INTELLIGIBILITY OF TIME-COMPRESSED SPEECH:
THREE WAYS OF TIME-COMPRESSION

Esther Janse
Utrecht institute of Linguistics OTS, Utrecht University, The Netherlands

ABSTRACT

Studies on fast speech have shown that word-level timing of fast speech differs from that of normal rate speech in that unstressed syllables are shortened more than stressed syllables as speech rate increases. An earlier experiment showed that the intelligibility of time-compressed speech could not be improved by making its temporal organisation closer to natural fast speech. To test the hypothesis that segmental intelligibility is more important than prosodic timing in listening to time-compressed speech, the intelligibility of bisyllabic words was tested in three time-compression conditions: either stressed and unstressed syllable were compressed to the same degree, or the stressed syllable was compressed more than the unstressed syllable, or the reverse. As was found before, imitating word-level timing of fast speech did not improve intelligibility over linear compression. However, the results did not confirm the hypothesis either: there was no difference in intelligibility between the three compression conditions. We conclude that segmental intelligibility plays an important role, but further research is necessary to decide between the contributions of prosody and segmental intelligibility to the word-level intelligibility of time-compressed speech.

1. INTRODUCTION

Artificial time-compression of speech, e.g. for the purpose of fast playback of long recordings, emails or voicemail messages, is normally performed in a linear way. Production studies on English and Dutch have shown that speakers shorten segment durations in a non-uniform way when they speak faster than normally (Van Santen 1994, Janse et al. 2000). On the word-level, the duration of unstressed syllables is reduced more than that of stressed syllables such that the relative difference in duration between stressed and unstressed syllables increases, thereby making the prosodic pattern more prominent. These production data then raise the question whether the intelligibility of time-compressed speech might be improved if its temporal organisation is closer to natural fast speech.

The idea to base time-compression on fast speech timing was worked out by Covell, Withgott & Slaney (1998). They developed a time-compression algorithm, called Mach1, which imitates the natural timing of fast speech. In this algorithm, word-level timing, phrase-level timing and discourse-level timing is taken into account. Their test results show that at equal overall speech rate, Mach1 offers significant improvements in comprehension over linear compression. However, it is not clear which characteristics of the algorithm are mainly responsible for the improvement over linear compression. By compressing pauses and silences the most, the remaining speech can be time-compressed to a lesser degree than the speech in the linear compression condition.

Several studies have shown that under difficult listening conditions, prosodic factors do play an important role in comprehension at sentence-level. Wingfield & Klein (1971) and Wingfield, Lombardi and Sokol (1984) showed that correct sentence level phrasing is helpful in the understanding of time-compressed speech.

The present study only focuses on the word-level intelligibility of bisyllabic nouns, and whether taking into account the changed timing at word-level can contribute to the intelligibility of time-compressed speech. In an earlier experiment the intelligibility of natural fast speech was compared with that of normal rate speech which is time-compressed linearly to that same fast rate, and with the intelligibility of normal rate speech which was electively compressed by adjusting the durations of the test words in a segment-by-segment way, based on the duration changes the speaker applied going from normal to fast rate. The results showed that speech in the linear time-compression condition was more intelligible than speech in the selective time-compression condition, at the same overall speech rate. The natural fast speech had the worst intelligibility. Apparently, the fast speech timing at word-level makes intelligibility worse, even more so when combined with the heavy coarticulation that comes naturally with very fast speech. The need for natural timing patterns is in conflict with the need for segmental intelligibility.

These results suggest that segmental intelligibility may be more important than the rules of fast speech timing in listening to time-compressed speech. Although it may be more natural to compress the unstressed syllables more than the stressed syllables, polysyllabic words can often not be recognised on the basis of the stressed syllable alone. On the one hand, a number of studies have shown that stress plays a role in word recognition (cf. Cutler et al. 1997 for references), and duration is an important perceptual cue for lexical stress in Dutch (Sluiter 1995). On the other hand, however, if segmental intelligibility is in conflict with prosody, intelligibility of time-compressed speech might be improved by an entirely unnatural compression strategy. This strategy involves...
compressing the lexically stressed syllable relatively more than the unstressed syllable, to retain the segmental intelligibility of both syllables.

To test the hypothesis that segmental intelligibility is more important than prosodic factors, we set up a study to investigate intelligibility of bisyllabic words in three time-compression conditions. Intelligibility of bisyllabic words will be compared in the conditions Linear Time-Compression (LC), Selective Time-Compression (SC), and Unnatural Time-Compression (UC). Selective Compression is a global imitation of fast speech timing whereby the unstressed syllable is compressed more than the stressed syllable. Unnatural Compression is the reverse condition of SC: the stressed syllable is compressed more than the unstressed syllable.

In auditory word recognition, the incoming signal is mapped onto a stored mental word template. This mental template also contains information about stress and thus about word-level timing. For the identification of non-words, only segmental intelligibility counts, as there is no such mapping between signal and template. Therefore, the intelligibility of words and non-words will be tested in the three compression conditions to study the effect of the compression conditions on word recognition and on non-word identification.

For the real words and non-words, the hypothesis is that Unnatural Compression condition yields the best intelligibility. Furthermore, the Selective Compression condition is expected to give the worst intelligibility.

2. METHOD

Forty-eight bisyllabic words were chosen to test the effect of the three ways of time-compression on intelligibility: half of them with initial stress and half of them with final stress. The intelligibility of forty-eight phonotactically legal non-words was also tested in the three compression conditions.

The words and non-words were embedded in two types of carrier phrase: normally the carrier phrase "[e moet ... typen]" (You must ... type) was chosen. In case the word started with /t/ the carrier phrase "[e moet... schrijven]" (You must ... write) was chosen. A male native speaker of Dutch was asked to read the words and non-words in the carrier phrases. Below the three compression conditions are given.

1. linear time-compression (LC): compress lexically stressed and unstressed syllable to the same degree
2. selective time-compression (SC): global imitation of fast speech timing: compress stressed syllable less (to 65% of normal rate duration) than unstressed syllable (to 40%)
3. unnatural time-compression (UC): opposite to (SC): compress stressed syllable more (to 40%) than unstressed syllable (to 65%)

After the selective or unnatural time-compression was applied, the phrases were linearly compressed even further to 35% of their original duration (for the LC condition speech was time-compressed linearly to 35% of the original duration). For the non-words, the overall compression rate was 45% in all three compression conditions, because an earlier experiment had shown that identification scores were extremely low at a compression to 35%. The carrier phrase was compressed linearly (in all conditions) to 35% for the real words, and to 45% for the non-words.

First, all real words were presented. To conceal that all test items were bisyllabic, twenty-four monosyllabic filler words were added to the test material. A practice set of twelve sentences preceded the actual test. Second, the non-words (plus monosyllabic fillers) were presented, and these were also preceded by twelve practice sentences. Items were presented in random order to control for a possible adaptation effect during the experiment.

The carrier phrase was first presented visually on the screen without the target word. Then the carrier phrase, including the target, was presented over closed headphones, and subjects were asked to type in the missing word. After they had pressed the Enter key, the next stimulus was presented. In both experiments, the three compression types were balanced over the 48 test words, and over stress position (Latin square design). There were three experimental lists for both experiments. Forty-two subjects participated in the listening experiments: fourteen on each list. The subjects, all students of Utrecht University, were paid Dfl.10 for their participation.

3. RESULTS

The raw percentages correct recognition are shown in Table 1. Note that the real words were compressed (overall) to 35% of their original duration and the non-words to 45%.

<table>
<thead>
<tr>
<th></th>
<th>% correct LC</th>
<th>% correct SC</th>
<th>% correct UC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real words</td>
<td>54</td>
<td>54</td>
<td>53</td>
</tr>
<tr>
<td>Non-words</td>
<td>37</td>
<td>25</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 1: Percentages correct identification in conditions Linear Compression (LC), Selective Compression (SC) and Unnatural Compression (UC).

The data in Table 1 suggest that there is no effect of the compression conditions on the recognition of the real words. A different picture is presented in Figure 1, which shows the results for the real words, broken down by Stress position.
Analyses of variance (Repeated Measures) were carried out on the percentages correct recognition (which were first transformed by an arcsine transformation), with Compression Type and Stress position as fixed factors, and either subjects or items as a random factor. There was no main effect of Compression Type \((F_1(2,40)<1,\ n.s.;\ F_2(2,45)<1,\ n.s.)\). The subject analysis showed a significant main effect of Stress position \((F_1(1,41)=13.86,\ p=0.001)\), but this was not significant in the item analysis (with items nested under Stress position: \(F_2(1,46)=1.23,\ n.s.)\). The interaction between Compression Type and Stress position, which can be seen in Figure 1, is not significant in the item analysis \((F_2(2,45)=2.44,\ p=0.098)\) but is highly significant in the analysis on subjects \((F_1(2,40)=6.02,\ p=0.005)\). When the Selective Compression condition is left out of the real word analyses, the interaction between Compression Type and Stress position is significant in both analyses \((F_1(1,41)=11.54,\ p=0.002;\ F_2(1,46)=4.98,\ p=0.031)\).

A separate analysis was carried out for the real words with initial stress to test whether the difference between UC and LC was significant (SC was left out of the analysis). Despite the weak trend seen in Figure 1, intelligibility in the Unnatural Compression condition did not significantly differ from that in the Linear Compression condition \((F_1(1,41)=8.87,\ p=0.056;\ F_2(1,23)=1.52,\ n.s.)\). The fact that intelligibility of words with final stress was worse in the UC condition than in the LC condition might have been due to a minor artefact of the stimulus set. Still, it seems highly unlikely that the UC condition would have improved recognition scores of the words with final stress in the absence of this artefact. Figure 1 also shows that intelligibility in the Linear Compression hardly differs from that in the Selective Compression condition, regardless of the stress position.

Thus, the differences in intelligibility in the three types of time-compression were very small indeed. An analysis was carried out to test whether intelligibility was mainly affected by word frequency. However, an item analysis with Compression Type as the fixed factor, items nested under Stress position, and word frequency (log) as a covariable did not show any main effect of word frequency \((F_2(1,45)= 2.16,\ n.s.)\).

The identification results of the non-words, broken down by stress position, are shown in Figure 2.

In order to have a closer look at the effects of the three compression types, the identification results per syllable are shown in Tables 1 and 2. Certain confusions within the syllable were allowed for the non-words: place of articulation of nasals, and voicing value of fricatives and plosives. The percentages correct (and almost correct) for the entire word are given.

1 A closer look at the stimuli showed that the mean word duration of the words and non-words with final stress was accidentally somewhat shorter in the UC condition than in the LC condition (real words: 165 ms compared to 172 ms in LC condition; non-words: 232 ms compared to 240 ms in LC condition). Although this difference in duration is only 3-4% (and the JND of duration differences is probably not better than 6% (cf. Nooteboom & Doodeman (1980))), this may have caused

<table>
<thead>
<tr>
<th>Words and non-words with Initial stress</th>
<th>Lexical Status</th>
<th>Compression Type</th>
<th>% Syllable correct</th>
<th>% Syllable correct</th>
<th>% word correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 1: Percentages correct per syllable for real words and non-words with initial stress, per Compression Type (for the non-words, the percentages almost correct in brackets).

The non-words most clearly show the effect of the three compression types. Table 1 shows that Selective Compression improved the identification scores of the stressed syllable (albeit at the cost of the unstressed syllable), and Unnatural Compression improved the identification scores of the unstressed syllable (at the cost of the stressed syllable). The net result was that neither SC nor UC improved the intelligibility of the non-words over LC. Although the results for the real words are not entirely predictable from the length of the syllable due to lexical factors, the same overall trend applies that the longer the syllable remains, the higher the segmental intelligibility. The UC condition does improve the intelligibility of the real words as a whole, but not of the non-words.

Table 2 shows the intelligibility results per syllable for words and non-words with final stress.

Table 2: Percentages correct per syllable, for real words and non-words with final stress (% almost correct in brackets).
Again, for the non-words it is obvious that segmental intelligibility is higher when the syllable remains longer. Still, for the non-words with initial and final stress, the positive effect on the one syllable is always outweighed by an equally large or even larger negative effect on the other syllable. The real words with final stress show that the UC condition cannot improve the intelligibility of the unstressed (first) syllable, whereas UC does improve the percentages correct for the non-words. As the intelligibility of the stressed (second) syllable is lower for UC than for LC, the net result is then that Linear Compression results in the best intelligibility.

4. DISCUSSION AND CONCLUSION

The different ways of time-compression did not appear to have important effects on the intelligibility of the real words. Lexical factors, such as lexical neighbourhood density might play the key role in the recognition of time-compressed speech, such that the effects of the different ways of time-compression are too small to be found.

Another possible reason why the different ways of time-compression had no effect was that the carrier phrase in which the target words were embedded provided a context which was prosodically not strong enough for prosodic manipulations to have an effect. Firstly, the carrier phrase was compressed linearly in all three compression types. Secondly, words spoken in short carrier phrases almost behave as if they are spoken in isolation: the normal timing relations within a bisyllabic word are then strongly affected by final lengthening on the word-level. If the three compression conditions had been applied to normal meaningful sentences as a whole, the differences between the three types of conditions might have been clearer.

As was found before, Selective Compression, which reinforces the prosodic pattern of the real words, does not improve intelligibility over linear compression. Contrary to our earlier findings, however, we did not find a negative effect of Selective Compression. There was hardly any difference between intelligibility in the LC vs. SC condition. In our former experiment correct identification rates in the Selective Compression condition were significantly lower than in the Linear Compression condition. In that experiment, we adjusted all durations in a segment-by-segment way, based on the speaker way of speeding up. Following the global strategy of the speaker is obviously less harmful for intelligibility rates.

The results of the non-words and the real words in the present experiment show that segmental intelligibility is affected by syllable duration. This syllable duration has a greater contribution than the prosodic pattern of normal rate speech (LC) or that of fast speech (SC). The data presented in Tables 2 and 3 also showed that the intelligibility of the real words cannot be predicted from the results of the non-words, due to lexical factors. Whereas the unnatural compression condition did not improve the identification rate of the non-words and of the words with final stress, it slightly improved the recognition scores of real words with initial stress. This interaction between the effect of UC and Stress position for the real words might indicate that the segmental intelligibility of the unstressed syllable is more important for the recognition of words with initial stress than for words with final stress. Figure 1 shows that recognition scores for words with final stress were higher overall than for words with initial stress. As initial stress is the default stress pattern in Dutch, words with final stress have fewer neighbours. Following the definition of neighbourhood by Luce & Pisoni (1998), the mean number of neighbours for our set of words with initial stress was 1.5, whereas it was only 0.5 for the words with final stress (more than half of them did not have any neighbours at all). Identifying the stressed syllable alone in words with final stress may more often result in correct recognition than for words with initial stress. The ratio of the number of segments in the stressed syllable divided by the number of segments in the unstressed syllable is also higher in the words with final stress (mean ratio is 1.4 for words with final stress vs. 1 for initial stress). This might also give more weight to the identification of the stressed syllable for words with final stress than for words with initial stress.

Obviously, segmental intelligibility and prosodic factors play conflicting roles in the intelligibility of time-compressed speech. Although we still expect that segmental intelligibility at word-level may be more important than prosody, the results of the present experiment do not warrant this conclusion. Further research is necessary to illuminate this point.

5. REFERENCES


