STOCHASTIC PERCEPTUAL MODELS OF SPEECH

Nelson Morgan, Hervé Bourlard, Steven Greenberg, Hynek Hermansky, and Su-Lin Wu

International Computer Science Institute, Berkeley, California
1947 Center St, Suite 600, Berkeley, CA. 94704
Tel: (510) 642-4274, Fax: (510) 643-7684
& University of California at Berkeley (for Morgan, Greenberg, and Wu)
& Faculté Polytechnique, Mons, Belgium (for Bourlard)
& Oregon Graduate Institute (for Hermansky)
{morgan, bourlard, steveng, hynek, sulin}@icsi.berkeley.edu

ABSTRACT

We have recently developed a statistical model of speech that avoids a number of current constraining assumptions for statistical speech recognition systems, particularly the model of speech as a sequence of stationary segments consisting of uncorrelated acoustic vectors. We further wish to focus statistical modeling power on perceptually-dominant and information-rich portions of the speech signal, which may also be the parts of the speech signal with a better chance to withstand adverse acoustical conditions. We describe here some of the theory, along with some preliminary experiments. These experiments suggest that the regions of acoustic signal containing significant spectral change are critical to the recognition of continuous speech.

1. INTRODUCTION

In [8], we proposed a perceptual model of speech as a sequence of Auditory Events (**Avents**), separated by relatively stationary periods (ca. 50-150 ms). We hypothesize that avents occur when the spectrum and amplitude are rapidly changing (as in [4]). These speech dynamics are precisely those likely to generate enhanced activity in the upper stations of the auditory pathway, and 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 the more stationary regions. In order to focus statistical power on the rapidly-changing portions of the time series, all of the stationary regions are tied to the same non-**avent** class. Markov-like recognition models use Avents as time-asynchronous observations. Discriminant models are trained to distinguish among all classes, including the non-Avent class. In the simplest scheme, the training data are automatically aligned using dynamic programming, and a discriminant system (e.g., a neural network) is trained on the new segmentation. These two steps are iterated, as discussed in [1]. Section 3 introduces an approach that actually minimizes the probability of errors at the global or utterance level [2]). This process focuses modeling power on the perceptually-dominant and information-rich portions of the speech signal, which may also be the parts of the speech signal with a better chance to withstand adverse acoustical conditions. A statistical framework that is more commensurate with higher-level auditory function is a better match to front-end modules that attempt to incorporate properties of the *auditory periphery* [6], particularly when similar temporal auditory properties are incorporated [7]. We have named this new framework the Stochastic Perceptual Auditory-event-based (Avent) Model, or SPAM. We have preliminary results with the recognition of isolated digits over the telephone that appears to support the notion of increased acoustical robustness for such models, and have run some preliminary perceptual experiments that also suggest the existence of related mechanisms in human listeners. We also have recently developed some related theory (REMAP) that suggests how a SPAM-based system could be trained to maximize the global posterior probabilities for the correct models of an utterance (the MAP estimate) [2]. This will minimize the probability of utterance error, taking advantage of the perceptually-based assumptions of SPAM.

2. THEORETICAL FRAMEWORK

We first define notation and basic terms:

- A set of **avents** (auditory events):
  $\mathcal{Q} = \{q_0, q_1, \ldots, q_K\}$. This set is currently initialized to correspond to truncated diphones; that is, phone boundaries with the local region of the time series associated with them. Given such an initialization, the **avents** would be determined automatically in an embedded Viterbi-based dynamic programming procedure (as is currently accomplished with phone-like subword models).

  Each q_k, $k = 1, \ldots, K$, represents an auditory event on which recognition will be based. q_0 represents a non-**avent** or **non-perceiving state**.

- A sequence of acoustic vectors that is associated with an utterance: $X = \{x_1, x_2, \ldots, x_N\}$.

  Ideally, these acoustic vectors should be chosen to optimize detection.

- $X_{n-d}^{n+c} = \{x_{n-d}, \ldots, x_n, \ldots, x_{n+c}\}$.

This is a sub-sequence of acoustic vectors that is local to the current vector x_n, extending d frames into the past and c frames into the future.

- An **utterance model** M_i is then represented as a sequence of **avents** with looped non-perceiving states in between (see Figure 1).
- q^n = **avent** perceived at time n.
- q_k^n means that **avent** q_k has been perceived at time n.

The goal of recognition is to find the most probable word or sentence j maximizing the **a posteriori probability** of M_j given the data (X), i.e.,

$$M_j = \underset{M_i}{\text{argmax}}\ P(M_i|X) \tag{1}$$

In the discriminant HMM approach that is described in [1], a local acoustic probability estimate is required, namely,

$$p(q_\ell^n|q_{\ell_{n-1}}^{n-1}, q_{\ell_{n-2}}^{n-2}, \ldots, q_{\ell_1}^1, X) \tag{2}$$

for all $n = 1, \ldots, N$ and all possible states q_ℓ ($\ell = 1, \ldots, L$) making up M_i. In the case of SPAM, these states are **avents**.

A reasonable simplifying assumption would be to ignore the dependence on states prior to the last perceived **avent**. In this case, the time to the previous **avent** is the only significant information concerning the intervening non-**avent** states. Therefore, the **avent** sequence $\{q_{\ell_{n-1}}^{n-1}, q_{\ell_{n-2}}^{n-2}, \ldots, q_{\ell_1}^1\}$ appearing in the conditional of (2) simplifies into

$$\{q_k^{n-\Delta(n)}, \Delta(n)\} \tag{3}$$

in which $n - \Delta(n)$ corresponds to the previous time index for which an **avent** had been perceived, i.e., the last time index $n - \Delta(n)$ for which a $q_k^{n-\Delta(n)}$ was perceived with $k \neq 0$. Note that this assumption is in principle less unrealistic than the typical first-order conditional independence assumption of HMMs, since the former implies only that an **avent** is independent of any before the previous one, which on the average might occur 100 ms before the current **avent**. This is reminiscent of an approach proposed in [9]. In that method, similar consecutive acoustic frames were dropped, leaving only the first frame and the length of the dropped sequence as input variables for the training and recognition process. The approaches differ in that SPAM is a recognition model, and that the choice to assign frames to "non-perceiving states" q_0 (essentially, which frames to drop) is based on a global criterion and not on local decisions. However, both approaches emphasize dynamic portions of the speech signal, incorporate implicit duration modelling and remove correlation between successive frames. These effects make the recognition process more consistent with HMM assumptions. In [9], this was shown to improve recognition performance.

Taking these assumptions into account, one can do SPAM recognition based on the following local acoustic probabilities, in order of decreasing complexity:

$$p(q_\ell^n|q_k^{n-\Delta(n)}, \Delta(n), X_{n-d}^{n+c}), \quad \left\{ \begin{array}{l} \forall \ell = 0, 1, \ldots, K \\ \forall k = 1, 2, \ldots, K \end{array} \right\} \tag{4}$$

If we assume that the probability of an **avent** is independent of the previous **avent**, we can also use:

$$p(q_\ell^n|\Delta(n), X_{n-d}^{n+c}), \tag{5}$$

or, if we disregard the $\Delta(n)$:

$$p(q_\ell^n|X_{n-d}^{n+c}). \tag{6}$$

During dynamic programming (or the REMAP procedure - see [2]), these probabilities will have to be estimated for all possible q_ℓ (according to the topology of the models) and all possible $\Delta(n)$ (if used).

Estimation of (4)-(6) can be done, for instance, using a Multi-layer Perceptron (MLP) trained with X_{n-d}^{n+c} at the inputs, (complemented by inputs for $\Delta(n)$ for (5) as well as $q_k^{n-\Delta(n)}$ for the case of (4)) and $K+1$ outputs, with one output for q_0 and the other K outputs associated with the classes of **avents**. Alternatively one net can be used to estimate avent class probabilities, and another to estimate the probability of avent/non-avent. See [1] for related approaches to MLP training for the case of phone-like units.

Figure 1 shows a SPAM consisting of three **avents** with intervening non-perceiving states.

3. IMPROVED MAP ESTIMATION

As discussed in [1], training of transition-based models has a number of inherent difficulties. This led us to develop the theory for an approach to the training of local posterior probability estimators that maximizes the estimate of global posteriors for the correct models [2]. In this approach, we use a forward-backward-like recursion to generate targets for the local network that push the network towards improved global discrimination. The simplest form of this approach uses a local probability that incorporates first-order dependence, namely $p(q_\ell^n|q_k^{n-1}, X_{n-c}^{n+c})$. However, the technique can be generalized to dependence on M previous states, and it is shown in [2] that the approach can further be used to train estimators of SPAM probabilities such as (4).

Other than the general advantage of modifying local training to maximize globally optimal criteria, this approach removes (in principle) a number of practical difficulties with SPAM training. In general, it is difficult to learn to uniquely locate transition frames, since spectral transitions often occur over more than one frame. Furthermore, during recognition we will consider all possible previous avents for every input feature vector, while during training the previous avent will be assumed to be that which is given by the word or phone level transcription for the training data. This results in undertraining through a lack of negative examples. Both of these problems can be alleviated by training with "soft" targets, so that all possible previous avents are considered, even during training (since they will generally have non-zero probabilities). If the soft targets represent "better" estimates of the desired posterior probabilities, then their use also implies a smooth transition over several frames (in a probabilistic sense).

Figure 1: A schematic of a three **avent** SPAM, with tied non-perceiving states separating the **avents**. This could be a model for a two-phone word, for instance, with the q_0 states corresponding to steady-state regions, and the q_i, q_j, and q_k states corresponding to the three phonetic transitions.

4. EXPERIMENTS IN PROGRESS

Much work needs to be done before we can apply this theory to a working speech recognition system, as we are proposing a fairly radical departure from existing systems. However, we are in the midst of two lines of experimental work.

4.1. Perceptual experiments

A series of preliminary perceptual experiments have been performed in our laboratory, designed to ascertain the temporal locations of the information-laden components of the speech signal. These experiments are a direct outgrowth of previous studies published by Furui [4] and Drullman et al. [3]. Furui has shown for isolated Japanese consonant-vowel syllables that the most significant perceptual information is concentrated in regions associated with a large amount of spectral change. Drullman et al. have demonstrated that speech intelligibility of spoken sentences is dependent on the integrity of slow temporal modulations (<8 Hz) in critical-band-like channels of the speech envelope correlated with syllable and phone segment boundaries. We have replicated Furui's and Drullman et al. 's results for naturally spoken English sentences, and have extended Drullman's methodological paradigm to enable us to more precisely pinpoint the locus of phonetically significant information in the speech signal. We have accomplished this by bandpass-filtering the temporal envelope modulations rather than using a low-pass filter. Our preliminary informal listening results indicate that temporal modulations between 2 and 6 Hz are the most important for maintaining speech intelligibilty, with intellibility approaching that of low-pass filtering the speech envelope below 8 Hz. This 2-6 Hz bandpass is similar to that used in RASTA [7], suggesting further confirmation of the auditory plausibility of this technique.

4.2. Machine classification experiments

In an initial recognition experiment we learned 44 types of phonetic transitions that occur in a telephone database of digits and two control words (yes and no). We trained an MLP with 8400 frames that had been labeled (by a phonetically-trained MLP) as being transitions, and also trained a second network to distinguish between transitions and non-transitions using an equal number of non-transition frames from the same database. Using the product of these two network outputs as an estimate of probability $p(q_\ell^n|X_{n-c}^{n+c})$, where each q_ℓ is a state that might either be an avent or a non-avent, we achieved the digit recognition results (on 50 speakers not used in the training set) shown in Table 1. While the results are thus far not as good as that achieved by a phone-based system with roughly the same number of parameters, we note the following:

- The SPAM system used a simple probability that did not include dependence on the previous avent or on the time between them (6); these characteristics may be critical to the idea [9].
- The phone-based system used automatically-generated alignments that had been optimized for that system; the SPAM system used the same alignments.
- We still have some difficulty estimating the probabilities of avent vs. non-avent. We are currently working on improving the estimator's performance.
- The avent classification network had significant difficulties learning onsets for two word-initial fricatives (at the start of "three" and "zero"). Inspection showed that in at least some of these cases there was exceptionally low energy, and the avent-based system had no silence class per se. This suggests to us that it may be useful to distinguish between non-transition speech and non-transition non-speech.
- We have used on-or-off targets here, rather than the REMAP-based soft targets that may be required for a transition-based system.
- Nonetheless, the initial SPAM implementation is able to do 95% accurate speaker-independent telephone digit recognition, and in the case of additive car noise, does about the same as the phone-based system. Thus, a relatively poor avent-based system (with three times the error for clean speech) achieves the same error level as a phone-based system in the case of severe additive car noise. Note that this performance was achieved with a recognition procedure that typically only used around 4 "distances" from avent hypotheses per word, as opposed to the roughly 50 or so frames contained in the average word.

5. SUMMARY

In this paper we have described a new statistical approach for speech recognition that is based on auditory perceptual criteria. In particular, we are proposing to focus statistical modeling power on regions of significant change rather than on relatively steady state regions, and to do so by using a single model to represent all possible stationary segments; discrimination is provided by modeling of categories of major spectro-temporal changes. Pilot perceptual experiments seem to be consistent with our view that certain

	phones	SPAM
Telephone digits	1.8	4.9
Added car noise (10 dB SNR)	35.1	34.3

Table 1: Recognition error in %. For 13-class isolated telephone word task (digits plus yes and no). Results are on 50 speakers saying each word once, after training on 150 other speakers also saying each word once. All trainings used the "clean" data, and the MLP used for avent discrimination used roughly 10,000 parameters; an additional 10,000 parameters were used to discriminate between avents and non-avents. The results suggest that the avents by themselves are sufficient to represent speech in noise as well as the full phones.

ranges of spectral change are essential for the intelligibility of speech. At the moment the avents can be thought of as truncated diphones, but as our research develops we expect our definition of these perceptually relevant regions to be refined. Related work for the case of likelihoods (as opposed to the posteriors of our model) can be found in a number of sources, e.g., [5].

We have just begun to explore the consequences of this hypothesis experimentally. Preliminary experiments with isolated digits suggest robust properties for the case of additive noise recorded in a moving automobile. This work is the first effort on our part to pursue the goal of acoustically robust recognition by modifying the fundamental statistical substrate, as opposed to merely improving the acoustical feature extraction of the speech input.

6. ACKNOWLEDGMENTS

We acknowledge the assistance of Mike Ward and Mike Shire for their help in generating the stimuli for our psychoacoustic experiments. Thanks to ESPRIT project 6487 (WERNICKE) and ICSI in general for supporting this work. Some of this work was also supported by NSF-ARPA grant IRI-9314959 to the Oregon Graduate Institute.

7. REFERENCES

[1] H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Press, 1994

[2] H. Bourlard, Y. Konig, and N. Morgan, REMAP: Recursive Estimation and Maximization of A posteriori Probabilities, ICSI Technical Report TR-94-064, 1994.

[3] R. Drullman, J. Festen, and R. Plomp, Effect of temporal smearing on speech reception, J. Acoust. Soc. Am. 95 (2), February 1994

[4] S. Furui, On the Role of Spectral Transition for Speech Perception *J. Acoust. Soc. Am.* **80**, (4), 1016-1025, 1986

[5] O. Ghitza and M.M. Sondhi. Hidden markov models with templates as non-stationary states: an application to speech recognition. *Computer Speech and Language*, 2:101–119, 1993.

[6] S. Greenberg, The Representation of Speech in the Auditory Periphery. *Journal of Phonetics*, 16:1-151, 1988

[7] H. Hermansky and N. Morgan, Towards handling the acoustic environment in spoken language processing. In *Proceedings ICSLP*, volume 1, 85–88, Banff, Alberta, Canada, 1992

[8] 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

[9] M.J. Russell, K.M. Ponting, S.M. Peeling, S.R. Browning, J.S. Bridle, and R.K. Moore, "The ARM Continuous Speech Recognition System", IEEE Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing, Albuquerque, NM, 1990.