International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)

International Conference on Information Systems and Computing (ICISC-2013), INDIA.

AN AUTOMATIC LANGUAGE IDENTIFICATION USING AUDIO FEATURES

Santhi.S[1], Raja Sekar[2]

[1,2]*Department of Computer Science and Engineering, Mepco Schlenk Engineering college, Sivakasi, TamilNadu, India;*
santhi31@gmail.com, jsrajasekar70@gmail.com

Abstract

An automatic Language Identification (LID) is the task of automatically recognizing a language from the given spoken utterance. Language identification is used to identify the language of the particular audio and reduce the complexity of the audio sample. LID systems that rely on multiple language phone recognition language modeling (PRLM) and n-gram language modeling produces the best performance in formal LID evaluations. By contrast, Gaussian mixture model (GMM) systems, which measure acoustic characteristics, are far more computationally efficient but tended to provide inferior levels of performance. We have described here the efficiency of an LID system for two different languages namely English and Hindi. The evaluation of languages is done on the standard recorded databases, from which features are extracted using Mel-frequency cepstral coefficients (MFCC). The language models are done using PRLM and classification is done using Gaussian mixture model (GMM). The obtained results ensure that accuracy of LID is efficient for the chosen languages and the system performance is evaluated on both PRLM and GMM.

***Keywords*-- Language Identification, PRLM, GMM, MFCC accuracy.**

I. INTRODUCTION

Automatic spoken Language Identification (LID) is the process of categorizing an utterance as belonging to one of a number of previously encountered languages. "Automatic", because the decision is carried out by device. It is implied that the process is independent of content, context, and task of vocabulary and robust with regard to speaker identity, sex, age as well as to noise and distortion introduced by the communication channel. Language identification plays a very important role in various speech related applications. The main challenge here is that speech of language is given as input and the system has to identify that language. If the person has familiarity with a given language it becomes easier for human beings to identify a language from its short utterance compared to machines. The characteristics that make one language differ from another language are phonology, morphology, syntax and prosody.

Audio LID is a mature technology, able to discriminate quite reliably between tens of spoken languages spoken by speakers that are unknown to the system, using just a few seconds of representative speech.

LID has various applications where one application could be a telephone based front-end system whose main work is to route the call to the appropriate caller who is fluent in that language. Other applications of language identification system would be in the speech-to-speech translation, shopping, airports and other commercial areas.

Additionally, LID is used to execute the speaker recognition operation based on speaker dependent and speaker independent way.

In this work we have explored theory and implementation of an automatic speaker dependent language identification system for two spoken languages: English and Hindi and believe the two models such as PRLM and GMM. Both have extracted set of features using MFCC, which is the most popular and efficient technique for feature extraction, the language models are constructed by using PRLM, classification of two languages is done using GMM. Finally testing will be done by both PRLM and GMM approach.

This paper is organized as follows: In Section 2 we have specified some background technique for Audio Language Identification (ALID). In section 3 we have done brief literature survey on identification of languages using PRLM and GMM. It has been observed that GMM is a very popular technique and is widely used for language classification along with various feature extraction techniques. In Section 4 we have given a description of language corpus i.e. the database used for LID system. Section 5 describes the language identification model consisting of detailed analysis of language model construction, classification, language identification and testing to recognize a particular language. In Section 6 we analyze the experimental results that are obtained and finally in section 7 we present our conclusions and scope for future work.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
International Conference on Information Systems and Computing (ICISC-2013), INDIA.

II. TECHNICAL BACKGROUND

2.1. Audio Language Identification Techniques:

Audio Language Identification is a mature field of research, with many successful techniques developed to achieve high levels of language discrimination with only a few seconds of test data. The main approaches make use of the phonetic and phonotactic characteristics of languages which are proven to be an identifiable discriminatory feature between languages.

2.1.1. Phone-Based Tokenization:

Spoken languages differ in many ways, and these differences have been studied closely by language experts and phoneticians. There are several approaches to LID which exploit the difference in phonetic content between languages to achieve language discrimination. Such techniques require the training of a phone recognizer, usually comprising a set of Hidden Markov Models (HMMs), which are used to segment input speech into sequence of phones.

In this approach called, Phone Recognition followed by Language Modeling (PRLM), phonotactics is the feature of language used for discrimination. It specifies different languages have different rules regarding the syntax of phones, and this can be captured in a language model.

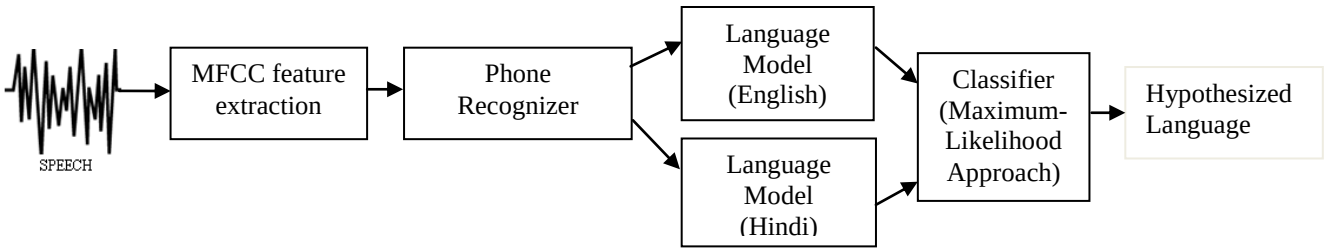

Fig. 1 Phone Recognition Followed By Language Modeling

Here, a single phone recognizer, employing either unilingual or multilingual phone units, "tokenizes" the acoustic input into a sequence of phone labels. This sequence is fed to a bank of parallel n-gram language models, one for each language to be identified as shown in Fig. 1.

Finally, a classifier determines the language of the input speech based on the language model score. For classification, simple maximum-likelihood approach can be used, or more complex back-end classifiers are used. The audio data is presented to a parallel bank of phone recognizer each have a different language called Parallel PRLM (PPRLM).

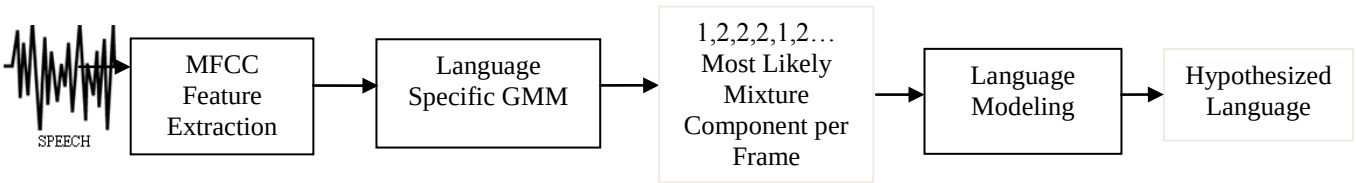

Fig. 2. Gaussian Mixture Model Tokenization

2.1.1. Gaussian Mixture Model (GMM) Tokenization:

A GMM is trained for each language from language-specific acoustic data. Each GMM can be considered to be an acoustic dictionary of sounds, with each mixture component modeling a distinct sound from the training data. The decoded sequence is then used to train the language models. The major components of the proposed system are a GMM tokenizer and a language model for each language of interest.

Additionally a Gaussian classifier can be used to combine the language model scores.

Given a Mel-frequency cepstral coefficient (MFCC) frame, the mixture component is found which produces the highest likelihood score, and the index of that component becomes the token for that frame. Finally, the language models are constructed in order to identify the particular language as shown in Fig. 2.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
International Conference on Information Systems and Computing (ICISC-2013), INDIA.

It has several advantages: firstly, the training tokenizer does not required transcribed data and no agreed protocol for transcriptions. Secondly, there is a reduction in computational cost compared with phone recognition.

III. Literature Review

Language identification is the task of identifying language spoken by a given speaker consisting of two phases training and testing. In the training phase feature extraction takes place and during testing phase feature matching is done. Here we discuss a few recent studies carried out in the particular area of language identification such as both PRLM and GMM.
These all recent trends are used to identify or recognize the particular language and used to extract the audio features of the recognized language and also done by speaker recognition concept.

3.1. PRLM Review:

PRLM approach [4], the dynamic classification is done by using the Hidden Markov Model and the phonotactics are identified by using a single-language phone recognizer. A single-language phone recognizer and n-gram analyzer form the two parts of the system. The phone recognizer act as the front end and n-gram model act as the back end of the system.

In Parallel PRLM (PPRLM) system [4], more than one single-language phone recognition front ends are used in parallel to tokenize the input speech. The phone sequences output by the front ends are analyzed and a language is hypothesized. The front end recognizers do not need to be trained in any of the languages to be recognized.

Parallel Phone Recognition (PPR) method [4], several single-language phone recognition front ends are used in parallel. The likelihoods of the Viterbi paths through each system are compared, from which a language is hypothesized. It is more efficient for identify the more than one language in a single system.

3.2. GMM Review:

Fusion of output scores [8] for identification of language was proposed. In this method, the input consisting of fusion of MFCC and PLP (Perceptual Linear Predictive Coding) feature vectors are increased the performance and decreases the error rate and the output consisting of GMM-UBM (Universal Background Model) LID system.

Gaussian Mixture Model tokenization was used for identification of more than one language spoken around the world [9]. Here speech is given as input to the system and after MFCC features are extracted and the languages are classified by using GMM classification.

For multiple GMM tokenizer it has been found that performance is better than PPRLM approach.

Self-splitting Gaussian Mixture Learning (SGML) algorithm [10] and fast SGML algorithm giving lower computational cost is proposed for GMM. It is based on Bayesian Information Criterion (BIC) and MFCC was used for feature extraction and CMC for channel normalization.

The Split and Merge Expectation Maximization (SMEM) algorithm [6], the features are extracted using 12-MFCC for each frame and CMS (Cepstral Mean Subtraction) is applied to remove the surrounding noise effects. It is more efficient compared to the normal EM algorithm in GMM.

IV. Language Corpus

For preparation of the database mono channel recording was done for fifteen speakers in the English and Hindi language in a closed and quiet noise free room. For digitization, 16 kHz of sampling frequency and 16-bit quantization were used. All the speakers were female speakers in the age group of 20-24.

Given below in the Table I are the sentences that were taken for both training and testing from 15 different speakers of each English and Hindi.

To train the system properly, each speaker is instructed to utter the same sentence 5 times from which 3 utterances were used for training and the remaining 2 utterances are used for testing. Thus for training we have total 90 samples and for testing we have 60 samples as shown in Table II. Each sample in the database is approximately 2-3 sec long in duration.

Table I.
Sentences considered for training and testing

Language	Sentence considered
English	“Let this day be a blessed day for you”.
Hindi	“Yae din ashirwad purvak ra haega”.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
International Conference on Information Systems and Computing (ICISC-2013), INDIA.

Table II.
Database Description

Langu-ages	Number of Speakers (N)	Number of times word repeated by each speaker (t)	Total training samples (N*3)	Total testing samples taken (N*2)
English	15	5	45	30
Hindi	15	5	45	30
Total	30	10	90	60

V. LANGUAGE IDENTIFICATION MODEL

The language identification model can be approached in two ways: Speaker Dependent Model and Speaker Independent Model. Fig. 3 shows the block diagram depicting the steps to identify a particular language. A LID system has following major components: database preparation, feature extraction, language model using PRLM, classification using GMM, identification and testing. LID model is divided into various phases according to these components. Here speech is taken as input for two languages English and Hindi. The recognition performance depends on the performance of the feature extraction.

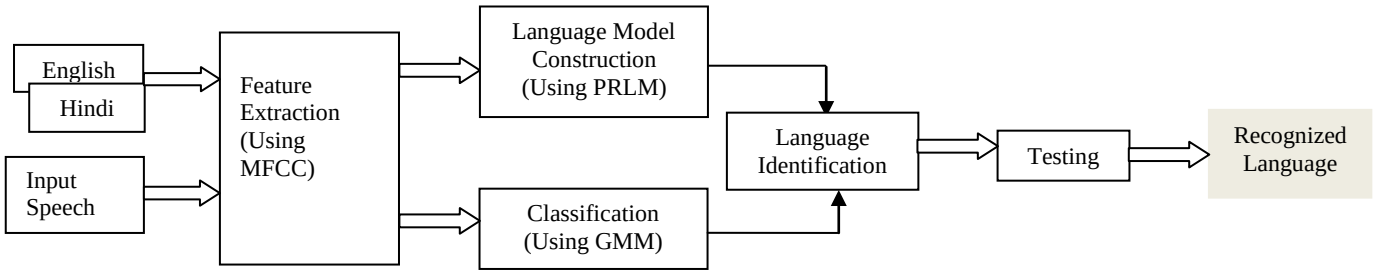

Fig. 3. Block Diagram of the language Identification model

5.1. FEATURE EXTRACTION:

After collecting speech data from various speakers which consist of sentences as given in Table I and database details described in Table II, next phase is to extract the features. There are various methods for feature extraction which include MFCC (Mel Frequency Cepstral Coefficient), PLP (Perceptual Linear Prediction), BFCC (Bank Frequency Cepstral Coefficients) and LPC (Linear Predictive Coding). For extracting features we have used MFCC calculation as shown in Fig. 4.

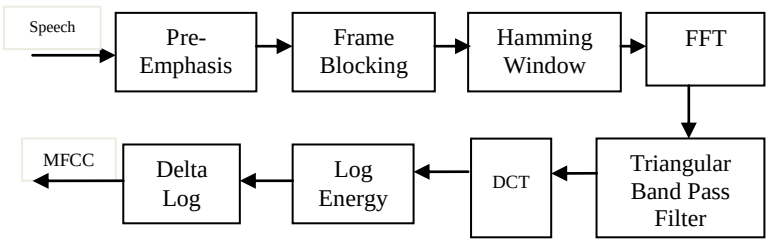

Fig. 4. MFCC Calculation

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
International Conference on Information Systems and Computing (ICISC-2013), INDIA.

5.2. LANGUAGE MODEL CONSTRUCTION:

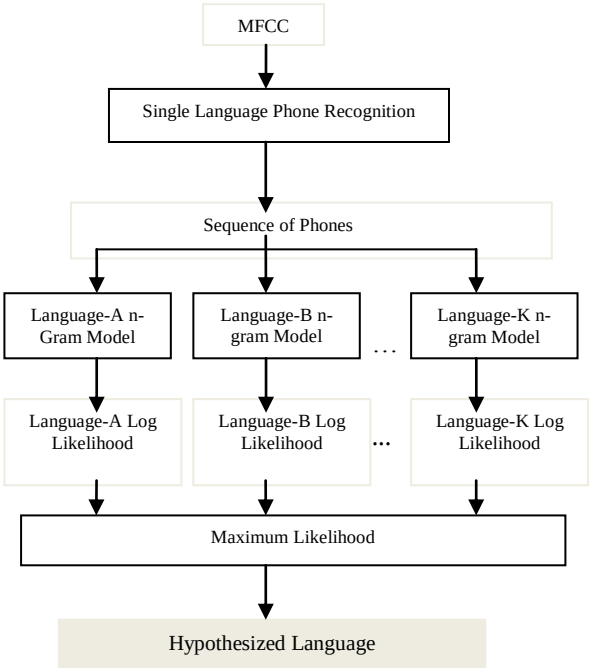

Fig. 5. Language Model Construction

As depicted in Fig. 5, single-language phone recognizer is used to tokenize the input speech, i.e., to convert the input waveform into a sequence of phone symbols. The phone sequences are then analyzed by n-gram analyzer and a language is hypothesized on the basis of maximum likelihood.

5.3. CLASSIFICATION:

After feature extraction, the classification is done using GMM. While doing classification using GMM-EM (Expectation Maximization) algorithm runs in the background for finding maximum likelihood parameter. For likelihood function with multiple maxima, convergence will be local maxima. Fig. 6. Shows the EM algorithm for GMM.

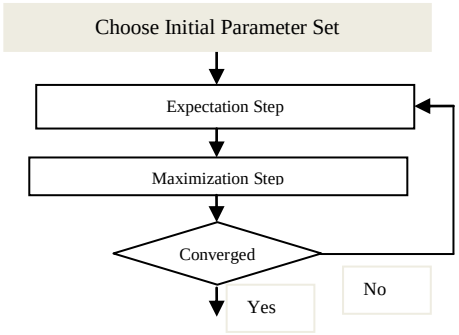

Fig. 6. EM Algorithm

5.4. LANGUAGE IDENTIFICATION:

In this phase the language is identified based on the threshold value, which is chosen as one. After finding the likelihood value of both methods (Lang. Model, Classification) have compared with the threshold value, if it is greater than one the language is identified as English otherwise the language is identified as Hindi.

5.5. TESTING:

The testing is done to find the accuracy of the recognized language. In this accuracy has compared with both methods. For Language Model, the accuracy is based on the construction of n-gram language model and it's likelihood value. For classification method, accuracy is done by using the number of cluster, number of iterations and it's likelihood value.

Accuracy = (correct/total)*100

where, correct=No. of samples correctly modeled or classified

total = Total number of samples given for testing.

VI. EXPERIMENTAL RESULTS

We have considered 15 speakers for LID as given in Table II. Here 90 training samples and 60 testing samples are taken. After doing feature extraction using MFCC, classification is done using GMM and finally testing of samples is done to find accuracy of recognized language. The experiment we have done consists of two phases where we have first tested the GMM system with increasing the iterations and then we have done testing for different number of speakers.

6.1. Iteration of GMM:

In this experiment while doing backend classification using GMMs, we have taken the cluster as 2 and the iteration starts from 2,4,6,…20. From the experimental results in Table III and Fig. 7 it has been found that accuracy is increasing with the increase in iteration for most cases with slight variation in some cases. For English language, accuracy is 100% for iteration 20. For Hindi, the maximum accuracy is 93.33% for iteration 20.

Table III.
Accuracy of two languages for different

Iteration	2	4	6	8	10	12	14	16	18	20
English	51.08	69.31	69.31	79.85	85.88	91.62	92.66	95.55	95.55	100
Hindi	59.78	51.08	69.31	69.31	75.6	80.78	79.85	86.75	93.33	93.33

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)
International Conference on Information Systems and Computing (ICISC-2013), INDIA.

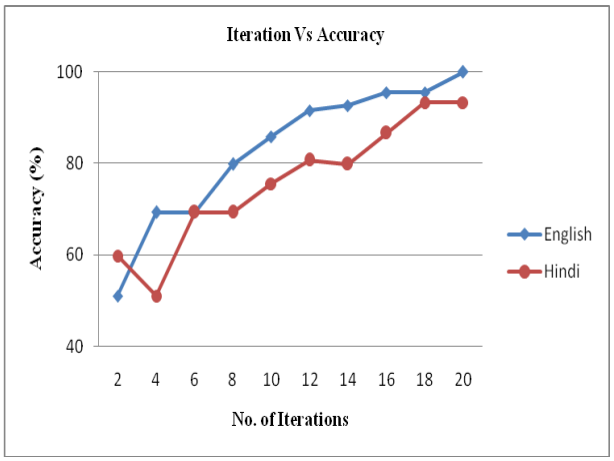

Fig. 7. No. of Iteration versus Accuracy

6.2. Number of Speakers:

Here we have taken highest iteration of 20 and compared the accuracy for number of speakers. From Table IV and Fig. 8, it can be observed that English is giving 100% accuracy for 15 numbers of speakers and Hindi is giving 94% accuracy for 15 numbers of speakers. Thus it is observed here that accuracy of LID increases with increase in the number of speakers.

Table IV.
Accuracy of two languages for different no. of speakers

No. of Speakers	3	6	9	12	15
English	55.65	69.86	78.88	89.6	100
Hindi	60.78	69.44	76.77	87.33	94

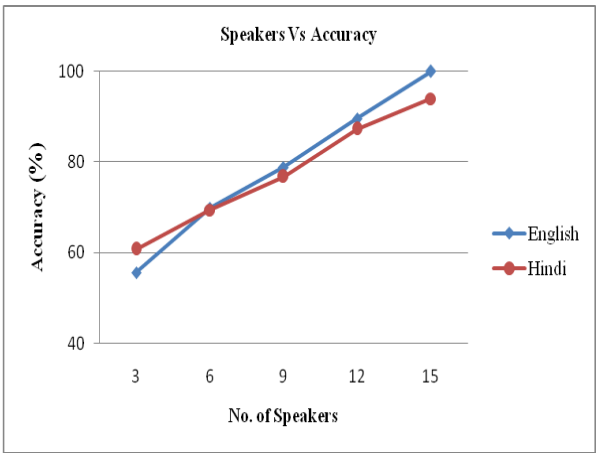

Fig. 8. No. of speakers versus Accuracy

6.3. Result Analysis

From experiment A it can be observed from Table III that after a certain iteration, the accuracy of LID becomes constant. The reason for this is that the database that is used for identification of language using GMM is comparatively small and it is giving best results for lower iteration and when we increase the number of iteration the accuracy becomes constant. The accuracy of LID is very good lowest up to 94% for Hindi and highest up to 100% for English.

From the Table 1V and Fig. 8 in case B, it is clear that as the number of speakers is increasing, accuracy of language identification is also increasing with a more variation for the English language and slight variation for the Hindi language. It can also be observed that accuracy is becoming constant after certain number of speakers. Reason for that difference is the number of speaker range is very less. Here another observation is that for lower number of speakers i.e. for 3 speakers accuracy is around 55% for English language and 60% for Hindi language. As the number of speakers is increased to 15 for iteration 20 the accuracy is reached above 90%. Thus 15 speakers give optimum accuracy and after that if we increase the number of speakers the accuracy becomes constant.

VII. Conclusion And Future Work

The results presented in this paper represent a step towards more flexible and adaptable LID systems. We began by comparing the performance of two approaches to automatic language identification of audio features: Phone Recognition followed by Language Modeling (PRLM), Gaussian Mixture Model (GMM). It is observed that PRLM runs slower because it requires phonetically labeled training speech for each language. GMM run faster compared to PRLM because it does not require any phonetically training speech for each language.

To conclude, GMM is efficient for two languages at GMM iteration 20. The system can be tested in better way if we can increase the number of speakers in the database. GMM gives best results when size of the database is large. Main obstacle that we are facing in designing a good LID system is the collection of good and bigger database. This work can be extended to more number of speakers and it can also be used for training and testing any other two languages except English and Hindi. Future works includes testing the system on a large sentence and large speech corpus and analyze the results for the same.

REFERENCES

[1] J.Newman and S.Cox, "Language Identification Using Visual Features," IEEE Trans. Audio, Speech, and Lang. Process., Vol. 20, No. 7, pp. 1936-1947, Sep. 2012.

[2] M.Zissman, "Automatic Language Identification using Gaussian Mixture and Hidden Markov Models," ICASSP-93.

[3] Y.Muthusamy, E.Barnad, and R.Cole, "Reviewing automatic language identification," IEEE Signal Process. Mag., vol. 11, no. 4, pp. 33-41, Oct. 1994.

[4] M.Zissman, "Comparison of four approaches to automatic language identification of telephone speech," IEEE Trans. Speech Audio Process., Vol. 4, no. 1, pp. 31-44. Jan. 1996.

[5] Todd K.Moon, "The Expectation-Maximization Algorithm", IEEE Signal Processing Magazine, pp. 47-60, November 1996.

[6] N. Ueda, R. Nakano, Z. Ghahramani and G. Hinton, "SMEM Algorithm for Mixture Models", Neural Computation, Vol.12, No.9, pp.2109-2128, 2000.

[7] M.A.Zissman and K.M.Berkling, "Automatic language identification," Speech Commun., vol. 35, no. 1-2, pp. 115-124, 2001.

[8] Eddie Wong and SridhaSridharan, "Fusion of Output Scores on Language Identification System", Workshop on Multilingual Speech and Language Processing, Aalborg, Denmark, September 2001.

[9] P.A. Torres-Carrasquillo, D.A. Reynolds, and J.R. Deller Jr., E. Singer, R.J. Greene, M.A. Kohler, "Language identification using Gaussian Mixture Model Tokenization," in ICASSP, Orlando, Fl., USA, 2002.

[10] Shih-Slan Cheng,Hsin-Min Wang and Hsin-Chia Fu,"A Model-Selection-based Self-Splitting Gaussian Mixture Learning with Application to Speaker Identification", EURASIP Journal on Applied Signal Processing:17, pp. 2626-2639, 2004.

[11] AldebaroKlautau, "The MFCC", November 2005. Available: http://www.cic.unb.br/~lamar/te073/Aulas/mfcc.pdf

[12] Haizhou Li, Bin Ma, and Chin-Hui Lee, "A Vector Space Modeling Approach to Spoken Language Identification," IEEE Trans. Audio, Speech, and Lang. Process., Vol. 15, No. 1, pp. 271-284, Jan. 2007.