Motivation and Preliminary Work

We propose to develop a landmark-based speech recognizer. We will experiment with the robustness and accuracy benefits provided by (1) diverse acoustic observations, and (2) multi-class label-sequence structural risk minimization.

As a preliminary baseline for this proposal, MFCC+d+dd 13-mixture 3-state HMMs were trained to distinguish stops, fricatives, nasals, vowels, glides, and silence. Depending on insertion penalty, six-class recognition accuracy varied from 52% to 74% (82% correct, 8% insertion). Binary recognition accuracy for the four manner-class distinctive features was 80% for [silence], 78% for [syllabic], 78% for [sonorant], and 77% for [continuant]. For this result, the HMM required 6310 trainable parameters.

Stevens proposed improving the accuracy and noise robustness of speech recognition by eliminating frame-based observation PDFs, and instead focusing the attention of the classifier on instantaneous phonetic landmarks. Phonetic landmarks are articulatory events with acoustic signatures easily detected at low SNR, including syllabic nuclei, intervocalic glides, and onsets and offsets of the distinctive features [sonorant] and [continuant]. Stevens proposed that acoustic observations should be optimized separately for each landmark type, e.g., measures of signal periodicity sampled once per 10ms for the detection of syllable nuclei, but energies sampled once per 1ms for the detection of stops.

Following Stevens’ proposal, Niyogi used kernel-based SVMs to detect stop consonants in TIMIT. Espy-Wilson and her students further developed SVM-based landmark detectors for onsets and offsets of the distinctive features [silence] (94% recognition accuracy), [syllabic] (79% accuracy), [sonorant] (93%), and [continuant] (94%). Six-manner-class recognition accuracy on TIMIT was 80%, using a total of 160 trainable parameters. Juneja and Espy-Wilson are currently testing a Viterbi-like algorithm for continuous speech recognition using SVM landmark detectors, together with a training algorithm comparable to segmental K-means re-estimation. These results form the foundation for the proposed workshop.

Proposed Experiments

1. Lattice rescoring. Landmark detectors comparable to Espy-Wilson’s will be trained using phonetically untranscribed Switchboard data, and tested using phonetically transcribed Switchboard data. Evaluation will include (1) six-class manner class recognition, as described under “preliminary work,” (2) lattice rescoring. Lattice rescoring experiments will use the word lattice output of an existing Switchboard recognizer. Each word will be rescoring using \( \sum \log p(y_i|x, t_i^*, w)p(t_i^* - t_{i-1}^*) \), where \( y_i \) is the \( i \)-th landmark in the dictionary entry of word \( w \), \( x \) is the set of observations, and \( t_i^* \) is the MAP alignment time of \( y_i \). The evaluation metric will be WER. Error analysis will tabulate landmark differences between error and correct path as a function of syllable position, stress, and distinctive features.

2. Embedded SVM re-estimation. Parameters of all landmark detectors will be simultaneously re-estimated using a maximum-margin multi-class label sequence training algorithm. Altun and Hoffman (2003) proposed maximum-margin training for hidden Markov models, using an empirical error based on Viterbi alignment, regularized by an SVM margin. SVM parameters estimated in experiment (1) will be re-estimated using an adapted Altun-Hoffman method, and tested in the tasks of manner-class recognition and lattice rescoring.

3. Noise. Landmark detectors based on different acoustic observations should respond differently to additive noise. We will test the hypothesis that word recognition accuracy benefits from SNR-dependent confidence weighting of the individual landmark detectors. By adding babble to the database, we will compute SNR-dependent landmark detection scores, \( p(y_i|x, t_i, SNR) \). First, SNR dependence of different landmark detectors will be compared to the human phoneme misperception rates reported by Miller and Nicely (1955), in order to evaluate the landmark detector as a model of speech perception. Second, SNR-dependent lattice rescoring will be evaluated by using \( p(y_i|x, t_i, SNR) \) in the lattice rescoring algorithm.