MODELING DYNAMICS IN CONNECTIONIST SPEECH RECOGNITION - THE TIME INDEX MODEL

Yochai Konig Nelson Morgan

International Computer Science Institute
1947 Center St. Suite 600
Berkeley, CA 94704, USA*

Abstract

We are experimenting with an approach to connectionist speech recognition that models the dynamics within a speech segment using temporal position as an explicit variable. Currently, the most common model for human speech production that is used in speech recognition is the Hidden Markov Model (HMM). However, HMMs suffer from well known limitations; most notably, the assumption that the observations generated in a given state are independent and identically distributed (i.i.d.). As an alternative, we are developing a time index model that explicitly conditions the emission probability of a state on the time index, where time index is defined as the number of frames since entering a state till the current frame. Thus, the proposed model does not require the i.i.d. assumption. Our pilot results suggest that the time-index approach can greatly reduce error if we have good information about the phoneme boundary location.

1 INTRODUCTION

We briefly review the main approaches to acoustic modeling in continuous speech recognition and prepare the ground for our time-index model.

1.1 What is wrong with traditional HMMs?

First, we point out some limitations of traditional HMMs. A Hidden Markov Model (HMM) generates a random sequence of observation vectors $X = \{x_1, x_2, \ldots, x_N\}$. These vectors depend on the unobserved random sequence of states $Q = \{q_1, q_2, \ldots, q_N\}$ according to the Markov chain. In most implementations of HMMs for speech recognition it is assumed that the probability that observation x_k was generated at time instance k depends only on the state q_k (in which x_k is generated). Hence the observations generated in a given state (phone) are independent and identically distributed (i.i.d). Therefore, given that the underlying process has remained in state q_j from t to $t+T$, the probability that it has generated a sequence of observation $X_t^{t+T} = \{x_t, x_{t+1}, \ldots, x_{t+T}\}$ is:

$$P(X_t^{t+T}|q_j) = \prod_{i=t}^{t+T} p(x_i|q_j) \quad (1)$$

The assumption that speech observation vectors are identically distributed might be reasonable for a short enough segment of 20-30 ms in certain situations, for example in the middle of a relatively steady-state vowel. However, when the state represents parts of sounds that are changing significantly, which is more like the rule than the exception for natural speech, associated observation vectors have statistics that are dependent on position in the segment. Furthermore, the independence assumption is inaccurate for all segments of speech, as there is strong correlation between nearby observation vectors.

1.2 Segment-Based Approaches

An alternative approach to the HMM's is the segment-based approach. In segment-based models the basic unit is a sequence of acoustic vectors emitted in a given speech unit (a "segment"), as opposed to a single acoustic vector as used for HMMs. The production of the acoustic vectors in a segment may be described as a three step procedure[1]:

1. Generate a fixed length segment M according to the distribution $P(y_1, y_2, \ldots, y_M|s_k)$, where s_k is a particular speech unit. The distribution models the trajectory of the sound in the feature vector space. $Y = \{y_1, y_2, \ldots, y_M\}$ is called the *hidden* sequence of acoustic vectors.

2. Select the length of the segment according to $P(L|s_k)$, where L is the random variable that denotes the length of the segment.

3. Down-sample Y using a time-warping transformation T_L and output the observed sequence of acoustic vectors $X = \{x_1, x_2, \ldots, x_L\}$. This transformation can be either linear or non-linear depending on the specific segmental model.

Stochastic segment models are not inherently subject to the constraints of the i.i.d. assumptions discussed earlier. An early stochastic segment model was developed by Ostendorf and Roukos [2]. A later model was introduced by Ghitza and Sondhi [3]. However, there are some practical difficulties:

1. The stochastic segment models explicitly assume a particular parametric form for the hidden observation distribution $P(y_1, y_2, \ldots, y_M|s_k)$, e.g., multivariate Gaussian. This can lead to many free parameters that must be estimated reliably from the data, e.g., large covariance matrixes. As a result, independence assumptions are often made, leading to less powerful models.

2. All the models assume a given segmentation, e.g., the knowledge of the boundaries between the basic speech units, that is known to be a difficult task. One solution is to do an exhaustive search of all the reasonable segmentations.

3. Warping the data to a fixed length segment may delete or obscure relevant information.

1.3 Preview

In the following section we introduce a time-index based model. In Section 3 we describe our implementation and experiments. We conclude by describing our thoughts for future work.

2 THE TIME-INDEX MODEL

We start by describing Deng's trended HMM, followed by our time-index model description.

2.1 Deng's Trended HMM

Deng described a model that explicitly conditioned the emission probability of a state on the time index, i.e., on the number of frames since entering a state till the current frame.. for example, if the Markov chain has two states and we assume a specific realization that alternates every two time steps between the states, the time index for a given state will be . . . 1,2,0,0,1,2, . . . as described in Figure 1 (note that the figure does not

*{konig,morgan}@icsi.berkeley.edu

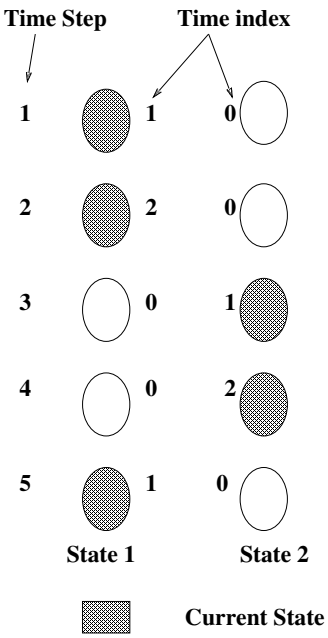

Figure 1: Time-index

show all the "machinery" of the HMM). Deng has coined his model the "trended HMM" [4]. In this model, a sequence of observation vectors generated in a given state is a combination of a stationary process and a deterministic function of time, as illustrated in the following equation for the multivariate normal distribution:

$$p(x_t|state, ti) = \frac{exp(-(x_t - g_{state}(ti))^T(\Sigma)^{-1}(x_t - g_{state}(ti)))}{(2\pi)^{\frac{n}{2}}(det\Sigma)^{0.5}} \quad (2)$$

Where ti is the time index as defined above, $g_{state}()$ is a deterministic function of the time index and has parameters that may differ from state to state. In this simplified example $g_{state}()$ shifts the mean vector of distribution as a function of the time index, while the stationary part is the variance-covariance matrix Σ. In principle this model explicitly conditions the emission probability on the time index, and a sequence of observations emitted from a given state are no longer assumed i.i.d. However, we don't know the optimal form of $g_{state}()$ for each unit of speech; For example, one would expect a different time index dependence in vowels from stops. Overall, the idea of changing the emission probability as a function of time index seemed to be innovative and potentially useful. We have incorporated this idea in a connectionist context.

2.2 An Introduction to the Time-Index Model

We are proposing a time index model that differs from an HMM in that the observations emitted in a given phone are no longer i.i.d.; and that differs from the Deng's model and others by its use of posterior probabilities as estimated with a connectionist network. In the time-index model, the realizations of the state process are no longer sequences of values taken from the phone set, but are rather chosen from a set of pairs consisting of a phone and a time index. The time index is defined as the number of frames since entering a state till the current frame. For this model, the probability of generating a sequence of observations $X = \{x_i, x_{i+1}, \ldots, x_{i+T}\}$ in a given phone $phone_j$ is:

$$P(X_t^{t+T}|phone_j) = \prod_{i=t}^{t+T} P(x_i|(phone_j, (i-t+1))) \quad (3)$$

We can see that the $q's$ in the HMM equation 1 are replaced by a phone and time index pair, as the state process is defined differently.

2.3 An Example

Figure 2 shows the topology of a basic unit of speech. Only the last state in the model has a self loop. For states with indices smaller than the minimum duration for that phone, only a transition to the next state (corresponding to a time-index increment of one) is permitted. For all other states, transitions are permitted either to the next state or to the exit state. This model differs from a traditional HMM (assuming a similar representation for duration) primarily in that the emission probability for each state (i.e. for each time associated with a phone or subphone unit type) is not constrained to be equal. Specifically, the emission probability of a state in the Markov chain is $P(x|(phone_j, ti))$, where ti is the time index. Note phones are used here as the basic speech unit. Similar equations could be used for multi-state HMMs that are also commonly used, in which the basic speech unit is smaller than a phone. While certainly one could define a standard HMM with the kind of model shown in Figure 2, and with a separate emission probability for each state, the basic problem is how to share parameters between the estimates for the separate densities. One solution would be to assume a parametric form for the trajectory, as was done by Deng. In our case, we have chosen to use a multilayer perceptron (MLP) approach, which in our previous work at ICSI, has proved useful for such estimates [5].

3 The Time Index Model - Implementation and Experiments

3.1 An Implementation of the Time Index Model

In our model we define the emission probability of a state as $P(x|phone_j, ti)$. While such a quantity can always be defined, the important question is how to estimate it. We can use the following decomposition according to Bayes' law:

$$\frac{P(x|phone_j, ti)}{P(x)} = \frac{P(phone_j|ti, x)P(ti|x)}{P(ti, phone_j)} \quad (4)$$

Where ti is the value of the time index, x is the acoustic vector, and $phone_j$ is a specific phone. Alternatively, we can decompose as follows:

$$\frac{P(x|phone_j, ti)}{P(x)} = \frac{P(ti|phone_j, x)P(phone_j|x)}{P(ti, phone_j)} \quad (5)$$

Each of the the terms conditioned on x can be estimated by an MLP with an acoustic vector (or a local neighborhood of acoustic vectors) as input, as well as any additional conditioning terms as input (for instance, an additional input representing time index ti in order to estimate $P(phone_j|ti, x)$). The targets correspond to a discrete binary coding of

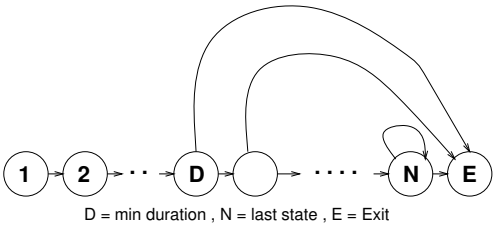

Figure 2: The topology of the time-index model

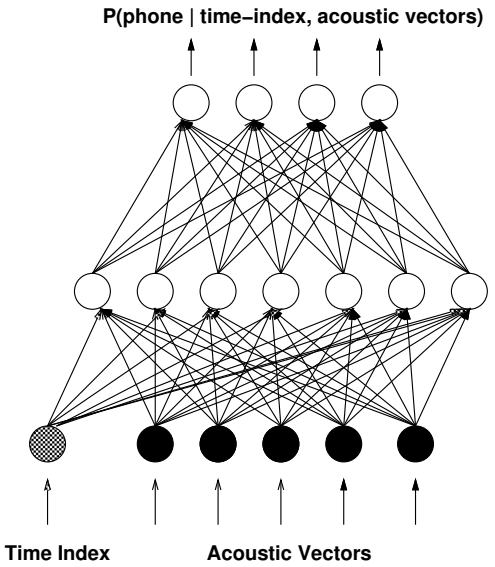

Figure 3: The time index net

the class identity that is to the left of the condition bar (e.g., $phone_j$ for estimating $P(phone_j|ti, x)$, or ti for estimating $P(ti|x)$). We have currently chosen to represent the ti inputs with a continuous-valued input as a smoother representation that requires fewer parameters. The first form of the equations given above requires the estimation of $P(phone_j|ti, x)$, and this can be done with the MLP shown in Figure 3.

$P(ti, phone_j)$ can be estimated by counting the relative frequencies in the training set. The most difficult probability to estimate is $P(ti|x)$ since this implicitly requires an estimate of the phone boundaries. Given the inertia of the articulators and the effects of co-articulation, these boundaries between adjacent phones are blurred. As a result, a reliable estimation of this probability is a still an open challenge. In the experiments reported below we have used pre-segmented data, so we could test the other parts of our model independently of the task of boundary detection. However, practical use of the time-index model will require good estimates of the probabilities of boundary positions. Some possible solutions are discussed later in this report (section 4).

3.2 Experiments

We used the Resource Management (RM) speaker independent task [6] and the TIMIT database for our initial experiments. In our RM experiments our training data consisted of 3990 read continuous sentences, and the 300 sentence Feb89 test set for development and cross-validation for the network training. The time index net (as shown in Figure 3) had 1000 hidden units, 61 outputs (the size of phone set). There were 235 inputs to the net, including 234 that consisted of 9 frames of 26 features each (PLP12 + log gain + delta features for each of these 13) [7], and a final time index input. With the exception of this final input feature, this net was the same as the hybrid HMM/MLP system as described in [5]. For the preliminary tests, we assumed knowledge of the boundaries between the phones as produced by an automatic alignment (Viterbi) procedure on the known word string [8]. These initial time-index results serve as a lower bound on error, as we can expect little improvement over the boundary detection found by the Viterbi procedure with a known word sequence. Note that this side information about the word sequence is used only to generate boundaries, and that no explicit phonetic information is preserved. Without the time-index input, the standard MLP system had 4.8% word error on this task (including insertions, deletions, and substitutions), while the incorporation of the knowledge of phoneme boundaries in the time-index network reduced the error to 1.1%.

We chose the TIMIT corpus for our second set of experiments because it is phonetically balanced, and in addition there are time-aligned phonetic transcriptions of all the sentences in the database. Our goals were to verify the potential of the model on a different test set and also to answer a potential criticism that the reduction of error is due to restricting the recognizer to utterances with the same number of phones as in the answers (it is done implicitly by suppling the known boundaries).

The experiments were done on a 200 sentence development set, that was selected from the official training set and were not used for the training. The size of the nets and the features were the same as in the RM task experiments. We used 3300 sentences for training and 396 sentences for cross-validation (the 200 sentence development set is a subset of the cross-validation set). No language model was used in these experiments. All our results are on the full 61 TIMIT phone set. Our standard system had 36.4% phone errors on this task, while the incorporation of the knowledge of phoneme boundaries in the time-index network reduced the error to 25.0%. When we restricted our standard system to sentences that have the same number of phones as in the known answers, the error rate was still 36.4%, but with a different mix of insertions, deletions, and substitutions.

These results suggest that the time-index approach can greatly reduce error if we have good information about the phoneme boundary location. This was a necessary result for the time-index approach to be ultimately useful; but it is certainly not sufficient. We are still left with the difficult and currently unsolved problem of either specifically locating boundaries, or getting reliable estimates of the probabilities of an acoustic observation corresponding to a particular temporal region of

a segment.

In the following section we discuss possible ways to address this problem.

4 Discussion and Future Work

4.1 Segmentation - How to Find Transitions?

If we could explicitly and reliably find the boundaries between phonetic segments, the preliminary result from the previous section would seem to indicate that we can greatly reduce errors. However, this problem does not have an easy solution, see for example [9]. We consider two possible styles of approach: first, try to learn smooth probability densities for the boundaries, as per the equations from the previous section; and second, use the time-index model as a second pass, where a previous pass will generate possible alternate segmentations to be considered using the new model.

To estimate probabilities such as $P(ti|phone_j, x)$ or $P(ti|x)$, we must train an MLP classifier to discriminate between different temporal regions of each segment, for instance between frames that are boundary and non-boundary frames. A critical issue here is the features used for this discrimination, both in terms of the signal analysis chosen and the frame rate and window size used. In one comparative study of signal representations [10] , Bark auditory cepstral coefficients (BASC) achieved the lowest deletion error rate (the percentage of the transcription boundaries not found by a boundary detector) when used with a frame rate of 5 ms and window size of 28.5 ms.

Using an unconstrained temporal estimation probability estimator has several inherent problems:

- Due to the inertia of the articulators, the boundaries between phones are blurred and ambiguous in continuous speech.
- Getting accurate targets for training an MLP through automatic procedures is difficult. Other sites that have been successful at nearly eliminating boundary deletions have done so at the expense of many insertions; this suggests that the temporal probability estimates may risk being too inaccurate to be useful in our model. However, global constraints may be used to eliminate at least some of the spurious boundaries.

However, if we can overcome these problems, the potential payoff is high (as noted in our preliminary experiment), and computational considerations may make such a method preferable over the N-best approaches described below. Furthermore, an explicit single-pass approach may find some correct segmentations that a two-pass N-best approach with finite N may eliminate.

An alternative approach to explicit segmentation is the N-best paradigm. Considering all possible segmentations is computationally infeasible. However, as many researchers have noted [11], recognizers that are already fairly good can yield a list of the most likely segmentations, such that all other segmentations are highly improbable. If only these reasonable segmentations are considered, a recognition score can be obtained for each one using the time-index model and the boundary information from the segmentation. For a more detailed discussion of the two-pass approach see [12].

4.2 Summary

This report describes an early stage in our research on time index models as potential representations of speech production that can be used for speech recognition. Our initial results on pre-segmented data are encouraging, showing that strong knowledge of the phonetic boundaries can improve the recognition accuracy. However, we still face the problem of either explicitly or implicitly finding the boundaries between the phones. We have discussed two possible solutions: estimating temporal probabilities (which implicitly requires learning where the boundaries are), and using the boundaries obtained from a first pass with a simpler recognizer using an N-best search.

Acknowledgement

We thank software guru Phil Kohn for BoB (our neural network simulator) and for many helpful discussions. Hervé Bourlard shared his wisdom, insight into stochastic models, and Belgian humor. We thank Gary Tajchman and Nikki Mirghafori for their help and advice. We also thank the many researchers who were so open with their discussions of their own segmental models, particularly: Mari Ostendorf, Oded Ghitza, and Li Deng. We gratefully acknowledge the support of the Office of Naval Research, URI No. N00014-92-J-1617 (via UCB), ESPRIT project 6487 (WERNICKE) (through ICSI), and ICSI in general for supporting this work.

References

[1] V.V. Digialakis. *Segment-Based Stochastic Models of Spectral Dynamics for Continuous Speech Recognition*. PhD thesis, Boston University, 1992.

[2] M. Ostendorf and S. Roukos. A stochastic segment model for phoneme-based continuous speech recognition. *IEEE ASSP trans.*, 37(12):1857–1869, December 1989.

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

[4] L. Deng. A generalized hidden markov model with state-conditioned trend functions of time for the speech signal. *Signal Processing*, 27:65–78, 1992.

[5] H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[6] P. Price, W. Fisher, J. Bernstein, and D. Pallet. The darpa 1000-word resource management database for continuous spee ch recognition. In *Proceedings IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, volume 1, pages 651–654, New York, 1988. IEEE.

[7] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *JASA*, 87, 1990.

[8] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. on Information Theory*, 13(2):260–269, 1967.

[9] J.R. Glass. *Finding Acoustic Regularities in Speech Applications to Phonetic Recognition*. PhD thesis, M.I.T, May 1988.

[10] H.C. Leung, B. Chigier, and J.R. Glass. A comparative study of signal representations and classification techniques for speech recognition. In *Proceedings Int'l Conference on Acoustics Speech and Signal Processing*, Minnesota, USA, 1993.

[11] M. Ostendorf et al. Integration of diverse recognition methodologies through reevaluation of n-best sentence hypotheses. In *Proceedings DARPA Speech and Natural Language Workshop*, 1991.

[12] Y. Konig and N. Morgan. Modeling dynamics in connectionist speech recognition - the time index model. Technical Report TR-94-012, International Computer Science Institute, 1947 Center St. Suite 600, Berkeley, CA 94704, April 1994.