Text Independent Language Recognition using Dhmm

M. Sadanandam
Assistant professor of CSE
Kakatiya University, Warangal,
Andhra Pradesh, India.

V. Kamakshi Prasad, PhD.
Professor of CSE
JNTUH, Hyderabad
Andhra Pradesh, India.

V. Janaki, PhD.
Professor of CSE
Ramappa Engineering College,
Warangal, Andhra Pradesh, India.

ABSTRACT

Spoken Language Identification is a task of recognizing the language from an unknown utterance of speech. The ability of machines to distinguish between different languages becomes an important concern with the emerging trends in global communications which are multilingual nature. This paper describes a text independent language recognition system using a common code book and discrete hidden Markov models (DHMM) to achieve a very good LID recognition performance with less computation time comparing with that of a state of art phone based systems available in literature. This approach includes generation of a common codebook and training of DHMM, one for each language. The experiments are carried out on the database of Indian language consists of six languages namely Telugu, Tamil, Hindi, Marathi, Malayalam and Kannada.

Keywords

Language Identification (LID), Mel frequency cepstral coefficients (MFCC), vector quantization (VQ), discrete hidden Markov model (DHMM).

1. INTRODUCTION

Spoken language Identification is the most popular and important research topic in the area of speech technologies. Text independent Language Identification (LID) is the task to recognize the language of an unknown speech utterance being spoken by an unknown speaker using machine. Several applications use LIDs including global communications, call routing systems, multilingual dialog systems, multilingual translation systems etc. All LIDs are categorized into two categories namely Text-dependent language identification systems and Text-independent language identification systems. To implement Text-dependent LID, labeled corpus is required whereas Text- independent LID requires only speech signal without knowing the underlying text of spoken utterance.

Zissiman [1] expressed several cues for including phonology, morphology, syntax structure and prosody. Muthusamy [2] proposed several methods to design LID including HMMs, expert systems, clustering algorithms and artificial neural networks etc. Algorithms for LID can be divided into two groups namely phonotactic modeling in which a tokeninzer translates the input speech into phones and then scoring is performed. The second group deals with Acoustic modeling in which input feature vectors are modeled directly by specified methods.

Nakagawa [3] proposed a method to identify the language using VQ distortion method and discrete HMM on phonotactics approach and 43.4 % of performance was achieved. Ka-keung Wong [5] tried to implement the LID by altering the phonotactics approach using discrete HMM and tokenizer. Takashi seino [4] developed an LID to recognize the digits using VQ and HMM on phonotactics and achieved performance about 90%.

Muthusamy [6] said that the differences between languages are present at phoneme level hence it can be exhibited at frame level. T. Nagarajan [7][8] designed a LID system using VQ based system using the cue that the frequency of phoneme is different in different languages. He proposed different methods to identify the languages using codebook and statistical formulas.

In this paper, we proposed a method to implement text independent LID for Indian languages on acoustic features using vector quantization and discrete hidden Markov model, consists of six languages, which gives significantly high recognition performance.

2. VECTOR QUANTIZATION

Vector quantization is a method of automatically partitioning a feature space into different clusters based on training data. It maps n-dimensional vectors in the vector space into a finite number of vectors. It can be implemented using clustering algorithms. In this paper we have used k-means clustering algorithm. Each vector, which is the centroid of the vectors in the cluster under consideration, is called as a code word and set of k code words is called codebook. MFCC feature vectors are grouped into clusters and each cluster is represented by a code word and in turn each code word is represented by a code book index. In the Vector Quantization phase, each feature vector is represented by a codebook index. The codebook size is very important as it influences the performance of the system [7] [8]. In this paper, the code book size 32 and 64 are considered for LID task.

3. HIDDEN MARKOV MODEL

Hidden Markov models (HMM) are the most popular and successful acoustic models for automatic speech recognition. These models provide the likelihood for the unknown test sample, given the sequence of feature vectors as input. These are double stochastic models with a finite set of states.

HMM can be described by three parameters namely states, state transition probabilities, and state symbol probabilities. Each state is associated with a discrete probability value. The model is represented by $\lambda= (\pi, A, B)$. Discrete HMM can be described as

I. Initial probability of states $\pi= \{\pi_i\}$.
II. State transition probability $A= \{a_{ij}\}$.
III. Output Probability distribution in each of states $B= \{b_j(k)\}$.

HMM are associated with 3 problems

I. Evaluation problem, given an HMM (λ) and given an observation sequence $o_1, o_2, \ldots\ldots.,o_t$ compute the probability of observation time sequence.

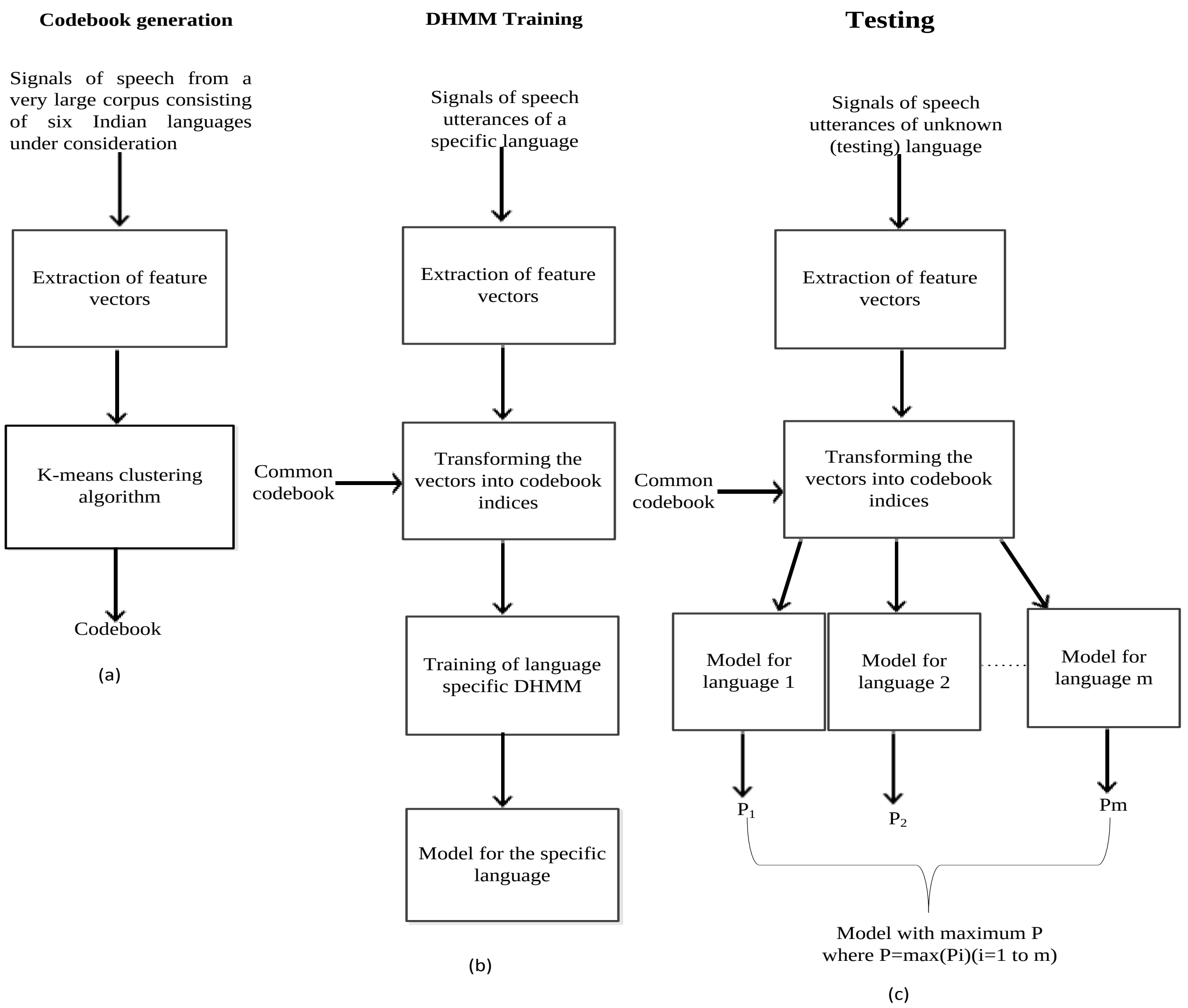
Codebook generation
DHMM Training
Testing
Signals of speech from a very large corpus consisting of six Indian languages under consideration
Extraction of feature vectors
K-means clustering algorithm
Codebook
(a)
Signals of speech utterances of a specific language
Extraction of feature vectors
Common codebook
Transforming the vectors into codebook indices
Training of language specific DHMM
Model for the specific language
(b)
Signals of speech utterances of unknown (testing) language
Extraction of feature vectors
Common codebook
Transforming the vectors into codebook indices
Model for language 1
Model for language 2
Model for language m
P1
P2
Pm
Model with maximum P where P=max(Pi)(i=1 to m)
(c)

FIG 1: Text Independent LID System

II. Decision problem to compute most likely sequence given the model λ and probability.
III. Optimization problem to optimize the π, A, B parameters of HMM.

In this paper, third problem is used to train the discrete HMM for the given the observation sequences using Expectation Maximization (EM) algorithm. In testing, first problem is used to evaluate the probability of utterance of speech using forward-backward algorithm.

4. PROPOSED SYSTEM FOR TEXT INDEPENDENT LANGUAGE IDENTIFICATION

The proposed language identification system (LID) consists of vector Quantization along with the discrete hidden Markov models one for each language. It consists of following steps

1. Code book Generation
2. Training of DHMM
3. Testing

4.1 Schematic diagram for LID

4.2 Codebook generation

In this step, MFCC feature vectors of each language are derived from the speech corpus of all languages for which the system is proposed to built. A common codebook is generated for both training and testing. 12 dimensional Mel frequency cepstral coefficients (MFCC) feature vectors are extracted from each sample speech signal. All extracted feature vectors are clustered into k clusters using k-means clustering algorithm and generated a common code book of size 32 and 64 as shown in Fig.1 (a).

4.3 DHMM Training

In the training phase of discrete HMM, the discrete HMM parameters are initialized randomly. The indices corresponding to each feature vector which are derived in code book are used as observation sequence to train discrete HMM. Using these parameters and EM-Optimal algorithm, discrete HMMs are trained one for each language as shown in Fig.1 (b).

The above steps are repeated for each language considered for training so that for each language $L_{1,}$ $L_{2...}$ $L_{m,}$ their corresponding discrete HMMs λ_{1}, λ_{2},……. λ_{m} are created.

4.4 Testing

During recognition phase, unknown test utterance is coded using the same common codebook which is used in training that gives sequence of code book indices and this sequence is treated as observation sequence. The probability of this observation sequence for each discrete HMM (λ_i) is calculated using forward- backward algorithm of discrete HMM. The model which gives maximum likelihood value is hypothesized as identified language as shown in Fig.1 (c).

5. PROPOSED ALGORITHM FOR TEXT INDEPENDENT SPOKEN LANGUAGE IDENTIFICATION

The steps in algorithm are summarized as follows:

5.1 Code book generation:

1. For each speech utterance of each language L_i, extract MFCC feature vectors.
2. Cluster all feature vectors of all the listed languages into k clusters using k-means clustering algorithm and get a common code book. For better performance huge training speech corpus is preferable.

5.2 Training Phase

For each language L_i (i is 1 to m) do

1. for each speech utterance in training Database
   a. Extract a sequence of MFCC feature vectors from sample speech utterance.

   b. For each MFCC feature vector, get the corresponding codebook index so that all feature vectors are represented by their corresponding code book indices. One sequence of indices is derived for one training speech utterance.
2. Initialize DHMM model for each language by taking parameters randomly and train DHMM using the code book indices as observation sequence.
3. Apply EM algorithm to obtain optimal solution.

5.3 Testing Phase

1. Extract MFCC feature vectors from unknown speech utterance.

2. For each MFCC feature vector, get the corresponding codebook index so that all feature vectors are represented by their corresponding code book indices and use this sequence of indices as observation sequence.
3. For each language DHMM (λ_i), find the likelihood of this observation sequence.
4. Select the language which gives maximum likelihood for a given test utterance.

6. EXPERIMENTAL SETUP

The experiments are carried out using Matlab9.0 on Windows7 platform. The database of Indian languages was used for study [9]. It consists of six languages namely Telugu, Tamil, Hindi, Marathi, Malayalam and Kannada each with a duration of 25 minutes. 12 dimensional MFCC feature vectors of speech signal are extracted and vector quantization with different codebook size of 32 and 64 is implemented using MATLAB. Discrete HMMs are implemented and testing is performed for different utterances of 1s, 2s and 3s duration using forward-backward algorithm developed using MATLAB.

7. RESULTS

The performance of language identification for six Indian languages for different duration of test and varying code book size is depicted in the following Table 1 and Table 2.

Table 1. LID performance for varying test duration for codebook size 32

Language	Performance (in %)		
	3s duration	2s duration	1s duration
Tamil	92	90	90
Hindi	90	90	85
Telugu	80	75	70
Marathi	85	82	80
Malayalam	83	80	78
Kannada	83	82	82

Table 2. LID performance for varying test duration for codebook size 64

Language	Performance (in %)		
	3s duration	2s duration	1s duration
Tamil	100	100	100
Hindi	100	100	100
Telugu	94	93	90
Marathi	98	98	97
Malayalam	98	97	95
Kannada	99	99	99

The overall performance of this system for 3s, 2s and 1s duration of test utterance is 85.5%, 83.3% and 80.0% respectively with code book size 32 as given in the Table 1 whereas for the codebook of size 64, the overall performance of this system for 3s, 2s and 1s duration of test utterance are 98.16%, 97.83% and 96.83% respectively as given in the Table 2.

It is observer that the LID performance for codebook size 32 compared to code book size 64 is inferior. This is due to the fact that number of distinct sound units in any Indian language are much more than 32, and this result in assigning same code book index to more than one phoneme as the code book size is 32. In case of code book size equal to 64, each different phoneme is distinctly represented by different code book index, hence the performance is improved.

8. JUSTIFICATION FOR THE EVALUATION PROCESS FOLLOWED IN LID TASK

One of the important language identification cues is the occurrence of different phonemes is different in different languages.

It is very likely that there is a one to one correspondence between the phoneme and code-book index as the code-book size chosen is 64. Hence the frequency of occurrence of code-book indices is directly related to the frequency of occurrence of phonemes in language, as the common codebook is used for all the languages.

Sequence of phonemes is also an important language cue, as typically different language possesses different phoneme sequence patterns in framing words and sentences.

VQ based approaches cannot capture temporal information (sequence information of phonemes). However hidden Markov models can capture positional information (frequency of occurrence of phonemes) in addition to sequential/temporal information (phoneme patterns). Hence hidden Markov model (HMM) is found to be a best choice for LID. Since the discrete HMM comparing with continuous HMM, is a thin solution, computational overheads are relatively less.

9. CONCLUSIONS

In this paper, a novel approach is proposed for text-independent language identification which does not require annotated corpora. A LID system is developed using vector quantization and discrete hidden Markov model (DHMM). The performance of proposed system is improved significantly to previous system analyzed in the literature.

10. REFERENCES

[1] M.A.Zissman "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech" lEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 4, NO. 1, p.p 31-44 JANUARY 1996.

[2] Y. K. Muthusamy and N. Jain and R. A. Cole, "Perceptual benchmarks for automatic language identification", in Proceedings of the 1994 International Conference on Acoustics, Speech and Signal Processing VOL.1.pages 333-336, Apr, 1994.

[3] S. Nakagawa, H. Suzuki "A New Speech Recognition Method Based on VQ-Distortion Measure and HMM" Proc. Int. Conf. ASSP, pp.673-679 (1993.4).

[4] Takashi Seino, Seiichi Nakagawa "Spoken Language Identification Using Ergodic HMM with Emphasized State Transition" Proceedings of EUROSPEECH'93.3rd, pp.133-136 (1993.9).

[5] Wong, Kakeung, Siu Man-hung "Automatic language identification using discrete hidden Markov model", In INTERSPEECH-2004, pp.1633-1636.

[6] K.Muthusamy "Reviewing automatic language identification", IEEE Signal processing Magazine PP.33-41, Apr, 1994.

[7] Balleda Jyothsna, A. Murthy.Hema and T. Nagarajan (2000) Language identification from short segment of speech. Sixth International Conference on Spoken Language Processing (ICSLP 2000) .

[8] Nagarajan T, Murthy. Hema A. "A pairwise multiple codebook approach to implicit language identification", In WSLP-2003, 101-108

[9] Language Technology Research Center, IIIT Hyderabad.