Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 59 (2014) 1–9
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Determining the relevance of different aspects of formant contours to intelligibility ☆

Akiko Amano-Kusumoto [a,b,*], John-Paul Hosom [b,1], Alexander Kain [b], Justin M. Aronoff [a,2]

[a] Department of Human Communication Science Devices, House Research Institute, 2100 West Third Street, Los Angeles, CA 90057, United States
[b] Center for Spoken Language Understanding (CSLU), Oregon Health & Science University (OHSU), 20000 NW Walker Road, Beaverton, OR 97006, United States

Received 15 April 2013; received in revised form 11 November 2013; accepted 8 December 2013
Available online 18 December 2013

Abstract

Previous studies have shown that "clear" speech, where the speaker intentionally tries to enunciate, has better intelligibility than "conversational" speech, which is produced in regular conversation. However, conversational and clear speech vary along a number of acoustic dimensions and it is unclear what aspects of clear speech lead to better intelligibility. Previously, Kain et al. (2008) showed that a combination of short-term spectra and duration was responsible for the improved intelligibility of one speaker. This study investigates subsets of specific features of short-term spectra including temporal aspects. Similar to Kain's study, hybrid stimuli were synthesized with a combination of features from clear speech and complementary features from conversational speech to determine which acoustic features cause the improved intelligibility of clear speech. Our results indicate that, although steady-state formant values of tense vowels contributed to the intelligibility of clear speech, neither the steady-state portion nor the formant transition was sufficient to yield comparable intelligibility to that of clear speech. In contrast, when the entire formant contour of conversational speech including the phoneme duration was replaced by that of clear speech, intelligibility was comparable to that of clear speech. It indicated that the combination of formant contour and duration information was relevant to the improved intelligibility of clear speech. The study provides a better understanding of the relevance of different aspects of formant contours to the improved intelligibility of clear speech.

Keywords: Speech intelligibility; Vowel perception; Speech synthesis

1. Introduction

People often adopt a particular speaking style to better communicate with hard-of-hearing or cognitively-impaired individuals, referred to as *clear* speech. Previous research has found that clear speech is more intelligible than *conversational* speech, which is spoken in regular communication (Picheny et al., 1985). The acoustic–phonetic characteristics of clear speech have been shown to differ from those of conversational speech (Picheny et al., 1986). For example, clear speech has slower speaking rate, expanded vowel space, increased energy at higher frequency regions, and lengthened phoneme durations (Picheny et al., 1986; Ferguson and Kewley-Port, 2002; Krause and Braida, 2004).[3] Given the various acoustic–phonetic characteristics of clear speech, what makes clear speech more intelligible is still questionable.

☆ Part of this study was presented in Amano-Kusumoto and Hosom (2009).

* Corresponding author at: Department of Human Communication Science Devices, House Research Institute, 2100 West Third Street, Los Angeles, CA 90057, United States. Tel.: +1 213 353 7084.
E-mail address: akiko.amano@gmail.com (A. Amano-Kusumoto).
[1] Currently at Sensory Inc.
[2] Currently at Department of Speech and Hearing Science, University of Illinois at Urbana-Champaign, 901 South Sixth Street, Champaign, IL 61820, United States.

[3] For a more complete review, see (Amano-Kusumoto and Hosom, 2011).

0167-6393/$ - see front matter
http://dx.doi.org/10.1016/j.specom.2013.12.001

To address the question, Kusumoto et al. (2007) and Kain et al. (2008) examined the contribution of certain acoustic features to improved intelligibility of clear speech by using a *hybridization* algorithm. The hybridization algorithm is a tool for creating hybrid stimuli that have certain features of clear speech utterance, and complementary features of conversational speech. The hybrid speech makes it possible to examine the perceptual relevance of certain features to the improved intelligibility of clear speech. Perceptual experiments indicated that hybrid speech with short-term spectra and phoneme durations modeled after clear speech yielded significant improvement in intelligibility over unprocessed conversational speech (Kain et al., 2008). Thus, the authors concluded that the combination of short-term spectra and phoneme durations caused the improved intelligibility of clear speech. However, the short-term spectrum contains a large number of features, therefore, the goal of this study was to investigate which specific features of short-term spectra, including the temporal aspect, can yield better intelligibility. The outcome of this study may be useful for assistive listening devices, which could provide speech with increased intelligibility for listeners with hearing impairment or in adverse listening conditions by enhancing relevant features.

Formant frequencies are the spectral peaks of the short-term spectrum, resulting from the positions of the articulators (e.g. tongue and lip positions) in the oral cavity. The formant contours, the time course of formant frequencies, reflect the movement of articulators from one position to another over time. Previous studies showed that the formant contour over the course of a vowel is perceptually important (Hillenbrand and Nearey, 1999) and that word intelligibility is significantly correlated with mean word duration and the difference between the F2 values of /i/ and /u/ (Hazan and Markham, 2004). Therefore, in the first hybrid condition (HYB-C), we examined whether the formant contour was relevant to improved intelligibility. We excluded other features from the short-term spectra such as formants over 4 kHz, formant residuals, and formant bandwidths that were not reported to be different between conversational and clear speech in most cases (Krause and Braida, 2004).

However, the entire formant contour may not be responsible for the improved intelligibility of clear speech, and a subset of the formant contour may be sufficient. Furui (1986) showed that the formant frequencies at the maximum spectral transition are sufficient to identify vowels. Additionally, Moon and Lindblom (1994) showed that the steady-states at the middle point of the vowel are shown to be different between conversational and clear speech, and that F2 at the transition moves faster in clear than in conversational speech. In the second hybrid condition (HYB-MT), we examine the relevancy of formant frequencies at the middle point and formant transition, and removing other features (points between transition and middle point, and duration of clear speech) from the first condition, HYB-C.

Finally, it may be sufficient to have formant frequencies from clear speech at the middle point of the vowel but not the formant transition of clear speech. Previous studies showed that the vowel space representing F1 and F2 values was expanded along both dimensions in the clear speaking style (Bradlow et al., 2003; Krause and Braida, 2004), and that formant steady-states at the middle point of the vowel in conversational speech do not reach their targets as much compared to clear speech (/w*V*l/ context) (Moon and Lindblom, 1994). The extreme steady-state values of clear speech may be the cause of improved intelligibility. Therefore, the third hybrid condition (/HYB-M/) was used to examine the relevancy of the formant values at the middle point of the vowel (defined as steady-state values). In summary, this study examined the relative contribution of formant contour, a combination of steady-state formant values and formant transition, or steady-state values alone, and whether these features resulted in an improved intelligibility compared to conversational speech.

2. Speech corpus

Speech materials containing conversational and clear speech were recorded for this study, using a previously developed word list (Moon and Lindblom, 1994). This list contains /w*V*l/ words designed to achieve (1) large consonant–vowel formant transitions, (2) the same degree of stress on the test word, and (3) systematic change of the vowel duration. The four front vowels (/i/, /ɪ/, /ɛ/ and /ei/) surrounded by the consonants /w/ and /l/ were recorded.

Four words (*wheel, will, well*, and *whale*) in a carrier sentence were repeated 16 times each. The carrier sentence "it's easy to tell the size of a *WORD*" was used to facilitate the use of fundamental frequency (F0), duration and intensity as necessary, upon the elicitation of conversational and clear speech. A total of 128 tokens were recorded (4 words × 2 speaking styles × 16 repetitions).

The speech signals were recorded digitally at a sampling rate of 16 kHz with 16-bit resolution. One male, a native speaker of North-American English with no professional training in public speaking, was recruited as a speaker because of his previous experience producing conversational and clear speech in the laboratory (Kain et al., 2008). When recording conversational speech, the speaker was instructed to recite the materials in a way that he would use to communicate in his daily life. When recording clear speech, he was instructed to speak clearly, as he would talk when communicating with elderly listeners or hearing-impaired listeners. The recording of conversational speech was followed by clear speech in the first and second recording sessions, respectively, using the speaker's own distinction between conversational and clear speech production.

The average speaking rates in words per minute (wpm), measured excluding pause durations, were 366 wpm and 179 wpm for conversational (CNV) and clear (CLR) speech,