PILOT PROJECT ON SPEECH RECOGNITION USING LAYERED ABDUCTION AND MULTIPLE KNOWLEDGE TYPES

J. Josephson, M. Beckman, R. Fox, A. Krishnamurthy,
T. Patten, L. Feth, B. Chandrasekaran
The Ohio State University, Columbus, OH 43210

Our long-term goal is to test the feasibility of basing a recognition system on a model of human speech processing. Key hypotheses are: (1) Computational models for reasoning from incomplete knowledge provide a useful metaphor for many aspects of human speech understanding even at levels where highly automatic perceptual processes are at work; (2) A computational model of human cochlear processing must be used as the signal-processing front end; (3) Some representation of articulation should mediate between the acoustics and the phonology in order to accommodate contextual variation of various sorts; (4) The phonological representation must encode the prosodic structure and intonation pattern as well as the phoneme string. Accordingly our specific short-term goal is to build a small prototype system that uses a layered-abduction architecture, in which there are many stages of processing, corresponding to different levels of knowledge. The common information-processing task at each stage is to form a coherent, composite (multi-part) hypothesis that explains the data presented from the preceding levels. The input signal will be the digitized speech processed by Patterson's Stabilized Auditory Image system. An articulatory representation based on Browman and Goldstein's gestural score will mediate between the auditory representation and the phonology, and Pierrehumbert and Beckman's prosodic tree and tone string will be used for the phonological organization and intonational melody.

We are nearing completion of an intial two-level system incorporating a stratum of acoustic features, and a stratum of phonological features, for a small vocabulary of CV monosyllables produced by one speaker. An abduction machine, with distinct elements of recognition, classification, and composite hypothesis formation, controls all the processing beyond some rudimentary initial signal processing; it even does the formant tracking. Our next step is to incorporate the Stabilized Auditory Image system as the signal-processing front end for the acoustic stratum in our initial two-level system. We have also begun to write gestural scores for a small vocabulary of monosyllables and disyllables on the basis of articulatory data that we recorded for our one speaker at the University of Wisconsin's X-ray microbeam facility. We will use these scores in building a separate system with three strata representing articulation, phonology, and a lexicon. This will provide a first test of a three-level machine; it will incorporate some top-down processing, and will provide the first opportunities to test the computational efficiency of the model. We are using these articulatory data also to determine the mapping between auditory features and articulatory gestures in preparation for merging the two systems into a single four-level system.