MULTI-TIME RESOLUTION ANALYSIS OF INTEGRATED ACOUSTIC INFORMATION IN REDUCED SPEECH

Megan M. Willi & Brad H. Story

Speech, Language, and Hearing Sciences, University of Arizona, Tucson, AZ;
mkittles@email.arizona.edu, bstory@email.arizona.edu

ABSTRACT

Listeners reliably extract meaning from spontaneous conversational speech even though the acoustic signal may lack many of the typical characteristics of citation speech. How listeners perceive these reduced signals is not well understood. The proposed demonstration presents a novel signal processing technique capable of extracting reliable acoustic information related to place-of-articulation information from reduced, stop consonant variants. A multi-time resolution analysis approach will be used to define the relative acoustic contributions of overlapping vowel and consonant segments. The acoustic patterns generated by analyzing the relative acoustic contributions of the overlapping vowel and consonant segments are called relative formant deflection patterns. Relative formant deflection patterns are a proposed invariant cue for place-of-articulation developed using a computational model of speech production. The demonstration will illustrate the extraction of relative formant deflection patterns from reduced stop consonant audio stimuli. The stimuli will be generated using a computational model of speech production. The demonstration will include animations of the tiers of the computational model used to simulate the reduced speech signals, graphics of the linear predictive coding (LPC) process used to estimate the relative vowel and consonant components of the overlapping speech signal, and validation of the acoustically extracted relative formant deflection patterns.

1. INTRODUCTION

The acoustic signal produced during natural conversational speech is highly impoverished. In comparison to the information native listeners extract from conversational speech, the acoustic signal appears highly degraded, overlapping, or even absent. Although spoken language comprehension is known to be a complex process mediated by listeners' contextual and linguistic knowledge, the influence of the acoustic information remaining in these reduced speech signals is not well understood. The proposed demonstration presents a novel signal processing technique capable of extracting reliable acoustic information related to place-of-articulation information from simulated, reduced speech signals lacking other known acoustic cues capable of explaining participants' perceptions.

2. COMPUTATIONAL MODEL OF SPEECH PRODUCTION

The presentation will highlight an invariant acoustic cue for place-of-articulation developed using a computational model of speech production and demonstrate a new methodology for extracting this type of invariant acoustic cue from simulated, reduced speech. The computational model of speech production used to investigate the proposed acoustic cue is based on magnetic resonance imaging data (MRI) of the vocal tract and involves perturbing the vocal tract shape in multiple hierarchical tiers [1]. The first tier changes the overall shape of the vocal tract simulating the area functions needed to create a vowel-to-vowel (VV) transition and the second tier perturbs the vocal tract at a particular location simulating the area functions needed to create a consonant constriction (C). When an acoustic wave resulting from the combined area functions is produced, a vowel-consonant-vowel utterance is simulated. For example, in the vowel-consonant-vowel (VCV) utterance [ədi], the first tier would change the overall shape of the vocal tract to produce the underlying vowel-to-vowel (VV) transition [əi] and the second tier would perturb the vocal tract shape created by the vowel-to-vowel (VV) transition at a specific location to simulate the consonant production in [ədi].

3. RELATIVE FORMANT DEFLECTION PATTERNS

By calculating the formant transitions produced in each tier of the model, the relative acoustic contribution of the formant transitions resulting from the consonant constriction (C) can be "demodulated" from the relative acoustic contribution of the formant transitions resulting from the underlying vowel-to-vowel (VV) transition. The acoustic patterns generated by analyzing the relative acoustic contributions of separable vowel and consonant vocal tract modulations are called *relative formant deflection patterns*.

Unlike formant transitions, relative formant deflection patterns are highly predictive of participants' perceptions and invariant across vowel contexts [2].

Previous studies of speech perception using formant synthesizers and unnatural manipulations of recorded speech could not guarantee that the resulting signals could be produced by a human vocal tract. The method used to identify relative formant deflection patterns is unique because it is based on an anatomically and physiologically guided model of production. The advantage of the current speech signal processing technique is that it will theoretically be present in natural speech signals, potentially even reduced conversational speech.

4. DEMONSTRATION

The purpose of the demonstration is to illustrate the extraction of relative formant deflection patterns from coarticulated speech. A novel, multi-time resolution analysis of speech will be used to define the relative acoustic contributions of overlapping vowel and consonant segments and generate the resulting relative formant deflection patterns. For the purposes of the demonstration, multiple place-of-articulation continua of approximant-like, voiced English stops will be simulated using a computational model of speech production and then will be analyzed using the proposed acoustic analysis. Approximant-like or acoustically continuous variants of voiced stop consonants occur frequently in conversational speech, but how listeners perceive the intended phoneme category from these highly reduced signals is poorly understood [3]. The demonstration will present a potential invariant acoustic cue for place information capable of categorizing these reduced, voiced stop consonants. The demonstration will include animations of the tiers of the computational model used to simulate the reduced speech signals, graphics of the linear predictive coding (LPC) processes used to estimate the formant tracks from the acoustic signal that correlate to each of the modulation tiers, and validation of the acoustically extracted relative formant deflection patterns using the known relative formant deflection patterns produced by the model.

5. CONCLUSION

The proposed signal processing technique has strong implications for automatic speech recognition (ASR). Natural, conversational speech contains highly integrated acoustic information. Processing the speech signal using a multi-time resolution approach that is guided by a theoretical framework of how the speech signal is encoded may help establish new acoustic regularities in the speech signal that could be used to improve automatic speech recognition.

6. REFERENCES

[1] B.S. Story, "A parametric model of the vocal tract area function for vowel and consonant simulation," *The Journal of the Acoustical Society of America,* pp. 3231-3254, 2005.

[2] B.S. Story and K. Bunton, "Relation of vocal tract shape, formant transitions, and stop consonant identification," *Journal of Speech, Language, and Hearing Research,* pp. 1514-1528, 2010.

[3] N. Warner and B.V. Tucker, "Phonetic variability of stops and flaps in spontaneous and careful speech," *The Journal of the Acoustical Society of America,* pp. 1606-1617, 2011.