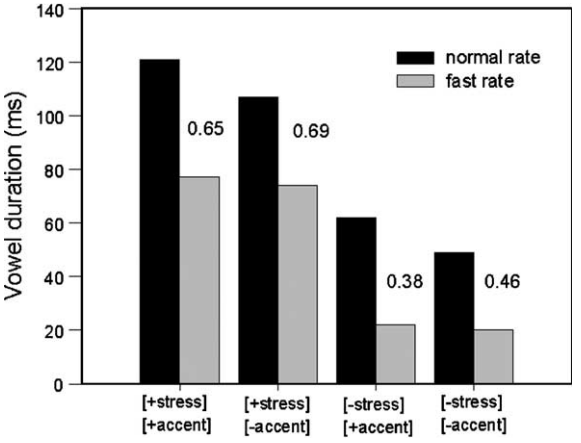

Fig. 2. Mean vowel durations (in ms) of stressed [+stress] and unstressed vowel [−stress] in both [+pitch accent] and [−pitch accent] condition, at both speech rates. Fast/normal ratios are given above the paired bars.

The analyses of variance on the fast/normal ratios with stress and sentence accent as fixed factors show that the main effect of accent does not reach significance ($F_1(1,3) = 7.07$, $p = 0.076$; $F_2(1,31) = 5.62$, $p = 0.024$). However, there is a trend for [+accent] vowels to be compressed relatively more in fast speech than the [−accent] vowels, which is the opposite of what we had expected. Fig. 2 shows a tendency for the lexically unstressed vowels to be affected more by the factor accent, but this interaction between Stress and Accent was not significant in the item analysis ($F_1(1,3) = 30.5$, $p = 0.012$; $F_2(1,31) < 1$, n.s.). In sub-analyses on the fast/normal ratios of the unstressed vowels only, with accent and vowel type as fixed factor, the effect of accent was not significant either ($F_1(1,3) = 10.0$, $p = 0.051$; $F_2(1,30) = 2.71$, n.s.). So, whereas other studies have found that accentual lengthening of vowels plays an important role at normal rate, our present results show that this duration difference becomes relatively smaller at a fast speech rate. The durational correlate of lexical stress becomes more prominent at

faster speech rates, but the durational correlate of sentence or pitch accent seems to become less prominent with increasing speech rate.

2.4. Discussion

As argued above, speakers tend to reduce the duration of unstressed syllables more than that of sentence-stressed syllables when speech rate is increased. Our aim was to find out whether speakers indeed show a selective compression behaviour which is completely along the lines of the H&H theory, namely that they speed up most during the parts which are least informative. The first expectation was that speakers show a greater relative reduction in the duration of the unstressed syllable than of the stressed syllable. In a study on Dutch by den Os (1988), increasing speech rate had a greater effect on stressed vowels than on unstressed vowels. However, den Os (1988) may have underestimated the relative shortening of unstressed vowels because the fastest unstressed vowels, which were too short to be measured, were disregarded. The present results show that unstressed vowels are reduced more than stressed vowels, and that this is not an artefact of unstressed vowels often being schwa. The relative amount to which a vowel is reduced with increasing tempo depends mainly on the stress level of the syllable, and not on the quality of the vowel.

A similar non-uniform compression was expected for the sentence-accented vs. unaccented words. We expected unaccented words to shorten relatively more than accented words so that the most informative parts of the sentence are preserved. We measured vowel durations of disyllabic words, which either did or did not have a pitch accent on the lexically stressed syllable. Contrary to our hypothesis, we found a trend for the relative duration difference between vowels in accented vs. unaccented conditions to become smaller at faster speech rate. In other words, the durational correlate of pitch accent becomes less prominent at faster speech rates. The durational correlate of pitch accent may be sacrificed when the speaker is pressed for time. For lexical stress, the duration cue is the most important one. However, to indicate which words are accented in a sentence, the pitch excursion itself is a much more important cue than duration (Sluijter, 1995). One should note that the results concerning the durational aspect of accent are strongly linked to the design of our duration study: we compared the duration reduction of disyllabic content words in [+pitch accent] and [−pitch accent] condition. If we had compared the reduction of these content words with the reduction of function words, such as articles or auxiliary verbs in the same phrase, we might have found some important changes in phrase level timing.

In summary, increasing speech rate is accompanied by important changes in word-level timing in Dutch. The next section deals with the question whether taking these timing changes into account can improve the word-level intelligibility of artificially time-compressed speech. In the Introduction section, several studies were mentioned that showed that the role of prosodic factors becomes more important under difficult listening conditions because prosodic information is preserved better than segmental information (Wingfield et al., 1984; Wingfield, 1975; van Donselaar and Lentz, 1994). Thus, in the next section a perception experiment is described to test the hypothesis that the more salient word-level prosodic pattern found in natural fast speech is helpful to listeners who are presented with artificially time-compressed speech.

3. Word-level intelligibility at a very fast speech rate

3.1. Introduction

The duration study described in the previous section shows that the prosodic pattern at word-level is made more prominent with increasing speech rate. These production data then lead to the expectation that the intelligibility of artificially time-compressed speech will be improved if its temporal organisation is closer to that of natural fast speech.

Experiments in our laboratory have shown that speech remains intelligible at rates that are much faster than can ever be attained in natural fast speech. Speech that is artificially time-compressed to the fastest rate which human speakers can

achieve is still almost perfectly intelligible. It would seem reasonable to evaluate the perceptual effects of applying fast speech timing to time-compressed speech at the fast rate which is produced by the speakers. However, we have two reasons not to do this. First, a practical reason is that intelligibility is very high, even at rates twice the normal rate. This ceiling effect would make any difference in intelligibility between linearly time-compressed and non-linearly time-compressed speech fairly difficult to find. Second, a more fundamental reason is that the role of prosody is expected to become more important as the listening situation becomes more difficult. The information carried by the more salient prosodic pattern might be exploited in difficult listening situations. For these two reasons, the rules of fast speech timing were extrapolated to even faster rates.

As argued above, on the one hand, one might expect that if temporal prosody is assigned a more prominent role in speech production at fast rates, such fast speech timing will also become more helpful in perception of fast speech. Speakers of fast speech are expected to speed up most during the parts that are least informative, and preserve the more important parts. Yet, on the other hand, at very fast rates of speech, prosody and segmental information play conflicting roles. Prosody requires that some syllables are longer and more prominent than others. Weak unstressed syllables will therefore be the first to become highly unintelligible after time compression, even more so when these syllables are compressed more than stressed syllables. Cutler and van Donselaar (2001) show that, although Dutch listeners make use of the suprasegmental cues in word recognition, the contribution of segmental information probably outweighs that of suprasegmental information. We should also consider the possibility that the speakers' non-linear way of speeding up is not so much caused by a communicative strategy, but rather results from articulatory factors. It may be the case that speakers simply cannot speed up in an approximately linear fashion. Although speakers may normally tailor their utterances to internal representations of the listeners' needs, there is some evidence that speakers do *not* do this under time or task pressure (Horton and Keysar, 1996). This could then mean that natural prosodic patterns do not always contribute to speech intelligibility. Thus, an alternative possibility is that segmental intelligibility plays such an important role that listeners are helped more, paradoxically, by an entirely unnatural compression strategy, namely by compressing the lexically stressed syllable relatively more than the lexically unstressed syllable (which is short already). This would preserve the segmental intelligibility of both syllables. Thus, three strategies for time compression need to be considered to evaluate these possibilities. An experiment was set up to compare the intelligibility of speech after linear compression (compressing all syllables to the same degree); after selective compression based on natural fast timing (compressing unstressed syllables relatively more than stressed syllables); and after unnatural compression (the reverse of selective compression: compressing stressed syllables more than unstressed syllables).

The hypothesis was that the word-level intelligibility of strongly time-compressed speech can be improved by taking into account natural fast speech timing which assigns more importance to the most informative parts in the speech stream (selective compression).

A competing hypothesis is that word-level intelligibility is not improved by making its timing more like that of natural fast speech because the change in timing is not a communicative strategy, but due to articulatory restrictions.

3.2. Method

Short sentences were constructed containing a target word which was to be identified in an intelligibility test. The short sentences were often the first clause of a longer sentence. The target words were of low semantic predictability in the sentence.

The PSOLA time-scaling technique (Charpentier and Stella, 1986), as implemented in the speech-editing program GIPOS (version 2.3) was used to time-compress the speech fragments (http://www.ipo.tue.nl/ipo/gipos/).

The three types of compression were applied to the entire sentences: each syllable was assigned a

plus or minus stress mark, and was then time-compressed accordingly. The broad distinction between function words and content words was used as a criterion to assign a stress level to each syllable. Auxiliary verbs and articles were assigned [−stress], whereas the main verb and other content words were assigned [+stress]. For the polysyllabic words with [+stress], only the stressed syllable received a [+stress] mark. Example sentences are presented in (2): the target word is in bold.

(2) Hij+had-de-**par-tij**+moe+ten-ver-nie+ti-gen-
(*He should the **batch** have destroyed*)
Het-pak-ket+bleek+**me-taal**+te-be-vat+ten-
(*The package appeared **metal** to contain*)
Ook+is-er-een-**mo-del**+te-vin+den-
(*Also is there a **model** to be found*)

For the selective compression condition, syllables with [+stress] marks were compressed less (to 65%) than [−stress] syllables, which were compressed to 45% of their original duration. In the speech-eding program GIPOS, selected parts of a speech waveform can be time-compressed, while, at the same time, the rest of the waveform remains unaffected. Thus, the sentence can be time-compressed syllable by syllable.

For the unnatural compression the compression strategy based on the plus and minus marks was reversed such that the [−stress] syllables were compressed less (to 65%) than the stressed syllables (to 45%). After this non-linear time compression, the entire word and sentence durations were measured and the word duration and the rest of the sentence were compressed further to attain a compression rate of 35% (a pilot experiment with this material had shown that only at compression to 35% of the original duration the intelligibility scores would be around 50% correct). This was done separately for the target words, such that the target word duration would be the same in all three compression conditions. For the linear time-compression condition, all syllables were compressed to the same degree.

There were 144 monomorphemic disyllabic targets (all nouns); embedded in sentences. Half had initial stress, and half had final stress. Since each subject could be presented with the same item only once, there were three experimental lists. On each list, the three compression conditions were distributed across the 72 initially stressed targets and across the 72 finally stressed items.

3.2.1. Subjects

Thirty three subjects were assigned to the three experimental lists: 11 for each list. They were tested individually in a sound-proof booth. Subjects were all students at Utrecht University. The material was presented to them over closed earphones. They were paid for their participation.

3.2.2. Procedure

A practice session of 20 items preceded the actual test session so that the subjects could adapt to the fast speech rate before the test began. The order of the items was randomised for each subject in order to cancel out a possible learning effect during the test. Monosyllabic fillers and filler targets with three syllables were interspersed in the material, so that subjects would not notice that all test words were disyllabic. First the entire sentence was presented on the screen, with a blank at the position of the target word plus its accompanying article. The article was also left out because the definite article in Dutch provides information about the grammatical gender of a word. Subjects were given sufficient time to read the visual presentation. After 3 s, the whole time-compressed sentence was presented to them auditorily, including the target word. Subjects were asked to fill in the missing word by typing on a keyboard. There was no time pressure: only after they had hit the Enter key, the following sentence would appear on the screen. After the practice session, subjects could still ask questions if anything was unclear. The entire experiment lasted about 30 min.

3.3. Results

The correct responses per condition were computed: the raw percentages correct are shown in Fig. 3.

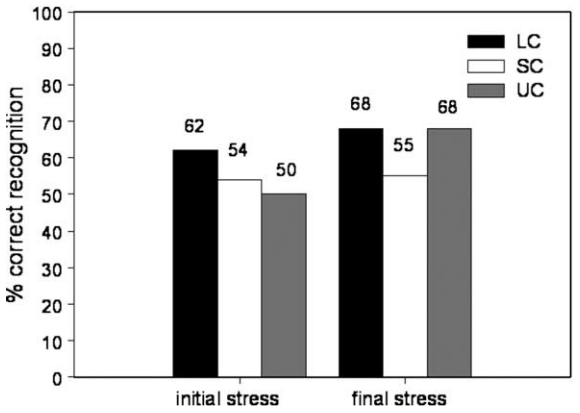

Fig. 3. Percentage of correct recognition, broken down by compression type and stress position.

The percentages correct were entered into analyses of variance (after an arcsine transformation), treating either subjects or items as random factors. The analyses of variance show that the main effect of stress position does not reach significance ($F_1(1,32) = 63.20$, $p < 0.001$; $F_2(1,142) = 3.35$, $p = 0.069$). Fig. 3 suggests that words with final stress are, on the whole, somewhat easier to recognise. This may be due to a sparse neighbourhood effect: words with final stress tend to have fewer close neighbours in the Dutch lexicon and hence may be recognised more easily. The main effect of compression type is highly significant ($F_1(2,31) = 15.4$, $p < 0.001$; $F_2(2,141) = 13.2$, $p < 0.001$). Fig. 3 shows that linear compression yields the highest intelligibility. Furthermore, there is a significant interaction between the effect of compression type and stress position ($F_1(2,31) = 10.33$, $p < 0.001$; $F_2(2,141) = 6.89$, $p = 0.001$). Making the prosodic pattern less pronounced (i.e. UC compression) decreases the intelligibility of words with initial stress, but it does not have a negative effect on the intelligibility of words with final stress. This pattern of results for finally stressed words can be explained as the outcome of two effects, working in opposite directions. The first effect is that of temporal alignment. The second effect is related to 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. Misstressing initially-unstressed Dutch words is more disrupting than misstressing initially stressed words (van Leyden and van Heuven, 1996). 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 in the UC condition, compared to the LC condition. Secondly, the duration of the stressed syllable is shorter in the UC condition than in the other two conditions. So, 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 intelligibility pattern observed for initially stressed words. Initial alignment with possible word candidates is more difficult because the first syllable is shorter. As the first syllable is in this case also the stressed syllable, intelligibility is only affected negatively.

Further confirmation for these two tendencies can be found in the distribution of incorrect responses. A closer study of the false responses shows that in the majority of cases either the correct stress pattern was reported, or subjects responded with a monosyllabic (stressed syllable) answer. The selective compression condition is expected to make correct alignment of word onset rather difficult for words with final stress because of the very short duration of the first unstressed syllable. The percentages of monosyllabic responses for the three types of compression are shown in Table 3.

The percentages monosyllabic responses, in all three compression conditions and in both stress conditions, were arcsine transformed. These transformed data were fed into analyses of variance with either item (nested under stress position) or

Table 3
Percentages of monosyllabic responses, broken down by stress position and compression type

	LC	SC	UC
Initial stress	15	17	14
Final stress	12	23	8

subject as random variable, and compression type and stress position as fixed factors. There was no significant effect of stress position on the percentage of monosyllabic responses ($F_1(1,32) = 2.99$, $p = 0.093$; $F_2(1,142) < 1$, n.s.). The effect of compression type, however, was significant ($F_1(2,31) = 22.1$, $p < 0.001$; $F_2(2,141) = 14.7$, $p < 0.001$), and so was the interaction between stress position and compression type ($F_1(2,31) = 9.15$, $p = 0.001$; $F_2(2,141) = 7.56$, $p = 0.001$). Pairwise t-tests showed that the percentage of monosyllabic responses was significantly higher in the SC condition than in the LC condition ($t_1(32) = -5.57$, $p < 0.001$; $t_2(143) = -4.13$, $p < 0.001$). The SC time-compressed speech elicits significantly more monosyllabic, or truncated, responses than the linearly compressed condition. In the SC condition, the unstressed syllable is reduced to such a short duration that, in some cases, it may be perceptually obliterated. Acoustically, there is something left of the syllable, but perceptually these very short syllables may almost 'fall out' of the signal. This is mainly the case for words with final stress. Alignment with word candidates clearly fails here because of the very short duration of the unstressed first syllable.

Lastly, the question remains whether the UC condition was successful in preserving the segmental intelligibility of both syllables. In an earlier pilot study, intelligibility of disyllabic non-words was evaluated in these same three artificial time-compression conditions. Segmental intelligibility was evaluated per syllable, such that these non-word results might give some insight into the effect of the UC condition on the stressed and on the unstressed syllables. The results showed that the percentage correct identification for the unstressed syllable is higher in the UC than in the LC condition. The longer the syllable duration remains after compression, the higher the segmental intelligibility. Still, UC's positive effect on the unstressed syllable is counterbalanced by a negative effect on the identification of the stressed syllable. The same two effects might apply to the present real word results: the identification of the unstressed syllable is higher, but at the expense of that of the stressed syllable. The net results is, obviously, that word intelligibility cannot be improved by time-compressing in this 'unnatural' way.

Overall, one can conclude that linear compression wins. For both words with initial and final stress, making the prosodic pattern of the disyllabic target words more like natural fast speech timing (SC) does not improve intelligibility over LC. Giving priority to the segmental intelligibility of both the stressed and the unstressed syllable (UC) does not improve word intelligibility either.

4. Discussion

Given the results of our production study, namely, that speakers speed up along the lines of the H&H theory, we expected that imitating the more salient prosodic pattern found in natural fast speech would improve word-level intelligibility over linear time compression. The alternative option was also investigated, namely, that intelligibility would be improved by a very unnatural timing: if segmental intelligibility of all stressed and unstressed segments outweighs the contribution of the prosodic pattern, intelligibility might be better if prosodic timing differences are made smaller than found in natural speech. This question was addressed in an intelligibility test with highly time-compressed speech material. The idea was that in these extreme listening conditions (time compression to about three times normal rate) speakers would benefit most from the more salient word-level prosodic pattern.

The results of the perception study did not confirm our hypothesis: the word-level intelligibility of artificially time-compressed speech could not be improved by using either type of non-linear time compression over linear compression. The fact that intelligibility is generally lower in the unnatural compression condition than in the linear compression condition shows that making the prosodic pattern less pronounced for the sake of the unstressed syllable's intelligibility does not help word identification either. This might be due to the fact that now the segmental intelligibility of the most informative syllable suffers. A longer duration of this most informative syllable in selective

compression did not improve word intelligibility over linear compression either. The temporal pattern found in natural normal speech rate, and its linear transforms at faster rates, gives the best overall word intelligibility. Note, however, that this may not be the only optimal pattern: words with stress on the final syllable were identified equally well in the unnatural compression condition.

Now, two issues need to be explained. First, how do our results fit in with the results obtained with the time-compression algorithm Mach1 (Covell et al., 1998)? Second, what about our particular interpretation of the H&H theory, namely that speakers make prosodic patterns more pronounced in order to help the listener?

Covell et al. showed that it is possible to obtain a significant increase in intelligibility over linear compression at *heavy* rates of time compression. Mach1 is based on the compression strategies found in natural fast speech timing, such as compressing pauses most and compressing stressed (i.e. sentence-accented) vowels least. Moreover, their algorithm was built so as to avoid overcompressing already rapid sections of speech. This suggests that it is beneficial for intelligibility if the prosodic pattern is not entirely at the expense of the segmental information. The fact that Mach1 could improve intelligibility of strongly time-compressed speech by imitating fast speech timing whereas we could not may be due to the following. The increase in intelligibility of Mach1 could mainly be caused by those aspects of fast speech timing that exceed the word level (such as pause reduction). In order to achieve the same fragment duration, the remaining speech can be time-compressed less after the pauses have been removed than in case of linear time compression. The positive effect of the slower articulation rate on intelligibility outweighs the perceptual importance of speech pauses. Furthermore, at sentence level one might find that applying the speaker's strategy of reducing function words more than content words improves sentence-level intelligibility, both at moderately fast, and at very fast rates of speech. Thus, the results obtained with Mach1 cannot really be compared with our results. The positive effect of pause removal, and thus of preserving segmental information, may be so important that it outweighs all other possible negative effects of imitating natural fast speech timing.

There are at least two possible explanations why the natural way of speeding up in the present experiment does not lead to improved intelligibility in heavily time-compressed speech. The first possible explanation is that selective compression would in fact have improved intelligibility at speech rates humans can achieve, but not at the very fast rate employed in our study. This would then still fit in with our expectation that the more prominent prosodic pattern is helpful for perception. Remember that we extrapolated the changes in timing from the moderately fast speech rate reached by our speakers to a much faster speech rate. Apart from the practical reason of avoiding ceiling effects in intelligibility, this was also done because we expected a degraded speech signal to cause listeners to rely more on prosodic cues than when speech quality is high. This very fast rate cannot be attained by human speakers, but listeners are still quite capable of filling in the missing words. However, by extrapolating fast speech timing to very fast rate, the unstressed syllable may become so short that its segmental information is badly damaged or even obliterated. Some evidence for this 'perceptual obliteration effect' was found in the fact that the selective compression condition elicited higher percentages of monosyllabic responses to the disyllabic target words than the linear compression condition. Disyllabic words can often not be recognised on the basis of the stressed syllable alone. Results reported in Cutler and van Donselaar (2001) show that segmental mismatch weighs rather more heavily than suprasegmental mismatch in word recognition under normal listening conditions. Both prosodic and segmental information contribute to recognition of the disyllabic words. So, when fast speech timing is extrapolated to extremely fast rates, the prosodic pattern may cause the segmental information of unstressed syllables to be perceptually obliterated.

A second explanation is that our interpretation of the H&H theory is wrong. According to our particular interpretation, speakers make prosodic

patterns more pronounced in order to help the listener. However, the non-linear way in which speakers speed up at word level may not be as strategic and communication-driven as we thought. It turns out that natural prosodic patterns do not contribute to speech intelligibility of fast speech. The attempt to preserve the segmental intelligibility of the unstressed syllable at the *expense* of the prosodic pattern even turned out to be more successful than enhancing the prosodic pattern. The more salient prosodic pattern is not helpful for listeners after all: it may just be easier for speakers to speed up in the selective fashion, or it may perhaps be even impossible to speed up in any other way. Even though non-linear speed up is harmful for intelligibility, speakers are not 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. Stressed vowels are closer to their citation form (Lehiste, 1970; van Bergem, 1993; Moon and Lindblom, 1994). In the mental representation, the target values for stressed segments may be more strictly specified than for unstressed segments. Consequently, the speaker is forced to spend more energy on coming close to the specified targets for stressed syllables than for the more loosely defined unstressed syllables. A fast articulation rate is inevitably accompanied by undershoot of the pre-defined targets (Lindblom, 1963; Moon and Lindblom, 1994). One of the reasons for this undershoot is the inertia, or stiffness, of the speech organs. Articulatory structures such as the jaw are relatively slow. So, if more precision or effort 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.

Current research in our laboratory suggests that even at the rate of speech that speakers can attain, selective compression does not improve perception over linear compression. We therefore argue that this second explanation fits the data best: the changes in timing that accompany faster articulation rates are not so much intended to make perception easier. They are rather the consequence of restrictions on articulation. Intelligibility of artificially time-compressed speech is not improved by applying the temporal pattern of fast speech: time compression threatens the identifiability of unstressed segments, and selective time compression only makes this worse. Obviously, a natural prosodic pattern is not beneficial for the intelligibility of fast speech. Prosody should not be at the expense of the segmental intelligibility of the speech signal: both syllables contribute to the identification of polysyllabic words.

5. Conclusion

We set up a production and perception experiment to investigate what speakers do when they are forced to speak faster, and secondly, to test whether the way in which speakers speed up can improve intelligibility of artificially time-compressed speech over linear time compression. Our first expectation was that speakers, in line with the H&H theory of speech, speed up most during the least informative parts of speech. This expectation was confirmed: lexically unstressed syllables were reduced more, relatively, than stressed syllables.

The second expectation was that vowels in words bearing a pitch accent on the stressed syllable would reduce relatively less, with increasing speech rate, than vowels in words without a pitch accent. The results did not confirm this hypothesis. This was attributed to the fact that duration is an important cue for lexical stress but not for sentence accent in Dutch (Sluijter, 1995). However, we did find that function words, such as articles and auxiliaries, were often heavily reduced or almost completely absent in fast speech. These changes in sentence-level timing are in line with our H&H-based predictions.

Our production study showed important changes in word-level timing. As we expected this non-linear compression behaviour of human speakers to be driven by a strategic communicative principle, our third expectation was that applying a more salient prosodic pattern would improve intelligibility over linearly time-compressed speech.

Other studies had shown that prosody can be exploited in difficult listening situations, so this is why we investigated this third hypothesis at a rather heavy rate of time compression. The results of the perception experiment proved this hypothesis to be wrong. The changes in timing were extrapolated from the moderately fast speech rate to a much faster speech rate than human speakers can ever attain. At this very fast rate, making the temporal pattern of artificially time-compressed speech more like that of natural fast speech had a negative effect on the intelligibility, relative to linear compression. Although we do not have conclusive results yet on whether selective time compression would improve intelligibility at a moderately fast speech rate, we argue that the non-uniform way in which speakers speed up a message may be caused by the fact that they simply cannot do it otherwise, even though this may be harmful for perception.

The balance between segmental information and prosodic information turns out to be important in speech perception. Even though the stressed syllable is the most informative one, our results show that at a very fast speech rate, perception is not helped by making the prosodic durational pattern more pronounced than at a normal rate. Natural prosodic patterns do not contribute to speech intelligibility of fast speech. The attempt to preserve the segmental intelligibility of the unstressed syllable at the expense of the prosodic pattern even turned out to be more successful than enhancing the prosodic pattern.

We conclude that prosody and segmental intelligibility both contribute to word recognition in fast speech. Putting too much emphasis on either distorts the balance between a natural prosodic pattern and an intelligible speech signal.

Acknowledgements

Thanks are due to Anke Sennema and Anneke Slis for their help in the production study, and to Theo Veenker for his technical assistance. We thank the reviewers for their useful comments and suggestions.

Appendix A. List of 32 target nouns in production experiment

	Unstressed vowel schwa		Full unstressed vowel	
	initial stress	final stress	initial stress	final stress
Long vowel in stressed syllable	beker 'beaker'	bedrijf 'company'	specie 'mortar'	kopij 'copy'
	schade 'damage'	betoog 'argumentation'	havik 'hawk'	saucijs 'sausage'
	code 'code'	getij 'tide'	sofa 'sofa'	octaaf 'octave'
	pater 'father'	bereik 'reach'	foto 'photo'	pastei 'pie'
Short vowel in stressed syllable	stekker 'plug'	gebod 'command'	ghetto 'ghetto'	kopie 'copy'
	ticket 'ticket'	bestek 'cutlery'	toffee 'toffee'	schavot 'scaffold'
	stakker 'poor wretch'	gezag 'authority'	asbest 'asbestos'	pakket 'parcel'
	bakkes 'mug'	gebit 'set of teeth'	sabbat 'sabbath'	effect 'effect'

References

Altmann, G., Carter, D., 1989. Lexical stress and lexical discriminability: Stressed syllables are more informative, but why? Computer Speech and Language 3, 265–275.

Charpentier, F., Stella, M.G., 1986. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing 86, pp. 2015–2018.

Covell, M., Withgott, M., Slaney, M., 1998. Mach1: Nonuniform time-scale modification of speech. In: Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, Seattle WA.

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: Proceedings of the 6th International Conference on Spoken Language Processing, Vol. 1, pp. 593–596.

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

den Os, E.A., 1988. Vowel reduction in Dutch and Italian, with special reference to fast speech rate. Doctoral dissertation Utrecht University.

Gay, T., 1978. Effect of speaking rate on vowel formant movements. Journal of the Acoustical Society of America 63 (1), 223–230.

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

Kager, R., 1989. A metrical theory of stress and destressing in English and Dutch. Foris, Dordrecht.

Kozhevnikov, V.A., Chistovich, L.A., 1965. Speech articulation and perception. Joint Publications Research Service, Washington.

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

Lindblom, B., 1963. Spectrographic study of vowel reduction. Journal of the Acoustical Society of America 35 (11), 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, pp. 403–439.

Max, L., Caruso, A.J., 1997. Acoustic measures of temporal intervals across speaking rates: variability of syllable- and phrase-level relative timing. Journal of Speech, Language, and Hearing Research 40, 1097–1100.

Moon, S., Lindblom, B., 1994. Interaction between duration, context and speaking style in English stressed vowels. Journal of the Acoustical Society of America 96 (1), 40–55.

Peterson, G.E., Lehiste, I., 1960. Duration of syllable nuclei in English. Journal of the Acoustical Society of the America 32 (6), 693–703.

Port, R.F., 1981. Linguistic timing factors in combination. Journal of the Acoustical Society of America 69 (1), 262–274.

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

Sluijter, A.M.C., 1995. Phonetic Correlates of Stress and Accent. Doctoral dissertation Leiden University.

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

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? Language and 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. Journal of the Acoustical Society of America 78, s21.

van Heuven, V., Hagman, P., 1988. Lexical statistics and spoken word recognition in Dutch. In: Coopmans, P., Hulk, A. (Eds.), Linguistics in the Netherlands. Foris, Dordrecht, pp. 59–68.

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.

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, pp. 146–156.

Wingfield, A., Lombardi, L., Sokol, S., 1984. Prosodic features and the intelligibility of accelerated speech: syntactic vs. periodic segmentation. Journal of Speech and Hearing Research 27, 128–134.