Contents

1 Introduction **1**
- 1.1 Introduction to Stochastic Speech Modeling . . . 1
- 1.2 What is wrong with traditional HMMs? . . . 2
- 1.3 Preview . . . 3

2 Related Work **3**
- 2.1 Introduction to Non-stationary Modeling . . . 3
- 2.2 Segment-Based Approaches . . . 4
  - 2.2.1 Introduction . . . 4
  - 2.2.2 Segmental Models . . . 4
  - 2.2.3 A Stochastic Dynamic System Approach . . . 5
- 2.3 Related Work in Stochastic Modeling . . . 6
  - 2.3.1 Deng's Trended HMM . . . 6
  - 2.3.2 Hidden Control Neural Architecture . . . 7

3 The Time Index Model **7**
- 3.1 Introduction . . . 7
- 3.2 Definition and an Example . . . 8

4 The Time Index Model - Implementation and Experiments **8**
- 4.1 An Implementation of the Time Index Model . . . 9
  - 4.1.1 Experiments . . . 11

5 Discussion and Future Work **11**
- 5.1 Segmentation - How to Find Transitions? . . . 11
  - 5.1.1 Explicit Segmentation . . . 12
  - 5.1.2 The N-Best Paradigm . . . 12
- 5.2 Summary . . . 16

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

Modeling Dynamics in Connectionist Speech Recognition - The Time Index Model

Yochai Konig and Nelson Morgan

TR-94-012

March 1994

Abstract

Here, we introduce an alternative to the Hidden Markov Model (HMM) as the underlying representation of speech production. HMMs suffer from well known limitations, such as the unrealistic assumption that the observations generated in a given state are independent and identically distributed (i.i.d). We propose a time index model that explicitly conditions the emission probability of a state on the time index, i.e., on the number of "visits" in the current state of the Markov chain in a sequence. Thus, the proposed model does not require an i.i.d. assumption. The connectionist framework enables us to represent the dependence on the time index as a non-parametric distribution and to share parameters between different speech unit models. Furthermore, we discuss an extension to the basic time index model by incorporating information about the duration of the phone segments. Our initial results show that given the position of the boundaries between basic speech units, e.g., phones, we can improve our current connectionist system performance significantly by using this model. However, we still do not know whether these boundaries can be estimated reliably, nor do we know how much benefit we can obtain from this method given less accurate boundary information. Currently we are experimenting with two possible approaches: trying to learn smooth probability densities for the boundaries, and getting a set of reasonable segmentations from an N-Best search. In both cases we will need to consider the effect of incorrect boundaries, since they will undoubtedly occur.

1 Introduction

1.1 Introduction to Stochastic Speech Modeling

Human speech production is usually modeled in statistical recognition systems by two stochastic processes: the state process, and the observation process. The state process is used to represent the underlying sequence of speech sounds that were said, which in principle could be viewed as a crude representation of possible configurations of the vocal tract and the articulators during the production of the phones. The observation process models the generated time-series of the feature vectors (one feature vector for every short segment of speech, typically around 10ms). The realizations of the state process are hidden from us, and we have to estimate them through the realizations of the observation process.

The goal of speech recognition is to find the most likely word sequence, which sometimes is mediated (for practical reasons) by first finding the hidden realization of the state process. Specifically, we want to find the most likely word sequence given a sequence of feature vectors, i.e., to maximize the probability $P(M|X)$ over all legal word sequences, where M stands for the word sequence (model), and X denotes the feature vector sequence. This probability is commonly decomposed using Bayes' law:

$$\max_{M} \; P(M|X) = \frac{P(X|M)P(M)}{P(X)} \tag{1}$$

$P(M)$ is typically estimated by a separate language model (assuming no interaction between the acoustic and language models), and $P(X)$ is generally ignored since it is constant over all hypothesized models during recognition and thus does not influence the selection of the recognized word sequence.

$P(X|M)$ can be computed by summing over the set of all the possible state sequences Γ that are valid for a particular word sequence M as given in the following equation:

$$\sum_{Q \in \Gamma} P(X, Q|M) \tag{2}$$

Where Q is a valid state sequence for the word sequence M. Given the heavy computation needed by the above equation (and the need to determine a state sequence) the following approximation known as the Viterbi assumption is often used:

$$P(X|M) \approx \max_{q_{l_1}, q_{l_2}, \ldots, q_{l_N}} P(X, q_{l_1}, q_{l_2}, \ldots, q_{l_N}|M) \tag{3}$$

Where $q_{l_1}, q_{l_2}, \ldots, q_{l_N}$ is a valid state sequence for the given word sequence M.

In the following section we describe the traditional HMM and the related assumptions about the state and observation processes. Relaxing some of these assumptions will then motivate a statistical model that explicitly incorporates the time index for each phonetic segment.

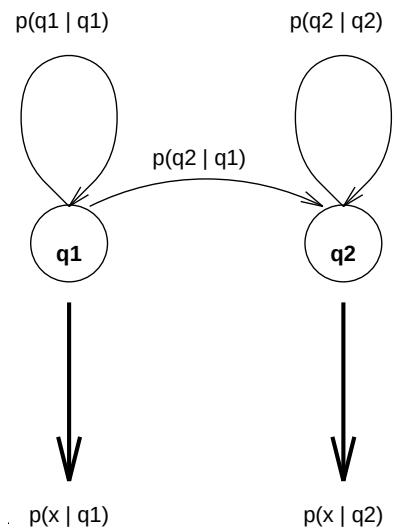

Figure 1: An HMM example

1.2 What is wrong with traditional HMMs?

In this section we give a brief introduction to HMMs that are used in current speech recognition systems and point out their limitations. For a more detailed description see [Lee89, XH90]. By using standard HMMs, we assume that the possible values for the state process at every time step are the basic speech units, usually phones. Specifically, let us consider a Markov chain with M states, $Q \equiv \{q_1, q_2, \ldots, q_M\}$, and associated transition probability matrix a_{ij} where a_{ij} is the probability of taking a transition from state i to state j, i.e., $a_{ij} = prob(q_j|q_i)$, and $b_i(x)$ the output probability set - the probability of emitting the feature vector x while in state i. Note that these probabilities are not dependent on the time within a state.

A Hidden Markov Model (HMM) based on this Markov chain 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 in 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 $X = \{x_i, x_{i+1}, \ldots, x_{i+T}\}$ is:

$$P(X_t^{t+T}|q_j) = \prod_{i=t}^{t+T} p(x_i|q_j) \tag{4}$$

(See Figure 1 for an example.) The assumption that speech observation vectors are identically distributed might be reasonable for a short enough segment of (20-30ms) 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.

1.3 Preview

In the following section we describe related work in stochastic modeling that addresses some of the problems mentioned above. In section 3 we introduce our time index model, and give its formal definition. We describe our particular implementation and initial experiments in section 4. We conclude with a discussion of the problem of finding the boundaries between phones (segmentation) that arise from our model. Two possible solutions to the segmentation problem are suggested: explicit segmentation, and the N-best paradigm.

2 Related Work

2.1 Introduction to Non-stationary Modeling

In this section we discuss a number of models that have been proposed to remedy some of the shortcomings of HMMs. A quick solution might be to represent each unit of speech by enough different states to approximate its non-stationary nature in a stepwise fashion. For instance, a vowel could be represented by ten different states. This solution has two major limitations:

- There are too many free and independent model parameters. This necessitates more training data, and also might be more prone to capture irrelevant sources of variance in the data than a simpler model.
- Such a model does not capture the correlation and dependence between the different states. For states with a short duration, this would be even more pronounced, since the change between two states in a sequence would correspond to only a small movement of articulators for a given speaker.

Several extensions to the basic HMM have been proposed in order to overcome some of these deficiencies. For example, autoregressive HMMs condition the emission probability of a given state on previous observations [JR85]. However, none of these extensions have explicitly modeled the emission in a given phone as a non-stationary process. In general this is too difficult to handle with a practical number of parameters.

A number of HMM alternatives model the sequence of frames emitted in a given sub-word unit as correlated and dependent on each other. The models differ in their assumptions about the nature of the correlation between the frames in the sequence. For instance, some assume that only consecutive frames are correlated, while others assume that all the frames in the sequence are dependent on each other. In general these models do not require the

HMM assumption of independent and identically distributed observations . In the following section we survey segment-based approaches that are in this family.

2.2 Segment-Based Approaches

2.2.1 Introduction

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 [Dig92]:

1. 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, and s_k is a particular speech unit.

2. A fixed length segment M is generated according to the distribution $P(y_1, y_2, \ldots, y_M|s_k)$. The distribution models the trajectory of the sound in the feature vector space. M is chosen to be greater than all the possible values of L. $Y = y_1, y_2, \ldots, y_M$ is called the *hidden* sequence of acoustic vectors.

3. Down-sample Y using the 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.

2.2.2 Segmental Models

To briefly illustrate the range of stochastic segment models, we review some specific examples and their implementations. These models differ in the form of the distribution $(Y|s_k)$ and in the time-warping transformation T_L. Ostendorf and Roukos [OR89] have used (among a number of methods) linear time sampling in their study, i.e., sampling Y in equal intervals along the time axis as their time warping procedure. Their specific implementation had ten 14-dimensional vectors of cepstral coefficients. They used a multivariate Gaussian to represent the entire segment, which can require a 140 by 140 full covariance matrix for each phone (assuming that feature dependence is accounted for).

Ghitza and Sondhi developed a model [GS93] that can also be viewed as a stochastic segment model with the following distinctions:

- Their warping procedure is a dynamic time warping technique, instead of the linear time warping method used by Ostendorf and Roukos.

- They used diphones as their sub word units, as opposed to to the phones in Ostendorf and Roukos' stochastic segment model [OR89].

- They maintained the HMM framework and assumed a semi-hidden Markov chain, i.e., each state has an explicit duration distribution.

These stochastic segment models are not inherently subject to the constraints of the i.i.d. assumptions discussed earlier. However, there are some practical difficulties:

1. There are 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. These methods explicitly assume a particular parametric form for the observation distributions, e.g., multivariate Gaussian. This assumption is already faulty for standard HMMs, but may be even a worse approximation once observation interdependencies are taken into account. (Nonetheless, it is a sensible place to start).

3. All the models assume a given segmentation, e.g., the knowledge of the boundaries between the basic speech units, which is known to be a difficult task. One solution is to do an exhaustive search of all the reasonable segmentations as discussed in section 5.

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

2.2.3 A Stochastic Dynamic System Approach

This model assumes a stochastic dynamic system with a continuous state process as the source for the observation process. To model an underlying dynamic system, some assumptions are required. For example, Digalakis has proposed two possible model constraints:

1. *Trajectory invariance*: It is assumed that the same sequence of states of the stochastic dynamic system is the source for all the possible realizations of the speech segment. The observed speech segment is a down-sampled version of the underlying trajectory of the feature vectors created by the system. Thus, long observation sequences have higher correlation between successive frames than short observation sequences.

2. *Correlation invariance*: It is assumed that the correlation between two observations depends only on the relative location of the observations in the segment, i.e., is invariant under the time-warping transformation. Thus, the sequence of the of the underlying model depends on the observed segment length.

In his study, Digalakis assumes that the observed segment of speech is the output of a piecewise time-invariant linear dynamical system. He uses up to five invariant regions for each model. The models based on the *correlation invariance* assumption outperformed the models based on the *trajectory invariance* assumption for the task of phone classification. For more details see [Dig92, DRO93]. The stochastic dynamic system approach appears to have more modeling power than an HMM, and can potentially capture the dynamics of acoustic vectors within a segment of speech. However, there are still open issues about the structure of the dynamic system.

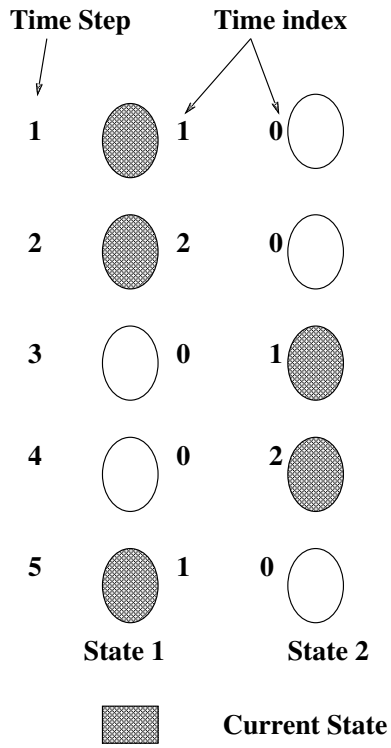

Figure 2: Time-index

2.3 Related Work in Stochastic Modeling

In this section we review stochastic models that try to capture the dynamics of speech, and that have also influenced the time index model that is the main topic of this report.

2.3.1 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 "visits" in the current state in a sequence in the chain. 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 2 (note that the figure does not show all the "machinery" of the HMM). Deng has coined his model the "trended HMM" [Den92]. 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:

$$p(x_t|state,t) = \frac{1}{(2\pi)^{\frac{n}{2}}(det\Sigma)^{0.5}} exp(-(x_t - g_{state}(t))^T(\Sigma)^{-1}(x_t - g_{state}(t))) \quad (5)$$

Where t is the time index as defined above, $g_{state}()$ is a deterministic function of the time index and has parameters which 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 than in 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.

2.3.2 Hidden Control Neural Architecture

Esther Levin suggested the idea of changing the mapping that is implemented by a multi-layered neural network as a function of additional control signals [Lev93]. Her model is called a "Hidden Control Neural Network" (HCNN). The advantage of the HCNN over a static neural network is that it can model signals generated by nonlinear dynamic systems with *time variability*, i.e., the mapping is a function of the state of the system. Before we can make use of this model we have to uncover the the hidden part of the model. In order to estimate the values of the hidden control input in training and recognition we must make assumptions about the nature of the system to be modeled. It is not clear what assumptions are suitable for the speech signal. Levin used her model for prediction of time series, and took the common assumption of modeling the prediction error as a white Gaussian noise. While Levin used her model in a non-discriminant model, hidden control can also be used in a discriminant fashion. This leads to a connectionist time index method.

3 The Time Index Model

3.1 Introduction

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 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 the number of "visits" in the current state in a sequence in the chain. 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:

$$P(X_t^{t+T}|phone_j) = \prod_{i=t}^{t+T} P(x_i|(phone_j, (i-t+1))) \tag{6}$$

We can see that the $q's$ in the HMM equation 4 are replaced by a phone and time index pair, as the state process is defined differently.

3.2 Definition and an Example

The time index model is defined as follows:

- The state at each time step is the ordered tuple $q = (\delta, ti) \in \Delta \times T$, where Δ is the set of basic language units and T is a random variable from the set $\{1, 2, \ldots, \eta\}$, where η is the maximum allowable duration. Then let Q_t^{t+T-1} represent the following state sequence:
$$\{(phone_j, 1), (phone_j, 2), \ldots, (phone_j, T-1)\}.$$
- Represent the sequence of acoustic vectors emitted between time step t and $t+T-1$ as X_t^{t+T-1}. Then define a conditional probability for the observation sequence given the state sequence corresponding to each $phone_j$ (Assuming output conditional independence)
$$P(X_t^{t+T-1}, Q_t^{t+T-1}) = P(T|phone_j) \prod_{i=t}^{t+T-1} P(x_i|phone_j, (i-t+1))$$

Figure 3 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 primarily differs from a traditional HMM (assuming a similar representation for duration) 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 3, 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 multi-layer perceptron (MLP) approach, which in our previous work at ICSI has proved useful for such estimates [BM94].

4 The Time Index Model - Implementation and Experiments

We present here a specific implementation of the time index model and discuss initial experiments with the Resource Management Task [PFBP88].

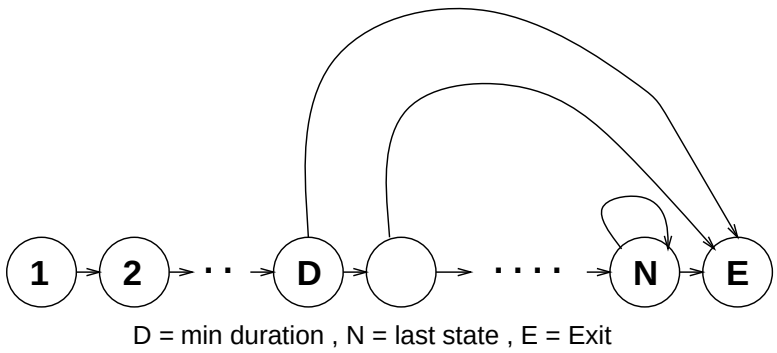

Figure 3: The topology of the time-index model

4.1 An Implementation of the Time Index Model

In section 2, we described 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:

$$\frac{P(x|phone_j, ti)}{P(x)} = \frac{P(phone_j|ti, x)P(ti|x)}{P(ti, phone_j)} \tag{7}$$

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:

$$\frac{P(x|phone_j, ti)}{P(x)} = \frac{P(ti|phone_j, x)P(phone_j|x)}{P(ti, phone_j)} \tag{8}$$

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) at the input, as well as any additional conditioning terms at the input (for instance, an input representing time index ti in order to estimate $P(phone_j|ti, x)$); the targets correspond to a discrete binary coding of 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 4.

$P(ti, phone_j)$ 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

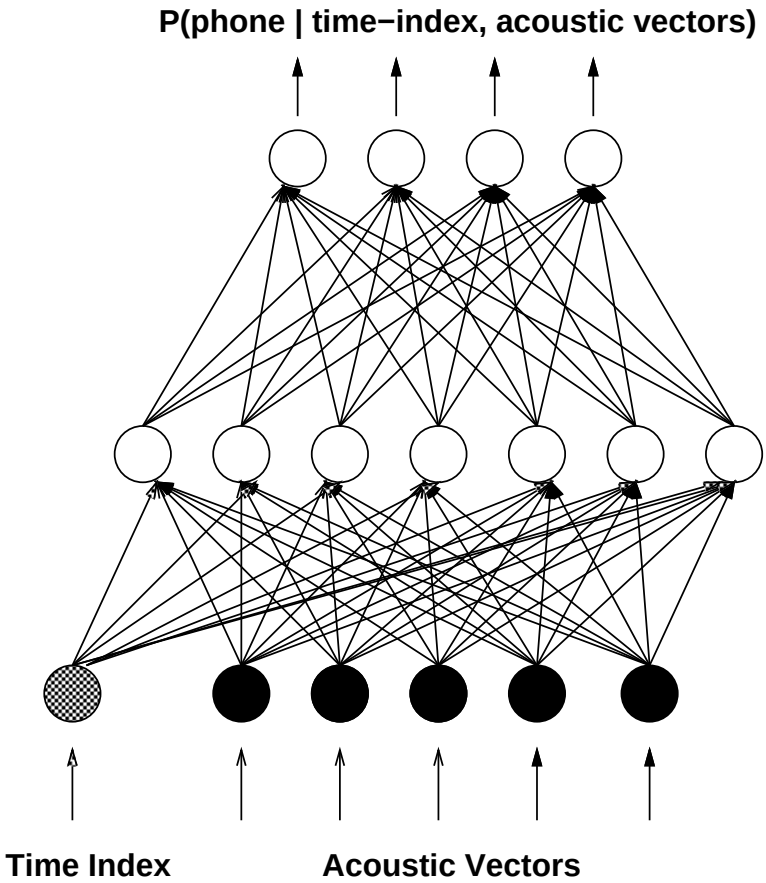

Figure 4: The time index net

model will require good estimates of the probabilities of boundary positions. Some possible solutions are discussed later in this report (section 5).

4.1.1 Experiments

We used the Resource Management speaker independent task [PFBP88] for initial experiments. 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 4) 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), 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 [BM94]. 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. These initial time-index results serve as a lower bound on error, as we can expect little improvement over the boundary detection found by Viterbi procedure with a known word sequence. Note that this side information about the word sequence is only used to generate boundaries, and that no explicit phonetic information is preserved.

Without the time-index input, the standard MLP system had 4.8% word errors 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%. This suggests 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, and also suggest several extensions to the basic time index model.

5 Discussion and Future Work

5.1 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 [Gla88]. 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.

5.1.1 Explicit Segmentation

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 a comparative study of signal representations[LCG93] it has been found that 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 5ms and the size of the analysis window of 28.5 ms. We are currently experimenting with these features to see how well we can get the required temporal probabilities. We may also try to use RASTA-PLP features for a similar reason. Another idea currently being considered is the use of broad phone categories for estimating quantities like $P(ti|phone_j, x)$, since the temporal effects are likely to be fairly similar for broad classes of phones, such as the one used by Leung and colleagues [LHZ91] to find transitions between vowels, nasals, liquids, fricatives, stops, and silence.

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

- Due to 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 2-pass N-best approach with finite N may eliminate.

5.1.2 The N-Best Paradigm

Considering all possible segmentations is computationally infeasible. However, as many researchers have noted, 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 obtaine for each one using the time-index model and the boundary information from the segmentation. Defining a segmentation S as follows:

$$S \stackrel{\text{def}}{=} \{(t_1, t_2), (t_2 + 1, t_3), (t_{n-1} + 1, t_n)\} \tag{9}$$

where

$$1 = t_1 < t_2 < t_3 < \ldots < t_n = N.$$

Given a segmentation we can compute the exact time index for each frame, and by this bypass the need to estimate probabilities such as $P(ti|x)$ explicitly. The goal of the N-best search is to generate a list of N candidate hypotheses, as well as the segmentations, i.e., the boundaries between the phones for each hypothesis. R. Schwartz and Y. Chow describe an exact algorithm for finding the N mostly likely sentence hypotheses in [SC90]. The main idea was to keep separate records for paths with different word sequences histories, adding the probabilities for multiple paths with the same history that transition into the same state. An approximation of this algorithm is to keep only the previous word as the history of the path (as described in [SA91]). Typically the N-Best paradigm is used in the following way:

1. Find the N-best sentence hypotheses with a relatively simple and fast system.
2. Rescore the sentences with a more powerful system.
3. Combine the scores of the two systems, and output the sentence with the best score.

See [ea91] for more details.

Chapter 3 in [BM94] describes statistical speech recognition in terms of a series of assumptions that lead to the standard HMM formulation. Here we describe the potential use of the N-Best procedure as a modification of the derivation given in that chapter, as a result of the knowledge of the segmentation and the time-index model. We can now rewrite equation 1 in the following way:

$$\max_M \; P(M|X) = \max_M \sum_{S \in \Theta} P(S, M|X) \tag{10}$$

Where Θ is the set of the segmentations obtained from the recognized N-best sentences, and where we assume that the $P(S, M|X)$ terms are negligible for $S \notin \Theta$. This equation can be approximated in the following way:

$$\max_M \sum_S P(S, M|X) \approx \max_{M,S} P(S, M|X) \tag{11}$$

Using Bayes law we use the following decomposition:

$$P(S, M|X) = \frac{P(X|M, S)P(S|M)P(M)}{P(X)} \tag{12}$$

As before we assume that $P(M)$ is estimated by a separate language model, and that $P(X)$ is constant during recognition. $P(S|M)$ can be computed by:

$$P(S|M) = \sum_{l_1, l_2, \ldots, l_n \in \Gamma_{M,S}} P(q_{l_1}, q_{l_2}, \ldots, q_{l_N}|M) \tag{13}$$

where $\Gamma_{M,S}$ is the set of the all the valid state sequences given the model M and the segmentation S. A reasonable approximation for the above sum is the Viterbi approximation, i.e., the sum becomes a max as shown below:

$$P(S|M) = \max_{l_1,l_2,\ldots,l_n \in \Gamma_{M,S}} P(q_{l_1}, q_{l_2}, \ldots, q_{l_N}|M) \tag{14}$$

The probability $P(X|M,S)$ is calculated using the forward recursion of the Forward-Backward algorithm [Bau72] as described in chapter 3 in [BM94] with the following set of assumptions:

- H1: The transition probabilities are independent of the observation vectors as described in the following equation:

  $$P(q_l^n|q_k^{n-1}, X_1^{n-1}, S, M) = P(q_l^n|q_k^{n-1}, S, M) \tag{15}$$

  Note that in the case of the phone models as described above, using single pronunciation word models, the state sequence becomes deterministic given the model and the segmentation. In the case of multi-pronunciation word models, the transition between phones is still stochastic, with probabilities given as part of the word models.

- H2: Taking the observation-independence assumption we get:

  $$P(x_n|q_l^n, q_k^{n-1}, X_1^{n-1}, M, S) = P(x_n|q_l^n, M, S) \tag{16}$$

  Note that our definition of the states is different from the standard HMM, as described above. Also note that in the estimation of the emission probability, as described below, we incorporate some dependence between observations by using multiple feature vectors as an input to the estimation network.

- H3: Assuming that the emission probability is *context independent* we get:

  $$P(x_n|q_l^n, M, S) = P(x_n|q_l^n, S) \tag{17}$$

We estimate $P(x_n|q_l^n, S)$ using Bayes law:

$$P(x_n|q_l^n, S) = \frac{P(q_l^n|x_n, S)P(S|x_n)}{P(x_n)} \tag{18}$$

Writing the q as a pair of two random variables we get:

$$P(x|phone, ti, S) = \frac{P(phone|ti, S, x)P(ti|x, S)P(S|x)P(x)}{P(phone, ti, S)} \tag{19}$$

We estimate $P(phone|ti, S, x)$ by the net described in figure 5. Note that we provide as an extra input to the time index net described above the duration of the segment that the

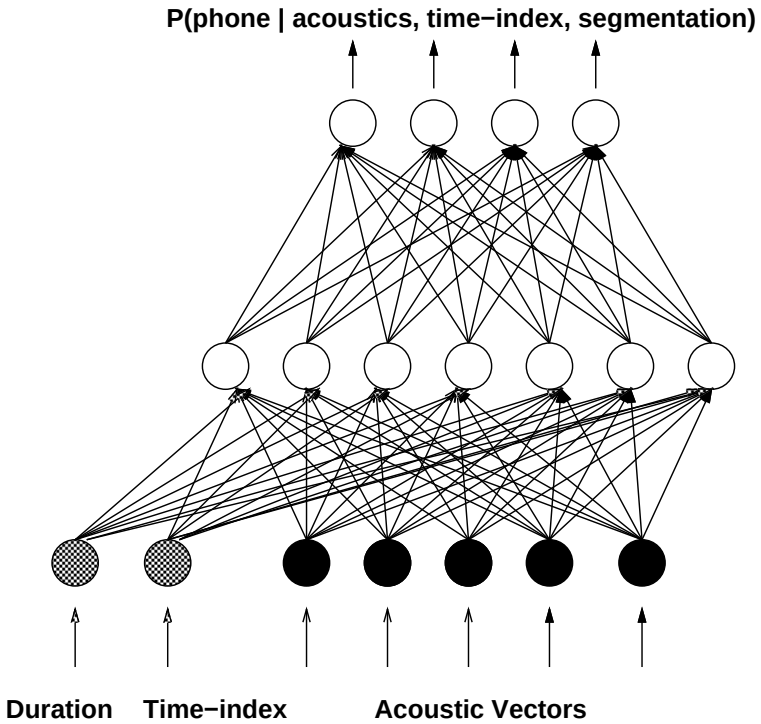

Figure 5: The time index net with a duration input

current acoustic vector (the center of the temporal window) belongs, as this in addition to the time-index input provides the relevant local information about the segmentation as implied by the conditioning on S. The duration of the adjacent segments could also be used as an input. The term $P(ti|x, S)$ becomes a deterministic term, a the correct time index is just the distance in frames from the previous boundary. The following can also be assumed:

$$P(phone, ti, S) = P(phone, ti) \tag{20}$$

And can be computed from the relative frequencies in the training set. The probability $P(S|x)$ is independent of the model and only influences the likelihood score. Currently we assume uniform priors for all segmentations independent of the acoustic vectors. However, we will be experimenting with plausible data dependent segmentation probabilities, such as the score of the recognized sentence from the first pass.

5.2 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 is almost as powerful as knowing the phonetic identities themselves. 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. Herve 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

[Bau72] L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3:1–8, 1972.

[BM94] H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[Den92] L. Deng. A generalized hidden markov model with state-conditioned trend functions of time for the speech signal. *Signal Processing*, 27:65–78, 1992.

[Dig92] V.V. Digialakis. *Segment-Based Stochastic Models of Spectral Dynamics for Continuous Speech Recognition*. PhD thesis, Boston University, 1992.

[DRO93] V.V. Digalakis, J.R. Rohlicek, and M. Ostendorf. Segment-based stochastic models of spectral dynamics for continuous speech recognition. *IEEE trans. on Speech and Audio Processing*, 1(4):431–442, October 1993.

[ea91] M. Ostendorf et al. Integraration of diverse recognition methodologies through reevaluation of n-best sentence hypotheses. In *Proceedings DARPA Speech and Natural Language Workshop*, 1991.

[Gla88] J.R. Glass. *Finding Acoustic Regularities in Speech Applications to Phonetic Recognition*. PhD thesis, M.I.T, May 1988.

[GS93] 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.

[JR85] B. H. Juang and L.R. Rabiner. Mixture autoregressive hidden markov models for speech signals. *IEEE ASSP Magazine*, 6(33):1404–1413, 1985.

[LCG93] 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.

[Lee89] K. F. Lee. *Automatic Speech Recognition: The Development of the SPHINX System*. Kluwer Academic Publishers, 1989.

[Lev93] E. Levin. Hidden control neural architecture modeling of nonlinear time varying and its applications. *IEEE trans. on Neural Networks*, 4(1):109–116, January 1993.

[LHZ91] H.C. Leung, I.L. Hetherington, and V.W. Zue. Speech recognition using stochastic explict-segment modeling. In *Proceedings European Conf. on Speech Communication and Technology. (EUROSPEECH)*, Genova, Italy, 1991.

[OR89] M. Ostendorf and S. Roukos. A stochastic segment model for phoneme-based continuous speech recognition. *IEEE ASSP trans.*, 37(12):1857–1869, December 1989.

[PFBP88] 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.

[SA91] R. Schwartz and S. Austin. A comparison of several approximate algorithms for finding multiple (n-best) sentnce hypotheses. In *Proceedings Int'l Conference on Acoustics Speech and Signal Processing*, Toronto, Canada, 1991.

[SC90] R. Schwartz and Y. Chow. The n-best algorithm: An efficient and exact procedure for finding the n most likely sentence hyoptheses. In *Proceedings Int'l Conference on Acoustics Speech and Signal Processing*, New Mexico, USA, 1990.

[XH90] M.A. Jack X.D. Huang, Y. Ariki. *Hidden Markov Models For Speech Recognition*. Edinburgh University Press, 1990.