SPAM: Experiments with Digit Recognition

Nelson Morgan [‡,*], Su-Lin Wu [‡,*], Hervé Bourlard [†,‡]

[‡] International Computer Science Institute, Berkeley, CA
[*] University of California at Berkeley, Berkeley, CA
[†] Faculté Polytechnique de Mons, Mons, Belgium

Recently, we have proposed the use of Stochastic Perceptual Auditory-event-based Models (SPAM) for machine speech recognition [1]. In this conceptual framework, speech is modeled as a sequence of Auditory Events (**Avents**), which are partial decisions separated by periods of time that are typically 50 to 150 msec. We hypothesize that avents occur most often when the spectrum and amplitude are rapidly changing, and that they may be fundamental components for the perception of continuous speech. The statistical model uses these **avents** as fundamental building blocks for words and utterances, separated by states corresponding to periods of time when decisions have not yet been made. In order to focus statistical power on the rapidly-changing portions of the time series, all of the non-decision states are tied to the same non-**avent** class.

In recent work, we have explored this notion with a set of experiments with a small isolated word recognition task, using a vocabulary consisting of the digits (including both "zero" and "oh"), as well as the control words "no" and "yes." The database, which originally came from Bellcore (and which has been reported in a number of our papers on RASTA processing) was recorded over the public-switched telephone network from 200 speakers.

For this task, we have been able to train a simplified SPAM-based system to perform comparably to a more conventional phone-based system, both for the natural telephone recordings and for the same recordings that had been artificially degraded by the addition of automobile noise. However, the distribution of errors as indicated by a confusion matrix is quite different for the two recognizers. A combined system incorporating models based on phones and models based on avents has yielded significantly improved results over either system alone.

In this presentation, we will discuss these results and explore how SPAM approaches may be used to improve robustness of existing recognizers.

References

[1] N. Morgan, H. Bourlard, S. Greenberg, and H. Hermansky, Stochastic Perceptual Auditory-Event-Based Models for Speech Recognition. In *Proceedings ICSLP*, pp 1943-46, Yokohama, Japan, 1994