VARIABILITY OF THE CONSONANT MODULATION SPECTRUM ACROSS INDIVIDUAL TALKERS

Pamela Souza[1] and Frederick Gallun[2]

[1]Dept. of Speech & Hearing Sciences, University of Washington, 1417 NE 42nd Street, Seattle, WA, USA, 98105, psouza@u.washington.edu

[2]National Center for Rehabilitative Auditory Research, Portland VA Medical Center, 3710 SW US Veterans' Hospital Road, Portland, OR, 97239, USA, Frederick.Gallun@va.gov

1. INTRODUCTION

Our previous work (Gallun & Souza, 2008) demonstrated that vowel-consonant-vowel syllables with similar modulation spectra were likely to be confused with one another. This implied that each consonant is identified by its modulation spectrum. However, if the modulation spectrum for a particular phoneme varies substantially across talkers, that would argue against use of those cues for consonant identification. This follow-up study investigated variability of consonant modulation spectra across individual talkers.

2. METHOD

2.1 Materials.

Speech recordings were drawn from a database (Markham & Hazan, 2002) which included adult male and female, and child male and female talkers. Each talker produced the same set of vowel-consonant-vowel tokens. This included 23 consonants in three vowel contexts (/ɑ/, /ɹ/, and /i/). The tokens were recorded on digital audio tape and transferred to a computer hard drive for analysis.

2.2 Modulation spectrum and spectral correlation index.

Details of the processing are available in Gallun and Souza (2008). Briefly, the modulation spectrum of each signal was calculated by (a) filtering the VCV into six octave-wide bands centered at .25, .5, 1, 2, 4, and 8 kHz (b) half-wave rectifying each of the filtered signals (c) low-pass filtering the rectified signal at 50 Hz (d) downsampling at 1000 Hz and completing a Fast-Fourier Transform (FFT). Prior to the FFT, the signal was zero-padded such that the duration of the signal was extended to five seconds, thus allowing a frequency resolution in the FFT of .2 Hz (e) the energy in each .2 Hz bin between .2 and 64 Hz was summed with the energy in adjacent bins. The choice of which bins to sum was made such that the summed energy was obtained for the equivalent of six, 1-octave wide rectangular filters with center frequencies stretching from 1 Hz to 32 Hz (f) the summed energy value was divided by the energy in the 0 Hz bin to provide a normalized modulation index value, which indicates the relative amount of modulation. Thus, each phoneme was represented by a matrix of thirty-six values (six modulation frequencies x six carrier frequencies). Note that there is no information about the relative phases of the modulation across channels. Figure 1 shows an example of the modulation spectrum for the /aza/ (for figure clarity, only three of the six bands are shown).

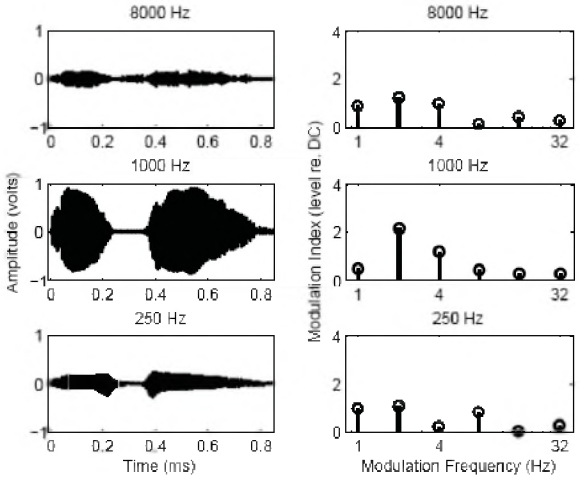

Figure 1. Example of modulation spectra (the .25 kHz, 1 kHz and 8 kHz carrier bands are shown) for the syllable /aza/. Left panels show the output of the first filtering stage; right panels show the amount of modulation at each modulation frequency from 1 to 32 Hz, with larger modulation index values indicating greater modulation.

Next, Spectral Correlation Index (SCI) (Gallun & Souza, 2008) values were calculated across subsets of the stimulus set. The SCI between two signals is obtained by concatenating the 36 modulation spectrum values into a single vector and correlating (Pearson *r*) the vectors. In this way, a single SCI value can be obtained for each pairing of phonemes.

3. RESULTS

3.1 Consonant similarity

Consistent with Gallun and Souza (2008), there was a wide range of SCIs across different consonant tokens produced by the same talker. Figure 2 shows an example for a single adult male talker, with all consonants compared to /aza/. High SCI values indicate the consonants that have the most similar modulation spectrum to /aza/ (e.g., /aʃa/, /adʒa/) and the least similar modulation spectrum to /aza/ (e.g., /apa/, /afa/). The phoneme /aza/, correlated with itself, has SCI = 1.0.

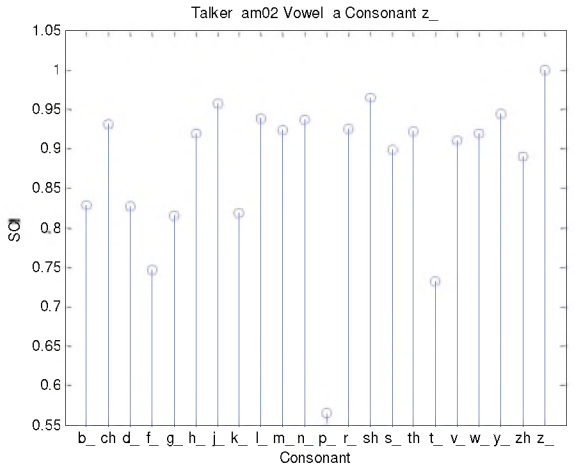

Figure 2. SCI values for a single male talker for all consonants compared to /z/, for the /a/ vowel context. SCI ranges from about 0.5 (for /p/) to 1.0 (for /z/, correlated with itself).

3.2 Talker variability

Results indicated that the modulation spectrum for a single VCV token was very similar across talkers. Typical results are shown for /aza/ in Figure 3. This shows how similar the modulation spectra are (expressed as SCI) for each adult male talker relative to adult male talker #11 (who has a SCI value of 1 with himself). Despite the range in voice pitch, vocal quality, and the fact that all talkers were untrained and produced the VCVs spontaneously, SCI values for all talkers are 0.9 or higher. This supports the idea that consonant identification is based on modulation characteristics which are maintained across individual productions of the consonant.

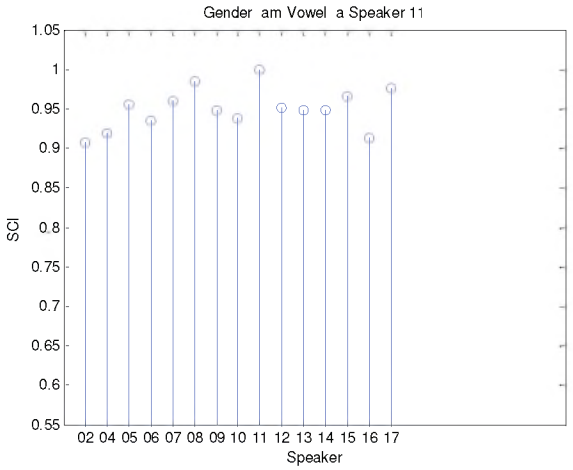

Figure 3. SCI values for all adult male talkers compared to adult male #11, for /aza/. The high (> 0.9) values indicate that all talkers produced /aza/ with a similar modulation spectrum.

4. DISCUSSION

Our previous work showed that similarity of the modulation spectrum, as expressed by the SCI, is significantly related to consonant error patterns. That suggests that modulation spectrum can be used to characterize the temporal (modulation) cues to consonant identity. This study examined how the modulation spectrum varies as the same utterance is produced by a variety of talkers. In general, modulation spectra for the same consonant across speakers are very similar, compared to modulation spectra for different consonants.

REFERENCES

Gallun F., Souza P. (2008). Exploring the role of the modulation spectrum in phoneme recognition. Ear and Hearing, in press.

Markham D., Hazan V. (2002). Speaker intelligibility of adults and children: The UCL Speaker Database. Speech, Hearing and Language: work in progress (pp. 1685-1688).

ACKNOWLEDGMENTS

The authors gratefully acknowledge the assistance of Valerie Hazan and University College London in providing the acoustic recordings and Eric Hoover for his help with stimulus processing. This work was supported by NIDCD (DC 006014 and DC04661), the National Center for Rehabilitative Auditory Research, and the Bloedel Hearing Research Center.