# METRICAL STRUCTURE AND THE PERCEPTION OF TIME-COMPRESSED SPEECH

Duncan Young[1], Gerry T.M. Altmann[1], Anne Cutler[2], Dennis Norris[2]

[1]*Experimental Psychology, University of Sussex, Brighton BNl 9QG, England*
[2]*MRC Applied Psychology Unit, 15 Chaucer Road, Cambridge CB2 2EF, England.*

## ABSTRACT

In the absence of explicitly marked cues to word boundaries, listeners tend to segment spoken English at the onset of strong syllables. This may suggest that under difficult listening conditions, speech should be easier to recognise where strong syllables are word-initial. We report two experiments in which listeners were presented with sentences which had been time-compressed to make listening difficult. The first study contrasted sentences in which all content words began with strong syllables with sentences in which all content words began with weak syllables. The intelligibility of the two groups of sentences did not differ significantly. Apparent rhythmic effects in the results prompted a second experiment; however, no significant effects of systematic rhythmic manipulation were observed. In both experiments, the strongest predictor of intelligibility was the rated plausibility of the sentences. We conclude that listeners' recognition responses to time-compressed speech may be strongly subject to experiential bias; effects of rhythmic structure are most likely to show up also as bias effects.

Keywords: compressed speech, rhythm, plausibility.

## 1. INTRODUCTION

The English language has a stress-based rhythmic structure; it displays wide variation in syllable structures and a marked contrast between "strong" syllables, which contain full vowels, and "weak" syllables, which contain reduced vowels. For example, in the words *final, Christmas, over* and *sprinkler* the first syllable is strong and the second weak. Cutler & Norris [1] have demonstrated that native English speakers segment speech input at the onset of strong syllables in the absence of explicitly marked cues to word boundaries. Cutler & Norris suggest that this is because the onsets of strong syllables provide the most efficient starting point for lexical access in English. Indeed, statistical studies show that lexical items which start with strong syllables are approximately six times more frequent in written English usage than words which start with weak syllables [2].

Evidence from experimental studies suggests that listeners are highly sensitive to metrical stress when listening conditions are difficult. For example, Lieberman [3] found that phonetically trained listeners could discriminate between strong and weak syllables when all segmental information was removed from electronically manipulated speech, although they were unable to distinguish degrees of lexical stress. Cutler and Butterfield [4] showed that subjects presented with utterances at just above their listening threshold made more word-boundary insertion errors before strong syllables than before weak syllables. Thus, listeners can distinguish between strong and weak syllables when speech input is non-optimal, and they make segmentation judgements that appear to accord well with the metrical segmentation strategy proposed by Cutler and Norris [1].

We can postulate, therefore, that utterances in which the strong syllables are word-initial should prove easier to recognize, when the recognition system is pushed towards its limits, than utterances in which the strong syllables are word-medial or word-final. The attendant mis-segmentations associated with word-medial or word-final strong syllables are likely to hamper the recognition process, which under difficult listening conditions will have a harder task in any case. To test these predictions we explored the intelligibility of utterances with different metrical structures under conditions of time-compression. Time-compression has the effect of increasing the perceived speech rate while preserving the segmental and prosodic distinctions of the original speech. This allows us to produce speech rates (in terms of syllables per minute, for instance) which are far in excess of those that can be produced naturally, and which should render the recognition system more prone to mis-segmentation errors.

Alternative distortions of the speech signal are of course available to achieve a similar effect, such as the addition of noise or application of a filter, however, these affect some aspects of the signal more than others because they must be applied over a certain frequency range. This in turn means that different kinds of speech sounds will be distorted to varying extents depending on the particular parameters applied. Time-compression, using an algorithm which selectively averages across adjacent pitch periods, or in the absence of a periodic signal, across adjacent 5 msec. windows [5], delivers a smooth signal and affects all portions of the signal equally. Furthermore, the effects are limited to duration rather than involving the masking or exclusion of frequency-based diagnostic features of the speech signal. (Of course, durational distinctions remain in the time-compressed signal, but the extent of the distinctions is diminished.)

## 2. EXPERIMENT 1

This experiment contrasted the intelligibility of sentences with two different metrical structures. One set of sentences had a preponderance of words starting with strong syllables, while the other had a preponderance of words starting with weak syllables (and hence, had a preponderance of strong syllables that were not word-initial). If mis-segmentation at strong syllables hampers the recognition process, this second set of sentences should be less intelligible than the set in which strong syllables do occur at word onsets.

*Materials.* 40 sentences, each 18 syllables long (mean length in words 8.5), were devised in which the initial portions of the content words were either of the form "strong-weak" ("The question of capital punishment occupied the morning session.") or "weak-strong" ("The apprentice refused to resign or accept enforced redundancy."). These sentences were recorded by a male speaker, digitised at 16kHz, and then compressed to (a) 50%, (b) 40% and (c) 35% of their original duration. These time-compressed sentences were then recorded onto Digital Audio Tape. Three tapes were made, one for each compression rate; on each tape the sentences occurred in a random order.

*Method.* The sentences were presented via headphones to listeners who were asked to write down after each sentence as many of the words as they had heard, in the order in which they had heard them. Fifteen listeners heard each tape. Before hearing the test sentences, subjects listened to three uncompressed sentences in order to familiarise themselves with the speaker's voice and the volume level used in the experiment. Each sentence was preceded by a warning tone and there was an 18-second response interval between sentences. The responses were scored in terms of number of words correcdy reported. A separate group of 22 subjects were asked to rate the plausibility of each sentence on a 7-point scale.

*Results.* Figure 1 shows the mean recognition performance at each compression rate for the two sentence types.

Contrary to expectations, recognition (defined as number of words correctly reported) did not differ significantly across sentence type, either overall or at any compression rate (all Fs <1).

The mean rated plausibility of the strong-initial sentences was 3.75, of the weak-initial 3.71, an insignificant difference ($F < 1$). A correlation analysis was conducted on mean recognition performance and rated plausibility for each sentence; this correlation was significant for each compression rate ($r [39] = 0.41$ at 50% compression, 0.56 at 40% and 0.59 at 35%, all $p < .01$). Thus the more plausible a sentence as a whole, the better recognised were the words comprising it.

Listeners' mis-segmentation errors were tabulated and analysed according to the criteria used by Cutler and Butterfield [4]. The metrical segmentation hypothesis predicts that erroneous word boundary insertions before strong syllables and deletions before weak syllables should be common, while erroneous word boundary insertions before weak syllables and deletions before strong syllables should be rare. There were 79 errors of the predicted types and 16 of the non-predicted types, a significant difference ($z = 6.36$, $p < .001$).

The mis-segmentation errors suggest that listeners were indeed applying the segmentation procedures observed in earlier studies [1,4]. However, this did not translate into an advantage for those sentences in which word onsets were indeed strong.

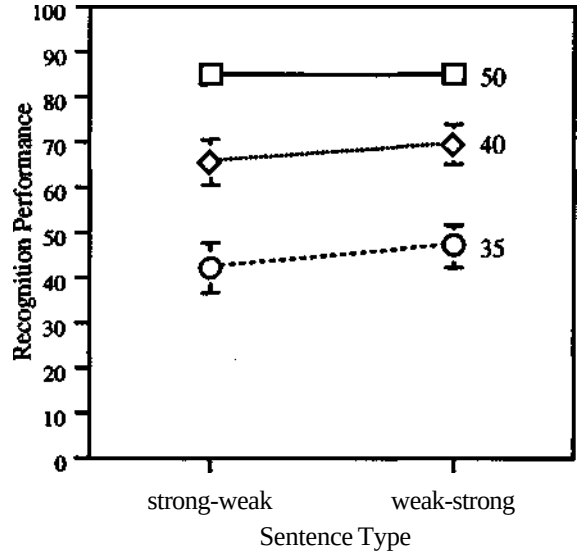


Figure 1
Mean recognition performance (% words correct), with standard error bars, at each of three compression rates: 50%, 40%, and 35%

To explore possible reasons for this, we performed a number of *post hoc* explorations of the data. First, in an attempt to assess rhythmic regularity, we calculated the total number of different types of foot structure in each sentence (the lower the number, the greater the regularity). This measure was also not related to recognition performance. The only other measure with which our isochrony measure correlated was the number of weak syllables - the fewer weak syllables, the greater the regularity ($r [39] = -0.49$, $p < 0.01$). In the course of analysing the rhythmic structure, however, we observed mat the best-recognised sentences tended to contain one particular rhythmic sequence: strong-weak-weak-strong. A correlation analysis showed that the more such sequences a sentence contained, the better it was recognised ($r [39] = 0.35$ at 50%, 0.33 at 40% and 0.50 at 35%, all $p < 0.05$).

We also conducted a regression analysis in which the strong effects of plausibility were first factored out. No relation was observed between recognition performance and sentence length (measured in words), the number of function or content words, or the number of word-initial strong syllables (note that not all the strong syllables in the "strong-weak" sentences were word-initial). At 50% compression, there was a very weak effect of the number of weak syllables present in each sentence; at higher compression rates (i.e. compression to 40% and 35% of original duration) the only residual effect was of the number of strong-weak-weak-strong sequences in each sentence.

Although the underlying reason for this effect is not immediately apparent, it does suggest that listeners' performance in recognising time-compressed speech may somehow be affected by rhythmic factors. Accordingly in our second experiment we undertook an explicit investigation of the role of this aspect of rhythmic structure. We used only the highest compression rate (35%) from the first study.

## 3. EXPERIMENT 2

In this experiment we explicitly manipulated the number of "strong weak weak strong" (SWWS) sequences in the experimental sentences. We predicted on the basis of our earlier findings in Experiment 1 that recognition performance would vary as a function of this factor.

*Materials.* A new set of 70 sentences was constructed, each 18 syllables long (average length in words 11.5). We varied the rhythmic structure in seven steps ranging from sentences containing no SWWS sequences at all ("I have consistently complained about the treatment I have received here"), to a highly rhythmic WWSWWSWWSWWSWWSWWS pattern ("Her amazing collection of animal paintings was put on display").

Plausibility ratings for these 70 sentences were collected from a group of 24 subjects. On the basis of the ratings, five sentences for each step in our rhythmic series were selected such that the resulting 35 sentences occupied the narrowest possible range of plausibility values (mean 3.33, standard error 0.09). Although the rhydimic structure of the present sentences was actually defined in terms of number of SWWS sequences, this measure of course correlated with the isochrony scores received by these sentences when we analysed them in the same way as the sentences of Experiment 1 ($r\ [34] = 0.54$, $p < 0.01$). The sentences were recorded by the same male speaker as in experiment 1, digitised at 16kHz, and time-compressed to 35% of their original duration. They were pseudo-randomly allocated to five new sets, each set containing one sentence from each rhythmic group, and these sets were arranged into five different experimental orders, each set of five sentences appearing in each possible position. The sentences were then recorded onto Digital Audio Tape in these five different orders. Each order began with a further five filler sentences, also compressed, and rated highly plausible by a separate group of subjects. These five sentences were designed to facilitate the subjects' adaptation to the compressed speech.

*Method.* The sentences were presented via headphones to 25 listeners, all of whom had taken part in a previous experiment involving time-compressed speech, although none had seen or heard the present sentences before. The subjects were asked to report the content of the sentences. As in experiment 1, each sentence was preceded by a warning tone and 18 seconds were allotted after each sentence for subjects to write down what they had heard. The responses were scored for number of words correct.

*Results.* The mean recognition score was 73% (note that this is much higher than the recognition scores at 35% compression in Experiment 1). Figure 2 shows mean recognition performance as a function of rhythmic structure.

A one-way analysis of variance showed that there was no significant effect of rhythmic structure on recognition performance ($F\ [6,28] = 1.37$). Rated plausibility again correlated significantly with recognition performance across sentences ($r\ [34] = 0.48$, $p < 0.01$), even though we had attempted to restrict the variation in plausibility across the different sentences.

The measure of isochrony which we had used in Experiment 1 was also uncorrelated with recognition performance in the present results (unsurprisingly, since it was strongly correlated with the present manipulation of rhythmic structure). No other factor which we analysed showed a significant correlation with recognition performance.

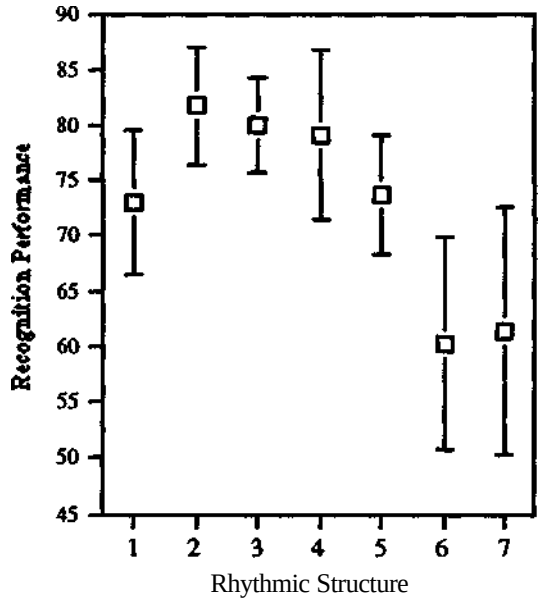


Figure 2
Mean recognition performance (% words correct) with standard error bars

*Discussion.* Although this second experiment re-affirmed the relationship between plausibility and the intelligibility of time-compressed speech, it did not show a clear, unambiguous relationship between intelligibility and rhythmic structure. This result was unexpected given our *post hoc* analysis of Experiment 1. In that experiment we found that the larger the number of sequences of a particular rhythmic type, the better the recognition performance. However, the results of the two experiments together now clearly show that that earlier finding cannot be interpreted as an effect simply of regularity of rhythm. The measure of rhythmic regularity, or isochrony, which we used failed to correlate with recognition performance in either experiment, and when the number of SWWS sequences was manipulated in such a way that it directly mirrored rhythmic regularity, it too failed to correlate with recognition performance.

## 4. GENERAL DISCUSSION

Overall, the only consistent predictor of intelligibility across the two experiments was plausibility. The more plausible a sentence, the better recognized are the component words. Of course the present results do not allow us to decide whether this is because a word can be better predicted, and hence better recognized, on the basis of the preceding words as the sentence is heard, or because during the subsequent transcription of me sentence it is easier to reconstruct words which had not been recognized originally. The latter explanation, invoking reporting biases based on listeners' experience, is certainly consistent with the finding that the effect of plausibility remains essentially constant across compression rates.

One salient difference between the results of Experiments 1 and 2 is that the sentences compressed to 35% in Experiment

1 were recognised at a rate of approximately 45% correct, whereas in Experiment 2 recognition performance at the same compression rate was 73%. A reason for this difference may be that subjects in Experiment 2 had had previous exposure to compressed speech whereas those in Experiment 1 had not. In other studies we have demonstrated that subjects show long-term adaptation effects in processing compressed speech, even across intervals of several months [6]. However, note also that the present experiments were quite long: 40 and 35 test sentences respectively. Given that even for naive subjects, adaptation is fast and takes place within the first few sentences, we believe that any advantage that might accrue due to previous experience is unlikely to persist over the entire course of the present test sets.

A second reason for the overall performance difference might involve possibly relevant differences between the two sets of stimuli. For instance, the sentences in Experiment 1 contained many more polysyllabic words than those in Experiment 2 (2.12 syllables per word in Experiment 1, compared with 1.57 in Experiment 2). If segmentation is syllable-based, there will be fewer mis-segmentation errors in sentences containing more monosyllabic words. Again, however, doubt is cast on this explanation by the fact that recognition performance did not differ as a direct function of number of words in the sentence, in either experiment.

With respect to the different pattern of results across the two experiments, it is possible that this could be parasitic upon the overall performance difference in that it could reflect as yet unrecognised factors which play no role at higher intelligibilities, but come into play when listeners are processing speech at relatively low intelligibilities. For instance, some aspects of rhythmic structure may be of assistance to listeners precisely in the case when listening conditions are difficult. The results of the two studies together show, of course, that this could not be an effect simply of rhythmic regularity. On the other hand, if reporting biases are operative in these studies (as suggested above), this result may indicate that just such a bias towards particular rhythmic structures could be available in cases of greater uncertainty. Further research on the relative frequency of particular rhythmic structures in English speech corpora could illuminate this possibility.

## 5. CONCLUSION

Studies of normal speech recognition have concluded that metrical structure plays an important role in the segmentation of fluent connected speech [1, 4]. The present study sought to explore further the role of rhythmic structure in the perception of compressed speech. Our motivation was to use a form of speech whose recognition would be particularly hampered by the mis-segmentations predicted by the metrical segmentation strategy postulated by Cutler and Norris [1]. Although there was clear evidence in Experiment 1 that listeners were indeed using the segmentation principles postulated by Cutler and Norris, the predicted effects on overall recognition performance did not materialise. Contrary to our expectations, sentences in which the strong syllables were word-medial or word-final were no less intelligible than sentences in which the strong syllables were predominantly word-initial. This was true across both the experiments reported above. It is possible that this result simply reflects unsuitability of compressed speech for testing the present hypothesis. If acoustic distinctions between strong and weak syllables are eroded at high levels of time-compression, for instance, then it would be unrealistic to expect clear effects of the strong/weak distinction in listeners' performance. Alternatively, it may be that weak-initial words are in general not mis-segmented, since application of metrical segmentation may be coupled with lexical access procedures based on strong syllables rather than on strictly left-to-right structure; such a proposal has been put forward, for instance, by Grosjean and Gee [7]. In any case, one clear result emerges from the present study: the more plausible a sentence, the more likely listeners are to recognise it correctly under time-compression. Experiments using compressed speech may be highly subject to effects of reporting bias; indeed, the rhythmic effects which we observed in Experiment 1 may themselves have been bias effects. This suggests that further exploration of the perception of time-compressed speech should address possible effects of bias at all levels of linguistic structure. Prediction of bias towards, for example, particular prosodic or phonological structures is feasible, given that the increasing availability of spoken language corpora now enables realistic estimates of listeners' prosodic and phonological experience.

## 6. ACKNOWLEDGEMENTS

This research was funded by the Human Frontier Science Program. We are grateful to Jacques Mehler and to CNET for making the compression algorithm available to us, and we further thank Ian Nimmo-Smith for statistical advice.

## 7. REFERENCES

[1]: *Cutler, A.; Norris, D.:* The Role of Strong Syllables in Segmentation for Lexical Access. Journal of Experimental Psychology: Human Perception and Performance, Vol.14, pp. 113-121, 1988.

[2]: *Cutler, A.; Carter, DM.:* The Predominance of Strong Initial Syllables in the English Vocabulary. Computer Speech and Language, Vol. 2, pp. 133-142.1987.

[3]: *Lieberman, P.:* On the Acoustic Basis of the Perception of Intonation by Linguists. Word, Vol.21, pp. 40-54, 1965.

[4]: *Cutler, A.; Butterfield, S.:* Rhythmic Cues to Speech Segmentation: Evidence from Juncture Misperception. Journal of Memory and Language, Vol. 31, pp. 218-236, 1992.

[5]: *Charpentier, F.:* Traitement de la Parole par Analyse-Synthese de Fourier Application a la Synthese par Diphones. CNET, Lannion, 1988.

[6]: *Altmann, G.T.M.; Young, D.:* Factors Affecting Adaptation to Time-Compressed Speech. Eurospeech '93, Sept. 1993.

*[7]: Grosjean, F.; Gee, J.P.:* Prosodic Structure and Spoken Word Recognition. Cognition, Vol. 25, pp. 135-155, 1987.