Acoustic Modeling in Continuous Speech Recognition - A Connectionist Approach

Yochai Konig

October 1994

Abstract

We are currently studying acoustic models for continuous speech recognition. These models represent the stochastic generation of feature vectors for an utterance, where a feature vector represents a short interval of speech, typically around 20ms. Our current emphasis is on devising stochastic models that overcome some of the principal limitations of Hidden Markov Models (HMMs). Some of these limitations include an inability to model non-stationarities within a speech segment, assumed independence between acoustic vectors within a state, and lack of local discrimination between similar sounds.

Our work is related to that of a number of researchers who are attempting to overcome these limitations. Like a number of these researchers, we are interested in non-stationary modeling within a segment corresponding to a general speech category (e.g., phoneme). Our work differs in that we are considering stochastic models that use a posteriori probabilities, which are generated by a Multi-Layer Perceptron (MLP). In one study, we have modeled the dynamics within a speech segment using temporal position as an explicit variable. This time index model 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 until the current frame. The lesson from the pilot experiments is that the approach can greatly reduce error if we have good information about the phoneme boundary location. Currently we are experimenting with other discriminant models. We report here on the limitations and theoretical solutions, and suggest a line of research to investigate their practicality.

1 Introduction

1.1 Motivation and Goal

The goal of speech recognition is to map speech waveforms into a sequence of words. Given a specific utterance there will be a lot of different realizations of it when spoken by different speakers, due to different speaking rates, gender, accent, age, etc. Even the same speaker may pronounce a sentence with some variation given his stress level, time of day, amount of background noise, etc. All of these introduce a great deal of uncertainty that we have to overcome in our quest for the mapping between a speech waveform into a sentence. We have three kinds of information (assuming no visual images are available) to help us: the acoustic waveform, and the previous words or sentences that were recognized, and any knowledge or expectation about what is being said. Traditionally in speech recognition the research is divided accordingly into *acoustic modeling* and *language modeling*. This study is mainly devoted to acoustic modeling.

Usually acoustic modeling is divided into two steps: feature extraction, and feature vector generation modeling. In the feature extraction step we transform the samples of a short segment of speech (typically around 20-30 ms) into a small number of features, which can be either continuous or discrete. Besides the need for dimension reduction, we want to filter out redundant information. In the second step we model the stochastic generation of a sequence of feature vectors while speaking an utterance. We concentrate our efforts in this step.

Our goal is to devise stochastic models that overcome some of the principal limitations of HMMs. These limitations include the assumption that observations within the same speech segment (generated by the same state) are independent and identically distributed (i.i.d), an unrealistic assumption given the non-stationary nature of speech. Furthermore the training procedure for HMMs is based on Maximum Likelihood Estimate (MLE) criteria, so there is a lack of local discrimination between similar sounds. We further discuss the limitations of HMMs in section 1.3.

In order to try to overcome some of the HMM's limitations, assumptions about the nature of the non-stationarity within a speech segment corresponding to a general speech category (e.g., phoneme) have to be made. One approach taken by other researchers is to model the generation of all the acoustic vectors in a speech segment as one stochastic event. Thus, the model captures the correlations and dependencies between the observations. We describe this approach (known as segment-based models), and other related work in non-stationary modeling in section 2.

One difference in our line of research from the existing work is that we consider stochastic models that use a posteriori probabilities, which are generated by Multi-Layer Perceptron (MLP). In one study we model the non-stationarity within a speech segment as dependent on the temporal position. The time index model 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. We describe this model and its relation to the existing work in section 3.

Currently we are experimenting with models that attempt to overcome the lack of discrimination between similar sounds and utterances in HMMs. Specifically we return to the original work on hybrid models that combine HMMs and MLPs [1] and reexamine the assumptions that have been made to make it work. The original formulation has the advantage of being globally discriminant between competing utterances and thus being closer to the ideal Bayes classifier. We report about our experiments in section 5.

Below we give a short introduction to stochastic modeling of speech and point out some of the limitations of HMM's.

1.2 Introduction to Stochastic Speech Modeling

In statistical speech recognition systems we model utterance production by two stochastic processes: the state process, and the observation process. The state process is used to represent the underlying sequence of speech sounds, 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 20ms segments, overlapped every 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:

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

$P(M)$ 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. To reduce the amount of computation and storage 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. Below we show that the Viterbi assumption is a reasonable one at least for the corpus that we have used in our study.

In the following section we describe the traditional HMM and the related assumptions about the state and observation processes.

1.3 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 [12, 16]. 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. See Figure 1 for an example.

A Hidden Markov Model (HMM) based on this Markov chain can generate sequences of observation vectors $X = \{x_1, x_2, \ldots, x_N\}$. These vectors depend on the unobserved random sequence of states $\Gamma = \{q_{l_1}, q_{l_2}, \ldots, q_{l_N}\}$ 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_{l_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:

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

In order to analyze the i.i.d assumption mentioned above we have to consider what is the unit of speech that is modeled by a state of the Markov chain. Usually the states model a sub-phone or a phone. On the one hand as we have more states per phone, we reduce the effect of the identically distributed assumption, because it 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. On the other hand as we have more states per phone we have to estimate more parameters from the same amount of training data. Furthermore, the independence assumption is inaccurate for all segments of speech, as there is strong correlation between nearby observation vectors. And even if there is a separate state for every acoustic vector we will suffer from the independence assumption unless we share parameters between the different states.

Another limitation of the HMM is that its standard training criterion is non discriminant. For every training sentence we maximize the likelihood of the correct model, without reducing the likelihoods of all the wrong models. Brown [3] has suggested a discriminant

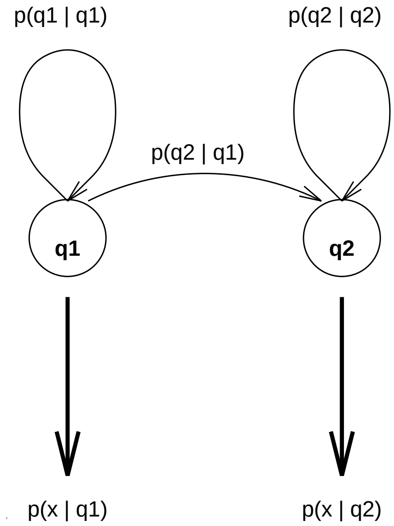

Figure 1: An HMM example

training criteria, to maximize the mutual information between the observation sequence to the correct model.

1.4 Preview

In the next section we review related work in non-stationary modeling that attempt to overcome the i.i.d assumption of the HMMs. Later we describe a non-stationary model, i.e., the time-index model and point out the common and different between it and the existing models. Then we experiment with models that are discriminant in nature and attempt to overcome the training criterion limitation of HMMs as mentioned above. We conclude by describing the line of research that we intend to pursue in the future.

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 all 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.

A number of HMM alternatives model the sequence of frames emitted in a given speech unit as correlated and dependent on each other. The models differ in their assumptions about the nature of the correlation between the acoustic vectors in the sequence and in their basic modeling unit. For instance, some assume that only consecutive frames are correlated, e.g., autoregressive HMM [9], while others assume that all the frames in the sequence are dependent on each other, e.g., segment-based models. In general these models attemp to overcome the HMM assumption of independent and identically distributed observations. In the following section we survey segment-based approaches.

2.2 Segment-Based Models

An alternative approach to HMMs 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[5]:

1. Generate a fixed length M that is the trajectory of the sound in the feature vector space. This trajectory is internal to the model and only serves as "baseform" to create the sequence of observations. Specifically the segment is generated according to the distribution $P(y_1, y_2, \ldots, y_M | s_k)$, where s_k is a particular speech unit, and $Y = \{y_1, y_2, \ldots, y_M\}$ is the trajectory.

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.

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 [13] 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 [6] 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 the phones in Ostendorf and Roukos' stochastic segment model [13].

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

Another segment-based model is the Segmental Neural Net (SNN). The Segmental Neural Network (SNN) models all the frames of a phonetic segment simultaneously. It does so by warping the variable number of frames in a phonetic segment to a fixed length by using the discrete cosine transform (DCT). Thus, the net estimates the probability P(phone | warped-acoustic-segment). The utterance is segmented into phonetic segments by an HMM-based system. The overall architecture uses the N-best paradigm, and the SNN is used as a post-processor [17].

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 HMM's, 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 6.

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

In the following section we discuss non-stationary modeling by frame-based models.

2.3 Frame-Based Models

In frame-based models generation of each acoustic vector is a separate event, in contrast to segment-based models that model generation of a sequence of acoustic vectors as one event. Different models take different approaches about the relation between these events. In HMMs, for instance, the assumption is that these events are independent of each other. But in autoregressive HMM there is a dependency between the current acoustic vector and the previous acoustic vectors [9].

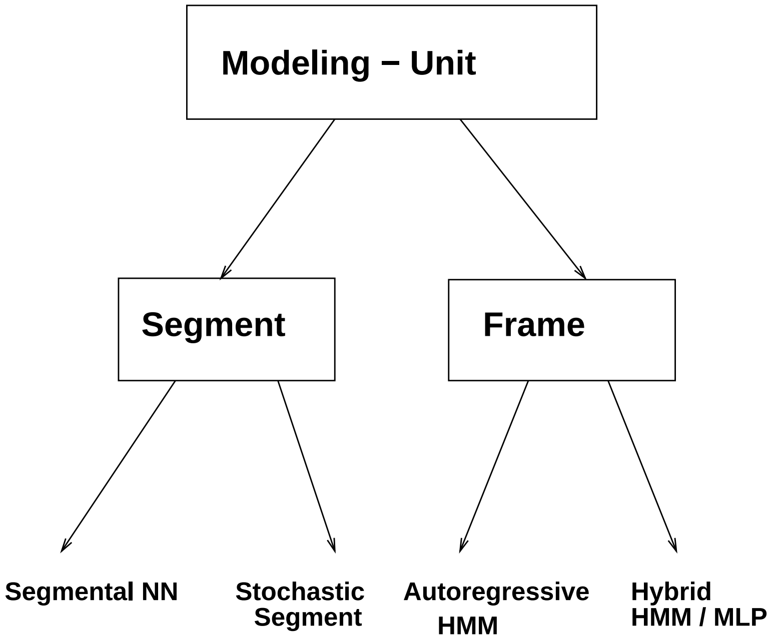

Figure 2: Division according to modeling unit

In the hybrid HMM/MLP framework the MLP estimates a posteriori probability of the phone given the acoustic data, i.e., $P(q|x)$, where q is speech sub-unit (e.g. phone) and x is the acoustic data. Then we divide the posterior probability by the relative frequencies of the phones in the training set to get the scaled likelihoods $\frac{P(x|q)}{Px}$ that serve as emission probabilities in the HMM. Besides the advantage of the MLP being discriminant between similar sounds, we present each acoustic vector to the net accompanied by several adjacent vectors from the past and the future. Typically we use temporal windows of nine frames, so we present four vectors into the past and four into the future. Thus, we capture some of the dependencies between neighboring acoustic vectors and relax the HMM's i.i.d. assumption.

In Figure 2 we show a division of the different models according to their basic unit of modeling.

3 THE TIME-INDEX MODEL

In the following sections we describe our study of models that attempt to overcome the HMM's limitations. In this section we describe the time-index model that attempts to overcome the HMM's i.i.d. assumption by conditioning the emission probability on the number of frames since entering the current state.

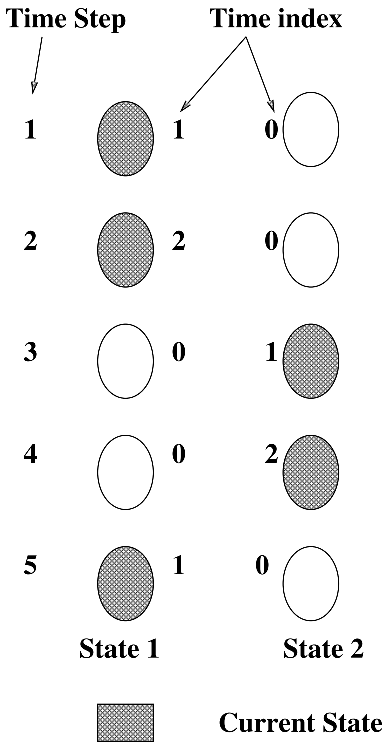

Figure 3: Time-index

3.1 Introduction

The main idea behind the time-index model is that all the trajectories of a phone in the acoustic vector space share a stochastic dependency on the time elapsed since the beginning of the phone. This dependency can be modeled by a parametric distribution as in Deng's model or by using an MLP in our model. We start by describing Deng's trended HMM, followed by our time-index model description.

3.2 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 3 (note that the figure does not 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:

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

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.

3.3 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 identically distributed; 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 q_j is:

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

We can see that the $q's$ in the HMM equation 4 are replaced by a a q and time index pair, as the state process is defined differently. [1]

3.4 An Example

Figure 4 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

[1] Formally the values of the state process are ordered pairs of the phone and the time-index

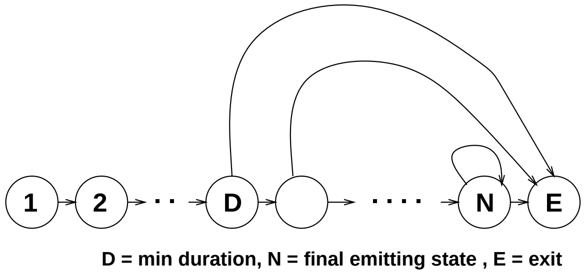

Figure 4: The topology of the time-index model

a phone or subphone unit type) is not constrained to be equal. Specifically, the emission probability of a state in the Markov chain is $P(x|(q_j, ti))$, 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 4, 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 [1].

4 The Time Index Model - Implementation and Experiments

4.1 An Implementation of the Time Index Model

In our model we define the emission probability of a state as $P(x|q_j, ti)$. 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:

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

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

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

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) as input, as well as any additional conditioning terms as input (for instance, an additional input representing time index ti in order to estimate $P(q_j|ti, x)$). The targets correspond to a discrete binary coding of the class identity that is to the left of the condition bar (e.g., q_j for estimating $P(q_j|ti, x)$,

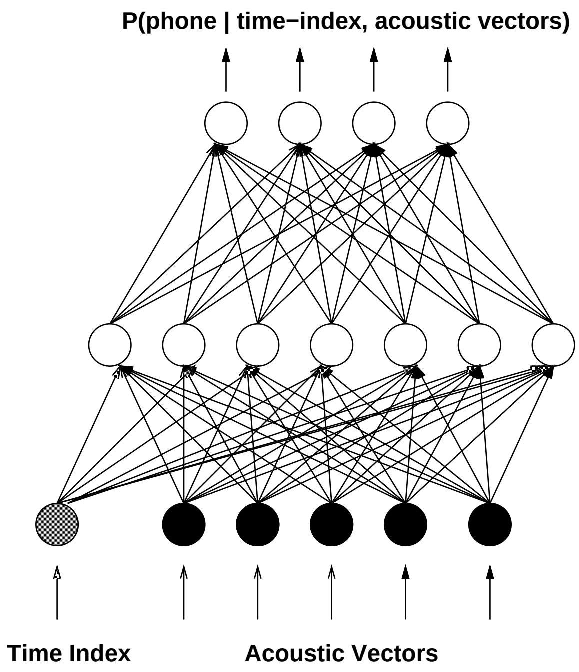

Figure 5: The time index net

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 $P(q_j|ti, x)$, and this can be done with the MLP shown in Figure 5.

$P(ti, phone_j)$ 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 6).

4.2 Experiments

We used the Resource Management (RM) speaker independent task [14] 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 5) 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) [8], 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 [1]. 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 [15]. 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.

5 Discriminant Models

In the following section we discuss models that attempt to overcome the lack of discrimination in HMMs. These models have properties that are somewhat complementary to the models discuss above, in the sense that they try to overcome a different limitation of the HMMs. Eventually we plan to have a merged model that can overcome both limitations.

5.1 Introduction

The optimal criterion for training acoustic models is the Maximum A Posteriori (MAP) criterion. According to it we should maximize $P(M|X)$, where M is the model associated with the acoustic information X. Implicitly from this criterion the following constraint has to be met:

$$\sum_i P(M_i|X) = 1 \tag{9}$$

Here lies the difference between the MAP criterion and the Maximum Likelihood Estimate (MLE) criterion. When we change the parameters of one model we have to modify the parameters of all other models according to the MAP approach, unlike the MLE approach. In this work, we investigate the model (referred to as Discriminant HMM in [1]) that is trained according to the MAP criteria. In [1] work it is also shown that the global MAP probability $P(M_i|X)$ can actually be expressed in terms of local acoustic probabilities like

$$p(q_{\ell_n}^n|X, q_{\ell_{n-1}}^{n-1}, q_{\ell_{n-2}}^{n-2}, \ldots, q_{\ell_1}^1) \tag{10}$$

in which $q_{\ell_n}^n$ represents the state at time n along a particular path γ_i of length N in M_i. Assuming that the conditional in equation 10 is restricted to the previous state (1st order Markov model) and $X_{n-c}^{n+c} = \{x_{n-c}, \ldots, x_n, \ldots, x_{n+c}\}$, a sub-sequence of X of length $2n + 1$ centered around x_n, $P(M_i|X)$ can then be calculated as

$$P(M_i|X) = \sum_{\gamma \in \Gamma_i} \prod_{n=1}^{N} p(q_{l_n}^n|X_{n-c}^{n+c}, q_{l_{n-1}}^{n-1}) p(M_i|\gamma) \tag{11}$$

in which Γ_i is the set of all possible paths in M_i, and $P(M_i|\gamma)$ is the contribution of the language model. In the case of Viterbi approximation, the sum in (11) is replaced by the max operator. Local probabilities $p(q_k^n|X_{n-c}^{n+c}, q_\ell^{n-1})$ can easily be estimated with an MLP.

In the following section we describe our specific implementation of the model and our pilot experiments.

5.2 Implementation

We estimate the local discriminant probabilities (Equation 10) by an MLP as represented in Figure 6. In this case, the previous state is coded as extra binary inputs, one for each possible previous state. For every hypothesized previous state we set the corresponding input to one and the rest to zero. In training we know the correct previous state. either by having a hand-segmented database as TIMIT, or by running an automatic alignment (Forced-Viterbi) on the training data. During recognition, the MLP outputs will have to be hypothesized for every possible previous state (possibly constrained by a particular HMM topology or a language model).

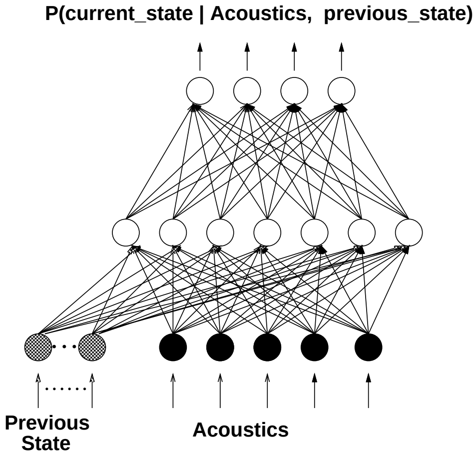

Figure 6: **Estimating Local Discriminant Probabilities**

5.3 Experiments

We chose the TIMIT corpus for our experiments because it is phonetically balanced, and in addition there are time-aligned phonetic transcriptions of all the sentences in the database. The experiments were done on a 200 sentence development set, that was selected from the official training set and was not used for the training. 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 phone models are a simple one state per phone model.

The net that estimates the local discriminant probabilities (as shown in Figure 6) had 1000 hidden units, 61 outputs (the size of phone set). There were 295 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) [8], and 61 binary inputs that represent the possible previous state. With the exception of these binary inputs, this net was the same as the hybrid HMM/MLP system as described in [1].

Our reference HMM/MLP system [1] (with optimized phone transition penalty) had 36.3% phone error on this task. When using MLP conditional transition probabilities (Equation10) on this task, without division by prior probabilities (as done with the standard HMM/MLP approach) and without any phone transition penalty, the error rate was 41.0%. Investigating the reasons for this performance degradation, we assume that it was mainly due to a disparity between the training input space of the MLP and the input space used in recognition. Specifically in training we present to the net only "correct" pairs of acoustic vectors and correct previous state, while in recognition we expect the net to generalize to all possible combinations of acoustic vectors and previous states, although some of them can be completely wrong like acoustics of a middle of a vowel and a previous state that is equal to a plosive. The net is not trained on anything close to these "impossible" pairs (but which could lead to the best MLP outputs during recognition). This could be solved by using a negative training procedure, i.e., teach the net that these pairs are improbable by forcing all the outputs of the net to zero, we discuss it in detail in section 6.

In order to test this assumption we have computed the frame level performance of the net on the development set for the following three cases:

1. Presenting the correct previous state. We got 90.15% frames correct.

2. Presenting all possible previous states and taking as the winner the output with the highest activity, i.e., for every frame we are taking the maximum of 61 by 61 probabilities (61 outputs for each possible previous state). We got 39.0% frames correct.

3. To check the assumption that the source of the problem lies in the interaction between the acoustic vector and the previous state, we have computed from the training set a biphone grammar and took the maximum over all the legal transitions according to this grammar. The performance remains as in the previous case, 39.0% frames correct.

So our current best hypothesis remains that the problem is the mismatch between the training and recognition conditions. We discuss possible ways to deal with it in the following section. Another potential source of improvement is to introduce a phone transition penalty as was done in our standard system. We experimented with this and with optimized phone transition penalty we get 40.4% phone error, still worst than our baseline system.

Another potential problem is missing transitions, i.e., implicitly the net is a transition detector because when it says that the current state is different from the previous it signals a transition. And transition detection between phonemes is a known hard problem, see [7]. Thus, we can completely get off from the correct path given our dependency on the previous state.

6 Discussion and Future Work

Our goal is to construct an acoustic model that overcomes the basic limitations of HMMs. To this end we have discussed two models: the time-index model and the discriminant HMM, each deals with a different limitation. The time-index model deals with modeling non-stationarity within a speech segment and the discriminant HMM deals with the lack of discrimination. Before we can think of a merged model that combines all the "good" properties of both models, we have to solve the theoretical and practical problems that we found in our study.

The two models that we have discussed above (the time-index model and the discriminant HMM) can be viewed as estimating two approximations of the following probability: P(current state | past, acoustic vectors). One difference between the two models is with their approximations of the past. The discriminant HMM assumes the previous state versus the time interval from the previous transition in the time-index model. Both models should be able to detect transitions between phones in order to perform well. In the time index model this happens when the time index is equal to one, and in the discriminant HMM when the current state is different from the previous one. However transitions in continuous speech are not sharp.

As noted earlier, another general problem that has surfaced from the pilot experiments is the disparity between the distribution of the input space during training and during recognition. This problem seems to effect all our efforts to estimate conditional probabilities with MLP. In all our previous experiments to add knowledge as conditional variables such as gender, context [2][10][11] we did not explicitly answer this problem. Although in the context-dependent neural network we implicitly solve it through a smoothing procedure that essentially involve initializing the weights of the context-dependent net with the weights of the context-independent net. Our experiments with a similar solution for the discriminant net did not produce the desirable results.

A possible solution to the problems mentioned above is an Estimate - Maximize algorithm. In the Estimate phase we compute the posterior probabilities for every transition. Thus, we will get smoother posterior distribution for the transitions. Furthermore we can learn these probabilities in a fully connected model, i.e., every phone can follow any other phone, and by this we will potentially solve the disparity problem. In the Maximize step we will train the MLP to estimate these probabilities. The theoretical framework for this approach has been started, and we plan to experiment with this idea.

In conclusion the line of research that we suggest is to study acoustic models that potentially overcome the HMM limitations. Then by analyzing their theoretical and practical problems, we hope to come with the appropriate solutions and eventually make progress towards constructing a better model than HMMs.

Acknowledgment

I would like to thank Nelson Morgan and Hervè Bourlard for their guidance and close collaboration in this research. Thanks to Gary Tajchman for his help with this report.

References

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

[2] Hervé Bourlard, Nelson Morgan, Chuck Wooters, and Steve Renals. Cdnn: A context dependent neural network for continuous speech recognition. In *Proceedings IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, San Francisco, California, March 1992. IEEE.

[3] P. F. Brown. *The Acoustic-Modelling Problem in Automatic Speech Recognition*. PhD thesis, CMU, Pittsburgh, PA, May 1987.

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

[6] 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.

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

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

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

[10] Y. Konig and N. Morgan. Supervised and unsupervised clustering of the speaker space for connectionist speech recognition. In *ICASSP*, Minneapolis, Minnesota, April 1993. IEEE.

[11] Y. Konig, N. Morgan, C. Wooters, V. Abrash, M. Cohen, and H. Franco. Modeling consistency in a speaker independent continuous speech recognition system. In J.S. Hanson and J.D. Cowan znd C.L. Giles, editors, *Advances in Neural Information Processing Systems 5*, volume 5, San Mateo, 1993. Morgan Kaufman.

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

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

[14] 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.

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

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

[17] G. Zavaliagkos, Y.Zhao, R. Schwartz, and J. Makhoul. A hybrid segmental neural net/hidden markov model system for continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2(1):151–160, January 1994.