Word perception in natural-fast and artificially time-compressed speech

Esther Janse
Utrecht institute of Linguistics OTS, Utrecht University
The Netherlands
E-mail: Esther.Janse@let.uu.nl

ABSTRACT

Word and sentence level timing in natural fast speech differs from that at a normal speech rate. However, previous research has shown that making the temporal pattern of artificially time-compressed speech more similar to that of natural fast speech did not improve intelligibility over linear time compression. Furthermore, pilot studies suggested that naturally produced fast speech is more difficult to perceive than artificially time-compressed speech because of its somewhat slurred articulation. In the present study, word processing speed and listeners' subjective preference were compared for three equally fast conditions: (1) naturally produced fast speech, which is perfectly intelligible; (2) artificially time-compressed speech, which has the temporal pattern of naturally produced fast speech; and (3) artificially linearly time-compressed speech (having the temporal pattern of normal-rate speech). The results suggest that linearly time-compressed speech has a temporal and a segmental processing advantage over naturally produced fast speech.

1. INTRODUCTION

The temporal pattern of natural fast speech has been shown to differ from that of normal-rate speech. When speakers increase their speech rate, they shorten unstressed syllables more than stressed syllables, relatively, making the prosodic pattern more prominent [1]. It seemed that this enhanced prosodic pattern could 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. Listeners make use of prosody in lexical processing [2], and they rely even more on prosodic information in adverse listening conditions [3]. Therefore, in a previous study [1], we investigated whether the intelligibility of time-compressed speech could be improved over linear compression by making its timing pattern more like that of natural fast speech. This was investigated with very fast speech, time-compressed to almost three times the original rate. However, making the timing pattern more similar to that of natural fast speech had a negative effect on intelligibility, relative to linear time-compression. Two possible objections can be made against those previous results. The first is that the negative results might have been due to the fact that the speech was time-compressed to a much faster rate than speakers can produce. The non-linearities found at the moderately fast rate that speakers can attain were extrapolated to much faster rates. This extrapolation may have been unwarranted: it is conceivable that imitating what speakers do would improve perception over linear compression at the rate at which this timing pattern was actually observed, but not at unnaturally fast rates. Secondly, it is possible that the nonlinear type of time compression applied in that previous study is not typical of 'normal' natural fast speech: it was based on a study of normal-rate versus very fast and slurred speech. The time pressure that was imposed on the speakers in that duration study may have made them lose sight of the listeners, and may have had a larger effect on the relative timing pattern than one would normally encounter in natural fast speech. If we are not sure that the speakers actually intended to be understood, we cannot exclude the possibility that the speakers just chose the easiest, but not necessarily the only possible, way to speed up the message. Therefore, the focus in the present study will be on moderately fast speech that is still perfectly intelligible. Again, the question is whether making artificially time-compressed speech more natural will make processing easier for the listener. The prediction is now that making its timing pattern (below phrase level) more similar to that of natural fast speech will not improve perception over linear time compression. We expect that the changes in temporal pattern are not due to a communicative and strategic principle, but to the fact that speakers cannot speed up in an approximately linear way. Lexical stress is specified in the mental lexicon, and target values for stressed segments may be more strictly specified than for unstressed segments. As a result of this specification, stressed syllables are produced with more articulatory precision: linguistic stress can be seen as localised hyperarticulation [4]. If more precision is required for the stressed 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 does not necessarily imply that the nonlinear type of speed-up is also beneficial for the listener. Secondly, the increased segmental overlap that inevitably occurs in relatively fast speech rates (faster than 1.2 times normal rate) is expected to hinder perception. Although spoken word processing is not hindered by assimilation when speech is articulated at a normal rate [5], speech presented at fast rates seems to be helped by a more redundant speech

 ISBN 1-876346-48-5

signal [6]. Pilot studies with several speakers' materials have suggested that naturally produced fast speech is more difficult to perceive than artificially time-compressed speech because of its somewhat slurred articulation. The prediction is that the more natural the fast speech is, with respect to its temporal pattern, or both temporally and segmentally, the longer the processing times. Word processing speed in three fast conditions will be investigated in experiment 1.

Additionally, the quality of the speech conditions was compared by investigating listeners' subjective preference. Experiment 2 will be set up to investigate which type of fast speech listeners find most agreeable to listen to.

The following questions will be addressed in this paper:

1. Can word processing be improved by making the timing pattern of artificially time-compressed speech 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?
3. Which type of speech is preferred by listeners in a subjective preference test?

2. EXPERIMENT 1: PROCESSING SPEED

This experiment investigates whether speakers speed up in a nonlinear way when they produce intelligible fast speech. If so, how does it affect perception, relative to linear compression of normal-rate speech? The increased segmental smearing that accompanies a faster speaking rate is expected to make word perception more difficult, relative to artificial time compression. To put this to the test, word perception is compared in three experimental conditions:

1. linear time-compression
2. copy-fast-speech-timing compression (all syllable durations of the normal-rate condition are set to the syllable durations of the natural fast condition)
3. natural fast speech

In this way, the effect of nonlinear time compression (copying the timing pattern of the natural fast condition, but preserving the segmental information of the normal-rate condition) and of the combination of nonlinear time compression and increased segmental overlap (i.e., natural fast speech) can be studied, relative to linear compression.

Material

Because the intelligibility of the three speech conditions was perfect, phoneme detection speed was used to measure word processing speed. News bulletin texts were collected and 84 sentences or sentence fragments were selected that had nouns in them starting with a plosive (mean sentence length of 23.4 syllables). The target-bearing words were polysyllabic nouns: half with initial stress; half with non-initial stress. The nouns were never compounds, but some were morphologically complex (as in 'gardener'). An example sentence fragment is given in (1) below (target word underlined):

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

One male native Dutch speaker, known for his clear speaking style, read the sentence material at normal and fast rate. It was stressed that the fast version should still be perfectly intelligible. The mean articulation rate in the sentence fragments was 6.1 syllables/second in the normal-rate condition, and 8.5 syll./sec in the fast condition. The overall fast-to-normal ratio was 0.72 (i.e., speed-up factor 1.4). The sentence fragments were labelled manually: markers were placed in the waveform at all syllable boundaries, in both 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 the respective duration of the same syllable in the normal-rate condition. Then, all syllables in the normal-rate condition were time-compressed one by one, according to these 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 syllable level. For the linear compression condition, the overall fast-to-normal ratio was computed for each sentence. This overall ratio was then applied 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. 80 Catch trials, also taken from the news bulletin items, were interspersed with the test material to keep subjects from pressing the button randomly. These catch trials did not contain the assigned plosive. In a pilot study, the intelligibility of the fast and time-compressed conditions was found to approach 100%.

Design and procedure

The 84 test sentences, in 3 experimental conditions, were distributed over 3 lists (Latin square design). Subjects were seated in sound-treated booths, wearing closed-ear headphones. Subjects saw a letter-sound on the computer screen in front of them. They were told to press a button as fast as possible whenever they detected a word-initial occurrence of the assigned phoneme during the upcoming auditory stimulus sentence. There were 10 practice items, after which additional instruction was possible. Test and filler items were presented in random order. Ten subjects were assigned to each list (all students at Utrecht University). They received €5 for their participation.

Results

Reaction times were computed from the start of the silent interval of the target plosive (or from start of voice bar for voiced plosives). All subjects agreed that the fast speech conditions were all highly intelligible. The raw mean phoneme detection times are presented in Table 1 (missing values are excluded). Miss rates are also given.

ISBN 1-876346-48-5

	Mean DT (ms)	Miss rate (%)
Linear time-compression	572	3
Copy-fast-speech-timing	600	3
Natural fast speech	624	3

Table 1: Mean raw detection time (in msec) and miss rate for the three fast conditions.

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 in the item analysis. Statistical analyses (ANOVAs with either subjects or items as repeated measures) were run on the inverse reaction times (1/RT) so that the data distribution is less skewed. Condition and Stress position were analysed as fixed factors (in the item analysis, items were nested under Stress position). The effect of Condition was significant ($F_1(2,28)=7.4$, $p=.002$; $F_2(2,81)=4.3$, $p=.039$). The effect of Stress position was not significant ($F_1(1,29)=1.0$, n.s.; $F_2(1,82)<1$, n.s.); and neither was the interaction between Condition and Stress position ($F_1(2,28)<1$, n.s.; $F_2(2,81)<1$, n.s.). To allow post-hoc testing, Univariate ANOVAs were run on the same data. The effect of Condition was again significant ($F_1(2,28)=5.5$, $p=.007$; $F_2(2,82)=3.1$, $p=.048$). Post-hoc pairwise comparisons (Scheffé) showed that only the linear time-compression and the natural fast condition differed significantly from each other (subject analysis $p=.03$; items $p=.009$). The two time-compressed conditions did not differ significantly, and the copy-fast-speech-timing condition did not differ from the natural fast speech either (all analyses $p>0.1$).

The small and insignificant difference between the copy-fast-speech-timing and linear compression condition may be attributed to the relatively small rate difference between the normal and fast articulation rate, which induced only small changes in word- and sentence-level timing. In the previous study [1], speakers were pressed to speak very fast. They increased their speech rate to 10.5 syllables/sec, which was accompanied by a fast-to-normal ratio of 0.65 for stressed and 0.45 for unstressed syllables. In the present material, the fast rate is 8.5 syll./sec, and stressed syllables had a mean fast-to-normal ratio of 0.77; unstressed syllables had a mean fast-to-normal ratio of 0.71. This suggests that a change in relative timing does not only occur in fast and sloppy speech: intelligible fast articulation is also accompanied by a shift in the temporal pattern. However, given the small timing shift, it is not surprising that listeners hardly show any processing difference between the linear and copy-fast-speech-timing condition.

Phoneme detection responses may be based on a merge of lexical and pre-lexical information [7]. Stimulus material and the task may influence which information source contributes most. As the use of meaningful sentences may induce listeners to focus mainly on lexical information, the results are taken to reflect lexical processing.

In a way, 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 1 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. Given that the most natural condition is obviously not the easiest one to process, it is interesting to find out whether listeners have a clear preference for either of the fast conditions. This is tested in the next experiment.

3. EXPERIMENT 2: PREFERENCE TEST

The three fast conditions of Experiment 1 were evaluated perceptually using the Comparative Mean Opinion Score (CMOS) test [8]. The question was whether listeners could indicate whether one version 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. There is evidence that the results of such judgment (or opinion) tests converge with those of functional evaluations [9]. Listeners' 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. The prediction is that, analogous to the previous results, listeners will judge the naturally produced fast version as less agreeable than the two artificially time-compressed versions. Hardly any preference was expected for either of the two artificially time-compressed versions because of the small difference between them.

Material, Design and Procedure

A selection of the speech material of experiment 1 was used (45 test and 5 practice sentences). 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). A complementary (Latin square) design was set up in which these comparison pairs were rotated over the sentence pairs and over three different lists. Subjects listened to the material over headphones while seated in a sound-treated booth 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). Subjects then indicated their preference by clicking on a 7-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). Each condition within a comparison pair appeared about equally often as version A or B. After they had indicated their preference judgment, they could click a button labelled 'Next' in order to hear the next sentence pair. Before subjects started with the actual experiment, they were presented with 5 practice sentences, after which additional feedback or instruction was given, if necessary. To each of the three experimental lists, 6 subjects were assigned. They were all students at Utrecht University, and were paid €5 for their participation.

ISBN 1-876346-48-5

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 2.

	Mean CMOS	s.e.
Linear& Copy-fast	-0.27	0.05
Linear & Natural-fast	-0.50	0.09
Copy-fast & Natural-fast	-0.02	0.09

Table 2: Mean perceptual scores of Comparative Mean Opinion Score (CMOS) test, on a scale from +3 to –3, plus standard errors.

A negative CMOS value indicates that the second member of the pair (as indicated in Table 2) was judged as less agreeable than the first. Statistical analyses of these CMOS values take the form of one-sample t-tests (on subject and item means) to test the hypothesis that the mean CMOS value per pair differs significantly from zero. The t-tests for the first comparison pair shows that, despite the small CMOS value, the Copy-fast (nonlinear) time-compressed condition is judged as significantly less agreeable than the linearly time-compressed condition ($t_1(17)$=-3.4, p=0.003; $t_2(44)$=-4.6, p<0.001). The difference between Linear and Natural fast is also significant ($t_1(17)$=-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. A significant difference between the two artificial time-compression conditions was found (in favour of linear compression), whereas this was not found in experiment 1. The direction of this preference agreed with the tendencies seen in experiment 1. Lastly, listeners did not prefer the copy-fast (nonlinear) time-compression condition over natural-fast speech. In sum, even at a rate at which all three fast conditions are still perfectly intelligible, listeners have a slight preference for the condition which also proved easiest to process in experiment 1. The results of the previous study [1] and those 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 normal-rate speech which is time-compressed linearly afterwards. This can be attributed partly to increased segmental overlap and to a changed temporal pattern.

Practically, this suggests that the only aspects of naturally produced fast speech that should be imitated in order to make time-compressed speech more intelligible are to be found at levels higher than the phrase level (such as compressing pauses more than the remaining speech).

4. CONCLUSION

Natural fast speech timing rules do not seem to improve intelligibility nor ease of processing, not even at the rate of speech at which they were observed. Secondly, listeners prefer artificially linearly time-compressed speech over naturally produced fast speech, even when the natural fast speech is still perfectly intelligible. This strengthens our belief that both the segmental and the timing changes that accompany natural fast speech rates are due to articulatory restrictions, and do not serve a communicative purpose. More carefully articulated items add redundancy to the speech signal which is beneficial for perception in difficult listening situations. 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. 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. Speakers will therefore only choose to do this when the communicative situation allows it.

REFERENCES

[1] E. Janse, S. Nooteboom and H. Quené, "Word-level intelligibility of time-compressed speech: prosodic and segmental factors", *Speech Communication*, in press.

[2] A. Cutler, D. Dahan and W. van Donselaar, "Prosody in the comprehension of spoken language: a literature review", *Language and Speech,* vol. 40 (2), pp. 141-201, 1997.

[3] W. van Donselaar and J. Lentz, "The function of sentence accent and given/new information in speech processing: different strategies for normal-hearing and hearing-impaired listeners?", *Language and Speech,* vol. 37 (4), pp. 375-391, 1994.

[4] K.J. De Jong, "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation", *Journal of the Acoustical Society of America,* vol. 97 (1), pp. 491-504, 1995.

[5] M.G. Gaskell and W.D. Marslen-Wilson, "Phonological variation and inference in lexical access". Journal of Experimental Psychology: Human perception and performance, 22, 144-158, 1996.

[6] H. Quené and J. Krull, "Recognition of assimilated words in normal and fast speech", Proceedings of the 14th ICPhS, San Francisco, pp. 1831-1834, 1999.

[7] D. Norris, J.M. McQueen and A. Cutler, "Merging information in speech recognition: Feedback is never necessary", *Behavioral and Brain Sciences,* vol. 23 (3), pp. 299-370, 2000.

[8] ITU-P.800, "Methods for subjective determination of transmission quality", Recommendation P.800 International Telecommunication Union (ITU), 1996.

[9] C. Pavlovic, M. Rossi and R. Espesser, "Use of magnitude-estimation technique for assessing the performance of a text-to-speech system", Journal of the Acoustical Society of America, vol. 87, pp. 373-381, 1990.