International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

Spoken Language Identification System using MFCC features and Gaussian Mixture Model for Tamil and Telugu Languages

N. Athiyaa[1], Dr. Grasha Jacob[2]

[1]M.Phil. Scholar, Department of Computer Science, Rani Anna Government College for Women, Tamilnadu, India
[2]Associate Professor, Department of Computer Science, Rani Anna Government College for Women, Tamilnadu, India

---

Abstract - *Spoken language identification is a major research area these days under speech processing. Speech processing refers to the process of analysing the speech signal. People in the different regions of the world speak different language. Hence, communication between people from different parts of the world is difficult and ineffective. Language translator software plays an important role in eliminating the communication barrier. Spoken language identification plays a key role in language translation. This paper proposes a Spoken language identification system using the Mel-Frequency Cepstral Coefficient (MFCC) features of the speech signal and Gaussian Mixture Model (GMM) that distinguishes between two different speech signal in Tamil and Telugu which are South Indian languages of India.*

Key Words: Spoken Language Identification, Gaussian Mixture Model, MFCC Features.

1. INTRODUCTION

Spoken language identification is the process of identifying the spoken language of a speech signal uttered by an anonymous speaker independent of the accent, speaker and gender. The Proposed Spoken language identification system can be used in combination with forthcoming innovative language translation technologies. The Spoken language identification system proposed in this paper, distinguishes the speech signal between two South Indian languages, Tamil and Telugu. The Proposed System can also be constructed for any number of other languages.

Different types of features are available to represent the input speech signal. The Features are - Syntactic, Lexical, Phonetic, Prosodic, and Acoustic. Spoken language identification systems provide best results when speech signals are represented using the acoustic features. Features are the parametric representation of the speech signal.

Language identification is a classification problem. In order to discriminate between the spoken languages, the languages must be accurately modeled with parameter estimation. Language modeling is broadly classified into generative models, and discriminative models. Generative classifiers learn a model of joint probability p (x, y) of the input x and label y, and make their predictions by using Bayes rules to calculate p(y|x), and then picking the most likely label y. Discriminative classifiers model the posterior probability p(y|x), or map from inputs x to the class labels. Gaussian Mixture Model (GMM) is a generative classifier used for language modeling.

2. RELATED WORK

In a country like India where a large number of languages are being spoken, there is a need to completely study the features which serves as a cue to distinguish between languages. In [1] and [2], Muthusamy has provided a detailed review about the acoustic, prosodic, phonetic and phonotactic features. He also carried out experiments for language identification which implements the combination of spectral features and pitch. He has suggested that phonetic information is sufficient to distinguish between languages with high accuracy.

In [3], Li and K.P., Edwards gives a brief overview of methods for Language Identification. They also discuss about the various measures that are used for evaluating the performance of the language identifications systems. Language modeling variations such as spectral feature modeling, phonotactic modeling, lexical modeling are discussed.

In [4], Julien Boussard discusses various machine learning based classifiers for spoken language identification. The Author has also constructed various algorithms which uses the spectral features derived from the English and Mandarin phone call utterances. He also discusses various feature vectors and language modeling approaches. The Various modeling approaches such as Music-genre motivated approach, Feed forward neural network, convolutional neural network, recurrent neural network and Gaussian Mixture Model (GMM) are discussed. He declares that Gaussian Mixture Modeling (GMM) in combination with Shifted Delta Cepstral (SDC) features provide greater accuracy in Spoken language Identification system.

Spectral features efficiently distinguish the language of the speech signal by representing the variation in acoustic-phonetic information. Linear Predictive Cepstral coefficient (LPCC), Linear Predictive coding (LPC) feature, Shifted Delta Cepstrum (SDC) features, Mel-frequency Cepstral Coefficient (MFCC) are the spectral features used in the area of speech processing. In [5], Quatieri has explained that among the other features of the speech signal, the

MFCCs are quite efficient in spoken language identification system since they use the auditory principles of the speech signal. Delta MFCCs capture the dynamics in a speech signal and represent the acceleration values of the MFCC feature using first order derivatives.

In [6], Zissman compares four approaches for Automatic Language identification of Telephone speech. Four different algorithms for automatic language identification are discussed. The Algorithms discussed are Gaussian Mixture Model (GMM) classification, Phone recognition followed by language modeling (PRLM), Parallel PRLM and Parallel Phone recognition. The Author has proved that the higher performance is achieved by using MFCCs with parallel phone recognition.

In [9],Torres-Carrasquillo proposed a different approach for spoken language identification system which implements PRLM in combination with order of mixture model, phone tokenizers and n-gram classifier. This approach proves that language modeling through Gaussian Mixture Model (GMM) tokenization has greater identification performance with minimal error rate.

In [10], Torres-Carrasquillo discusses the various approaches to language identification using Gaussian Mixture Models (GMM) for classification. The Proposed approach uses the Shifted cepstral delta features in combination with GMM. In Gaussian Mixture Model (GMM), each Gaussian density represents the phonetic information of the speech signal. The GMM based language modelling does not characterize the Contextual information of the speech.

In [11], Pellegrino discusses an alternative approach to phonetic modelling for automatic language identification system. He has explored a spoken language identification system based on the MFCC features and discusses the influence of vowels in identifying the spoken language.

In [12], Nagarajan and Murthy have proposed a spoken language identification system based on Vector Quantization (VQ). The System uses several statistical methods using MFCC feature vectors and it uses the usefulness parameter to improve the performance of the language identification system.

In [13], Cimarusti and Eves discusses the various approaches for pattern classification in Spoken language identification system. The System uses Gaussian Mixture Model (GMM). The Author has conducted several experiments using 100 dimensional feature vectors derived from linear prediction coefficients (LPC).

3. SPEECH CORPUS COLLECTION

The Proposed spoken language identification system uses the speech samples from Microsoft Speech Corpus for Indian languages [14] (Data provided by Microsoft and SpeechOcean.com). The Speech corpus provides training and test data set with large number of speech samples for Tamil, Telugu and Gujarati languages. The Data sets contain speech samples in .wav format. The Proposed system uses subset of the datasets of the Tamil and Telugu languages. The Number of speech samples used for training and testing the language models are 60 and 20 respectively.

The Proposed Spoken Language Identification System uses the acoustic feature to discriminate between languages, since it represents the speech sound and plays a vital role in forming the linguistic structure. Acoustic features are used to distinguish the structure of the speech. Vowels are the sounds that are produced with open vocal tract. Consonants are the sounds that are produced by the constriction in the vocal cord.

Tamil is the oldest classical language from the Dravidian family of languages. It is the official language of the South Indian state Tamil Nadu. It has 30 letters/phonemic characters which includes 10 vowels, 18 consonants and 2 diphthongs. In Tamil, vowels and consonants are represented as *uyireluttu* and *meyyeluttu* respectively. The vowels are classified as *kuril* (short), *nedil* (long) and *kuuriyl* (shortened) vowels. The consonants are divided into *vallinam* (hard), *mellinam* (soft or nasal) and *idayinam* (medium). It also has a special character which is referred as *aytham* in classical Tamil. 10 vowels are further classified as long and short vowels.Consonants in Tamil Language further include 5 nasals, 4 stops and 2 affricates.

Besides the generic features of the Dravidian Languages, Tamil Language has its own features that distinguish it from other languages. The Long vowel in the Tamil Language is twice long as short vowels. Tamil Language does not have any aspirated or voiced stop like other Indian Languages. It has six distinct nasal sounds. At word level, there is no stress and has same emphasis for all the syllables.

Telugu Language also belongs to the Dravidian Language family and it is the official language of the South Indian States Andhra Pradesh and Telangana. Telugu Language is derived from the ancient languages Sanskrit and Prakrit. It has 51 letters/phonemic characters which are classified into 10 vowels and 35 consonants. In Telugu language, Vowels are referred as *achchulu* or *swaralu* and consonants are referred as *hallulu* or *vyanjan*. *Hallulu* (Consonants) combines with *achchulu* (vowels) to form sentence. A unique feature which discriminates it from other Indian Languages is that the words are ended with vowels. Vowels are also classified into short and long vowels. Consonants are classified as 3 nasals, 15 plosives, 5 fricatives and 7 affricates.

Telugu language also has its unique features that distinguish it from other languages. The Long vowels are almost twice long as the short vowels. Therefore, the ratio of the duration of short vowels to the duration of long vowels is 1:2.1. Telugu language exposes unique characteristic called vowel harmony phenomenon which is not a usual characteristic of other Dravidian Languages. Stop consonants can be classified into velar, retroflex, dental and labial.

4. METHODOLOGY

The Proposed Spoken language identification system involves three phases

i) Pre-processing
ii) Feature extraction
iii) Language modelling

The Pre-processing phase makes the speech sample suitable for feature extraction. The Speech sample is resampled at 44.1 KHz. The Feature extraction phase extracts the MFCC (Mel- Frequency Cepstral Coefficient) from the pre-processed input speech signal. Acoustic feature serves as a cue to distinguish between the languages.

The Feature extraction phase involves five steps –

1) Windowing
2) Discrete Fourier Transformation
3) Mel-Scale warping
4) Computation of filter bank energies
5) Discrete Cosine Transformation
6) Mel Cepstrum.

The Mel-Frequency Cepstral Coefficient (MFCC) feature is represented using the vector c which is a set of values C_1, C_2, C_3, C_4, C_k, ... C_n. In the Feature vector C, k represents the frame number that contains the MFCC features of the speech signal and referred as C_k. The n value in feature vector represents the number of feature frames present in an input speech signal.

The Flowchart of the MFCC feature extraction is given in Fig -1.

In the proposed Spoken language identification system, the languages are modeled using the Gaussian Mixture Model (GMM). Gaussian Mixture model (GMM) is an unsupervised classification technique and is the widely used language modeling technique in Spoken language identification. Gaussian Mixture Model (GMM) is based on the probability density which is a weighted sum of multivariate Gaussian densities:

$$g\left(\frac{x}{\mu},\sigma\right) = (2\pi)^{\frac{D}{2}}|\sigma|^{\wedge}((-1)/2)\, e^{\wedge} - ((1)/2\,(x - \mu)^{\wedge}T\, \sigma^{\wedge}(-1)\,(x-\mu)) \quad \ldots\ \ldots\ \ldots\ (1)$$

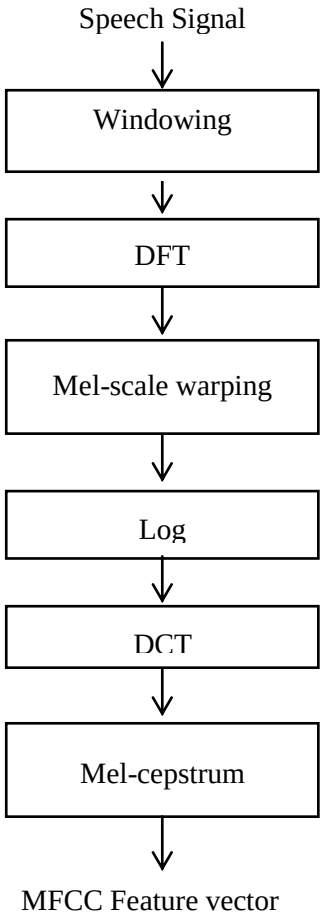


Fig -1: MFCC Feature extraction

The Proposed Spoken language identification system can be elaborated using the algorithms for training and testing the Gaussian mixture model (GMM).

The Training Algorithm trains the GMM language models with the MFCC features extracted from the training input speech signals. The Flowchart for training the Gaussian Mixture Model (GMM) is given in Fig-2. The Testing Algorithm tests the given test speech signal against the trained GMM language models. Feature vectors of the test input signal are extracted and evaluated against each GMM language models to calculate the log-likelihood. The Language model with the maximum likelihood is the detected language. The Flowchart for Testing the Gaussian Mixture Model of the proposed Spoken Language Identification System is given in Fig -3.

Training Algorithm:

Step 1: Pre-process the input speech signal.

Step 2: Divide the speech signal into short time window segments.

Step 3: Extract the MFCC (Mel-frequency cepstral coefficients) from the segmented signals.

Step 4: Estimate the parameters for the Gaussian Mixture Model (GMM).

Step 5: Construct the language models for each language.

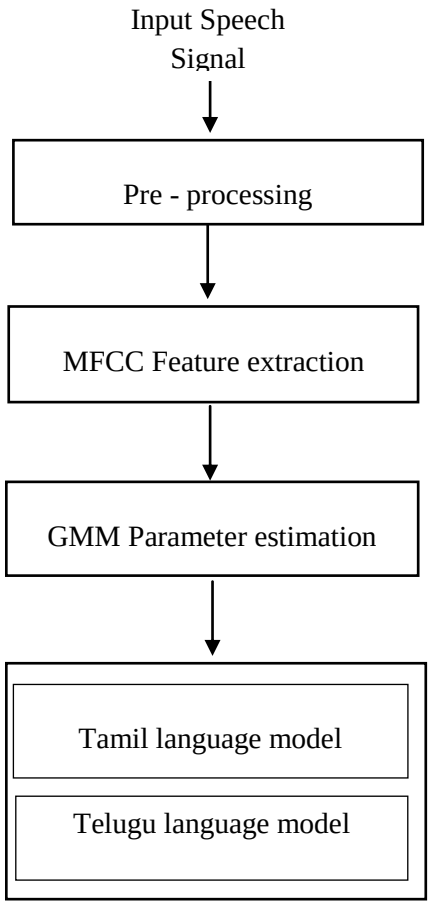


Fig -2: Training the GMM

Testing Algorithm:

Step 1: Pre-process the test speech signal of anonymous speaker.

Step 2: Divide the input speech signal into short time window segments.

Step 3: Extract the features of the window segments and it is represented using the vector C.

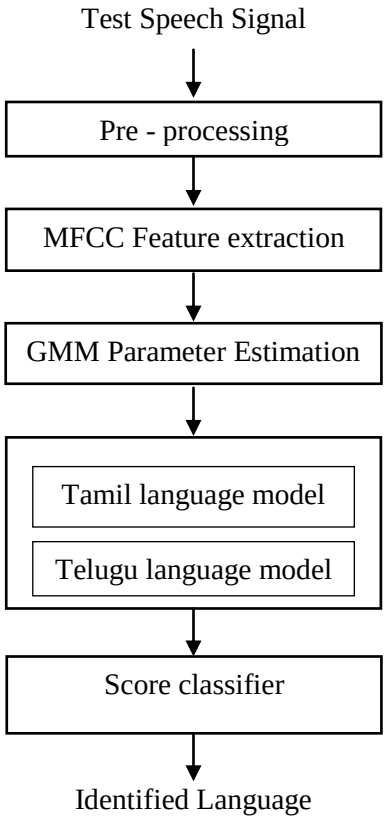
Step 4: The C vector is interpreted against the language model and log-likelihood is calculated,

$$pC|\lambda = \log pC_n|\lambda \ldots \ldots \ldots (2)$$

where λ represents the language model and C_n represents the feature frame of the test input speech signal.

Step 5: The Language model with maximum log-likelihood (score) is the detected language.

In the Proposed Spoken language Identification system, the language model with maximum score (i.e. maximum likelihood) is represented as the winner of the testing phase and it represents that the given test speech signal has the maximum possibility to correspond to that language model.



5. EXPERIMENTAL SETUP

The Experiments are carried out using PYTHON 3.6.6 on windows 7 platform with 2GB RAM and Intel Pentium(R) CPU with 2.10 GHz processor speed.

6. RESULTS AND DISCUSSION

The Performance of the proposed spoken language identification system is measured in terms of Accuracy, False Acceptance Rate (FAR) and False Rejection Rate (FRR). Accuracy is the percentage of speech signals in language classes that are classified as "true" for those language classes. False Acceptance Rate (FAR) is the percentage of speech signals that are not in particular language classes but classified as "true" for those language classes. False Rejection Rate (FRR) is the percentage of speech signals that are in particular language classes but classified as "false" for those language classes. This is also known as miss classification rate.

The Measurements for False Acceptance rate (FAR) and False Rejection Rate (FRR) is given by,

$$FAR = FP/ (FP + TN)$$

$$FRR = FN/ (FN+TP)$$

where,

True positive (TP) is the correct positive prediction.

False positive (FP) is the incorrect positive prediction.

True negative (TN) is the correct negative prediction.

False negative (FN) is the incorrect negative prediction.

The performance of spoken language identification system is evaluated based on accuracy with varying number of mixture components of Gaussian Mixture model (GMM) and is depicted in Table-1. It is observed that the performance of the system increases as the number of Gaussian Mixture Component increases.

Table -1: Language wise performance in terms of accuracy for varying number of mixtures

No. of mixtures	Language identification performance in %	
	Tamil	Telugu
32	70	85
64	85	90
128	90	100
256	90	100
512	100	100

The Language wise average performance of the Spoken language identification system is depicted in Table -2. From the average performances for varying number of mixture components, it is observed that the performance of the proposed system is higher for Telugu language when compared with the performance for the Tamil language. Best performances for both the languages are achieved by the proposed system with the number of mixture components ranging from 128 to 512.

Table -2: Average performance for varying number of mixtures

Language	Average Performance in % with various no. of mixture components				
	32	64	128	256	512
Tamil	70	85	90	90	100
Telugu	85	90	100	100	100
Average	77.5	87.5	95	95	100

The False Acceptance Rate (FAR) and False Rejection Rate (FRR) of the proposed Spoken Language Identification System for 32 and 64 number of mixture components are given in Table-3. It is observed that the False Acceptance Rate and False Rejection Rate reduce as the number of Gaussian mixture components increases.

Table - 3: FAR and FRR of the spoken language identification system for varying number of mixtures

No. of mixtures	FAR	FRR
32	15	30
64	10	15

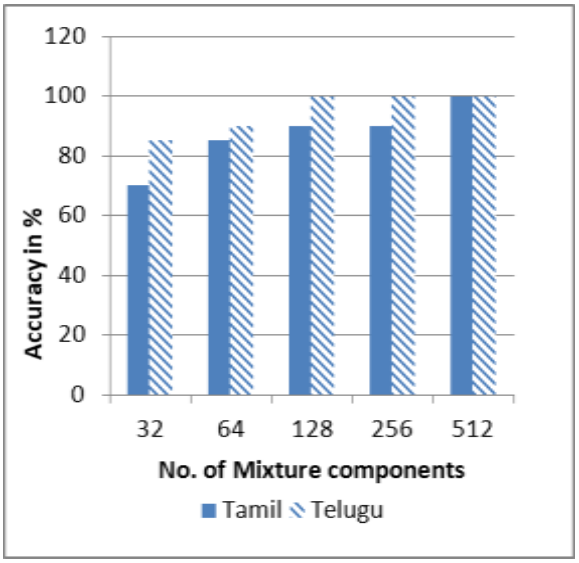


Fig -4: Accuracy graph for spoken language identification System

It is evident from the graph that increasing the number of mixture components results in greater accuracy. Both the Tamil and Telugu languages provide higher results when the number of mixture components is high.

7. CONCLUSION AND FUTURE SCOPE

In this work, speech signals of two South Indian languages Tamil and Telugu are used for the Spoken language identification task. The Proposed Spoken language Identification System achieves higher accuracy for both the languages with more number of mixture components. The Proposed system can be upgraded by including speech signals of any number of other languages. In the proposed system, Gaussian Mixture Model (GMM) is trained with the MFCC features extracted from the speech signals. The Performance may be improved by combining the prosodic features along with the MFCC feature. Language specific features may also be taken into consideration to improve the performance.

REFERENCES

1. Muthusamy, Y. K., Barnard, E., and Cole, R., 1994. "Reviewing Automatic Language Identification". Signal Processing Magazine, IEEE, 11(4),pp. 33-41.

2. Muthusamy, Y. K., October 1993."A Segmental Approach to Language Identification". Ph.D. thesis, Oregon Graduate Institute of Science and Technology.

3. Li, K.P., Edwards, T.J.,April 1980, "Statistical Models for Automatic Language Identification". In Proc. of ICASSP, pp. 884-887,

4. Julien Boussard, Andrew Deveau, Justin Pyron, December 2017, "Methods for Spoken Language Identification", Stanford University.

5. Quatieri, T. F., 2002:"Discrete-Time Speech Signal Processing: Principles and Practice". Engle-wood Cliffs, NJ, USA: Prentice-Hall.

6. Zissman, M. A., 1996, 31: "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech". IEEE Transactions on Speech and Audio Processing, 4(1).

7. Nakagawa, S., Ueda, Y., Seino, October 1992, "Speaker-Independent, Text-Independent Language Identification by HMM". In Proc. of International Conference on Language Sciences and Psycholinguistics (ICSLP), pp.1011-1014.

8. Sugiyama, M., May 1991. "Automatic Language Recognition using Acoustic Features". In Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.813-816.

9. Torres-Carrasquillo, P.A., Reynolds, D.A., and DellerJr, J. R., May 2002: "Language Identification using Gaussian Mixture Model Tokenization". In Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1, IEEE, pp. 1-757.

10. Torres-Carrasquillo, P., Singer, E., Kohler, M., Greene, R., Reynolds, D., and Deller, J. Jr., 2002. "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features". In Proc. of ICSLP, pp.89-92.

11. Pellegrino, F., and Andr-Obrecht, R., 2000. "Automatic Language Identification: An Alternative Approach to phonetic modelling". Signal Processing, 80(7), pp. 1231-1244.

12. Nagarajan, T., & Murthy, H. A., 2002. "Language Identification using Spectral Vector Distribution across the Languages". In Proc. of International Conference on Natural Language Processing.

13. Cimarusti, D., & Eves, R. B., 1982."Development of an Automatic Identification System for Spoken Languages",phase I. In Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1661-1663.

14. "Microsoft Speech Corpus for Indian Languages",https://msropendata.com/datasets/7230b4b1-912d-400e-be58-f84e0512985e.