Analysis and Symbolic Processing of Unrestricted Speech

XEROX PALO ALTO RESEARCH CENTER
M. Margaret Withgott and Ronald M. Kaplan, Principal Investigators

Our objective is to develop methods and computational models for the perception and understanding of sensory data. The theories and computational technologies we develop will result in information interpretation for multi-media documents and database retrieval.

- We have developed a theoretical account of variation and perceptual constancy in spoken language, and have developed tools and techniques for modeling phonetic variation. Machine learning and statistical techniques have been adapted for use in organizing data into structures representing the contextual factors associated with phonetic variation. This is useful for evaluating theories of variation and for the design of word models in recognition systems. We are able to automatically generate rules that create variant pronunciations, which we then can compare with our large multiple-pronunciation dictionaries. Studies of American English speech have also advanced our understanding of what information to include in recognition models. For instance, results from an investigation of palatal sounds (with M. Peet, MITRE) argue for a particular use of duration, since the acoustic characteristics of spectrally-similar sounds differ temporally as correlated with their distinct underlying sources.

- We have developed word prediction and verification techniques using unrestricted text as input. To take advantage of discourse context for word prediction, we use a dynamic cache of recently encountered words in a Markov model to characterize the phenomenon of word recurrence associated with the topic of discourse. This reduces the average rank of correct word hypotheses by 10 percent in our electronic mail corpus. A related implementation has been developed for automatically tagging text. To take advantage of local, intra-word context, we use morphological analysis tools and are investigating the use of dynamically-created word-segment models for verification.

- Most current approaches to the problem of speech recognition in the presence of competing speech are based on the assumption that the co-channel speech streams can be separated using local waveform characteristics such as amplitude or fundamental frequency, independent of linguistic content. In our alternative paradigm, target-interference separation and target recognition occur simultaneously, driven by a model of the recognition vocabulary. The basis of the method is a spectral similarity measure which allows a reference spectrum to match only a subset of the input spectral features. We have evaluated the method through a set of speaker-dependent isolated-word recognition experiments in which the co-channel interference consisted of sentences drawn at random from the DARPA TIMIT database. Results indicate a 50-70 percent reduction in recognition error rates at low signal-to-noise ratios relative to those observed with a conventional whole-spectrum cepstral distance metric.

In concert with work supported by Xerox and NSF, we plan to study the integration of explicit signal knowledge representation with information-theoretic approaches to recognition. We intend to use language models (both categorical and statistical) to constrain the recognition. The work will involve development of a speaker-independent phonetic classification system for continuous speech as well as text-based recognition studies.