PDF hosted at the Radboud Repository of the Radboud University
Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/15666

Please be advised that this information was generated on 2019-02-17 and may be subject to change.
Mетрическая структура и восприятие сжатого речи

Дункан Йоунг1, Гери Т. М. Альтман1, Анн Кутлер2, Денис Норрис2

1Экспериментальная психология, Университет Суссекс, Брайтон BN1 9QG, Англия.
2MRC Психологический Центр, 15 Чакерер Роуд, Кембридж CB2 2EF, Англия.

Абстракт

Абсenceee of explicitly marked cues to word boundaries, стeners tend to segment spoken English at the onset of strong syllables. This may suggest that under difficult listening conditions, speech should be easier to recognize because strong syllables are word-initial. We report two experiments in which listeners were presented with sentences that 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. The second study contrasted sentences in which all strong syllables were word-initial with sentences in which all strong syllables were word-final. The intelligibility of the two groups of sentences did not differ significantly. These results suggest that listeners are most likely to show up also as bias effects.

Ключевые слова: сжатая речь, ритм, реализм.

1. ВВЕДЕНИЕ

Английский язык имеет метрическую структуру; он варьируется в пределах сегментной структуры и маркирован различием между сильными и слабыми сегментами, которые содержат полные и сокращенные гласные. Например, в словах final, Christmas, over and tinkler, первый сильный сегмент является сильным, а второй слабым. Лейберман и Норрис [1] доказали, что носители английского языка разделяют наиболее эффективные моменты для сегментации сильных сегментов, чтобы обнаружить явления, которые содержат сокращенные гласные. Лейберман и Норрис предполагают, что это происходит из-за того, что слабые сегменты, которые начинаются с сильных сегментов, присутствуют в более чем шесть раз в множестве сегментов в тексте.

Исследования показали, что носители акустически-воспринимаемых ошибок более склонны к метрическим ошибкам, когда они слышатся на слух. Например, Лейберман [3] обнаружил, что фонально тренированные слушатели могут различать между сильными и слабыми сегментами, когда одна из метрических ошибок отсутствует в полностью сегментированной речи.

1 Experimental Psychology, University of Sussex, Brighton BN1 9QG, England.
2 MRC Applied Psychology Unit, 15 Chaucer Road, Cambridge CB2 2EF, England.
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 correctly 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.

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 that 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 (not 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 listener performance in recognising time-compressed speech may somehow be affected by rhythmic factors. Accordingly in a 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.

![Figure 1](image-url)
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 rhythmic 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. No other factor which we analysed showed a significant correlation with recognition performance.

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 the 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...
It is possible that this result simply reflects overall recognition performance did not materialise. Contrary to Cutler and Norris, the predicted effects on a form of speech whose recognition would be particularly 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 time-compressed speech. 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 the 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.

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 presence 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 on the strong/weak distinction in listeners' performance. Alternatively, it may be that weak-initial words are in general not mis-segmented, since application of metric segmentation may be coupled with lexical access procedure 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 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 the 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


