Tutorial 4

Modeling Speech Production and Perception Mechanisms and their Applications to Synthesis, Recognition, and Coding

Professor A. Alwan
University of California at Los Angelos (UCLA), USA

ABSTRACT

Quantitative models of human speech production and perception mechanisms provide important insights into our cognitive abilities and can lead to high-quality speech synthesis, robust automatic speech recognition and coding schemes, and better speech and hearing prostheses.

In this talk, I will describe some of our research activities in these two areas. Our speech production work involved collecting, and analyzing Magnetic Resonance Images (MRI), acoustic recordings, and Electropalatography (EPG) data from talkers of American English during speech production. The articulatory database is the largest of its kind in the world and contains the first images of liquids (such as /l/ and /r/) and fricatives (such as /s/ and /sh) for both male and female talkers. MR images are useful for characterizing the 3D geometry of the vocal tract (VT) and for measuring lengths, area functions, and volumes. EPG is used to study inter- and intra-speaker variabilities in the articulatory dynamics, while acoustic recordings are necessary for modeling. Inter- and intra-speaker characteristics of the VT and tongue shapes will be illustrated for various speech sounds, as well as results of acoustic modeling based on the MRI and acoustic data. I will also discuss the implications of our findings on vocal-tract normalization schemes and speech synthesis.

In the speech perception area, aspects of auditory signal processing and speech perception are parameterized and implemented in a speech recognition system. Our models parameterize the sensitivity to spectral dynamics and local peak frequency positions in the speech signal. These cues remain robust when listening to speech in noise.

Recognition evaluations using the dynamic model with a stochastic Hidden Markov Model (HMM) recognition system showed increased robustness to noise over other state-of-the-art representations. For example, our models of temporal adaptation, spectral peak isolation, an explicit parameterization of the position and motion of local spectral peaks, and the perception of pitch-rate amplitude modulation cues are shown to reduce the error rate of a word recognition system in noise by more than a factor of 4 over common Mel-Frequency Cepstral Coefficients (MFCC) representations.

I will also discuss the applications of auditory modeling to speech coding. As an example, we developed an embedded and perceptually-based speech and audio coder. Perceptual metrics are used to ensure that encoding is optimized to the human listener and is based on calculating the signal-to-mask ratio in short-time frames of the input signal. An adaptive bit allocation scheme is employed and the subband energies are then quantized. The coder is fully scalable--increasing the bit rates, improves the quality of encoded speech. The coder is variable-rate, noise-robust and suitable for wireless communications.

Work done jointly with Shrikanth Narayanan, Brian Strope, and Albert Shen.