RECENT ADVANCES IN SPEECH RECOGNITION

Yaakov Stein

EFRAT FUTURE TECHNOLOGY LTD.
23 HaBarzel St.
Tel Aviv 69710, Israel

ABSTRACT

Speech recognition systems tend to consist of the following modules:

- input (acquisition and acoustic processing),
- classification (phoneme/word recognition),
- language (lexical/syntactic/semantic) processing,
- output (actions to be performed);

although in many systems these modules are further subdivided, in some systems some of these modules may be omitted or several may be combined.

The input module is responsible for enhancement, segmentation, feature extraction, and speaker/channel adaptation. Classical features, based on non-speech-specific or speech-generation models are now being supplemented by perception based features. These utilize concepts from the psychophysics and neurophysiology of hearing. I will discuss in particular the RASTA-PLP features which have been shown to improve recognition of telephone quality speech.

There are three classical classifiers:

- *Dynamic Time Warping* (DTW),
- discrete symbol *Hidden Markov Models* (HMM),
- continuous distribution HMM,

with the third debatably having been the most successful. Various Artificial Neural Network (ANN) architectures, although originally static pattern classifiers, have become popular in speech recognition as well, and hybrid neural-classical systems are now thriving. I will consider the basic principles of ANN architectures, concentrating on the *Multi-State Time-Delay-Neural Network* (MS-TDNN) as a specific example.

Previously Hidden Markov Models were universally trained by *Maximum Likelihood*, and Artificial Neural Networks by *Least Squared Error* criteria. Various alternative training procedures including discriminative training methods have been developed over the past few years. I will discuss the principles of *Minimum Classification Error* (MCE) which attempts to directly minimize the probability of misclassification.

Language processing can be used in speech recognition systems to lower the average *perplexity*, or to construct the most probable word sequence given the classifier's outputs. Early systems suffered from primitive (word or word-pair frequencies) or over-restrictive (constrained grammar) language models, and even today this is the least well developed of the speech recognition related disciplines. One problem that is receiving much attention is that of the 'translating telephone'. I will mention modern techniques for syntactic, semantic, and pragmatic level analyses.

In an extreme case of module mixing, the acquired input is fed directly into a 'black box' whose output is the action to be performed. During training of such systems both acoustic and language levels must be automatically acquired. There has been some experimentation with such systems, and I shall describe a prototypical one.

Speech recognition has proven to be much less reliable in real-world applications than under laboratory conditions, due mostly to the much greater variance in speech styles and operating conditions. After over thirty years of intensive research in speech recognition, we are still surprisingly far from true speaker independent free speech interaction between man and machine. I will conclude with a short survey of state-of-the-art performance of actual speech recognition systems.