Comprehending speech at artificially enhanced rates

Lucia Da Silva, Adriano V. Barbosa and Eric Vatikiotis-Bateson*

*Corresponding author's address: Linguistics, U British Columbia, Vancouver, V6T 1Z4, British Columbia, Canada, evb@mail.ubc.ca

This study is part of a larger study comparing the production, perception, and long-term comprehension of natural fast speech and artificially sped-up speech. In the 1950's, Fairbanks and colleagues made the interesting claim that artificially compressing speech to half its duration (twice the rate) and listening to it twice is comprehended better than listening to the original natural speed utterance once. Since this finding has been disputed [T. Sticht, J. Expr. Ed., 37, 60-62, 1969], we focus here on providing baseline results for perception of naturally produced speech at conversational and fast rate and of speech sped up to twice its original speed. Specifically, we compare how well the different stimuli are perceived when presented once, twice in rapid succession, and twice with delays of hours and days between presentations. We evaluate perception for both audio-only and audiovisual stimuli.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

For the past 60 years at least, sound playback devices have facilitated different practical aims by simply slowing down (useful for manual transcription) or speeding up (useful for locating key points) the playback. Changing playback speed in this way shifts all acoustic parameters in linear proportion. By contrast, when a speaker changes speaking rate, the effects are nonlinear. For example, increased speaking rate reduces duration of vowels more than consonants [1, 2], unstressed syllables in stress-timed languages such as English are reduced more than stressed syllables [3], and pauses between words and phrases are shortened.

Experimental studies have compared the effects of artificial and natural speech signals on listeners’ ability to identify words or smaller sound units such as syllables and phonemes and to comprehend message content [4]. Fairbanks, Guttman & Miron [5] showed that artificially compressing speech to half its duration (twice the rate) and listening to it twice results in better comprehension than listening to it once at its original speed. However, this gain was observed in tests administered immediately after exposure to the speech, within the domain of short-term memory. A major goal of this research is to examine the longer-term retention of information delivered at different naturally and artificially induced rates. Presenting spoken material to perceivers and testing their comprehension either immediately or after some delay show what differences there are in long-term retention of information and comprehension extracted from normal and fast rate speech, where fast rate speech is either naturally or artificially produced. In addition, stimuli presented audiovisually or only acoustically at different rates and naturalness, shows the effects of sensory modality on comprehension and retention.

The motivation for pursuing this study comes from the great volume of online videos available on the Web and the demand for staying up-to-date with a large volume of multimedia information.

METHODS

Production

Digital video (HD 720p) recordings were made for two male and two female speakers of North American English. The image frame includes the head and shoulders of each speaker as he or she reads the first 150 sentences from the Harvard sentences [6]. Speakers read the sentences in fifteen blocks of ten sentences each, first at a comfortable or ‘normal’ rate and again at a much faster rate.

Artificially compressed speech

To create the artificially sped-up speech, the normal rate recordings are linearly time-compressed to match the duration of the naturally produced fast speech. This results in 12 recorded conditions for analysis and perceptual evaluation: normal, naturally fast, and artificially fast sentence sets for four speakers.

Perceptual evaluation

One hundred and twenty native speakers of North American English (university undergraduate and graduate students) with no reported hearing impairment participated in the perceptual evaluation.

The finding by Adank and Devlin [7] that comprehension of time-compressed speech improves over time influenced both the design and procedure of the perceptual evaluation. Regarding procedure, in order to enhance the comprehension performance, subjects are exposed to sixteen trial sentences before the start of the experiment.

Immediate test

Immediate testing is aimed at replicating previous results. Sixty perceivers are tested on 10 different sets of 12 sentences (3 sentences for each of 4 speakers) presented at normal, naturally fast and artificially fast speech rates for two modes of presentation: audio only and audiovisually. The stimuli are presented once and sentence set assignment to test conditions are balanced across perceivers. A second group of 60 perceivers are tested on the same stimuli, but in this case each stimulus sentence is presented twice.
Delayed test

Delayed testing addresses our primary interest in how different natural and unnatural manipulations of speaking rate affect long-term retention and comprehension. The 60 perceivers who were exposed in the immediate test to the stimulus sentences only were presented with the same stimuli again one day later. Again sentence set assignment are counterbalanced.

CONTRIBUTIONS

This study contributes to the availability of compressed multimedia materials for examining methods of enhancing learning. This and subsequent studies may guide development of new tools to facilitate distance- and other unusual learning. Another contribution of this research lies in the study of correlations between different dimensions of human communication. In the real world, channels other than verbal information (e.g. body language and facial expressions) contribute to the generation of sense and meaning. Until recently, technology has focused on text and audio, but there is now an overwhelming emphasis on audiovisual media where visible motions of the face and head influence perception and comprehension, for better or for worse. That is, auditory and visual streams have been shown to reinforce one another in redundant and/or complementary ways [8, 9, 10, 11], but little is known about how these modalities will interact under the sort of rate manipulation and long-term memory task proposed here. Furthermore, little is known about the ongoing processing of auditory and visual speech information in longer dialogues [12], which is a major goal of future studies.

ACKNOWLEDGMENTS

This research was supported in part by the Canada Tri-Council (SSHRC and NSERC) grants to Eric Vatikiotis-Bateson. We would like thank our production volunteers.

REFERENCES