Volume 2, Issue 11, November 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Language Identification Using Ergodic Hidden Markov Model

A.Nagesh[#1]
Associate Professor,
Dept .of CSE, MGIT,
Hyderabad-India

M.Sadanadam [#2]
Assistant Professor
Dept. of CSE, Kakatiya University
Warangal-India

Abstract— ***The objective of the language identification (LID) is to quickly and accurately identify the language being spoken. The language identification system require segmented and labelled speech corpus for identification. In this paper, a discussion has been carried out as the development of LID system which rely on features derived from speech signals, and do not need labelled speech data at any stage of the identification task. The existing language identification systems vector quantization (VQ) and Gaussian mixture model (GMM) perform primarily static classification, In that the system feature vectors are to be independent of each other and no use of feature vector sequences is made. The current language identification systems based on Hidden Markov model (HMM) use the sequential characteristics of speech signal. A language identification study using acoustic features for different HMM states and mixtures, has given a good performance. The language identification studies on OGI_MLTS database have been demonstrated.***

Keywords— ***Gaussian Mixture model (GMM), Hidden Markov model (HMM), Ergodic Hidden Markov model (EHMM), Language Identification (LID).***

I. Introduction

Among the various factors that define culture and communities, an important factor is language. Speech would thus be the most natural medium of interaction between humans and machines too, with advances in speech technology, scenarios where humans converse with computer-backed systems to accomplish complex tasks (such as airline ticket reservation, information retrieval), have become a reality. A desirable feature of such systems would be the ability to converse with any user in his or her native or most-preferred language. Language identification systems are classified into two main categories based on the way in which languages are modelled: namely Explicit LID systems and Implicit LID systems. Systems that require segmented and labelled speech corpus can be considered as Explicit LID systems and the systems that require only the digitized speech signal and the corresponding true identities of the language being spoken are termed as Implicit LID systems, in which languages models are derived only from the speech signal. Development of LID system based on Implicit LID systems has been discussed in the paper.

A significant technology shift from template-based to statistical modelling methods led to the highly successful model of Hidden Markov model. The most common approach to the problem of classifying speech patterns is by the use of Hidden Markov models. HMMs are predominantly used as acoustic models in most speech recognition systems. This is because speech signal is varied differently in both time and signal amplitudes and HMMs are designed to cope with this kind of signal. The hidden state property in HMMs can normalize the time-variation while the statistic parameters in each state can cover the signal amplitudes. The states of HMM correspond to acoustic units of a language and its state-transitions represent the variation in the spectral vectors distribution of a language.

HMM-based systems have shown better performance than the conventional template-based text-dependent language identification systems. The language-identification systems (VQ and GMM) perform primarily static classification, in that the feature vectors are assumed to be independent of each other and no use of feature vector sequences is made, in other system Hidden Markov models have used to model sequential characteristics of language. In this paper an automatic text independent language identification based on HMM for acoustic features has described. HMM-based language identification was first proposed by House and Neuburg [1]. They created a discrete-observation, Ergodic HMM (EHMM) that took sequences of speech symbols as input and produced a source language hypothesis as output. Training and test symbol sequences were derived from published phonetic transcriptions of text. Zissman, Nakagawa, and Savic [2][3][4] have applied HMMs to feature vectors derived automatically from the speech signal. In these systems, HMM training was performed on unlabeled training speech. In the reference [2] [3] it was found that HMM systems trained unsupervised manner did not perform as well as some of the static classifiers are found to perform. Nakagawa [5], however, eventually obtained better performance for his HMM approach than his static approaches. In related research, Li and Edwards [6] segmented incoming speech into six broad acoustic-phonetic classes. Finite-state models were used to model transition probabilities as a function of language.

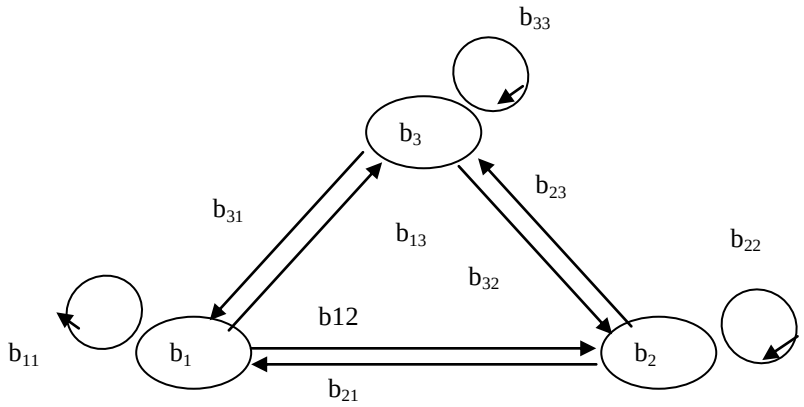

FIG.1. 3-state ergodic HMM

II. Feature Extraction from Speech Signal

Most of today's automatic speech recognition systems are based on Mel Frequency Cepstrum Coefficients (MFCCs), Because of its capability of capturing the phonetically important characteristics of speech. MFCC's are based on the known variation of the human ear's critical bandwidths with frequency. To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. This scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

The feature extraction component derives a set of language-specific vectors from the input signal. In this step, 12 dimensional Mel Frequency Cepstral Coefficients (MFCC) are extracted from the speech files of all languages under consideration.

III. Hidden Markov Model for Language Identification

By using left-right HMM can model temporal sequencing of the patterns only. Hence to capture the patterns of different types, ergodic HMM is used. An ergodic or fully connected HMM structure becomes more appropriate than L-R HMM structure because every state in the ergodic structure can be reached in a single step from every other state. Here an ergodic or fully connected HMM has been used to model the LID system.

The HMM is a doubly embedded stochastic process where the underlying stochastic process is not directly observable. HMMs have the capability of effectively modelling statistical variations in spectral features. HMM not only models the underlying speech patterns but also the temporal sequencing among the sounds. This temporal modeling is advantageous for text-dependent language identification system.

In the training phase, one HMM for each speaker is obtained (i.e., parameters of model are estimated) using training feature vectors. The parameters of HMM are state-transition probability distribution: It is represented by $A = [a_{ij}]$, where

$$a_{ij} = P(q_{t+1} = j \mid q_t = i) \quad 1 \le i, j \le N$$

defines the probability of transition from state i to j at time t. For a three state left-right model the state transition matrix is given as

$$A = \{ a_{ij} \} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}$$

The state transition matrix of three state ergodic model is given by

$$A = \{ a_{ij} \} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Observation symbol probability distribution: It is given by $B=[b_j(k)]$ in which

$$b_j(k) = P(O_t = V_k \mid q_t = j) \quad 1 \le k \le M$$

defines the symbol distribution in state j, j= 1,2,…N

The initial state distribution: It is given by $\pi = [\pi_i]$ where

$$\pi_i = P(q_1 = i) \quad 1 \le i \le N$$

Here, N is the total number of states, and qt is the state at time t, M is the number of distinct observation symbols per state, and 0_t is the observation symbol at time t. In testing phase, P(O|λ) for each model is calculated, where O=(o_1 0_2 o_3 …. o_T). Here the goal is to find out the probability for a given model to which the test utterance belongs to. The language whose model gives the highest score is declared as the identified language.
The model parameters can be collectively represented as λ= (A_i, B_i, π_i) for i = 1 … M. Each language in a language identification system can be represented by a HMM and is referred to by the language's respective models λ.
In the testing phase, p (O/λ) for each model is calculated, where O=($o_1o_2o_3…O_T$) is the sequence of the test feature vectors. The goal is to find the probability of given the model that the test utterance belongs to that particular model. The language model that gives the highest score is declared as the identified language.

IV. Baseline System

The probability of speech feature vectors generated from an EHMM is computed using the transition probabilities between states and the observation probabilities of feature vectors in a given state.

At the beginning of the training process, rough estimation of the λ values of EHMM should be computed. Viterbi algorithm is used to assign the initial values after the observation sequence is uniformly segmented with a segmental K-means algorithm. Using these initial values of the parameters, further improvements are achieved with a Baum-Welch or Expected-Maximization re-estimation procedure.

In the training, the tokens are obtained for all languages which are to be identified. For all languages, acoustic models are estimated using HMM modelling based on strings of derived tokens. HMM is estimated from all languages together.

An M-language LID is the task of classifying an input speech utterance as belonging to one of M languages L = (L_1, L_2, L_3,….,L_M). The EHMM system has M paths for a M language LID task. In testing, scores are evaluated using language-independent HMM models. The P(O|λ) for each model is calculated, where O=(o_1,o_2,….,o_T) is the sequence of the test feature vectors. For a given input utterance, EHMM yields M 'Viterbi likelihood' scores, one for each language L_i, which are obtained by a Viterbi decoding of the input utterance O by the EHMM λ_i of language L_i. The goal is to find the probability, given the model that the test utterance belongs to that particular model. The language model that gives the highest score is declared as the identified language.

V. Experimental Evaluations

A. Database Used for the Study

The Graduate Institute Multi-language Telephone Speech (OGI_MLTS) Corpus which is designed specifically for LID research is used for both training and testing. This corpus currently consists of spontaneous utterances in 11 languages. The utterances were produced by ~90 male and ~40 female, in each language over real telephone lines. In the present work, three languages namely English, German and Spanish are used. To maintain the homogeneity in training and testing across languages, for each language first 40 male speakers and first ten female speakers' with 45s duration utterances are used for training and next 50 male speakers' and next 30 female speakers' with 45s duration utterances are used for testing. All the training and test set speakers are different [11].

B. *Experimental Setup*

The system has been implemented in Matlab7. we have trained the model(HMM) using Gaussian components as 2, 4, 8 and 16 at each state for a training duration of 10min. Testing is performed using different test speech duration such as 1s, 2s and 3s.
The steps involved in the proposed algorithm for text independent language identification system is as follows:
Training Phase:

```
for each language Li from language list M do
  for each speech utterance Si of language Li
      Pre-process speech signal Si
      for each sample of Si from K samples do
          Extract MFCC features fk from Si
      end
  end
Initialize HMM model parameters λi=(A, B, πi)
Train λi for optimal solution using EM algorithm.
end.
```


Testing Phase:

for each language L_i from language list M do
for each speech signal S_i of language L_i
Pre-process speech signal S_i
for each sample of S_i from K samples do
Extract MFCC features f_k from S_i
end
end
for each model $\lambda_1, \lambda_1, \ldots \lambda_M$ do
using the Viterbi decoding process calculate $P(O|\lambda)$, where $P(O|\lambda)$ is the probability of the observation sequence $O(o_1 o_2 , \ldots o_T)$
end

Calculate 1-best result for a given testing speech utterance using

$$\arg\max_{1\le i\le M} p\left(O/\lambda_i\right)$$

end

C. *Results*

Determining the optimal number of mixture components needed to model a language adequately is an important task. There is no theoretical way to estimate the number of mixture components to model a language. The experiments are carried out for varying number of statues and Gaussian components to evaluate the performance of the language identification system.
In EHMM, the main parameters are the number of states 'N' and the number of Gaussian mixtures per state 'K'. In the experiments carried out by us, for each language, EHMM systems were designed for N=2,3 and 4 with K=2,4,8 and 16 Gaussian mixtures per states.

5.3.1 LID Performance for Varying Number of Gaussian Components
The performance of LID system for 2-state, 3-state and 4-state ergodic HMM for different Gaussian components such as 2,4,8 and 16 is shown in table1. The recognition performance increases significantly with increased in the number of states up-to 3 states. The minimum number states to achieve good identification performance seems to be 3 and thereafter the good identification performance is minimal.
D. LID Performance Varying Number of Gaussian Components

The experiments are carried out for 2-state, 3-state and 4-state EHMM for varying number of Gaussian components to evaluate the performance of the LID system. The percentage of identification of 3-state ergodic HMM for different Gaussian components such as 2,4,8 and 16 seems to be increasing uniformly. The minimum number of Gaussian components to achieve good identification performance seems to be 8 and thereafter the identification performance is minimal.

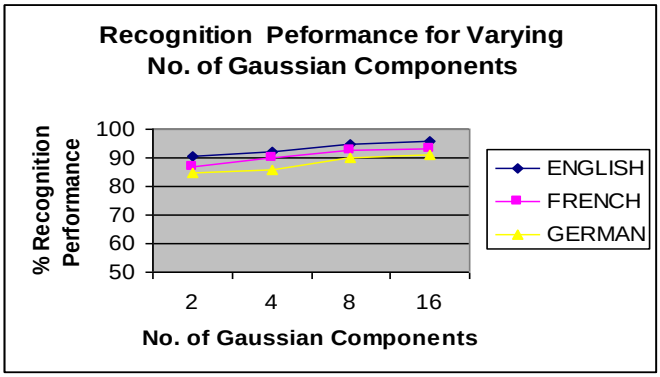

Fig.2: 2-state ergodic HMM performance for different number of Gaussian components.

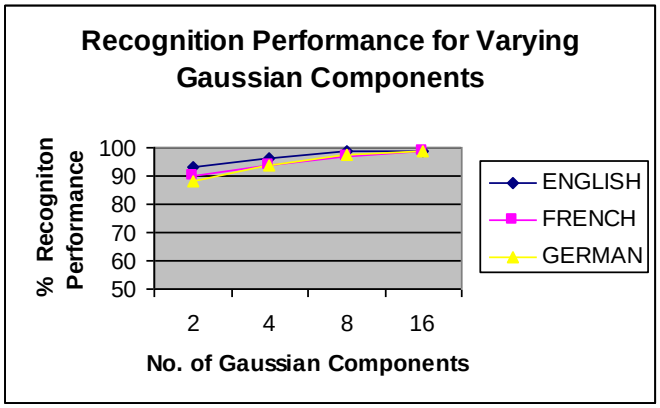

Fig.3: 3-state ergodic HMM performance for different number of Gaussian components.

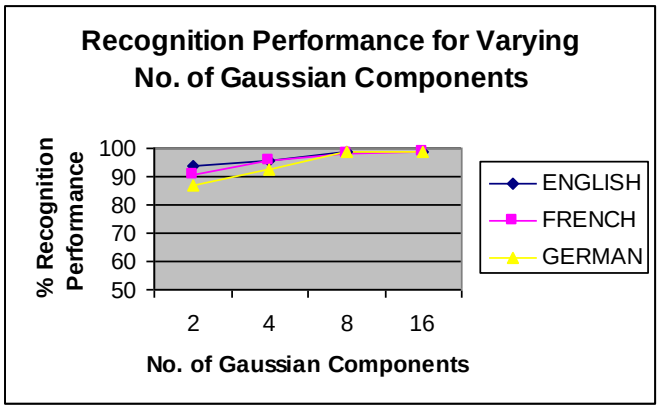

Fig.4: 4-state ergodic HMM performance for different number of Gaussian components.

VI. Conclusions

It is observed from the performance of above model that there is a considerable improvement in the Hidden Markov model when compare to VQ and GMM systems due to non consideration of feature vector sequences. Performance of the experiments shows that Ergodic Hidden Markov model can capture some acoustic-phonetic information from the speech signal for language identification.

References

[1] A. S. House and E. P. Neuberg. Toward automatic identification of the language of an utterance. *Journal of Acoustic Society of America*, 62(3):708–713, Sep 1977.

[2] M. A. Zissman. Automatic language identi.cation using Gaussian mixture and hidden Markov models. In *Proc. ICASSP*, pages 399–402, Apr 1993.

[3] S.Nakagawa, and Hashimoto. A method of continuous speech segmentation using HMM, In Proceedings IEEE Int. Conf. Acoustics, Speech, and Signal Processing. Pp.960-962,1988.

[4] M.Savic, E.Acosta, and S.K. Gupta, An automatic language identification system. In Proceedings IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 1991.

[5] S.Nakagawa, Y.Ueda, and T.Seino. Speaker-independent, text-independent language identification by HMM. In. Proceedings Int. Conf. On Language Processing, 1992.

[6] K.P Li and T.A.Edward , Statistical models for automatic language identification. In Proceedings IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 1980.

[7] K.P. Li, "Experimental improvements of Language ID system," *IEEE International Conference on Acoustic, Speech, and Signal Processing*, pp. 3515-3518, 1995.

[8] Y. K. Muthusamy, E. Barnard, and R. A. Cole. Reviewing automatic language identification. *IEEE Signal Processing Magazine*, 11(4):33–41, 1994.

[9] Y. Yan, E. Barnard, and R. A. Cole, Development of an approach to automatic language identification based on phone recognition. *Computer Speech and Language*, 10(1):37–54,1996.

[10] A. K. V. Sai Jayaram V. Rama Subramanian, T. V. Sreenivas, 2003, "Language identification using parallel sub-word recognition," *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Proceessing*, vol. I, pp.32-35.

[11] The Oregon Graduate Institute Multi-language Telepho- ne speech (OGI-MLTS) corpus.

