Volume 5, Issue 10, October-2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

GMM-UBM Modeling for New features Based Language Identification

Dr. A. Nagesh
Department of CSE, Mahatma Gandhi Institute of Technology
Telangana, India

Abstract*: *In-spite of similarity in phoneme sets, every language has its own influence on the phonotactic constraints (the phonemes that precede or follow another phoneme) of speech signal in that language. The frequency of occurrence of phonemes and phonotactic constraints are the important LID cues. The frequency of occurrence of phonemes between languages effectively captured using new feature extraction method. In this work, the phonotactic variations imparted by the different languages are modeled using Gaussian mixture modeling with a universal background model (GMM_UBM) technique. In this work, performance of the proposed GMM-UBM new feature vector based LID system is compared with conventional GMM new feature vector based LID system. The results are shown that the identification performance of GMM-UBM new feature vector based LID system superior to the conventional GMM new feature vectors based LID system.

Keywords - Language Identification, Gaussian mixture modeling(GMM), Universal background model (UBM), MFCC.

I. INTRODUCTION

Automatic Language Identification (LID) is the task of identifying the language from the short utterance spoken by the unknown speaker. There is lots of important application for automatic language identification. Due to global economic community expansions, it needs of automatic language identification services. In a multi-lingual country like India, automatic LID system has special significance. Today, the need for multi-language communication applications, which can serve people from different nations in their native languages, has gained an increasing importance. Automatic Language Identification has a significant role in the pre-process phase of multi-language systems. The main purpose of a language identification application includes the ability of automatically adapting a speech-based tool, such as online banking or information retrieval, to the native language of the user. With the growth of the Internet, we now live in a worldwide society communicating and doing business with people who use a wide variety of languages which makes language identification more important each day. Multilingual environments may have political, military, scientific, commercial or tourist context.

A detailed review of LID systems in terms of speech features and its modeling techniques are presented by Ambikairajah [1]. Various approaches for developing implicit language identification systems are described by Nagarajan [2]. By Using Spectral features as a language discriminative information in speech, a GMM based LID system is developed in Maityet [3]. Mary and Yegnanarayana [4] are developed language specific prosody information is used to develop a language identification system using GMMs. Performance of the LID systems mainly depends on the acoustic information used to represent the language discriminative information and the modeling technique used to develop the LID systems. The language discriminating acoustic information is represented as a new feature vectors. The language discriminating information is effectively captured using Gaussians. the phonotactic variations imparted by the different languages are modeled using Gaussian mixture modeling with a universal background model (GMM_UBM) technique[5][6][7][8]. Remaining paper is organized as follows: Section 2 describes the new feature extraction method. Baseline method used for the present work is presented in section 3. Details of the proposed method are provided in section 4.and Conclusion in section 5.

II. FEATURE EXTRACTION

In this work a new form of feature vector representation is described. The feature vectors are represented in the form of Gaussians. Using GMM as a front end the feature vectors are extracted from the speech signal. For any system the basic requirement is to obtain the feature vectors from the speech signal. In the literature is found that some attempts are made to explore the new way of representing the feature vector based on the GMM feature extraction. The feature vectors are represented in probability vectors form. Instead representing GMMs as scalar value, it is represented as a probability vector. So the system performance is improved. For LID task, the new feature vectors are obtained from the speech signal estimating using probability density function based on Gaussian mixture model. The underlying language specific discrimination information is represented as a Gaussians.

Beginning from the training data of language L_i, a 12 dimensional feature vectors are extracted with a frame size of 25ms and frame shift of 10ms. These feature vectors are grouped into clusters with 'R' Gaussian mixtures as shown in Fig.1.

Fig.1: R Gaussians for Language L_i.

2.1 Computation of New Feature Vectors

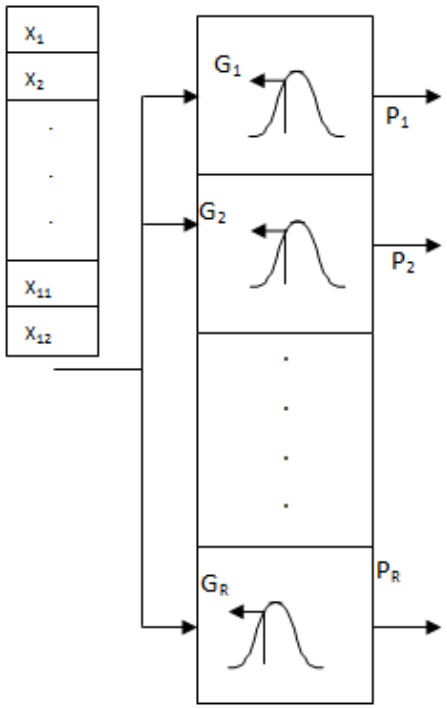

Fig.2.: Parameter estimation for new feature vector P. When R=20, the good identification performance has been achieved.

Once 'R' Gaussians, R clusters are formed. Each cluster represented as one Gaussian. The feature vector X=(X_1, X_2,…X_{12}) is passed through a Gaussian G_1 by calculating probability P_1 using probability density function of Gaussian G_1. This P_1 is first coefficient in the new feature vector. In the same way feature vector X is passed through R Gaussians by creating R feature vector coefficients namely $P_1, P_2, \ldots, P_R$, as shown in Fig. 2. These R coefficients create a new feature vector of dimension R.

In this way, all the feature vectors are passed through ' R ' Gaussians (G_1, G_2,……., G_R) generating new R dimensional feature vectors. In other words the 12 dimensional MFCC feature vectors of size 'N' are transformed to 'R' dimensional feature vectors of size N. The 12 dimensional MFCC feature vector is represented as a 'R' dimensional feature vector.

In the new feature vector, each Gaussian probability density represents one coefficient. Experiments are carried out to find the dimension of new feature vector for good language recognition performance. This is done by varying the number of Gaussians (coefficients) from 15 to 30, i.e number of coefficients in the new feature vector. When the number of coefficients are 12, the good recognition performance is achieved. The 12 dimensional MFCC feature vector is represented as a 12 dimensional new probability feature vector. The newly derived feature vectors are given to the HMM based classifier for language identification.

III. GMM-UBM MODELING FOR LANGUAGE IDENTIFICATION

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

In a GMM model, the probability distribution of the observed data takes the form given by the following equation,

$$p(\bar{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\bar{x})$$

where M is the number of component densities, $\bar{x}$ is a D dimensional observed data, $b_i(\bar{x})$ is the component density and p_i is the mixture weight for i = 1, .., M .

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)'\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)\right\}$$

Each component density $b_i(\bar{x})$ denotes a D-dimensional normal distribution with mean vector $\bar{\mu}_i$ and covariance matrix $\sum_i$. The mixture weights satisfy the condition $\sum_{i=1}^{M} p_i = 1$ and therefore represent positive scalar values. These parameters can be collectively represented as $\lambda = \{p_i, \bar{\mu}_i, \Sigma_i\}$ for i = 1, … M . Each language in a language identification system can be represented by one distinct GMM and is referred by the language models λ_i, for i= 1, 2, 3,…..N, where N is the number of languages.

3.1 Training the Model

Clusters are formed within the training data. Each cluster is then represented with multiple Gaussian probability distribution function (pdf). The union of many such Gaussian pdfs' is a GMM. The most common approach to estimate the GMM parameters is the maximum likelihood estimation, where $P(X|\lambda)$ is maximized with respect to λ. $P(X/\lambda)$ is the conditional probability and vector $X = \{x_1, x_2 \ldots x_t\}$ is the set of all feature vectors belonging to a particular acoustic class. An iterative approach is followed for computing the GMM model parameters using Expectation-Maximization (EM) algorithm . The aim of training is to obtain the mean, variance, and weighting of each Gaussian distribution (λ).Steps for training:

1. Begin with an initial model λ then calculate the new mean, variance, weighting for the new model $\bar{\lambda}$.
2. Check if the newly calculated parameters are more suitable to model the language by using the following formula.

$$p(i \mid \vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)}$$

3. If the $p(X \mid \bar{\lambda})$ is larger than the $p(X \mid \lambda)$, then the new model $\bar{\lambda}$ is used to do the training again. i.e.

$$p(X \mid \bar{\lambda}) \geq p(X \mid \lambda)$$

4. Continue to do the training by repeating step (2) and step (3).

when procedure is repeated to train the new model $\bar{\lambda}$, the new parameters are more close to the actual parameter for modeling the language. The error between the actual parameter for the model and λ become smaller and smaller through training. This procedure is repeated until the error is reached to certain threshold.

3.2 Adapting the universal background model

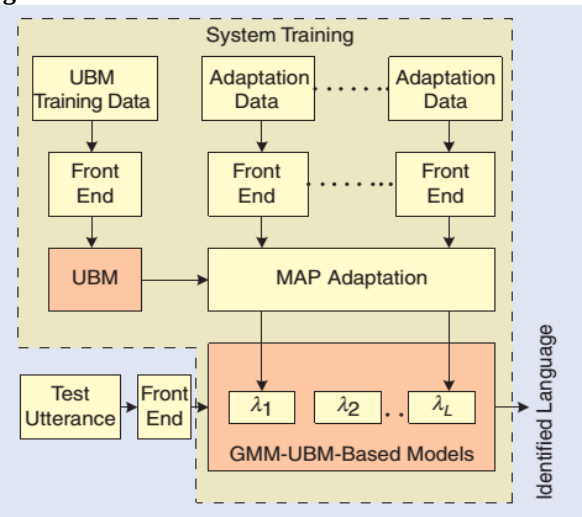

Fig3: A GMM-UBM based LID system the UBM contains the common characteristics shared by all languages and thus represents a more appropriate choice as the starting point to train language-specific models, as opposed to a random distribution used in simpler GMM based systems.

In an GMM-UBM approach, we derive the hypothesized language model by adapting the parameters of UBM to the language data. A block diagram of an adapted GMM-UBM based LID system is shown in Figure. In the conventional Maximum likelihood estimation method (used in GMM) training a 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 training phase of operation of this system occurs in two distinct stages. First a set of feature vectors taken from a number of different languages (typically data from all languages to be tested will be used) are used to train a single GMM. This GMM is referred to as the Universal Background Model (UBM) and is considered to represent the characteristics of all different languages. From the UBM, a GMM is then adapted for each of the languages in the system (using only data from that language) using Bayesian adaptation. In the training phase, the language models are adapted from the UBM using maximum a posteriori estimation. The idea behind maximum a posteriori estimation is that the parameters for the Gaussian mixtures which bear a high probabilistic resemblance to the language specific training data will tend towards the parameters of that training data whereas the parameters of the Gaussian mixtures bearing little resemblance to the language specific data will remain fairly close to their original UBM values. The adaptation procedure is described in and. maximum a posteriori estimation of GMM parameters is often only applied to the means of the mixture components rather than the means, mixtures and weights.

IV. BASELINE SYSTEM FOR LANGUAGE IDENTIFICATION

4.1 Database used during the present study

The Oregon Graduate Institute Multi-language Telephone (OGI_MLT) Speech Corpus, is used for the study[7]. The utterances were produced by ~90 male and ~40 female, in each language over the telephone lines. In the present work, ten languages are used for LID task. For each language, 25 male speakers and 15 female speakers' utterances are used for training and the remaining speakers utterances are used for testing.

In this study, new form of spectral vectors are extracted from the speech signal are used for the LID task. The conventional GMM technique is used to develop language models for language identification. For analyzing influence of length testing speech sample on the performance of LID is computed for varying test speech samples such as 3sec. 5sec and 10sec. 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 50 different test cases from the test data set and average of all the test cases is reported in table1. It is observed that the LID performance is increased with increase in the mixture components. A slight increase in performance is noted with an increase in length of a testing speech signal from 3sec to 10sec.

For dis-criminating a language using its Phonotactic information in the presence of similar phoneme sets needs 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.

This UBM model is adapted to all the classes to develop language models with 256, 512 and 1024. In the present work, performance of GMM-UBM with different number of mixtures is given in the Table 2. 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.

Table1. Comparing the LID Performance for varying length of testing speech sample and number of mixture components.

No of Components	3Sec	5Sec	10Sec
8	46	51	55
16	50	55	58
32	56	61	63
64	65	66	69

Table2. Comparing the Performance of the proposed LID system developed using GMM-UBM.

No of Components	3Sec	5Sec	10Sec
128	54	56	60
256	59	60	55
512	62	66	69

V. CONCLUSION

In this paper, GMM-UBM based modeling technique is used to develop the LID system. The performance of new feature based GMM-UBM LID system is superior when compare to conventional new feature based GMM LID system.

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] 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).

[4] 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)

[5] E. Wong, "Automatic spoken language identification utilizing acoustic and phonetic speech information," Ph.D. dissertation, Speech and Audio Research Laboratory, Queensland Univ. Technol., 2004.

[6] E. Wong and S. Sridharan, "Methods to improve Gaussian mixture model based language identification system," in Proc. Int. Conf. Spoken Language Processing (ICSLP-2002), 2002, pp. 93–96.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, vol.10, pp. 19–41, 2000.

[8] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Processing, vol. 3, pp. 72–83, 1995.

[9] Y.K.Muthusamy, R.A.Cole, and B.T.Oshika, The OGI-MLTS corpus, Proceedings of Int. Conf. Spoken Language Processing, pp. 895-898. Oct. 1992.

