PHONEME RECOGNITION IN CONTINUOUS SPEECH USING LARGE INHOMOGENEOUS HIDDEN MARKOV MODELS

R. N. V. Sitaram and T. V. Sreenivas

Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore-560012, India
Email: sitaram@ece.iisc.ernet.in; tvsree@ece.iisc.ernet.in

ABSTRACT

In this paper we present a novel scheme for phoneme recognition in continuous speech using inhomogeneous hidden Markov models (IHMMs). IHMMs can capture the temporal structure of phonemes and inter-phonemic temporal relationships effectively, with their duration dependent state transition probabilities. A two stage IHMM is proposed to capture the variabilities in speech effectively for phoneme recognition. The first stage models the acoustic and durational variabilities of all distinct sub-phonemic segments and the second stage models the acoustic and durational variability of the whole phoneme. In an experimental evaluation of the new scheme for recognizing a subset of alphabets comprising of the most confusing set of phonemes, spoken randomly and continuously, a phoneme recognition accuracy of 83% is observed.

1. INTRODUCTION

The recognition of phonemes accurately in continuous speech is yet a difficult problem. Acoustic variability due to coarticulation and high variability inherent to some phonemes are some of the important causes of difficulty in phoneme recognition. Phonemes such as stop consonants and semi-vowels are the commonly misclassified ones because of their short duration and non-stationarity. Stop consonants have different spectral properties in their sub-phonemic segments i.e. in the closure and release. Vowels are highly affected by context in their onset and cessation regions compared to the steady regions. Therefore, to characterize phonemes accurately, the distinct sub-phonemic segments of each phoneme have to be accounted separately by modeling their acoustic and durational variability. Further, the temporal relationships between the sub-phonemic segments have also to be captured to distinguish some difficult phonemes. Additional knowledge, such as bigram transition probabilities between phonemes and durations of phonemes if incorporated into the recognizer, it will contribute to achieve better recognition performance.

The problem of phoneme recognition in continuous speech using HMMs has been addressed earlier by several researchers. Levinson et al.,[1] have used a single large continuously variable duration hidden Markov model (CVDHMM) with the number of states equal to number of phonemes in the vocabulary. Here, each state models both the acoustic variability and durational variability of each phoneme. The maximum likelihood (ML) state sequence of this CVDHMM gives the recognized phoneme sequence. The limitation with this approach is that a single CVDHMM state is characterizing a whole phoneme. Earlier we have discussed that many phonemes have distinct subsegments, with unique temporal relationships. Through experiments, we have found that highly non-stationary phonemes cannot be effectively modeled by a single state, since a single state can model only a stationary signal with minor fluctuations. Also, the distinct temporal structure of the acoustic spectra of the phoneme is not characterized by a single state. Levinson et al, achieved a reasonable performance because of the external knowledge incorporated into the model, such as phoneme duration and phoneme bigram transition probabilities. In another experiment of phoneme recognition, Lee et al.,[2] have reported better performance for phoneme recognition by using a large network of separately trained phonemic HMMs. As before, the maximum likelihood state sequence of this network gives the recognized phoneme sequence. In this model, each HMM representing a phoneme had a multi-state fixed structure with three observation probability distributions, each modeling the acoustic variability of a sub-phonemic segment. Even though the acoustic variability is thus characterized better the durations of the phonemes and the temporal structure of phonemes are not modeled accurately. One of the reasons for this is the inherent geometric distribution of the state duration probability of the HMM states. This system if implemented using IHMMs in place of HMMs and also incorporate a phoneme duration postprocessing, it should yield better results.

In this paper, we present a novel scheme of phoneme recognition in continuous speech using a two stage inhomogeneous hidden Markov modeling approach. This scheme models the acoustic and durational variability of sub-phonemic segments and also of the whole phonemes in an unique way, capturing the temporal structure of speech effectively. The proposed new scheme is discussed in section 2. In section 3, we present the experimental evaluation of the scheme. Conclusions are given in section 4.

2. TWO STAGE IHMM MODEL OF SPEECH

The most successful HMM based phoneme recognizer [2] comprises explicit inter-connection as a large network of a large number of HMMs specifically trained for the individual phonemes. This large network does not characterize

0-7803-1775-0/94 $3.00 © 1994 IEEE

the continuous speech structure as effectively as a single large HMM trained using continuous speech sentences. This is because in a single large HMM the Markovian property models the coarticulation effects and talker variability much better when trained using continuous speech. However, a larger HMM should be better initialized to be sensitive to all distinct acoustic properties. Further, all the parameters of this HMM are determined automatically through ML training without manual intervention as done in other methods [1,2]. The ML state sequence of such a large HMM reflects the consistency of the phoneme segments in continuous speech. However, these state sequences could be easily mapped to the corresponding phonemes using a second HMM. The second HMM has the advantage of focusing on these differences in the state sequences that differentiate phonemes. The second HMM could also incorporate phonemic constraints, such as duration and bigram probabilities, whereas the first stage focuses on sub-phonemic properties. It is found that IHMMs are effective in realizing these properties.

2.1. Inhomogeneous HMM

Inhomogeneous HMMs [3], are more general and hence better suited to model the state duration probability distribution of the speech signal, compared to the ordinary (homogeneous) HMMs and hidden semi-Markov models (HSMMs) [4]. In homogeneous HMMs, the inherent state occupancy probability follows a geometric distribution, which is not appropriate for speech events such as phonemes. In HSMMs, the self loop transitions of the states are removed and a separate duration probability distribution is attached to each state to model the duration of the speech event to which the state is mapped. Inhomogeneous HMMs model the state occupancy using duration dependent state transition probabilities. The transition probabilities are not represented as $a_{ij}'s$ but as $a_{ij}(d)'s$, $(1 \leq i,j \leq N; 1 \leq d \leq D)$. The transition probabilities depend on the duration d already spent in the originating state i. Therefore, each state has a matrix of transition probabilities to all other states, for different time instants, until a maximum duration limit D. For duration $d \geq D$, the transition probability $a_{ij}(d) = a_{ij}(D)$. The probability of state occupancy is accurately modeled for $d \leq D$, and it follows a geometric distribution beyond the duration D. Therefore the limit D on transition probabilities is fixed greater than the average duration of the speech segments modeled by the state. IHMMs not only model the state occupancy more effectively than the HSMMs using duration dependent self loop transition probabilities $a_{ii}(d)$, but by using $a_{ij}(d)$, the duration dependent transition probabilities to other states, they can capture the temporal constraints between distinct speech events modeled by different states, (i.e., duration dependent bigram transition probabilities between various speech segments), which is not achieved by any of the earlier models.

2.2. Two Stage Modeling

In the present scheme, we use two large IHMMs to model the speech in two stages. This scheme effectively models the acoustic variations and the temporal structures within the phonemes and their temporal relationships with other phonemes. Fig 1. gives an overview of the new scheme. The first stage ergodic IHMM has number of states equal to the total number of distinct acoustic events (i.e., total number of distinct sub-phonemic segments), this will be about two to three times the number of phonemes in the vocabulary. Each state is designed to model a distinct acoustic segment: stop consonants have two distinct segments, closure and release; vowels have three distinct regions because of contextual effects in its onset and cessation regions compared to the steady region. Each state in this IHMM models the acoustic variability and duration of one distinct sub-phonemic segment i.e., a distinct acoustic event. This IHMM is trained by first initializing each of the state observation probabilities with the statistics of the VQ symbols corresponding to a distinct acoustic segment. The segments are identified using the labeled training data. The initial state probabilities and transition probabilities are initialized to uniform values. The Baum-Welch reestimation procedure is used to train the IHMM using the VQ coded continuous speech sentences. Because this model is trained using continuous sentences it represents the coarticulation effects accurately, compared to the case of a network of HMMs representing continuous sentences. After training, the observation symbol probability distributions learnt by each state tend to be peaky (with less variance), modeling accurately one acoustic segment for which it is initialized. Because of the specific initialization, the duration dependent transition probabilities and the initial state occupancy probabilities are learnt correctly reflecting the temporal relationship of the sub-phonemic segments. The maximum likelihood state sequence of this IHMM for a speech VQ symbol sequence, reflects the segmental structure of speech without the acoustic redundancies. Since the initialization of states is done by distinct sub-phonemic segments of all phonemes, the segmental commonality between several phonemes leads to more than one state modeling the same acoustic event. This will be reflected in the state sequences of the IHMM, for a single acoustic event different state indices may appear in state sequences, and state sequences of different phonemes may have some common state indices. Because of coarticulation and talker variability, one segment of a phoneme may get drastically affected, resulting in a different state sequence for the same phoneme. Thus different phonemes may share a few states. These intra-class and inter-class variabilities are taken care of by modeling them using a second IHMM following the first IHMM. The second IHMM processes the state sequences of first IHMM to absorb the variabilities and differentiate the phoneme classes by also using additional knowledge such as overall duration of phonemes and bigram transition probabilities.

The second stage IHMM is also ergodic with the number of states exactly equal to the number of phonemes in the vocabulary. Each state here models a single phoneme. The observation symbols of the second stage IHMM are the state indices of the first IHMM. The observation probabilities and the duration dependent transition probabilities of the second IHMM are found from the statistics of the first IHMM state sequences for the training data. Since the speech training set is labeled as phonemes, the statistics of the state indices of the first IHMM corresponding to each phoneme (i.e. for each state in the second IHMM) could be directly found. The duration dependent transition probabilities in the second IHMM, are determined by counting the event occurrences of transitions from one phoneme to another, after the first phoneme occurred for that duration. These transition probabilities model the overall duration of phonemes and also temporal relationships between them. There is no further training done for this IHMM. The state sequence of the second IHMM gives the recognized phoneme string. Even though in the second stage only one state is modeling a phoneme, the observation sequence coming as input to this stage is not as non-stationary as the original speech signal, because of the segmental decoding done in the first stage. Thus, different known causes of variability in phoneme representation is modeled without resorting to unduly large HMM networks, which provides the additional advantage of better trainability leading to better performance.

The idea of two stage modeling has been explored by Pepper and Clements [5] using homogeneous HMMs. The formulation of their first stage HMM has a different motivation than ours. In their HMM each state observation density function corresponds to a cluster in vector quantization, i.e., corresponding to a single VQ codeword, and not to any particular acoustic event such as sub-phonemes in our case. Therefore, the output ML state sequence of this model would have much more variability compared to our case. Also, both the stages of modeling do not characterize the durational properties of speech accurately because of the use of homogeneous HMMs. This resulted in only a moderate performance improvement compared to a direct VQ based finite state automaton [5].

3. EXPERIMENT

Two experiments were conducted in this work. (i) Recognition of a confusable subset of phonemes sliced from speaker independent TIMIT database; (ii) Phoneme recognition in speaker dependent continuously spoken random alphabet sequences, from the set of "B", "D", "G" and "T" alphabets.

Because of the huge computational and memory requirement to implement the new scheme on the full set of phonemes from TIMIT, an experiment on a subset of phonemes is devised. Thirteen phonemes were selected in this experiment: they are */b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, /ng/, /th/, /dh/, /ch/* and */jh/*. These are the most confusable phoneme subset containing voiced and unvoiced stop consonants, affricates and nasals. These phonemes are sliced out from training and test set of dialect 2 of the TIMIT database. Feature vectors of 18 LPC derived cepstral coefficients are found on 16ms analysis window every 8ms, for all the utterances. These utterances are coded using 128 VQ codewords. A 29 state first stage IHMM is trained, after each state is initialized to a distinct acoustic event as explained earlier. The state sequences of the entire training set using this IHMM are found. Since the training data is of sliced phonemes, and not continuous speech, the knowledge of phoneme bigram transitions to be used in second stage modeling cannot be found. Also within the selected set of phonemes, no two phonemes occur in sequence in practice. Therefore a second stage IHMM with uniform transition probabilities to other states and duration dependent self loop probabilities reflecting the durations of the individual phoneme slices, was designed. An accuracy of 43% is observed on the test set data. This performance is low because the coarticulation effects are not modeled well in the limited implementation using the sliced phonemes rather than continuous speech. Also, the reduced model does not incorporate additional knowledge such as bigram probabilities. But this experiment demonstrates the feasibility of using a single model for all the classes in the first stage. To demonstrate the capability of the two stage modeling fully, a different experiment is performed using a continuous speech database. The second experiment, which tests the entire scheme explained in the paper without any restrictions, is described below.

3.1. Task and Database

In this experiment the task consists of recognizing phonemes in continuously spoken alphabets from the set B,D,G and T. These alphabets are selected because they comprise of the most confusing stop consonants, whose temporal structure is important for their discrimination. There are totally 6 phonemes occurring in the database including silence, they are /b/,/d/,/jh/,/t/,/i/ and /sil/. The database consisted of 180 random alphabet sequences from the above set, sampled at 10 KHz, spoken by a single speaker. The entire database is labeled phonetically. Out of 180 utterances, 120 were used for training the VQ codebook and the first stage IHMM; remaining utterances are used for testing.

3.2. Preprocessing

The speech database is analyzed in frames of 20 ms with an overlap of 10 ms between frames. From each frame of speech 10 LPC derived weighted cepstral coefficients [6], were found, for the entire database. A VQ codebook having 32 codewords was designed using the LBG algorithm [7]. All the speech sequences are later coded using these 32 codewords.

3.3. Two Stage Modeling

An 18 state first stage IHMM is trained using Baum-Welch reestimation formulae, with proper initialization, as explained earlier by allocating one state for each of the distinct sub-phonemic event. The ML state sequences for the entire training set is found using this IHMM. Then a 6 state second stage IHMM is designed using these state sequences and the labeling information provided with the training set as explained before. The same scheme is also implemented using HMMs in each of the stages for comparison with the use of IHMMs. The results of all the experiments are given in Table-1. It is evident that IHMM is capturing the temporal structure of speech more effectively than HMMs. Especially, the two stage IHMM-IHMM combination models the speech most effectively.

4. CONCLUSIONS

A novel scheme of phoneme recognition in continuous speech using a two stage inhomogeneous hidden Markov modeling approach is presented. The first stage models the

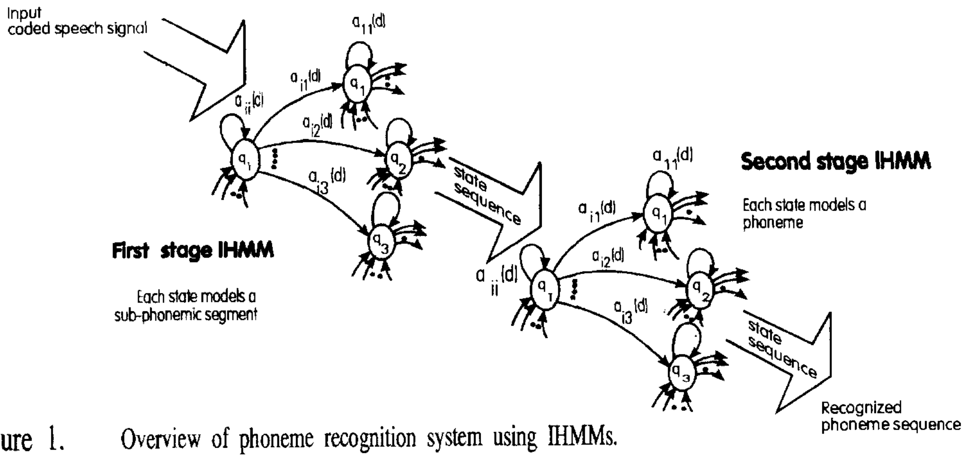

Figure 1. Overview of phoneme recognition system using IHMMs.

acoustic and durational variabilities of sub-phonemic segments and the second stage models the acoustic and durational variabilities of the whole phoneme. Thus, both the stages put together are modeling all the temporal relationships of the speech effectively. From the experimental results it is clear that the proposed scheme using IHMMs is performing better compared to using HMMs because the temporal structure of speech is captured more effectively, which is crucial for discriminating stop consonants.

Table 1. Results of experiments done with HMMs and IHMMs

Sl No.	I Stage	II Stage	Phoneme Recognition
1	HMM	HMM	72%
2	HMM	IHMM	78%
3	IHMM	HMM	80%
4	IHMM	IHMM	83%

REFERENCES

[1] S .E. Levinson, M. Y. Liberman, A. Ljolge and L. G. Miller: "speaker independent phonetic transcription of fluent speech for large vocabulary speech recognition", Proc. ICASSP89, pp 441-444, 1989.

[2] K. F. Lee and H. W. Hon: "Speaker independent phone recognition using HMMs", IEEE Trans. on ASSP, pp 1641-1648, November 1989.

[3] P. Ramesh and J. G. Wilpon: " Modeling state durations in Hidden markov models for automatic speech recognition", Proc. ICASSP92, pp I-381 to I-384, 1992.

[4] J. D. Ferguson: "Variable duration models for speech", Proc. Symposium on Application of Hidden Markov models to text and speech, pp 143-179, October 1980.

[5] Mark A. Clements and David J. Pepper: "Phoneme recognition using a large hidden Markov model", IEEE Trans. on Signal processing, pp 1590-1595, June 1992.

[6] B. H. Juang, L. R. Rabiner and J. G. Wilpon: "On the use of bandpass liftering in speech recognition", Proc. ICASSP86, pp 765-9, 1986.

[7] Y. Linde, A. Buzo and R. M. Gray: "An algorithm for vector quantizer design", IEEE Trans. on communication, vol COM-28, pp 84-95, January 1980.