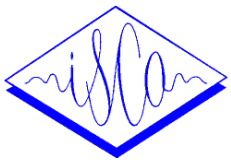

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

FACTORS AFFECTING ADAPTATION TO TIME-COMPRESSED SPEECH

Gerry T.M. Altmann and Duncan Young

Laboratory of Experimental Psychology,
University of Sussex,
Brighton BN1 9QG, UK.

ABSTRACT

Speech inputs vary widely in amount of background noise, speaking rate, and accent; yet listeners can quickly adapt to such variation and recognize the content of an utterance. The present study is aimed at understanding the mechanisms underlying this adaptation, and the processing unit which the human recognition system attempts to recover during adaptation. We examined adaptation to speech under various conditions of time compression. Specifically, we explored how the intelligibility of a target set of compressed English sentences varied as a function of prior exposure to three kinds of compressed stimuli: compressed English, compressed French, and in a separate study, compressed "nonsense" sentences. We also explored the degree to which the adapting effects of prior exposure to compressed speech persist over time. We conclude on the basis of the results that lexical level word recognition does not drive the adaptation mechanism. Instead, recognition of sub-lexical units, or suprasegmental regularities in rhythm, may form the basis for successful, and persistent, adaptation.

Keywords:compressed speech, adaptation, language-specificity

1. INTRODUCTION

Speech input varies widely for a number of reasons: speaking rate can vary, both within and across speakers; different speakers may speak in different accents; background noise may vary; and so on. Despite these changes, listeners can rapidly adapt to the variation in the speech signal and so arrive at the intended content of an utterance. The purpose of the present study is to understand further some of the mechanisms underlying this adaptation process. For instance, on a "fine-tuning" approach to adaptation, what drives the tuning process? What information causes this tuning to occur in the first place, and what information ensures that the system *converges* on the right setting (i.e. the adapted state)? Presumably, adaptation has evolved in order to ensure that different tokens of the same underlying (i.e. intended) form can nonetheless map onto the same internal structures. But what are the processing units which the human recognition system attempts to recover during adaptation? In principle, these units could be whole words, or sub-word units such as syllables or phonemes. The system may even attempt to recover *sequences* of units, perhaps using knowledge about the phonotactic properties of the language. If the information available to the adaptation process is language-specific, as would be the case if it concerned the phonotactic properties of the language, or the syllabic structures permitted in that language, then adaptation itself may be language-specific. In other words, the fine-tuning that might be achieved to variation in speech from one language may not necessarily serve equivalent variation in the speech from another language.

In a series of experiments we have explored factors affecting adaptation to speech which has been "time-compressed". Time compression is an automated process that results in a speech signal whose global characteristics remain the same as the original, but which sounds as if it had been uttered at a faster rate than the original. Briefly, the algorithm [4] identifies the pitch periods in the signal (and in an unvoiced portion of the signal it assumes a fixed-width period), and subsequently averages over adjacent periods to create a new signal which is shorter than the original, but which maintains the original pitch characteristics and perceived speaker identity. We have found that there is virtually no loss of intelligibility, or subjective "quality", when sentences are compressed to, for instance, 50% of their original duration.

Adaptation to time-compressed speech may not rely on the same mechanisms underlying adaptation to natural variation in speech rate, but it may nonetheless tell us something about the adaptive mechanisms available to the speech processing system. In the studies that follow, we investigated the ways in which prior exposure to time-compressed speech could affect the intelligibility of novel compressed speech presented subsequently. We had previously found that the intelligibility of a set of test sentences compressed to relatively high speech rates (measured in words per minute, for instance) was significantly improved if some prior exposure was given to a (different) set of compressed sentences [1]. This improvement, relative to a control condition with no prior exposure to compressed stimuli (or even relative to prior exposure to *un*compressed stimuli) is indicative of the processing system's ability to adapt to compressed speech.

2. EXPERIMENT 1

The first study replicates in part data reported in [1]. Compressed sentences, uttered in English, were presented to listeners under three experimental conditions: In the control

condition, the sentences were not preceded by any other compressed utterances; in the second condition they were preceded by other English sentences compressed to the same rate; and in the final condition, they were preceded by compressed sentences uttered by the same speaker in French. Although we expected the English condition to lead to higher intelligibility of the test sentences than the control condition, at issue was the advantage, if any, that would accrue from prior exposure to compressed sentences uttered in another language.

Stimuli A set of 15 sentences, spoken in English by a female English-French bilingual speaker, were recorded and subsequently digitized at 16Khz prior to being compressed to 37% of their original duration. Each sentence contained 11 words, and 15 syllables (e.g. "My grandparents were born nearly two years before the Great War"). French translations of these sentences, which maintained the number of words and syllables, were also recorded by the same speaker and compressed to the same rate (e.g. "Mes grandparents sont nes presque deux ans avant la Grande Guerre"). Prior studies had established that the intelligibility of highly compressed speech correlates highly with the plausibility of the sentences (measured on a 1 to 7 scale of "likelihood"). 20 English subjects rated the English stimuli for plausibility and on the basis of these ratings, the 15 sentences were divided into three groups of sentences which were matched as closely as possible with one another for plausibility.

Design The three groups of sentences appeared in two basic orderings: G1-G2-G3 and G3-G2-G1. The first ten sentences of these tapes (G1-G2 or G3-G2) were either in English (the "English" conditions) or in French (the "French" conditions). Thus there were four tapes in all, two in the English condition, and two in the French condition. In all four cases, the final 5 sentences (G3 or G1) were in English.

As stated at the outset, the purpose of this experiment was to explore three cases. The first, no prior exposure, is served by examining the intelligibility of G1 in G1-G2-G3, and the intelligibility of G3 in G3-G2-G1. The second, prior exposure to English, is served by examining the intelligibility of G3 in G1-G2-G3 and G1 in G3-G2-G1. The third, prior exposure to French, is as the second, but for those tapes in which G1-G2 or G3-G2 were in French. Thus, each sentence in G1 and G3 is heard (by a different subject) either with or without prior exposure to compressed speech.

Subjects Thirty-two University of Sussex students took part in this study, and were randomly allocated to one of four groups (corresponding to the four different tapes). None of the subjects could speak French (although they may have been familiar with the language and able to recognize it as French), and none had had any experience of listening to time-compressed speech.

Procedure Subjects were instructed to listen to each sentence and (in the English conditions) to write down as much of the sentence as they could. They were given approximately 18 seconds after each sentence to do this. In the French conditions, subjects were asked to listen to the first 10 French sentences and then to write down everything they could after each of the following five English sentences. A warning tone was played before each sentence. Subjects were played three uncompressed sentences in order to familiarise them with the procedure before proceeding to the experimental tape. The stimuli were presented over headphones.

Results The intelligibility of the target sentences was defined as the percentage of words in each sentence correctly recognised by listeners (alternative analyses based on percentage of syllables or content words correctly recognized produced the same overall patterns). Mean intelligibility rates for the sentences in groups G1 and G3 were calculated in order to examine the three conditions of interest:

Control condition: (G1 - G2 - G3; G3 - G2 - G1):
Prior English: (G1 - G2 - G3; G3 - G2 - G1):
Prior French: ($G1_{fr}$ - $G2_{fr}$ - G3; $G3_{fr}$ - $G2_{fr}$ - G1):

Figure 1 illustrates the appropriate means.

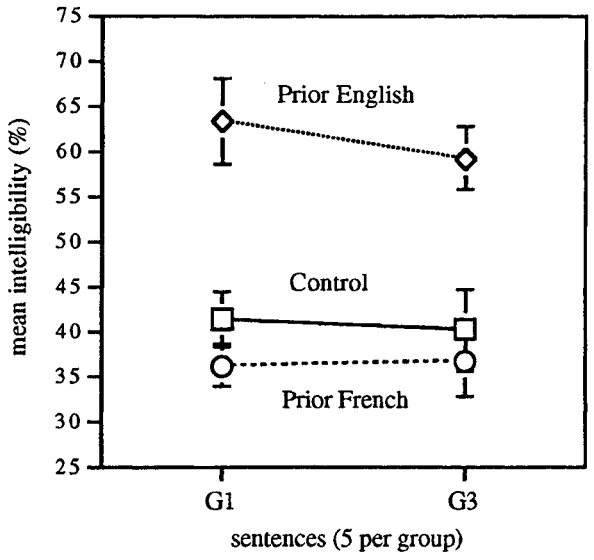

Figure 1
Mean intelligibility with standard error bars.

A two-way Analysis of Variance revealed that there was no difference in intelligibility between the two groups of sentences ($F<1$), although there was a significant difference between the three conditions ($F_{2,28}=28.5$, $p<0.0001$). Planned comparisons revealed that the difference between the control condition and the French condition was not significant ($F_{1,28}=1.67, p>0.2$), while the difference between the control condition and the English condition was significant ($F_{1,28}=55.4$, $p<0.0001$).

Discussion of Experiment 1 The fact that there was no significant adapting effect with prior exposure to compressed French leads us to suggest that the process of adaptation to time-compressed speech may be dependent upon language-specific factors. However, the design of this experiment does not allow us to distinguish between language-specificity due to phonotactic factors or due to rhythmic differences between the French and English sentences used in these studies.[1] It is conceivable that regularities in rhythm constrain the operations of the adaptation mechanism. The present experiment also fails to identify the nature of the speech unit whose extraction (or lack of extraction) drives the adaptation mechanism. For instance, if the "goals" of the adaptation mechanism are to fine-tune the system so that whole words can be extracted,

then the French stimuli may have been poor adaptors because no (English) words could be extracted. On the other hand, if the goal is to fine-tune the system so that permissible syllables can be extracted, then the French stimuli may have been poor adaptors because of differences in the syllabic structures of English and French.
In Experiment 2 we addressed this issue by exploring whether lexical level speech recognition is a necessary requirement for adaptation to compressed speech.

3. EXPERIMENT 2

This second experiment was similar in design to Experiment 1, and explored three conditions: In the control condition, the sentences were not preceded by any other compressed utterances; in the second condition they were preceded by other English sentences compressed to the same rate; and in the final condition, they were preceded by compressed "sentences" composed of "nonsense words" — that is, each nonword comprised a syllable or sequence of syllables which, although phonotactically permissible, did not constitute an English word. If successful adaptation requires the extraction of recognizable words, this latter condition should not differ from the control condition. However, if successful adaptation requires only the extraction of recognizable syllables, then this latter condition should not differ from that in which recognizable (compressed) words precede the test sentences.
Stimuli The same set of 15 sentences as used in Experiment 1 were uttered by a male English speaker. In addition, new versions of these sentences were created in which the content words were replaced by nonsense words with the same number of syllables (e.g. "Yai wannednearants were rorn venly waa grooz betock the woot graw"). The stimuli were recorded at a speech rate that was closely matched to the original stimuli from Experiment 1. Each sentence was compressed to 37% of its original duration.
Design The same design was employed as for Experiment 1, with the difference being that the French condition was replaced by a "nonsense" condition:

G1 - G2 - G3
G3 - G2 - G1
$G1_{nonsense}$ - $G2_{nonsense}$ - G3
$G3_{nonsense}$ - $G2_{nonsense}$ - G1

Subjects Thirty-two University of Sussex students took part in this study, and were randomly allocated to one of the four experimental groups. None of the subjects had taken part in Experiment 1.
Procedure See Experiment 1.
Results Figure 2 illustrates the appropriate means. A two-way Analysis of Variance revealed that there was no difference in intelligibility between the two groups of sentences ($F_{1,14}$=2.56,p>0.1), although there was a significant difference between the three conditions ($F_{2,28}$=23.67, p<0.0001). There was no interaction between sentence group and condition ($F_{2,28}$=1.63, p>0.1). Planned comparisons revealed that the difference between the control condition and the Nonsense and English conditions was significant ($F_{1,28}$=47.1, p<0.0001), while the difference between the Nonsense condition and the English condition was not significant (F<1).

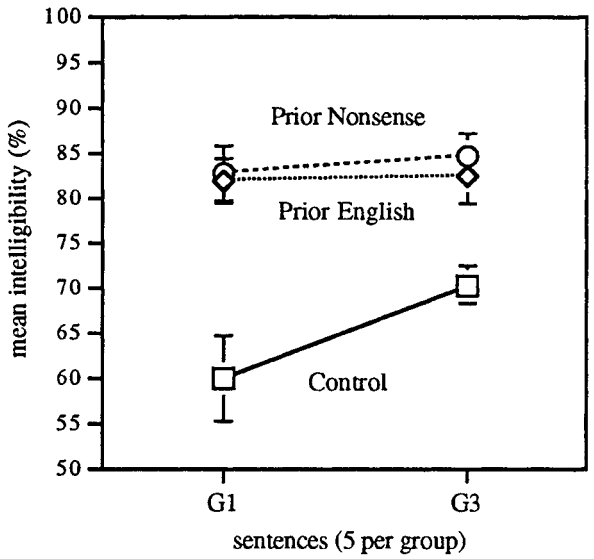

Figure 2
Mean intelligibility with standard error bars.

Discussion of Experiment 2 The finding that adaptation can occur to English nonsense utterances implies that lexical level speech recognition is not a necessary requirement for adaptation to time-compressed speech. These results, with those from Experiment 1, suggest that the recognition units underlying adaptation are pre-lexical and reflect language-specific properties of the language. Whether a single unit (e.g. the syllable) or a variety of different features (including rhythmic features) drive the adaptation process is at present unclear.
In Experiment 1, the baseline control condition was around 41% intelligible, compared to a baseline in Experiment 2 of 65%. This difference is probably due to the different voice characteristics of the speakers used in the two studies, which include differences in pitch — the sentences in Experiment 1 were uttered by a female speaker, and those in Experiment 2 by a male speaker. This difference may have contributed significantly to the different intelligibilities of the control conditions.
Although Experiments 1 and 2 allow us to rule out certain hypotheses concerning the nature of the perceptual units that drive the adaptation mechanism, they tell us little about the relationship between the pre- and post-adapted state. For instance, to what state does the system return when, following adaptation to compressed speech, re-adaptation to *un*compressed speech takes place? Is this re-adapted state identical to the state the system was in before adaptation took place? In order to explore this issue we conducted a further experiment in which we compared two populations of subjects — one which had had no exposure to compressed speech, and another which had been exposed briefly to compressed speech approximately one year earlier.

4. EXPERIMENT 3

In this third experiment we explored the degree to which subjects' adaptation to compressed speech is influenced by any previous experience of such adaptation.

Stimuli Fifteen new sentences (mean length = 11.1 words, each with 18 syllables) were uttered by the same male speaker from Experiment 2. They were all roughly matched for plausibility (see Experiment 1). Each sentence was compressed to 35% of its original duration.

Subjects Twenty-six University of Sussex students took part in this study. Fifteen of the subjects had previously taken part in one of a series of experiments that had involved listening to no more than 40 compressed sentences [5]. The experiment had taken place approximately one year previously. Eleven of the subjects had had no prior experience of compressed speech. All subjects were asked to listen to each sentence and to write down immediately afterwards as many of the words as they could remember hearing.

Results The intelligibility of each sentence was calculated for each subject, and the means for the two groups of subjects ("experienced" and "naive") are shown in Table 1:

subject group	intelligibility (%)	standard error
"experienced"	78	1.6
"naive"	67	3.3

Table 1

This difference was statistically highly significant ($t_{1,14}$=6.1, $p<0.001$).[2] Moreover, the difference was consistent across each of the 15 sentences ($p<0.001$). In other words, the difference was not located simply on the first few sentences.

Discussion of Experiment 3 The results from Experiment 3 demonstrate that adaptation to compressed speech is more than just a short-term re-tuning. Even though a year had elapsed since the "experienced" subjects had last listened to compressed speech, they found novel compressed stimuli consistently more intelligible than the control group. Whether this was because subjects could adapt very much more quickly on the basis of their prior experience, or because in some sense they could recall the precise tuning required is an open question. Nonetheless, some information relevant to the adaptation process must have persevered.

5. GENERAL DISCUSSION

The findings from Experiments 1 and 2 lead to two conclusions: First, the significant adapting effect of English nonsense utterances implies that lexical level speech recognition is not a necessary requirement for adaptation to time-compressed speech. Second, the fact that there was no significant adapting effect of the French leads us to suggest that the process of adaptation to time-compressed speech may be dependent upon language-specific phonotactic information, or that it may be sensitive to rhythmic differences between French and English sentences. Such a finding is unsurprising given, for instance, the data on syllabification in the two languages (e.g. [2,3]). If different segmentation strategies evolve in the different languages, and the role of adaptation is to ensure continued segmentation in the face of variation in the input, then it follows that the adaptation mechanism may well be sensitive to those same features of the language which give rise to language specificity in segmentation.

The third finding, that adaptation is long-lived (or that information pertinent to successful adaptation is long-lived), is perhaps more surprising. If adaptation is a matter of applying new parameters to the analysis and interpretation of the speech signal, we might instead expect that adapting to a new signal involves modifying the existing parameter set, and in effect replacing it with a new set of parameters that are optimized, via the adaptation process, for the analysis of the new signal. That is, we might expect to "un-learn" the earlier parameter set. We believe that the current results pose a challenge for both psychological and computational models of adaptation, and that the process of adaptation should be viewed, in such models, as the product of an extremely efficient and robust learning device.

6. REFERENCES

[1] *Mehler, J.; Sebastian, N.; Altmann, G.; Dupoux, E.; Christophe, A.; Pallier, C.:* Understanding Compressed Sentences: The Role of Rhythm and Meaning. The Proceedings of The New York Academy of Sciences workshop on "Temporal Information Processing and the Nervous System". pp. 272-282, New York, 1992.

[2] *Cutler, A.; Mehler, J.; Norris, D.; & Segui, J.:* The syllable's differing role in the segmentation of French and English. Journal of Memory and Language, **25**, pp. 385-400, 1986.

[3] *Cutler, A.; Mehler, J.; Norris, D.; & Segui, J.:* The monolingual nature of speech segmentation by bilinguals. Cognitive Psychology, **24**, pp. 381-410, 1992.

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

[5] *Young, D.; Altmann, G.; Cutler, A.; Norris, D.:* Metrical Structure and the Perception of Time-Compressed Speech. Eurospeech '93, Sept. 1993.

7. ACKNOWLEDGEMENTS

This research was part of a collaborative project involving Jacques Mehler, Emmanuel Dupoux, and Anne Christophe (Paris), and Nuria Sebastian (Barcelona). Our thanks to the Paris group for the effort that was put into refining the time compression algorithm kindly supplied by CNET. The design of Experiments 1 and 2 was conceived in conversations with Emmanuel Dupoux. This research was supported by the Human Frontier Science Program.

---

[1] Some caution is required given that the speech rates of the English and French sentences were not identical— the English sentences were uttered approximately 10% faster, measured in syllables per minute, than the French. An informal pilot study which used differential compression rates to correct for this difference nonetheless yielded the same pattern reported above.

[2] This same pattern was replicated with a slightly different design and with different stimuli.