STOCHASTIC PERCEPTUAL SPEECH MODELS WITH DURATIONAL DEPENDENCE

Jeff Bilmes †,‡ *Nelson Morgan* ‡ *Su-Lin Wu* †,‡ *Hervé Bourlard* ‡,*

‡ International Computer Science Institute (ICSI), Berkeley, CA
† Dept. of Computer Science, U. of California, Berkeley
* Faculté Polytechnique de Mons, Belgium
Email: {bilmes,morgan,sulin,bourlard}@icsi.berkeley.edu

ABSTRACT

In [6], we develop statistical model of speech recognition where emphasis is placed on the perceptually-relevant and information-rich portion of the speech signal. In that model, speech is viewed as a sequence of elementary decisions or Auditory Events (avents) that are made in response to loci of significant spectral change. These decision points are interleaved with periods during which insufficient information has been accumulated to make the next decision. We have called this a Stochastic Perceptual Avent Model, or SPAM. In the work reported here, we have extended our initial experimental implementation [7] to include other probabilistic dependencies specified in the original theory, particularly the dependence on the time from the current frame back to the previous hypothesized avent.

1. INTRODUCTION

Artificial speech recognition systems often model speech as a sequence of short-time stationary acoustic segments. Acoustic feature vectors within a segment are usually assumed to have been generated by an i.i.d. random process. This allows the segment to correspond to successive outputs of a Hidden Markov Model (HMM) state. In such systems, modeling power is usually focused on matching each locally stationary segment with a particular HMM state.

Regions of spectral transition, as noted by many researchers, notably Furui [3], can be important for phonetic discrimination. Decision points made in response to such acoustic evidence might be more invariant under different speaking styles and acoustics, and therefore we would like to build a structure for focusing modeling power on these decisions. We would also like to avoid the i.i.d. assumptions described above.

In [6], we introduced such a structure. Perceptual decisions made in response to significant spectral changes are called Auditory Events, or *avents*. Time points corresponding to avents are assumed to be interleaved with periods during which we have insufficient evidence for a new decision, and as such are presumed to have less importance for discrimination. The training stage of a Stochastic Perceptual Avent Model (SPAM) based recognition system is designed to focus modeling power on the identification of the transition decisions. Therefore, similar to certain segment-based models, a SPAM system is less handicapped by assumptions of intra-phonetic independence.

In [7], we reported on a simplified SPAM implementation. In the work reported here, we have extended that system to include the time and avent dependencies specified in the original theory.

In Section 2, we review the essentials of our earlier SPAM experiments. In Section 3, we describe our new SPAM implementation that includes durational dependence. In Section 4, we evaluate the word-error performance on an isolated digits database. In Section 5, we describe a SPAM implementation that includes both the durational and previous avent dependence. And finally, in Section 6, we describe possible future work.

2. SPAM CLASSIC

As discussed in [7], under the SPAM recognition model, speech is considered a sequence of avent points interleaved with non-decision regions that typically correspond to acoustic segments containing less spectral change. In our current implementation, avents, which correspond to states in HMM-like models, are approximated by left-context-dependent onsets, so that every transition between phonetic segments can correspond to an avent. Therefore, with a p phoneme lexicon, there could be at most p^2 possible speech units, though the actual number is always less because of the topological constraints of the word models.

The stationary segments within a speech signal are spectrally quite diverse, but pre-processing approaches such as delta computation or RASTA [4] will significantly reduce this diversity. Because we aim to focus modeling power on the decisions corresponding to phonetic transitions, we wish to suppress the detailed disparities between different non-transition segments. Therefore, we use one broad category, called a non-transitional state, to represent speech frames falling within these segments.

Recognition under the general SPAM model is based on a computation of global posteriors based on the following local acoustic probabilities:

$$p(q_\ell^n | q_k^{n-\Delta(n)}, \Delta(n), X_{n-d}^{n+c}), \quad \left\{ \begin{array}{c} \forall \ell = 0, 1, \ldots, K \\ \forall k = 1, 2, \ldots, K \end{array} \right\} \tag{1}$$

where q_ℓ^n refers to avent q_ℓ occurring at time index n, $q_i \;\; i \neq 0$ refers to avent type i, q_0 refers to the non-transitional state, $\Delta(n)$ is the number of states between the current state n and the previous avent, $n - \Delta(n)$ corresponds to the time index for the previous avent (i.e., the closest time index in the past where $k \neq 0$), and X_{n-d}^{n+c} is a sub-sequence of acoustic vectors local to the current vector X_n, extending d frames into the past and c frames into the future.

In [7, 8], we reported on the initial SPAM implementation using a telephone quality database of digits and control words from Bellcore that we call "digits+". Only 46 diphones and therefore avents (including the non-transitional state) actually occur in this database. Using a hybrid HMM/ANN approach [2], probabilities for an HMM-like decoding are estimated by training ANNs with acoustic feature vectors and target labels. As discussed in [7, 8], two networks are separately trained based on the factorization given in Equation 2, one for classifying avent states and one to distinguish between avents and non-avents (i.e., non-transitional states).

In this first implementation, the local probabilities in Equation 1 used the dependence only on the acoustics, i.e., $P(q_n^\ell | X_{n-d}^{n+c})$. We nevertheless found that in the presence of noise, a statistically significant improvement in word error rate could be achieved over a purely phone-based system when using a combined SPAM-phone based system with a comparable number of free parameters. These numbers are recalled in Section 4, along with newer results.

3. SPAM MODEL WITH DURATIONAL DEPENDENCE

The original SPAM theory [6] suggested that Equation 1 could be simplified in various ways. In particular, if we assume that the probability of an avent is independent of the previous avent, Equation 1 simplifies to:

$$P(q_n^\ell | \Delta(n), X_{n-d}^{n+c}).$$

Our second SPAM implementation therefore extends the basic SPAM model with durational dependence.

A multi-layer perceptron (MLP) architecture using one hidden layer, as described in [8], is our method for computing the local frame-based probabilities based on an acoustic frame. This architecture was extended with additional input units to encode the durational information. The initial (or boot) weights for this system were either obtained from the first SPAM implementation, with additional weights set to random values, or were entirely set to random initial values.

The durational dependence as specified in Equation 1 is a continuous variable that has various possible encodings in a neural network. We have often found it useful to represent the dependence of variables like time a with discrete encoding at the input of the MLP. We began with a very coarse quantization of the time variable — we quantized the duration between the current frame and the previous avent to one of three values, *short*, *medium*, and *long*. The MLP is therefore extended with 3 input units, where each unit corresponds to one of the three time ranges.

The quantized durations represent intervals measured from the training set's avent labels. Specifically, we divide the training set's inter-avent duration distribution into thirds (e.g., we associate the unit *short* with durations less than the 33rd percentile).

During network training, the additional durational MLP input units for each frame are set depending on the duration to the previous avent. During recognition, however, we do not actually know the duration to the previous avent without having a phonetic alignment of the correct utterance. Therefore, we hypothesize each duration for each frame and compute the corresponding probability vector. For each time frame of speech, we adjust the durational input units of both the avent classifying network and the avent/non-avent network setting them in turn to *short*, *medium*, or *long*, combine the resulting output probabilities based on Equation 2, and produce probability vectors of the form:

$$P(q_\ell^n | \Delta, X_{n-d}^{n+c}), \quad \left\{ \begin{array}{r} \ell = 0, 1, \ldots, 46 \\ \Delta \in \{\text{short}, \text{medium}, \text{long}\} \end{array} \right\} \tag{3}$$

Although the training procedure did not use negative training examples (i.e., each speech frame was trained only positively with the correct duration but not negatively with the incorrect duration), we rely on the interpolative properties of the MLP and the nonlocal constraints of the HMM to deal with incorrect duration/spectral combinations presented during recognition. For another approach to handling such a mismatch, see [1].

Our new HMM-like model uses the durational dependent probability vectors produced by the ANN. It incorporates three groups of non-transitional states and three avent states, each corresponding to a different durational dependence. The *short* (resp. *medium*, *long*) non-transitional states utilize the *short* (resp. *medium*, *long*) local probabilities and lead to the *short* (resp. *medium*, *long*) avent state. Figure 1 shows an example of this model. We decided on this topology after experiments with several alternatives, including one with a forced minimum duration and one with self loops.

$$p(q_\ell^n|q_k^{n-\Delta(n)}, \Delta(n), X_{n-d}^{n+c}) = p(q_\ell^n|q_k^{n-\Delta(n)}, \Delta(n), X_{n-d}^{n+c}, q_{\neg 0}^n)p(q_{\neg 0}^n|q_k^{n-\Delta(n)}, \Delta(n), X_{n-d}^{n+c}) + p(q_\ell^n|q_k^{n-\Delta(n)}, \Delta(n), X_{n-d}^{n+c}, q_0^n) \quad (2)$$

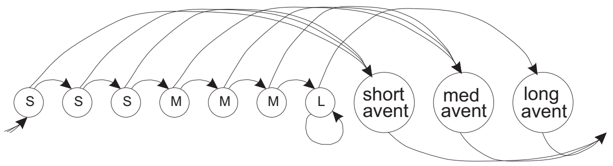

Figure 1: *The HMM-like SPAM model that includes a dependence on the duration from the current state to the previous avent. This figure corresponds to one non-transitional state and one avent from the original SPAM model. In this example figure, there are three avent states corresponding to* short, medium, *and* long *and seven non-transitional states, three corresponding to* short, *three to* medium, *and one to* long. *A* short *(resp.* medium*) duration corresponds to anything below 4 (resp. 7) speech frames. Greater than 7 non-transitional frames is possible, as indicated by the self loop, but greater lengths become decreasingly likely.*

4. RESULTS AND DISCUSSION

Figure 1 shows the experimental SPAM results. Rows 1 through 3 show error rates for the experiments described in [7]. Row 1 corresponds to a SPAM recognizer incorporating two 100 hidden-unit (HU) networks as described in Section 2, row 2 to a single 200 HU phone-based recognizer, and row 3 to a system that combines the scaled word-likelihoods of the two preceding systems. A 400 HU phone-based recognizer was shown to perform worse than the combined system in row 3. Each system has roughly the same number of parameters and uses the same input features (jrasta-plp-8, associated delta features, and delta jrasta model gain, with a context window of 9 frames and a frame step size of 12.5 msec).

Rows 4 and 5 show the new error rates. As in [7], all new results are trained and tested using the digits+ database. We tested each system using data from 200 speakers totaling 2600 examples from 4 jackknifed cuts. In other words, scores shown are the average of 4 tests in which 150 speakers were used for training and 50 were used for testing. We tested both using clean data and data with additive noise. The clean data was always used for training.

In all cases, the ANNs were trained on a subset of the input space used during testing. That is, we trained with only positive examples of the durational and previous avent ANN input units, but tested with hypothesized inputs never actually "seen" during training. To minimize over-fitting, for all testing cuts we employed an early epoch stopping criterion. In addition, all avent/non-avent networks were trained starting from random initial values rather than booting from an earlier system's weight matrix.

	clean	10dB SNR
1. SPAM classic	3.2%	10.6%
2. phone based	1.8%	10.9%
3. combined 1+2	1.6%	8.1%
4. SPAM + time dep.	2.2%	10.4%
5. combined 4+2	1.2%	8.1%

Table 1: Error rates for isolated digits plus "oh", "no", and "yes", recorded over a public-switched telephone network. The noisy case includes artificially added car noise resulting in a 10dB SNR.

As can be seen, the addition of durational dependence to the SPAM model reduces the error rate in the clean case while maintaining that rate in the noisy case. The clean case error reduction is significant at $p < 0.05$, assuming a normal approximation to a binomial distribution for the errors. The reduction in the error rate for the combined system, while not as statistically significant, is certainly in the right direction.

5. SPAM MODEL WITH DURATIONAL AND PREVIOUS AVENT DEPENDENCE

We are currently beginning to evaluate a SPAM implementation that uses local probabilities with all of the dependencies described in the original theory [6] and given in Equation 1.

Because of the limited size of the digits+ database, avents are clustered into broad classes rather than used themselves. Two avent classification methods based either on phonemic or phonetic-like information are being evaluated. We have been attempting to equalize the number of avents in each class to reduce the chance of inappropriately learning the relative class size distribution.

Recall that in the current SPAM methodology [8], avents correspond to left-context-dependent onsets. Our first classification method therefore uses phonetic attributes of the constituent phonemes, namely, the continuant/non-continuant property. This attribute should, in theory, separate phonemes into groups with significantly different spectral characteristics. Avents are therefore grouped into one of four classes depending on the attributes of the constituent phonemes. With digits+, however, this classification produces somewhat imbalanced classes. In particular, the continuant-continuant (CC) class is twice the size of the continuant/non-continuant (CN) and the non-continuant/continuant (NC) classes. In addition, the non-continuant/non-continuant (NN) class contains only one element. To create more balanced classes, the NN and CN classes are merged and the CC class is divided using the vowel/consonant phonemic attribute, presumably one that also corresponds to significant spectral difference.

A second method of avent clustering used phonetic-like information obtained from the training set. Specifically, we ran a K-means procedure directly on the acoustic vectors of each speech frame in the training set. This produces a K-clustering of the avent data. Each of the 46 avents are then mapped to the one K cluster that contains the majority of the instances of that avent. As usual with the K-means procedure, the optimal K is unknown, and different values must be evaluated. We optionally extend the resulting K-means classes with a separate class indicating that the previous avent is the beginning of an utterance.

We note that this second method resulted in avent broad classes that appeared to be quite "pure"; that is, a high percentage of the examples corresponding to each avent were mapped to a single cluster, even though there was no constraint enforcing this.

The ANN for this SPAM implementation has input units for both durational and avent category information. During training, the additional input units are set to values computed from the training labels. During recognition, we once again hypothesize the unknown information. In this case, however, the resulting number of probability values for each avent of each speech frame is $3c$ where c is the total number of avent classes.

The corresponding HMM-like model uses both the durational and previous class dependent probability vectors. Each non-avent/avent cluster as shown in figure 1 is duplicated c times, each replica using a different probability vector and is therefore dependent on a different previous class.

6. CONCLUSIONS AND SPECULATIONS

In this paper, we have described and reported error results for a SPAM implementation that includes durational dependence. We have shown that the addition of durational information to local probabilities can reduce the word error rate. We have also described the structure for future experiments with dependence on previous avent broad class.

Preliminary experiments with the full structure have not yet provided any improvements from incorporating the dependence on the previous avent. While we still have a number of modifications left to try (see below), we currently suspect that we need to change the database in order to profit from longer term dependencies on earlier (avent) states. Digits+ might be inappropriate for these more complex SPAM models — a larger database, such as the Numbers task distributed by OGI, or PhoneBook, distributed by LDC, may hold more promise. The number of quantized durations should be varied to experiment with resolution/complexity tradeoffs. The number of non-transitional states should also be varied — a single state type might represent too large and too diverse a feature space. Soft targets [1] would relax avent region boundaries — rather than one speech frame per avent, which our current system uses, an avent should be able to occur at any of several multiple frames and have decreasing likelihood at the edges (see [1] for an approach to determine such a probabilistic target). Negative training examples, in which the recognition phase uses input space regions within the training set, might also help performance. The stopping criterion and learning rate scheduler for the ANN training phase should use word recognition rather than frame-based errors on the cross-validation set.

SPAM might be ideally suited to domains with considerable speech rate variation. During very fast and very slow speech, more stationary regions (e.g., vowel nuclei) might be significantly compressed or expanded relative to median-rate speech [5]. Transitional regions (e.g., bursts), however, undergo relatively little transformation. The SPAM model, which if required can completely ignore or indefinitely extend states associated with the more stationary regions, might be well suited handling to such rate variations.

We wish to thank Steve Greenberg and Hynek Hermansky for useful discussions. This work was partly sponsored by JSEP Contract No. F49620-93-C-0014.

7. REFERENCES

1. H. Bourlard, Y. Konig, and N. Morgan. REMAP: Recursive estimation and maximization of a posteriori probabilities, application to transition-based conn ectionist speech recognition. Technical Report TR-94-064, International Computer Science Institute, Berkeley, CA, 1994.
2. H. Bourlard and N. Morgan. *Connectionist Speech Recognition - A Hybrid Approach*. Kluwer Academic Press, 1994.
3. S. Furui. On the role of spectral transition for speech perception. *JASA*, 80(4):1016–1025, 1986.
4. H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, September 1994.
5. I. Lehiste. Suprasegmental features of speech. In N.J. Lass, editor, *Principles of Experimental Phonetics*, chapter 6. Mosby, 1996.
6. N. Morgan, H. Bourlard, S. Greenberg, and H. Hermansky. Stochastic perceptual auditory-event-based models for speech recognition. *Proc. Intl. Conf. on Spoken Language Processing*, pages 1943–1946, 1994.
7. N. Morgan, S.-L. Wu, and H. Bourlard. Digit recognition with stochastic perceptual models. *Proc. Eurospeech'95*, September 1995.
8. S.-L. Wu. Properties of stochastic perceptual auditory-event-based models for automatic speech recognition. Technical Report TR-95-023, International Computer Science Institute, Berkeley, CA, May 1995.