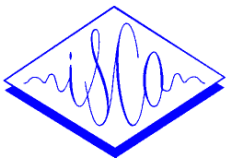

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

ON INCORPORATING PHONEMIC CONSTRAINTS IN HIDDEN MARKOV MODELS FOR SPEECH RECOGNITION

R. N. V. Sitaram and Thippur Sreenivas
Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore-560012, India

ABSTRACT

Phonemes have characteristic properties such as unique temporal structure, context sensitive behaviour and specific duration etc. Phoneme models should incorporate such constraints to provide better classification accuracy. In this paper these phonemic properties are incorporated into a HMM based phoneme recognizer with the addition of several degrees of freedom to the HMM state. The resulting models have shown improved performance on the TIMIT database.

1. INTRODUCTION

This paper addresses the problem of speaker independent phoneme recognition in continuous speech. Phonemes are generally identified based on their acoustic (spectral) evidence. The acoustic manifestation of phonemes have typical properties such as unique temporal structure, duration, context sensitive behaviour etc. All these properties have to be used together in order to achieve high performance phoneme recognition. Many HMM based phoneme recognizers [1,2] try to incorporate as many of these properties as possible into the system.

The temporal structure of some phonemes such as stops and diphthongs is not stationary and they have distinct acoustic sub-segments. In a phonemic HMM, these sub-segments have to be characterized properly to realize phoneme uniqueness. For this reason, Lee, et al, [1] use a HMM model with three observation distributions, where each sub-segment is characterized by a separate distribution. Each phoneme, therefore, requires multiple $pdfs$ to model its temporal structure.

One more well known fact about phonemes is that their manifestation varies with respect to the context. To counter this type of variability the knowledge of different contextual effects on each phoneme has to be used. For this Lee, et al, [1] treat each phoneme with a specific right phonemic context as a unit for modelling and recognition. They use a different HMM for each right phonemic context of a phoneme, and make a large network of all such diphone HMMs. With this network, a given test speech is phonemically decoded using the viterbi algorithm.

Apart from the spectral characteristic, one of the important properties of a phoneme is its duration. It has been shown by several researchers that if the duration information of phonemes is incorporated into the recognizer, it contributes considerably to the recognition. Levinson, et al, [2] use a 43 state continuously variable duration HMM (CVDHMM) for phoneme recognition, where each state models a phoneme and its duration explicitly. Here, recognizing a test sentence involves finding the maximum likelihood (ML) state sequence of the 43 state CVDHMM.

The system reported in [1] could successfully capture the temporal structure and context sensitivity of phonemes but it did not incorporate durational information of phonemes. The durational knowledge can be incorporated in their system as a post processor on the optimum selected path, but it does not get used in the forward search to find the best path. In [2], even though the duration of phonemes is modelled well, the contextual effects are not captured and the temporal structure of non-stationary phonemes is not well modelled because of a single observation distribution per state used in the model.

In this paper, all the phonemic properties i.e., temporal structure, duration and context sensitivity are incorporated into the HMM based phoneme recognizer by adding several degrees of freedom to the HMM state. The proposed phoneme recognizer consists of a large ergodic HMM (or improvised HMM), which has a number of states exactly equal to the number of phonemes, i.e., each state models a phoneme. Recognition of a test speech involves finding the ML state sequence through the model. In comparison to the existing methods, the new model has more powerful HMM states. Instead of using many states for a phoneme, a single state is used. This allows for better modelling of durational constraints of each phoneme. Also, higher linguistic constraints such as phoneme trigram probabilities can be better incorporated in the new model.

2. PHONEME DURATION MODELLING

In the new HMM, a single state models a phoneme and hence phoneme duration modelling is equivalent to HMM state duration modelling. In standard HMM the (implicit) probability of duration in a state decreases exponentially with time, which is not appropriate for phonemes. There are many solutions for better state duration modelling in HMMs [3,4,5]; of these, the inhomogeneous HMM (I-HMM) proposed by Ramesh and Wilpon [3] is more flexible for improvisation and requires less computation. In I-HMM, the state occupancy probability is better modelled using duration dependent state transition probabilities. Here, the transition probability is made dependent on the duration d spent in the state i; i.e., it becomes $a_{ij}(d)$, $1 \leq i,j \leq N$ and $1 \leq d \leq D$. Thus, each state will have a sequence of transition probability distributions, one for each time instant until

10.21437/Eurospeech.1995-177

a maximum duration limit of D. For duration $d \geq D$, the transition probabilities are truncated to $a_{ij}(d) = a_{ij}(D)$. Thus, an I-HMM can model state duration with arbitrary probability distribution in the interval $1 \leq d < D$, and for $d \geq D$ it is geometrically distributed.

$$\begin{aligned} P_i(d) &= [\Pi_{\ell=1}^{d-1} a_{ii}(\ell)](1 - a_{ii}(d)) \qquad 1 \leq d \leq D, \\ &= [\Pi_{\ell=1}^{D} a_{ii}(\ell)](a_{ii}(D))^{d-D-1}(1 - a_{ii}(D)), d > D \end{aligned}$$

Most of the phoneme durations can be modelled by either Poisson or Gaussian distributions [5]. In both of these distributions the shape of *pdf* becomes exponential after the peak. Therefore fixing D slightly greater than the mean duration of phonemes, an I-HMM can model the phoneme durations effectively. I-HMM is selected for the phoneme recognizer so as to characterize phoneme durations and for further improvement by incorporating rest of the properties of phonemes.

3. MODELLING TEMPORAL STRUCTURE

The acoustic manifestation of some phonemes (e.g., vowels) is quasi-stationary over most of the segment and hence it can be modelled using a single observation probability distribution. But phonemes such as stop consonants do not have any stationary region and they have different spectral properties in their sub-segments. Such non-stationary phonemes, however, have a specific sequence of acoustically distinct sub-segments. Their temporal structure being unique, if properly modelled will help in good discrimination. Modelling a non-stationary phoneme with a single state is to capture all its spectral variations using a single observation distribution. This single observation distribution can cause ambiguities because of general broadening of the *pdf* (increased variance) thus representing a phoneme coarsely, loosing its temporal uniqueness. The solution to this is to use multiple observation distributions through multiple states for each phoneme, where each state characterizes a distinct acoustic sub-segment. However, this contradicts the requirement of a single state per phoneme. This can be solved by adding another degree of freedom to the HMM state. The observation symbol distribution of the HMM state, $b_j(k)$, is made dependent on the duration d spent in the state j; i.e., it becomes $b_j(k, d)$, a different observation distribution after staying in state j for d clock intervals. Thus, each state will have a sequence of observation distributions, one for each clock interval until a maximum limit of D'. Beyond duration D', the distributions are truncated to $b_j(k, D')$. The value of D' has to be chosen so as to cover the whole length of the phoneme. This new improvisation, called Trend-HMM (T-HMM), is very similar to the duration dependent transition probabilities in I-HMM, where the upper limit on duration of transition probabilities is denoted by D. Incorporating both inhomogenity and trend into a HMM results in Trend-Inhomogeneous-HMM (TI-HMM). TI-HMM can model temporal structure and durations of the phonemes at the same time. To use TI-HMM for phoneme recognition, the viterbi algorithm of I-HMM [3] is modified, for decoding the ML state sequence as given below.

3.1 Viterbi Algorithm for TI-HMM

Let $\delta_t(i, d)$ be the joint probability of the best state sequence $q_1, q_2, ...q_t$ with $q_t = i$ and duration spent in state i is d for the partial observation sequence $O^t = O_1, O_2, ...O_t$; thus

$$\delta_t(i, d) = \max_{\mathbf{q}} P[q_1, ..., q_t = i, d_t(i) = d, O^t|\lambda],$$

where $d_t(i)$ is duration spent in state i at time t, and $\lambda = (A, B, \Pi, N, M)$, where Π, M, N, have the same meaning as in a HMM [4]. A denotes the set of duration dependent transition probability matrices $a_{ij}(d)$, $1 \leq i, j \leq N; 1 \leq d \leq D$, and B denotes the set of duration dependent observation probability matrices $b_j(k, d)$, $1 \leq j \leq N; 1 \leq k \leq M; 1 \leq d \leq D'$.

Initialization:- for $1 \leq i \leq N$

$$\begin{aligned} \delta_1(i, 1) &= \pi_i b_i(O_1, 1), \\ \delta_1(i, d) &= 0, \quad d > 1 \end{aligned}$$

Recursion:- for j=1,2,...,N and d=1,2,...,t; t=1,2,...,T-1.

$$\delta_{t+1}(j, 1) = \max_{1 \leq \tau \leq t} \max_{1 \leq i \leq N,\ i \neq j} [\delta_t(i, \tau) a_{ij}(\tau)] b_j(O_{t+1}, 1)$$

$$\begin{aligned} \delta_{t+1}(j, d+1) &= \delta_t(j, d) a_{jj}(d) b_j(O_{t+1}, d), \\ \Delta_t(j, i) &= arg \max_{1 \leq \tau \leq t-1} [\delta_{t-1}(i, \tau) a_{ij}(\tau)], \\ &\qquad for\ 1 \leq i \leq N,\ i \neq j \end{aligned}$$

$$\Psi_t(j) = arg \max_{1 \leq i \leq N\ i \neq j} [\delta_{t-1}(i, \Delta_t(j, i)) a_{ij}(\Delta_t(j, i))]$$

$\Delta_t(j, i)$ gives the best duration in state i before transiting to state j at time t along the best path; $\Psi_t(j)$ gives the best previous state before reaching state j at time t along the best path.

The probability of the best sequence of states is given by

$$P^* = \max_{1 \leq i \leq N} [\delta_T(i, \eta(i))]$$

where $\eta(i) = arg \max_{1 \leq \tau \leq T-1} [\delta_T(i, \tau)]$, $1 \leq i \leq N$, which gives the best duration spent in state i at time T along the best path. The final state of the best state sequence is given by:

$$q_T^* = \max_{1 \leq i \leq N} [\delta_T(i, \eta(i))]$$

The best state sequence is backtracked as follows:

Initialization:-

$$x = \eta(q_t^*),\ t = T,\ y = x;$$

The following equations are recursed until q_1 is obtained

$$\begin{aligned} q_{t-x+k}^* &= q_t^*, \quad k = 1, 2, \ldots, y-1 \\ q_{t-x}^* &= \psi_{t-x+1}(q_t^*) \\ y &= \Delta_{t-x+1}(q_t^*, q_{t-x}^*) \\ t &\leftarrow t - x \\ x &\leftarrow y \end{aligned}$$

the array $\mathbf{q}^*$ gives the best state sequence.

4. INCORPORATING PHONEME CONTEXT

It is well known that the consonants induce specific formant transitions in the adjacent vowels, which is an important cue in their recognition. The manifestation of some phonemes like vowels vary drastically because of neighbouring phonemes. These reasons have motivated many researchers [1] to use the context dependent HMMs. In the present model, since only one state is assigned for each phoneme, the state has to capture these contextual effects also. For this, we need to some how maintain unique observation distributions for a phoneme in each of its contexts. Left phonemic contextual effects are incorporated into the HMM based phoneme recognizer by making the observation probability distribution of a state j, dependent on the previous state i, from which transition took place into the present state, i.e., $b_j(k)$ of standard HMM becomes $b_{ij}(k)$. Once the transition takes place from state i to state j, the observation distribution used in state j will be $b_{ij}(k)$. This distribution $b_{ij}(k)$ is used as long as the state j is occupied, with the condition that the previous state was i. Every state will have $N-1$ context dependent observation distributions ($b_{jj}(k)$ distribution does not exist), and one non-speech left context distribution $b_{0j}(k)$. The distribution $b_{0j}(k)$ indicates the probability of observation symbols occuring whenever the phoneme represented by state j occurs as the beginning phoneme in the given speech data; i.e., the left context is non-speech. Every state thus captures contextual effects by catering one observation distribution for each left phoneme context. For phoneme recognition using this new model, called as Context-HMM (C-HMM), a sub-optimal hidden state sequence decoding algorithm is presented below.

4.1 State Sequence Decoding for C-HMM

Let $\delta_t(i)$ be the joint probability of the best state sequence $q_1, q_2, ...q_t$ with $q_t = i$ and partial observation sequence $O^t = O_1, O_2, ...O_t$. Thus,

$$\delta_t(i) = \max_{\mathbf{q}} P(q_1, ...q_t = i, O^t|\lambda)$$

where $\lambda = (A, B, \Pi, N, M)$. A, Π, M and N have the same standard meaning as in HMM [4]. B denotes the set of context dependent observation probability distributions, $b_{0j}(k), b_{ij}(k)$ for $1 \leq i, j \leq N$ and $1 \leq k \leq M$.

Initialization:- for $1 \leq i \leq N$

$$\begin{aligned} \delta_1(i) &= \pi_i b_{0i}(O_1), \\ \Psi_1(i) &= 0, \\ \ell_1(i) &= 0; \end{aligned}$$

$\Psi_t(i)$ is the back pointer array and $\ell_t(i)$ keeps track of the previous non-self state at time t such that the best path ends in state i at time t .

Recursion:- for $t = 1, 2, ...T-1; j = 1, 2, ...N$.

$$\delta_{t+1}(j) = \max_{1 \leq i \leq N} [\delta_t(i) a_{ij} b_{ij}(O_t), \delta_t(i) a_{ij} b_{\ell_t(i),j}(O_t)]$$
$$\Psi_{t+1}(j) = arg \max_{1 \leq i \leq N} [\delta_t(i) a_{ij} b_{ij}(O_t), \delta_t(i) a_{ij} b_{\ell_t(i),j}(O_t)]$$

In the above two equations, when $i = j$, i.e., a self transition occurs, the left context is a non-self state (phoneme) from where the transition took place to the present state in the recent past. The left context observation distribution to be used is given by $b_{\ell_t(j),j}(k)$.

$$\begin{aligned} If\ \Psi_{t+1}(j) = j\ then\ \ell_{t+1}(j) &= \ell_t(j) \\ else\ \ell_{t+1}(j) &= \Psi_{t+1}(j) \end{aligned}$$

Termination:- The probability of the best state sequence is

$$P^* = \max_{1 \leq i \leq N} [\delta_T(i)]$$

The final best state:

$$q_T^* = arg \max_{1 \leq i \leq N} [\delta_T(i)]$$

Backtracking:- for $t = T-1, T-2, \ldots, 1$

$$q_t^* = \Psi_{t+1}(q_{t+1}^*)$$

the array $\mathbf{q}^*$ gives the best state sequence.

5. EXPERIMENTS

To evaluate the improved models discussed in the previous sections for phoneme recognition, several experiments were conducted using the TIMIT database. To achieve speaker independent phoneme recognition the training set used consisted of 3696 sentences taken from all the 8 dialects of TIMIT training set. The two SA sentences provided for each speaker were not included in these experiments. A set of 36 sentences were chosen from the coretest set of TIMIT database for testing purpose. This test set consisted of speakers not spoken in the training set and it spanned over all the 8 dialects.

The sampling frequency of the signal is 16 KHz. The speech signal is analysed in frames of 16 ms with an overlap of 8 ms between frames. 18 LPC derived cepstral coefficients are computed every frame, after windowing with a Hamming window and pre-emphasizing with a factor of 0.95. The cepstral feature vectors are quantized using a 256 size codebook, which is designed using the LBG algorithm.

5.1 Design of Models for Phoneme Recognition

All the model parameters discussed before are determined by using the statistical information of the corresponding events in the speech database. The number of states N in the models of all the experiments is fixed as 60, i.e., one state per phoneme. The 60 phonemes used here are taken from 61 phoneme set of TIMIT, where phonemes */ng/* and */eng/* are treated as one phoneme.

To provide a benchmark for comparison with the new models the first experiment performed was phoneme recognition using a standard ergodic HMM. The parameters of the HMM $\lambda = (A, B, \Pi)$ are calculated as follows: As the speech database is phonemically labelled, the labelling information is first mapped onto the codeword sequences of the entire speech database. With this it is straight forward to construct histograms of codewords occuring in each phoneme and by normalizing them the observation probability distributions of all the states of HMM are found. By counting the number of times each phoneme occurs in the beginning of all sentences in the training set, the initial state occupancy probability distribution is determined. Similarly, by counting the number of times a phoneme (frame) transits

to another phoneme including itself, the transition probability distribution of the state characterizing that phoneme is found, and this is repeated for all states.

The design of phoneme recognizers using I-HMM, T-HMM, TI-HMM and C-HMM are similar to the design given above. In I-HMM all parameters are determined similar to HMM, except $a_{ij}(d)$ are determined by counting the occurrences of phoneme transitions from phoneme represented by state i to phoneme represented by state j, with the condition that the transition took place when the first phoneme duration is d frames. For the last duration D, all the phoneme transitions, where the first phoneme duration is $\geq D$ are included in determining the $a_{ij}(D)$. The value $D = 25$ is found to be optimum.

Determining the T-HMM and TI-HMM parameters are similar to that of HMM and I-HMM parameters respectively, except for the duration dependent observation probabilities. The $b_j(k, d)$ are determined by counting the codewords at different time instants in all the occurences of the phonemes. Again, $b_j(k, D')$ represents the observation distribution for all durations $\geq D'$ of the phoneme represented by state j. The values $D = 25$ and $D' = 10$ are found to be optimum in trend models.

C-HMM has only one parameter set different from standard HMM, i.e., $b_{ij}(k)$. This left phoneme context dependent observation distributions are determined by counting the occurrences of all the codewords in the phoneme occurence when it has occured with the specific left context. Thus, for all phonemes with all possible left contexts the codeword histograms are constructed and are normalized to get context dependent observation distributions.

The phoneme recognition on the test set is conducted using all the above models by the respective Viterbi algorithms. The results are given in Table-1. The scores given in this table are found by using the scoring package distributed by NIST, which matches the recognized phoneme string with the original string using dynamic programming and gives the number of phonemes correctly recognized, inserted, deleted and substituted. While reporting results some phoneme confusions are not serious and the reduced phoneme set has 40 phoneme classes which included 39 classes as in [1] plus the glottal stop /q/, which is treated as a separate class.

Table-1: Phoneme Recognition results.

Model used	% correct (40 classes)	Insertions
HMM	43.97 %	7.56 %
I-HMM	45.44 %	6.38 %
T-HMM	44.27 %	6.31 %
TI-HMM	46.32 %	5.13 %
C-HMM	43.02 %	20.55 %

From Table-1 it is clear that there is improvement in percentage of phonemes correctly recognized by the addition of each degree of freedom to the HMM. But the C-HMM has shown slight decrement in performance and the insertions have increased. We attribute this observation for the want of enough training data. Even the performance of rest of the models also can be improved by increasing training data. Various smoothing techniques have been tried to improve the models. Firstly the observation distributions have been restricted (floored) to a minimum value $\epsilon = 0.001$, to avoid zero probabilities. The duration dependent parameters of I-HMM, T-HMM and TI-HMM are averaged in time as follows:

$\bar{b}_j(k, d) = (b_j(k, d) + b_j(k, d+1) + b_j(k, d+2))/3$

Similar smoothing is done for $a_{ij}(d)$ also. The context dependent observation distributions are interpolated with context independent distributions, with varying proportions depending on the number of codewords occuring in each context. The results using these smoothing techniques are given in Table-2.

Table-2: Phoneme Recognition results after smoothing.

Model used after flooring	% correct (40 classes)	Insertions
HMM	44.86 %	6.9 %
I-HMM	46.32 %	7.12%
T-HMM	45.74 %	7.26 %
TI-HMM	47.43 %	6.9 %
C-HMM	45.96 %	10.50 %

The results show a good improvement and the C-HMM has performed better than standard HMM. There is no smoothing done on standard HMM parameters except for flooring. Lee, et al, [1] have reported an improvement of 9 % in phoneme recognition using context dependent models. We hope C-HMM can perform much better with newer smoothing techniques and addition of trend and inhomogenity.

6. CONCLUSIONS

In this paper we attempted to incorporate three properties of phonemes into HMM based recognizer by making its state more powerful. By using a bigger training set and newer smoothing techniques we feel the proposed models, i.e., TI-HMM and C-HMM could perform better. Presently Context-Inhomogeneous-HMM, Context-Trend-Inhomogeneous-HMM, incorporation of trigram probabilities and use of sex dependent models etc, are being investigated.

References

[1] K. F. Lee and H. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," *IEEE Trans. ASSP*, Vol.37, No.11, pp.1641–1648, Nov. 1989.

[2] A. Ljolje and S.E. Levinson, , "Development of an Acoustic-Phonetic Hidden Markov Model for Continuous Speech Recognition," *IEEE Trans. SP*, Vol. 39, No.1, pp.29–39, Jan. 1991.

[3] P. Ramesh and J.G. Wilpon, "Modelling State Durations in Hidden Markov Models for Automatic Speech Recognition," *Proc. ICASSP*, pp.I-381–I-384, 1992.

[4] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, Vol.77, No.2, pp.257–285, Feb. 1989.

[5] S. E. Levinson, "Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition," *Computer, Speech and Language*, Vol.1, No.1, pp.29–45, Mar, 1986.