ELSEVIER

CrossMark

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 54 (2015) 231 – 236

Procedia
Computer Science

Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

Significance of GMM-UBM based Modelling for Indian Language Identification

Ravi Kumar V., Hari Krishna Vydana and Anil Kumar Vuppala*

Speech and Vision Lab, IIIT Hyderabad, Hyderabad 500 032, India

Abstract

Most of the Indian languages are originated from Devanagari, the script of the Sanskrit language. In-spite of similarity in phoneme sets, every language its own influence on the phonotactic constraints of speech in that language. A modelling technique that is capable of capturing the slightest variations imparted by the language is a pre-requisite for developing a language identification system (LID). Use of Gaussian mixture modelling technique with a large number of mixture components demands a large training data for each language class, which is hard to collect and handle. In this work, phonotactic variations imparted by the different languages are modelled using Gaussian mixture modelling with a universal background model (GMM-UBM) technique. In GMM-UBM based modelling certain amount of data from all the language classes is pooled to develop a universal background model (UBM) and the model is adapted to each class. Spectral features (MFCC) are employed to represent the language specific phonotactic information of speech in different languages. During the present study, LID systems are developed using the speech samples from IITKGP-MLILSC. In this work, performance of the proposed GMM-UBM based LID system is compared with conventional GMM based LID system. An average improvement of 7–8% is observed due to the use of UBM-based modelling of developing a LID system.

Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

Keywords: Language identification; Phonotactics; Spectral features; Gaussian mixture modelling (GMM); Universal background model (UBM).

1. Introduction

Apart from the message to be conveyed, human speech has a lot of para-linguistic information regarding the speaker authenticity, language being spoken and emotional state of the speaker. The primary objective of an automatic language identification (LID) is to detect the language being spoken from the speech sample. LID has a wide range of real life applications in the area of multi-lingual services like voice operated information query systems and spoken dialogue systems. Every speech system operating on multi-lingual speech data needs a LID system in the front end system. Every language has its own set of rules and linguistic constraints which can be used to improve the accuracy of an automatic speech recognition system (ASR) for that, language of the speech sample has to be detected implicitly. Human computer interaction (HCI) through speech can be taken more deep in to human society if the interaction is through multiple regional languages, for that LID is the initial task to be addressed. A detailed review of LID systems in

*Corresponding author. Tel.: +0-8500727346.
E-mail address: anil.vuppala@iiit.ac.in

Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)
doi:10.1016/j.procs.2015.06.027

the perspective of speech features and models is presented in Ambikairajah *et al.*[1] Various approaches for developing implicit language identification systems are described in Nagarajan[2]. Spectral features are employed to represent the language discriminative information in speech and GMM based LID systems developed in Maity *et al.*[3] In Rao *et al.*[4] and Mary and Yegnanarayana[5] language specific prosody information is used to develop a language identification system using GMMs. Spectral and prosody cues extracted from speech at multi-levels are used for the task of language identification in Reddy *et al.*[6] Magnitude and phase components of excitation are explored for building a LID system in Nandi *et al.*,[7] a GMM based modelling technique is used to develop a language identification system using the features extracted using the information from the excitation. Most of the above LID systems are designed with GMM as their modelling techniques, but use of GMM-UBM for the task of language identification in Indian scenario has not been investigated. This gives motivation to study the use of GMM-UBM based modelling technique for developing the LID systems for the Indian scenario.

Performance of the LID systems depends on the acoustic correlate used to represent the language discriminative information and the modelling technique used to develop the LID systems. Though there is a lot of similarity between the Indian languages every language has its influence on the phonotactic constraints on that language. Owing to the similarity, a model with a large number of mixture components and a large amount of training data for each language is a pre-requisite for developing a LID system. In this work, the pre-requisite of a modelling technique i.e., need of large data to train a model with large number of mixture components is met by using GMM-UBM based modelling technique. Remaining paper is organized as follows: Section 2 describes the details of GMM-UBM modelling technique. Baseline method used for the present work is presented in section 3. Details of the proposed method are provided in section 4. Conclusion and future scope are described in section 5.

2. GMM-UBM Modelling for Developing a LID System

Language specific characteristics of speech can be attributed to the characteristics of the vocal tract system, excitation source and supra-segmental patterns[3]. In the present work, spectral features namely Mel-frequency cepstral coefficients (MFCCs) are employed to represent language discriminative phonotactic information in speech. In this work, spectral vector is obtained by block processing the whole utterance or speech segment using a 20 ms window with an overlap of 10 ms. From every 20 ms speech MFCC features are computed using 24 filter bands.

The spectral vector represented by X is given by

$$X = [x_1, x_2, x_3, \ldots, x_k, \ldots, x_T] \tag{1}$$

where k is the frame index and x_k represents N dimensional MFCCs from the k^{th} frame and T is the total number of features used to form spectral vector. After transforming the input speech into spectral vectors these vectors are used to develop a language model by training the Gaussian mixture models.

A Gaussian mixture density is a sum of M weighted component densities given by the equation

$$p(x_k|\lambda) = \sum_{p=1}^{M} w_p b_p(x_k) \tag{2}$$

where x_k is an N dimensional vector, $b_p(x_k)$, $p = 1 \ldots M$ are the component densities and w_p, $p = 1 \ldots M$ are the weights of the mixtures.

Each component of a D variate mixture function is given by:

$$b_P(x_k) = \frac{1}{(2\pi)^{\frac{D}{2}} \| \Sigma p \|^{\frac{1}{2}}} \exp^{-\frac{1}{2}\{(x_k-\mu_p)(\Sigma p)^{-1}(x_k-\mu_p)'\}} \tag{3}$$

where μ_p is the mean vector computed from the features and Σ_p is the covariance matrix intended to give information regarding the difference between features. Mixture weights are normalized and hence their sum is unity i.e., $\Sigma_p = 1$.

Component Gaussian is a function of mean vector μ_p and Σ_p. Component density is the product of the component Gaussian with its mixture weight i.e., $b_p(x_k)w_p$.

Sum of the component densities is given by Gaussian mixture density. Parameters of a Gaussian mixture density is defined as $\lambda\{w_p, \mu_p, \Sigma_p\}$, $p = 1 \ldots M$.

Further in this work, λ represents a model for language. Each language class is given one GMM i.e., λ. Computing the best possible values for the λ is the prime motive of training phase.

Training phase The main aim of training phase is to compute the best parameters of λ to match the distribution of the feature vectors. A maximum likelihood estimation technique is used to estimate the parameters of $\lambda\{w_p, \mu_p, \Sigma_p\}$, $p = 1 \ldots M$ which maximizes the likelihood of GMM for the given training data.

In the present work, X is the spectral vector obtained from the MFCC features of speech given in 1. The GMM likelihood can be written as:

$$p(X|\lambda) = \prod_{k=1}^{T} p(x_k|\lambda) \tag{4}$$

In the present work, maximum likelihood estimation is performed by an iterative process called expectation maximization. Expectation maximization model starts with a model λ and computes the new model $\hat{\lambda}$ such that $p(X|\lambda) < p(X|\hat{\lambda})$. The new model is considered as the initial model for the next step and the process is continued till a converging threshold is attained.

Testing phase During the testing phase spectral vector from the testing speech sample i.e., Y is given to all the GMM models i.e, $[\lambda_l | l = 1, 2, 3 \ldots L]$, where L is the total number of language classes. LID system computes the posteriori probability for the spectral vector obtained from the testing speech sample, to identify the GMM model that is most likely to produce the feature vector similar to testing spectral vector (Y). The posteriori probability is given by

$$\hat{l} = \arg \max_{1 \leq l \leq L} \Pr\left(\frac{Y}{\lambda_l}\right) = \arg \max_{1 \leq l \leq L} \frac{p(\frac{Y}{\lambda_l}) \Pr(\lambda_l)}{p(Y)} \tag{5}$$

All the language classes are equally likely to happen so $p(\lambda_l)$ is $1/E$ and $p(Y)$ is a constant. So they can be ignored from the equation:

$$\hat{l} = \arg\max \sum_{1 \leq l \leq L} \Pr\left(\frac{\lambda_l}{Y}\right) \tag{6}$$

The language class with the highest posteriori probability $(\hat{l})$ is assumed as language of the spoken speech sample

2.1 GMM with a universal background model

Universal background model is an improvement in the Gaussian mixture modelling technique. The method is initially to select a trained model and determine the likelihood ratio of testing speech sample with the trained model and the universal background model (UBM). The details of GMM-UBM and likelihood ratio are given in the following subsections.

2.1.1 Likelihood ratio

Given a segment of speech Y and hypothesized language of the speaker is l and the task of LID system is to detect the whether Y has the language l. $P(Y|l)$ is the likelihood that speech segment Y has the hypothesized language and $P(Y|\bar{l})$ is the likelihood that speech segment Y does not have the hypothesized language.

$$\text{Likelihood ratio} = \frac{P(Y|l)}{P(Y|\bar{l})} \tag{7}$$

where $P(Y|l)$, $P(Y|\bar{l})$ are refereed as the likelihood values of the hypothesis for the given speech segment. The basic goal of LID system is to determine the values of these likelihoods $(P(Y|l), P(Y|\bar{l}))$. Mathematically l is represented by a model denoted by λ_l which characterizes the of hypothesized language in feature space X. The alternative hypothesis is denoted by $\overline{\lambda_l}$.

Logarithm of the likelihood ratio is termed as log-likelihood ratio: $LLR(x) = \log p(Y|\lambda_l) - \log p(Y|\overline{\lambda_l})$

The model λ_l is well estimated using the training data but $\overline{\lambda_l}$ is not so well trained as needs large data to represent the all possible alternatives of the hypothesized language. One of the major suggested approach is to pool data from all the language classes and build a single model, this model is termed as universal background model. Mathematically the developed UBM model is equivalent to $\overline{\lambda_l}$. In the GMM-UBM based approach parameters of λ_l are estimated by adapting the UBM to language class l and the details of the adaptation process is given in the following subsections.

2.2 Adapting the universal background model

In an GMM-UBM approach, we derive the hypothesized language model by adapting the parameters of UBM to the language data. In the conventional Maximum likelihood estimation method (used in GMM) training an language model is performed independent of the UBM. But in the adaptation approach parameters of the language models are derived by updating the trained parameters of UBM. In GMM technique expectation maximization algorithm is used for training the models, similarly, in adaptation based models are trained by *maximum a posteriori estimation* (MAP). MAP algorithm is a two step process in the initial step the information about the parameters required to adapt the UBM to present class is estimated and in the latter step the new information regarding the parameters is mixed with old parameters and the models of UBM are updated using a data dependent mixing coefficient. The data dependent mixing is done such that the mixtures that are highly influenced by the language specific data in the present class are modified based on the newly estimated parameters and the mixtures that are less influenced by data in the present class retains the parameters from UBM. Due to the tighter coupling between the UBM and the trained models the performance of UBM based approaches are superior than the decoupled approaches. Due to the use of coupled approaches, the performance is not affected by unseen acoustic events i.e., when an unseen acoustic event occurs the mixture parameters of that unseen acoustic classes are directly copied from the UBM. So the during the testing phase unseen acoustic event produces almost zero likelihood ratio which does not contribute evidence either towards or away from the hypothesized language. Though GMM represents a distribution over a large space but a single vector can influence only a few components of the GMM, adaptation and estimation of likelihood can be done faster by considering best scoring mixture components among all the components.

3. Baseline System for Language Identification

3.1 Database used during the present study

Indian Institute of Technology Kharagpur - Multi Lingual Indian Language Speech Corpus (IITKGP-MLILSC) Maity *et al.*[3] is used during the course of present study. In this database 27 regional languages are collected from the radio broadcasts and television talk shows. A minimum of ten speakers including both male and female are present in each language. From each speaker, 5–10 minutes of data is recorded at 16 kHz sampling rate and 16 bits per sample, such that a minimum of one hour data is available for each language.

In this study, spectral features extracted from the speech data are used for the task of language identification. Conventional Gaussian mixture modeling technique is used to develop language models for language identification. The accuracy of the LID system not only depends on the feature vector but also on the parameters of the GMM such as dimensions of feature vectors, number of feature vectors and number of mixture components. In this work, performance of language identification system is analyzed in speaker independent case only i.e., data from different speakers is used for training and testing the language models. From the entire available data set, speech data from two speakers (1 male and 1 female) is omitted during the process of developing the language models and the two speakers (who are not involved in the training process) is used to test the LID system. For analyzing influence of length testing speech sample on the performance of LID performance of LID is computed for testing speech sample with various lengths such as 3 sec, 5 sec and 10 sec. Multiple LID systems are developed by varying the number of mixture components from 8 to 64 to analyze the influence of number of mixture components on the performance of LID. The performance of LID is computed for 75 different test cases from the testing data set and average of all the test cases is

Table 1. Comparing the performance of the LID system by varying length of testing speech sample and number of mixture components.

Performance of baseline (GMM) system			
	Length of testing speech sample		
No. of components	3 sec	5 sec	10 sec
8	36	39	41
16	39	44	45
32	**55**	56	58
64	55	**59**	**59**

Table 2. Comparing the performance of the proposed LID system developed using GMM-UBM.

Performance of GMM-UBM based LID system			
	Length of testing speech sample		
No. of components	3 sec	5 sec	10 sec
256	62	62	64
512	65	66	68
1024	**68**	**68**	**69**

reported in the Table 1. The performance of baseline language identification system is given in Table 1. Column 1 is the number of mixture components used in building the LID system. Performance of LID for testing speech samples of various lengths is given in column 2–6. A significant improvement in the performance of the system is noted with an increase in number of mixture components from 8 to 32. Though there is a slightest improvement with 64 mixture components but that increase is not acceptable due to the time complexity associated with it. A slight increase in performance is noted with an increase in length of a testing speech sample from 3 sec to 10 sec.

4. Proposed Method for Language Identification

Owing to decent from the same origin, most of the Indian languages have overlapping phoneme sets. Despite the similarity in the phoneme sets, every language has its influence on the phonotactic constraints of that language. For discriminating a language using its Phonotactic information in the presence of similar phoneme sets need a large amount of training data for developing a language model. The modelling technique should have a large number of mixture components to account for the slight variation in Phonotactics imparted by the language. Collecting and handling large amounts of data for each class to train a model with a large number of mixture components (i.e., large model) may not be possible always. In this work, GMM-UBM technique is used to develop the language models. In GMM-UBM based modelling technique certain amount of data from all the classes is pooled to build a universal background model with a large number of mixture components and this UBM model is adapted to all the classes. By this approach, a LID with a large number of mixture components can be developed though data in each class is inadequate to support a large model independently.

Speech data from 10 speakers per each language is present in the database. Speech data of 2 speakers from all the languages is is polled to develop a universal background model (UBM) and the data from other 6 speakers in a language is used to adapt the UBM model to develop the corresponding language model. Speech data from the remaining 2 speakers in each language is used to test the developed language models. In the present approach, a universal background model with 256, 512 and 1024 mixtures developed using 90–150 minutes of speech data comprising of all the languages. This UBM model is adapted to all the classes to develop language models with 256, 512 and 1024 respectively. During the process of adaptation 15–20 seconds of speech data from all the 6 speakers of a language i.e., 90–120 seconds of data per language is used to develop the corresponding language model.

In the present work, performance of GMM-UBM with different number of mixtures is given in the Table 2. In Table 2, column 1 is the number of mixture components considered during the present study. Columns 2–4 gives the performance of LID systems with testing speech samples of different length. From the results of Table 2, it is evident that the performance of the proposed approach is superior to the baseline system (GMM based approach). In the proposed approach, there is an average improvement of 7–8% compared to the baseline system. The improvement in the performance is due to the use of a model with a large number of mixture components.

5. Conclusion & Future Scope

In this work, GMM-UBM based modelling technique is used to develop the LID systems. In GMM-UBM based approach, a large model is initially developed by using data from all the languages which is later adapted to all the languages to develop language models. The performance of the proposed method is superior compared to the baseline method. Due to the use of GMM-UBM modelling technique, there is an improvement of 7–8% in the performance of the LID system compared to the baseline system. Performance of the baseline system can be further improved by considering more sophisticated language specific features from the speech. The task of improving the performance of the baseline system would be taken up for further study.

Though there is a significant improvement in the performance of the baseline system owing to the use of UBM based modelling, but this performance is very far from deploying a working model with reasonably good level of accuracy. More sophisticated language discriminative evidences from excitation and supra-segmental patterns (prosody) are to be developed and employed for this task. Along with the excitation and supra-segmental cues language specific cues from the knowledge of phonetics are to be detected, quantified and modeled to accomplish the task of language identification. Due to the existence multi-lingual culture a LID system is highly appreciated in India, but still a long way to go.

References

[1] E. Ambikairajah, H. Li, L. Wang, B. Yin and V. Sethu, Language Identification: A Tutorial, *Circuits and Systems Magazine, IEEE*, vol. 11(2), pp. 82–108, (2011).

[2] T. Nagarajan, Implicit Systems for Spoken Language Identification, Ph.D. Thesis, *IIT*, Madras, (2004).

[3] S. Maity, A. K. Vuppala, K. S. Rao and D. Nandi, IITKGP-MLILSC Speech Database for Language Identification, In *National Conference on Communications (NCC), 2012, IEEE*, pp. 1–5, (2012).

[4] K. S. Rao, S. Maity and V. R. Reddy, Pitch Synchronous and Glottal Closure Based Speech Analysis for Language Recognition, *International Journal of Speech Technology*, vol. 16(4), pp. 413–430, (2013).

[5] L. Mary and B. Yegnanarayana, Extraction and Representation of Prosodic Features for Language and Speaker Recognition, *Speech Communication*, vol. 50(10), pp. 782–796, (2008).

[6] V. R. Reddy, S. Maity and K. S. Rao, Identification of Indian Languages using Multi-Level Spectral and Prosodic Features, *International Journal of Speech Technology*, vol. 16(4), pp. 489–511, (2013).

[7] D. Nandi, D. Pati and K. S. Rao, Language Identification using Hilbert Envelope and Phase Information of Linear Prediction Residual, In *International Conference on Oriental COCOSDA held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013, IEEE*, pp. 1–6, (2013).

[8] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, Speaker Verification using Adapted Gaussian Mixture Models, *Digital Signal Processing*, vol. 10(1), pp. 19–41, (2000).

[9] M. A. Zissman and K. M. Berkling, Automatic Language Identification, *Speech Communication*, vol. 35(1), pp. 115–124, (2001).