Cognition 151 (2016) 68–75

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Some people are "More Lexical" than others

Mako Ishida [a,d], Arthur G. Samuel [b,c,d,*], Takayuki Arai [a]

[a] Sophia University, Tokyo, Japan
[b] Basque Center on Cognition, Brain and Language, Donostia, Spain
[c] IKERBASQUE, Basque Foundation for Science, Spain
[d] Stony Brook University, Dept. of Psychology, Stony Brook, NY, United States

ARTICLE INFO

Article history:
Received 6 June 2015
Revised 4 March 2016
Accepted 8 March 2016
Available online 14 March 2016

Keywords:
Degraded speech
Perceptual units
Lexical support for speech perception

ABSTRACT

People can understand speech under poor conditions, even when successive pieces of the waveform are flipped in time. Using a new method to measure perception of such stimuli, we show that words with sounds based on rapid spectral changes (stop consonants) are much more impaired by reversing speech segments than words with fewer such sounds, and that words are much more resistant to disruption than pseudowords. We then demonstrate that this lexical advantage is more characteristic of some people than others. Participants listened to speech that was degraded in two very different ways, and we measured each person's reliance on lexical support for each task. Listeners who relied on the lexicon for help in perceiving one kind of degraded speech also relied on the lexicon when dealing with a quite different kind of degraded speech. Thus, people differ in their relative reliance on the speech signal versus their pre-existing knowledge.

1. Introduction

People can understand speech under an impressively wide range of listening conditions. In fact, one difficulty that speech researchers face is that the process is so good that it is difficult to examine the system's operation because it works so quickly and accurately. For that reason, researchers have used various techniques to stress the system in order to be able to probe what it is doing. These techniques include presenting the speech in noise (e.g. Miller & Isard, 1963), filtering away different parts of the spectrum (e.g., Wilson, Zizz, Shanks, & Causey, 1990), vocoding the speech (e.g., Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005), compressing the speech in time (e.g., Dupoux & Green, 1990), dropping out pieces of the signal (e.g., Huggins, 1964), and other manipulations that impair perception enough to see when errors occur.

In the current study, we use a method that Saberi and Perrott (1999) introduced (see also Steffen & Werani, 1994) that we will call Locally Time-Reversed Speech, or LTRS. With LTRS, an utterance is first segmented into pieces of a fixed size (e.g., every 50 msec, or every 100 msec), each such segment is then reversed along the time axis (i.e., played backwards), and the segments are put back together. This is quite different than simply playing the whole utterance backwards, as it breaks up the speech every N msec. Now, if each segment were tiny, e.g., only 1 msec long, this manipulation would not harm the signal very much because there is not much change in the waveform on such a short scale. The surprising result reported by Saberi and Perrot was that the segments could be quite long before listeners thought that intelligibility was impaired – with segments as long as 130 msec, listeners only rated the speech as having lost half its intelligibility.

Saberi and Perrott's (1999) report has spawned about a half dozen other studies of LTRS, and has been cited in arguments about whether speech is decoded into syllable-sized units (suggested by the long segments that are tolerated) versus phoneme-sized units. The basic effect has now been shown in French (Magrin-Chagnolleau, Barkat, & Meunier, 2002) and in German (Kiss, Cristescu, Fink, & Wittmann, 2008), in addition to English (Greenberg & Arai, 2001; Remez et al., 2013). In all cases, researchers have shown that with very small segments (e.g., 10 msec) performance is quite good, and as the segments get longer, performance declines. There has been some variation across studies in how quickly the curve falls as a function of segment size, but this variation presumably mostly traces to differences in how the measurements were done. For example, Remez et al. pointed out that some studies, including the original Saberi and Perrot paper, presented a very small number of stimuli repeatedly (e.g., only a single sentence in the original paper), and asked subjects to rate intelligibility, whereas in other studies listeners were required to

* Corresponding author at: Department of Psychology, Stony Brook University, Stony Brook, NY 11794-2500, USA.

E-mail address: Arthur.Samuel@stonybrook.edu (A.G. Samuel).

http://dx.doi.org/10.1016/j.cognition.2016.03.008
0010-0277/

report the words that were presented. Remez et al. included both measures, and as expected, found that subjective intelligibility yielded much longer estimates of the tolerated segment size than actual intelligibility measures.

Having listeners transcribe what they hear is clearly a better measure of perception than getting subjective intelligibility ratings, but even the transcription approach has some significant limitations. Because transcription of sentences is often slow and difficult, in most studies researchers have had to present each stimulus repeatedly to allow listeners to write or type what they hear. This repetition interacts with the use of sentences in which the listeners clearly are using the context to guess quite a bit, making the reports less informative about what is actually being perceived. In the current study, we apply a technique based on signal detection procedures to measure what listeners perceive when they hear LTRS stimuli. This approach overcomes the existing problems, and allows us to address two important issues.

In Experiment 1 we use this method to examine whether the "critical" time window depends on the phonetic properties of the speech. As Remez et al. (2013) pointed out, some speech sounds (e.g., vowels and fricatives) have relatively static spectral properties, and on these grounds one might expect that perception of such sounds would be more robust across longer time windows than speech sounds like stop consonants that change rapidly over time. The first experiment thus assesses how perception of degraded speech depends on the details of the phonetic signal.

In Experiment 2, we use our technique to determine how much perception of such degraded speech relies on support from lexical representations. Prior research with speech degraded in other ways has demonstrated that listeners do use lexical context to guide perception of phonetic input (e.g., Grataloup et al., 2009; Samuel, 1981, 1996). In addition to looking for lexical influences on perception of LTRS stimuli, in Experiment 2 we ask whether the degree to which a given listener relies on the lexicon is a general property of that listener: If an individual brings to bear lexical information when trying to understand speech degraded via LTRS, does that same individual also tend to rely on the lexicon when confronting speech with a very different kind of challenge? For this test, we had the same group of listeners listen to LTRS stimuli, and to stimuli in which a single phoneme could be replaced by white noise. Warren (1970) discovered that listeners are generally unable to detect that a phoneme is missing when it is replaced by a loud noise, an effect he termed phonemic restoration. Samuel (1981, 1996) demonstrated that a significant source of the restoration is lexical knowledge – listeners restore missing phonemes more in words that they know than in matched pseudowords. In Experiment 2, we test whether the lexical influence on phonemic restoration correlates with any lexical influence on how listeners perceive LTRS stimuli.

In our new LTRS paradigm, on each trial listeners hear two items. Each item is either a word (e.g., "academic") or a pseudoword that differs by one phoneme (e.g., "acabemic"). The first item is a locally time-reversed stimulus in a male voice and the second item is normally produced, in a female voice. The task is to judge if the two items were the same (e.g. "academic" followed by "academic", or "acabemic" followed by "acabemic"), or different (e.g., "academic" – "acabemic", or "acabemic" – "academic"). If perception of the LTRS-modified first item is good, judging its phonetic similarity to the normal second item should be accurate, but if LTRS modification disrupts perception, making the same-different judgment will be difficult. As will be seen, this method provides a very sensitive signal detection based measure of perception of the degraded speech. Because this method uses a discrimination task in which listeners are never asked to report what word (or pseudoword) they hear, the results cannot be traced to any post-perceptual decision stage.

2. Experiment 1

Experiment 1 had three goals. First, we sought to establish that our signal detection test was an excellent way to assess perception of LTRS stimuli. Second, we wanted to determine whether lexical support played a significant role in perceiving such stimuli. Finally, we wished to see whether the critical time-window differed for words that differed in their relative proportions of more static sounds (fricatives) versus more dynamic ones (stops).

2.1. Methods

2.1.1. Participants

A total of 60 undergraduate students from Stony Brook University (12 males, 48 females) participated in Experiment 1. All were native American English speakers, age 18 or older, and had no known hearing problems. They received research credit for their participation.

2.1.2. Stimuli

Two sets of words were selected from the MRC Psycholinguistic database: 72 fricative-dominant stimuli and 72 stop-dominant stimuli. All words were 3, 4, or 5 syllables long, had a Kucera-Francis written frequency greater than 1/million, and a familiarity rating of 500–700 in the MRC database. Fricative-dominant words were chosen to have relatively many fricatives ($M = 1.72$) as compared to stops ($M = 0.72$), while stop-dominant words contained more stops ($M = 3.18$) than fricatives ($M = 0.79$). Working within the constraints of the overall occurrence rates of stops and fricatives in English words of this sort, these two sets provide a strong test of whether perceptual degradation via LTRS depends on the degree to which the speech has rapid spectral changes versus relatively steady-state segments: There are over twice as many steady-state fricatives in the fricative-dominant set as in the stop-dominant set, and over four times as many stops in the stop-dominant set than in the fricative-dominant set. The mean frequency of words in the two sets was matched (fricative-dominant words: 16.45; stop-dominant words: 18.32), based on the CLEARPOND database (Marian, Bartolotti, Chabal, & Shook, 2012).

After selecting the fricative- and stop-dominant words, matched pseudowords were created by changing the place of articulation of one phoneme in each word (e.g., "academic" – "acabemic"). This degree of change was designed to be small enough that if perception were to be impaired, errors would be likely. With this procedure, there were 72 fricative-dominant words, 72 fricative-dominant pseudowords, 72 stop-dominant words, and 72 stop-dominant pseudowords. All items were spoken by one male and one female native American English speaker, and recorded in a sound shielded room. The sounds were first digitized at 48 kHz (16 bit), and downsampled to 16 kHz (16 bit). The amplitude of stimuli was also normalized with GoldWave digital audio software, and each token was stored as a WAV file (16 Hz, 16 bits, mono). Words and pseudowords from the male voice were locally time-reversed by MATLAB with reversal window lengths of 10, 30, 50, 70, 90, and 110 ms. The MATLAB script imposed 5-msec linear onset and offset amplitude shaping at the points where successive segments were joined to prevent clicks from being generated at those points. The sets of words and pseudowords are shown in the Appendix.

2.1.3. Procedure

Participants were randomly assigned to either the fricative-dominant or stop-dominant stimuli group. Each trial included a locally time-reversed stimulus in the male voice followed by a