STOCHASTIC PERCEPTUAL AUDITORY-EVENT-BASED MODELS FOR SPEECH RECOGNITION

Nelson Morgan[†,‡], *Hervé Bourlard* [†], *Steven Greenberg*[†,‡], *and Hynek Hermansky* [†,*]

International Computer Science Institute (ICSI), Berkeley, California[†]
U. of California, Berkeley, California[‡]
Oregon Graduate Institute, Portland, Oregon[*]

ABSTRACT

We have developed a statistical model of speech that incorporates certain temporal properties of human speech perception. The primary goal of this work is to avoid 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. A focus on perceptual models may in principle allow for statistical modeling of speech components that are more relevant for discrimination between candidate utterances during speech recognition. In particular, we hope to develop systems that have some of the robust properties of human audition for speech collected under adverse conditions. The outline of this new research direction is given here, along with some preliminary theoretical work.

I. INTRODUCTION

Automatic speech recognition systems traditionally rely on an underlying model of speech as a sequence of stationary segments. For Hidden Markov Models (HMMs), practical considerations generally require the assumption that adjacent feature vectors are statistically independent, as well as being identically distributed within a segment corresponding to a Markov state. As a result, modeling power is sometimes focused on regions that are relatively unimportant for discrimination. These factors compromise the performance of recognition systems under conditions that are easily handled by human listeners, such as acoustic ambiguity and background noise. These observations have motivated us to develop a statistical model that incorporates some simple temporal properties of human speech perception that we believe to be crucial to human capabilities.

Statistical recognition models are potentially more discriminant than production-based models. The mathematical foundation for using these models is described in [1]. This earlier work has now been extended to constrain the underlying statistical model to consist of a sequence of Auditory Events or **avents**, separated by relatively stationary periods (ca. 50-150 ms). **Avents** occur during temporal intervals in which the spectrum and amplitude are rapidly changing (as in [2]). **Avents** are likely to generate enhanced activity in the upper stations of the auditory pathway, and may be fundamental components for the perception of continuous speech. During the intervals between **avents** the auditory system and higher nervous centers are likely to be engaged in some sort of perceptual integration analogous to "gap-bridging" [6]. In our approach, all of the stationary regions are tied to the same class. Markov-like recognition models can use **avents** as time-asynchronous observations. In this case, the models themselves consist of states that represent **avents** (on which recognition will be based) and other "non-perceiving" states that are responsible for processing the stationary segments (and which are not used directly for recognition). Discriminant models can be trained to distinguish among all classes (including the non-**avent** class). The training data can be automatically aligned using dynamic programming, and the discriminant system (e.g., a neural network) can be trained on the new segmentation.

These two steps can be iterated, as discussed in [1], and are guaranteed to converge to a local minimum of the probability of error (on the training set). This process should focus 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 survive in adverse environmental conditions.

A statistical framework that is more commensurate with higher-level auditory function should be a better match to front-end modules that attempt to incorporate properties of the *auditory periphery* [3], particularly when similar temporal auditory properties are incorporated [5]. We have named this new framework the Stochastic Perceptual Auditory-event-based (Avent) Model, or SPAM.

II. AUDITORY FOUNDATIONS

The present framework draws its inspiration from physiological and psychoacoustic studies of auditory processing. Speech and other complex vocal communication systems have evolved to exploit specific properties of sensory transduction and neural coding that enhance reliability of transmission for a wide range of acoustic conditions. Speech is readily understood by human listeners under many conditions with which the most advanced automatic speech recognition systems can not cope, including low signal-to-noise ratios, reverberant environments, unusually fast or slow speaking rates, and strong regional dialects and foreign accents.

Although it is not our intent to model all of the operations performed by the auditory system, we aim to capture those sensory properties which enable listeners to extract informationally salient elements of the speech signal under a wide range of adverse conditions. The relevant set of auditory properties would enable us to derive a more stable representation of the speech signal than is presently used in speech recognition applications.

This effort began a number of years ago with the development of Perceptual Linear Prediction (PLP) by Hermansky [4]. PLP transforms the speech signal into a spectral form commensurate with the resolving power and integration capabilities of the auditory system. More recently, Morgan and Hermansky [5] have extended this effort with RASTA-PLP, which de-emphasizes those portions of the signal undergoing little spectral change over time.

The current project extends this earlier work by focusing on the identification and description of **avents**, which signal the occurrence of perceptually significant information. **Avents** are associated with temporal intervals over which there are rapid changes in the spectrum. Such spectral transitions are associated in the speech signal with changes in the underlying articulatory configuration and thus can be used to infer the initiation of a new phonetic segment.

The auditory properties of relevance to **avent** detection are:

1. Rapid adaptation, in which the probability and magnitude of neural discharge is highest during the initial 5 ms of a new segment

2. The frequency resolution of the auditory pathway controlling the discharge probability of neural elements across frequency channels

3. Inhibitory circuitry that sharpens the activation to novel spectral features in both the time and frequency domains

The consequence of these properties is to increase the level of neural excitation at those points in time and frequency associated with the greatest magnitude of change, and to de-emphasize the more predictable, less content-laden portions of the speech signal. These properties are important for creating a representation of the speech signal that maximizes the probability of accurate phonetic segment identification under unpredictable conditions; the system is optimized to focus on information relevant to phonetic identification and to disregard most of the other properties of the signal. Although it is logical to assume that the most important portions of the speech signal are those which vary greatest as a function of time and frequency, there are surprisingly few perceptual studies that enable us to identify those segments with any degree of confidence. Furui's study [2] is perhaps the only experimental investigation that has so far convincingly shown associated spectral transitions with phonetic identification in natural speech. We have initiated some related experiments, which are briefly described in section IV.

III. THEORETICAL FOUNDATIONS

3.1 Definitions 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 be the set of boundaries between phones. In the future this choice may be modified as a consequence of psychoacoustic experiments now in progress. 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 defined to optimize detection of transitions or **avents** (e.g., RASTA-PLP, or measures of auditory neural synchrony).

- $X_{n-d}^{n+c} = \{x_{n-d}, \ldots, x_n, \ldots, x_{n+c}\}$.
  This is a a sub-sequence of acoustic vectors that is local to the current vector, extending d frames into the past and c frames into the future.

- A **word model** M_i is then represented as a sequence of **avents** with looped non-perceiving states in between.

- q^n = **avent** perceived at time n.

- q_k^n means that **avent** q_k has been perceived at time n.

- $\mathcal{L}$ represents our knowledge about the language (e.g., syntax, semantics, pragmatics).

3.2 Decoding

In this approach, the goal is to use recognition models for speech, as opposed to standard HMM production models. If M_i, $(i = 1, \ldots, I)$, represent the possible word or sentence models as defined above and X a sequence of N acoustic vectors, the goal of recognition is to find the most probable word or sentence j maximizing the **a posteriori probability** of M_j given what you hear (X) and what you know about the language ($\mathcal{L}$), i.e.

$$j = \operatorname*{argmax}_i \; P(M_i|X, \mathcal{L}) \tag{1}$$

This is referred to as the Maximum A Posteriori (MAP) criterion.

Taking into account all possible **avent**-based segmentations, $P(M_i|X, \mathcal{L})$ is computed in the following manner:

$$\begin{aligned} & P(M_i|X, \mathcal{L}) \\ & = \sum_{\ell_1=1}^{L} \cdots \sum_{\ell_N=1}^{L} P(q_{\ell_1}^1, \ldots, q_{\ell_N}^N, M_i|X, \mathcal{L}) \end{aligned} \tag{2}$$

for all possible $\{q_{\ell_1}^1, \ldots, q_{\ell_N}^N\} \in \Gamma$, the set of all possible paths in M_i.

If we are only interested in the best segmentation, a MAP approximation equivalent to the Viterbi maximum likelihood approximation used in standard HMMs can be obtained by replacing all summations in (2) by a max operator, yielding

$$\overline{P}(M_i|X) \;=\; \max_{\ell_1,\ldots,\ell_N} P(q_{\ell_1}^1, \ldots, q_{\ell_N}^N, M_i|X, \mathcal{L}) \tag{3}$$

where $\overline{P}(\cdot)$ represents the Viterbi approximation of the actual a posteriori probability.

Without any assumptions, each term of the sums in (2) or of the max operator in (4) can be factored into

$$\begin{aligned} & P(q_{\ell_1}^1, \ldots, q_{\ell_N}^N, M_i|X, \mathcal{L}) \\ & = P(q_{\ell_1}^1, \ldots, q_{\ell_N}^N|X, \mathcal{L}) P(M_i|X, \mathcal{L}, q_{\ell_1}^1, \ldots, q_{\ell_N}^N) \end{aligned} \tag{4}$$

In the above expression, the first factor represents the acoustic decoding, in that the acoustic vector sequence X is decoded in terms of a sequence of **avents**. For many purposes the second factor can be ignored, since the state sequence and language model may uniquely determine the utterance.

The global probability $P(q_{\ell_1}^1, \ldots, q_{\ell_N}^N|X, \mathcal{L})$ that is required to calculate $P(M_i|X, \mathcal{L})$ can be factored, without any assumptions, into local probabilities as

$$\begin{aligned} & P(q_{\ell_1}^1, \ldots, q_{\ell_N}^N|X, \mathcal{L}) \\ = \;& p(q_{\ell_1}^1|X, \mathcal{L}) p(q_{\ell_2}^2|X, \mathcal{L}, q_{\ell_1}^1) \ldots \\ & p(q_{\ell_N}^N|X, \mathcal{L}, q_{\ell_1}^1, \ldots, q_{\ell_{N-1}}^{N-1}) \\ = \;& \prod_{n=1}^{N} p(q^n|Q_1^{n-1}, X, \mathcal{L}) \end{aligned} \tag{5}$$

where Q_1^n represents the **avent** sequence associated with X_1^n. Probabilities $P(q_{\ell_1}^1, \ldots, q_{\ell_N}^N | X, \mathcal{L})$ can thus be calculated from local probabilities $p(q^n | Q_1^{n-1}, X, \mathcal{L})$ that will be referred to as **conditional transition probabilities**.

Given (5), calculation of $P(q_{\ell_1}^1, \ldots, q_{\ell_N}^N | X, \mathcal{L})$ for use in (2) (full MAP) or in (3) (Viterbi approximation) requires the calculation of the (local) conditional transition probabilities

$$p(q_\ell^n | q_{\ell_{n-1}}^{n-1}, q_{\ell_{n-2}}^{n-2}, \ldots, q_{\ell_1}^1, X, \mathcal{L}) \tag{6}$$

for all $n = 1, \ldots, N$ and all possible **avents** q_ℓ ($\ell = 1, \ldots, L$) making up M_i.

Local probabilities (6) may be simplified by relaxing the conditional constraints.

For initial experiments, we are ignoring the contribution of linguistic knowledge, for instance by using independent cost contributions from simple statistical grammars. However, ultimately the integration of the $\mathcal{L}$ term may be crucial to this approach.

A reasonable simplifying assumption can be to ignore the dependence on states prior to the last perceived **avent**. In this case, 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 (6) simplifies into

$$q_k^{n-\Delta n} \tag{7}$$

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 **avents** prior to the previous **avent**, which on the average might be 200 msec into the past. Ideally, the influence from earlier perceived **avents** is lumped into the contribution from the language model, which provides expectations of phonetic sequences to follow.

By also including the Δn factor in the conditional, we include the only remaining information about the non-perceiving state between **avents**. In this case 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 (6) simplifies into

$$\{q_k^{n-\Delta n}, \Delta n\} \tag{8}$$

Taking these assumptions into account, one can do SPAM recognition based on the following local probabilities, in order of decreasing complexity (ignoring for the moment the influence of the language model):

$$p(q_\ell^n | q_k^{n-\Delta n}, \Delta n, X_{n-d}^{n+c}), \quad \left\{ \begin{array}{l} \forall \ell = 0, 1, \ldots, K \\ \forall k = 1, 2, \ldots, K \end{array} \right\} \tag{9}$$

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

$$p(q_\ell^n | \Delta n, X_{n-d}^{n+c}), \tag{10}$$

or, if we disregard the Δn:

$$p(q_\ell^n | X_{n-d}^{n+c}). \tag{11}$$

During dynamic programming, these probabilities will have to be estimated for all possible q_ℓ (according to the topology of the models) and all possible Δn (if used).

Estimation of (11) is quite easy and can be done, for instance, using a Multi-layer Perceptron (MLP) trained with X_{n-d}^{n+c} at the inputs, and $K+1$ outputs, with one output for q_0 and the other K outputs associated with the classes of **avents**. See [1] for related approaches to MLP training for the case of phone-like units.

Figure 1 shows an **avent** with the non-perceiving states that surround it.

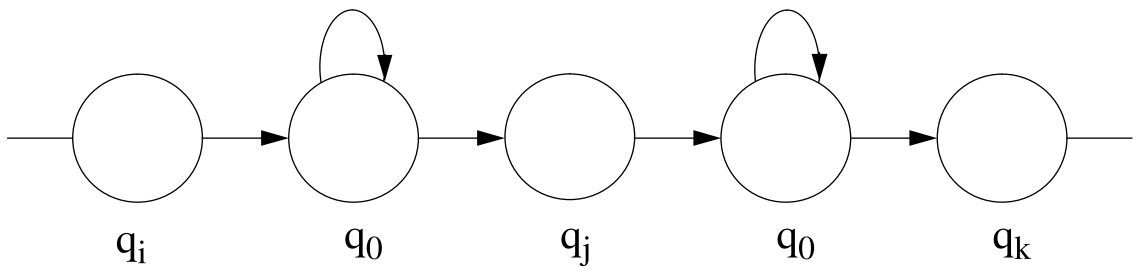

Figure 1: A schematic of a three **avent** SPAM, with tied non-perceiving states separating the **avents**. This could be a model for a two-phone word, for instance, with the q_0 states corresponding to steady-state regions, and the q_i, q_j, and q_k states corresponding to the three phonetic transitions.

IV. EXPERIMENTAL DIRECTIONS

We have developed an initial theory. However, we have also begun a series of experiments with human listeners and with machine classification systems.

We have initiated a series of perceptual experiments designed to extend Furui's results [2] to English spoken in sentence contexts (Furui's study focused exclusively on isolated consonant-vowel syllables in Japanese). Of particular interest is the minimum duration required for accurate phonetic identification, and the temporal location of the maximal information-bearing segments. In particular we wish to refine our definition of an **avent** based on human listening experiments. We are also interested in how spectral information pertaining to a specific phonetic entity is integrated with that of other phonetic entities to create a single perceptual object (e.g., syllable or word) from such dynamic spectro-temporal patterns.

We have also initiated some machine experiments with the TIMIT database to examine what features, windows, and classifier strategies work best for the classification of the **avent** categories we have initially chosen. To begin with we are using 408 categories of **avent**, where each category initially corresponds to 51 fine phonetic classes on the right of the phone boundary for each of 8 broad acoustic-phonetic categories on the left. A pilot experiment was done in which an MLP (used as in [1]) was trained on 136,667 phonetic boundary regions as marked in TIMIT. 8th order PLP features (and their time derivatives) were extracted every 10 msec. Each training pattern consisted of 9 of these feature vectors extracted from a local region of acoustic context (plus and minus 40 msec). The input included no duration or conditioning on prior context (i.e., (11) was estimated). 29% of an independent test set consisting of 7675 patterns corresponding to **avents** were correctly classified. Adding a simple unsmoothed **avent** bigram grammar improved this result to 34.5%. For the no-grammar case, roughly 60% of the **avents** were in the top 5 choices. For this large number of phonetic categories, we consider these to be reasonable initial scores, but we do not as yet know how they will translate to speech recognition.

V. SUMMARY

In this paper we have described a hypothesis, and have presented a theoretical foundation for its application to speech recognition. 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 several hundred categories of major spectro-temporal changes. At the moment such regions can be thought of as truncated diphones, but as the work develops we expect our definition of these perceptually relevant regions to be refined.

We have just begun to explore the consequences of this hypothesis experimentally. Ultimately our hope is to develop a speech recognition system that has some of the robust properties of human audition for speech perceived under adverse conditions. This paper represents the first effort on our part to pursue this goal by modifying the fundamental statistical substrate for this goal, as opposed to merely improving the acoustical feature extraction.

VI. ACKNOWLEDGEMENTS

Thanks to Gary Tajchman and Yochai Konig for their assistance in generating the pilot results. We acknowledge the support of the International Computer Science Institute for this work.

References

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

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

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

[4] H. Hermansky. Perceptual linear predictive (PLP) analysis for speech. *J. Acoust. Soc. Am.*, pp. 1738-1752, 1990

[5] 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

[6] A. Huggins, Temporally segmented speech. *Percept. Psychophysics*, 18: 149-157, 1975