DHMM Based Automatic Language Identification System

M. Sadanandam [1], V. Kamakshi prasad[2] and V. Janaki [3]
[1] Assistant professor of CSE, Kakatiya University, Warangal, Andhra Pradesh, India.
[2] Professor of CSE, JNTUH, Hyderabad , Andhra Pradesh, India.
[2] Professor of CSE, Vagdevi College of Engineering, Warangal, Andhra Pradesh, India.
sadanb4u@yahoo.co.in, kamaksiprasad@yahoo.com, janakicse@yahoo.com

Abstract: This paper focuses on the implementation of automatic language identification system (LID). Automatic language identification is a task to classify unknown utterance of speech into listed languages. LID is implemented using discrete hidden Markov models (DHMM). This system involves two phases and they are training phase and testing phase. In training phase, a common code book of MFCC features is created from huge speech corpus of all listed languages. Language specific DHMMs are created one for each language. In testing phase, MFCC feature are extracted from unknown speech and evaluated against each created DHMM. The language is hypothesized as identified language based on the likelihood value of sequence of feature vectors of unknown speech. The OGI database is used for the study. Even though we have used simple and easy method, the results are very impressive.

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 have proposed a novel approach to implement text independent LID for OGI database on acoustic features using vector quantization and discrete hidden Markov mode. In this work, a common code book is created for both training and testing of LID. In training phase, each language feature vectors are evaluated using common codebook and discrete HMMs are created one for each language. In testing phase, the observation sequence of unknown utterance of test speech is evaluated using the same code book and get the likelihood against each language specific DHMM. A model which gives maximum likelihood value is hypothesized as identified language. Even though it is simple method, the results are moderate.

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.

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

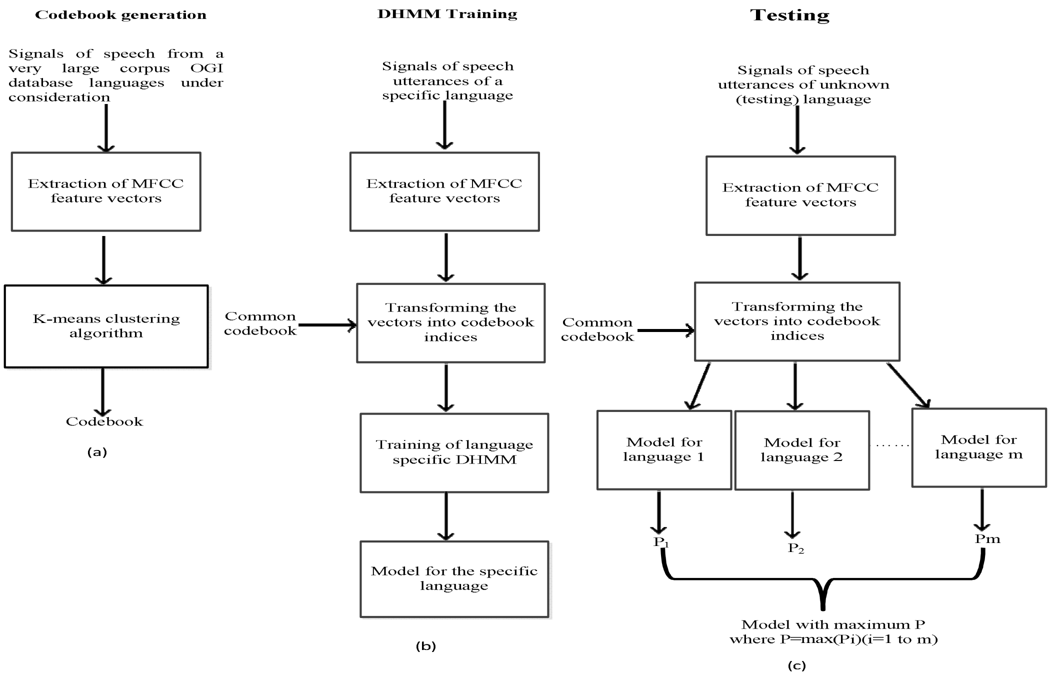
Codebook generation
Signals of speech from a very large corpus OGI database languages under consideration
Extraction of MFCC feature vectors
K-means clustering algorithm
Codebook
(a)
DHMM Training
Signals of speech utterances of a specific language
Extraction of MFCC feature vectors
Common codebook
Transforming the vectors into codebook indices
Training of language specific DHMM
Model for the specific language
(b)
Testing
Signals of speech utterances of unknown (testing) language
Extraction of MFCC 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: DHMM based Text Independent LID System

4.2 Codebook generation

In this phase, a common code book is generated which is the same for both training and testing. 12 dimensional Mel frequency cepstral coefficients (MFCC) are extracted from the huge speech corpus of all listed languages. These feature vectors are grouped into k clusters using k-means clustering algorithm and returned the cluster centers, called as codebook indices. In this work, code book size is chosen 32 and 64. The procedure for the codebook generation explained in Fig.1 (a).

4.3 Training phase

The training phase of LID involves two steps. In the first step, 12 dimensional feature vectors are derived from the huge speech of a considered language. The features are coded with the common codebook which is generated in the section 4.2 so that each feature vector is represented by a codebook index. the sequence of feature vectors are represented by the sequence of codebook indices.

In the second step, the sequence of codebook indices of feature vectors is used as observation sequence to train discrete hidden Markov model. Using EM-optimal algorithm and observation sequence of feature vectors, discrete HMM is trained one for each language. This procedure is repeated for each language considered for training so that for each specific language, the corresponding discrete HMM is created as shown in Fig.1. (b).

4.4 Testing phase

Testing phase of LID also involves two steps. In the first step, 12 dimensional MFFC feature vectors are obtained from the unknown test utterance of the speech. These features are coded with the common code which is also used for the training. These features are represented by the sequence of code book indices.

In the second step, the sequence of codebook indices of unknown utterance of speech is evaluated against each language specific discrete HMM using forward-backward algorithm of HMM and returned the likelihood value of observation sequence of indices as shown in Fig.1.(c). The language specific model which gives maximum likelihood value is hypothesized as identified language.

5. Experimental Setup

The experiments are carried out using Matlab9.0 on Windows7 platform. The OGI database was for study [9]. Each language consists of 25 minutes duration. 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 OGI database for different duration of test and varying code book size is depicted in the following Table 1 and Table 2.

TABLE1. LID performance for varying test duration for codebook size 32

Language	Performance (in %)		
	3s duration	2s duration	1s duration
English	62	68	74
French	65	75	75
Farsi	64	70	70
German	66	69	70
Mandarin	2	7	15
Spanish	35	48	57
Japanese	59	64	68
Korean	34	41	43
Tamil	70	72	74
Average	50.8	57.2	60.8

TABLE2. LID performance for varying test duration for codebook size 64

Language	Performance (in %)		
	1s duration	2s duration	3s duration
English	92	94	95
French	75	75	80
Farsi	76	78	79
German	76	77	77
Mandarin	25	27	25
Spanish	65	68	70
Japanese	79	79	82
Korean	50	51	51
Tamil	80	82	83
Average	68.7	70.2	71.4

The overall performance of this system for 3s, 2s and 1s duration of test utterance is 50.8%, 57.2% and 6.8% respectively with code book size 32 as given in the Table1 whereas for the codebook of size 64, the overall performance of this system for 3s, 2s and 1s duration of test utterance are 68.7%, 70.2% and 71.4% respectively as given in the Table2.

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 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 important cues of language identification is the frequency of different phonemes is different in different languages. In the vector quantization, feature vectors are represented by code book index. If the book size is moderate, there is one to one correspondence between the phoneme and code-book index. In this work common code book of all the listed languages is used so that the frequency of occurrence of phonemes in the languages is represented in terms of the frequency of occurrence of code-book indices.

Typically the different languages possess the different phoneme sequence patterns in framing words and sentence so that the sequence of phonemes also very important cue for LID.

The sequence information (temporal information) of phonemes cannot be captured by vector quantization approaches for automatic language identification. Hidden Markov models can capture sequential/temporal information (phoneme patterns) and temporal information (sequence information of phonemes). As discrete HMM comparing with continuous HMM, is thin solution, computational overheads are relatively less, discrete hidden Markov model (DHMM) is found to be a best choice for LID.

9. Conclusions

This paper has been focused on the new approach proposed for text- independent language identification using DHMM. DHMMs are trained each for a listed language. We have used thin variant solution to implement LID and the results are satisfactory. The study focused on the OGI database. The system evaluates unknown utterance using the common codebook and experiments are carried out with different codebook sizes. Performance of proposed system is improved significantly to previous system analyzed in the literature.

References

[1] 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.

[2] 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] OGI Multi Language Telephone Speech. *www.cslu.ogi.edu/corpora/mlts/*, January 2004.

[10] M. Sadanandam, V.Kamakshiprasad, V.Janaki " Text Independent Language Recognition Using DHMM.", International Journal of Computer Applications (0975 – 888) Vol.48– No.7, June 2012.,