Continuous Speech Recognition from Phonetic Transcription

S. E. Levinson
A Ljolje

Speech Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974

Previous research by the authors has been directed toward phonetic transcription of fluent speech. We have applied our techniques to speech recognition on the DARPA Resource Management Task. In order to perform speech recognition, however, the phonetic transcription must be interpreted as a sequence of words. A central component of this process is lexical access for which a novel method is proposed.

The new technique treats lexical access as simply a string-to-string editing problem. That is, we seek the grammatically correct sequence of words whose phonetic spelling, as given in a dictionary, is closest, in a well defined sense, to the computed phonetic transcription of an actual utterance. The distance between two phonetic units is taken as the Euclidean distance between the means of their respective spectral and durational densities. This measure is called the *rhobar* metric and is known to have many interesting properties. This measure is especially appealing here since the needed densities are already present in the hidden Markov model on which the phonetic transcriptions are based. We extend this metric heuristically to the case of insertions and deletions by using the distance between the inserted or deleted unit and silence. The string-to-string editing operation is then easily performed by a two-level dynamic programming algorithm in which the inner level performs lexical access and the outer level does the parsing.

The method has been applied to the DARPA continuous speech recognition task using the strict task grammar. The acoustic/phonetic model was derived from 3969 sentences recorded from 109 speakers. Model parameters were obtained from sample statistics computed on the basis of phonetic segmentation performed by a segmental k-means algorithm.

Rather than attempt to measure phonetic accuracy, we have synthesized the speech from the phonetic transcription augmented by computed duration and pitch. The result is remarkably intelligible, especially when considering that it is equivalent to a 120 BPS coder. Audio tapes of the synthesized speech are available from the authors. We also tested our system on the October 1989 DARPA Test Data. Although the best speaker (speaker #1) achieved an 82% word accuracy rate, The overall word accuracy on the 10 speakers was 57.2% comprising 27.3% substitutions, 12.9% insertions, 2.7% deletions and 70.2% correct classifications. The high error rate is the result of catastrophic failure of the recognizer in which occasional sentences are entirely misrecognized. We do not yet understand the cause of this failure mode.

Although these results are disappointing, the high quality of the phonetic transcription and the simplicity of the architecture are an incentive to continue the research.