Proceedings of the International Conference on Communication and Electronics Systems (ICCES 2018)
IEEE Xplore Part Number:CFP18AWO-ART; ISBN:978-1-5386-4765-3

Automatic Language Identification using Machine learning Techniques

Hariraj Venkatesan
School of Computing
SASTRA deemed
to be University,
Thanjavur – 613401

T. Varun Venkatasubramanian
School of Computing
SASTRA deemed
to be University,
Thanjavur – 613401

J. Sangeetha
School of Computing
SASTRA deemed
to be University,
Thanjavur – 613401

Abstract

Investigation in the area of spoken language identification on regional languages aids to broaden the outreach of technology to regional language speakers and also gives to the preservation of regional languages. In this paper, we report our work on identifying spoken data in four local Indian languages Kannada, Hindi, Tamil and Telugu. Automatic Language Identification systems take a speech signal as input and perform computations on the speech input to classify it into one of the natural languages. Mathematical computations performed on the properties of a speech signal such as frequency or amplitude can be used to derive information about the audio and its speaker. In this paper, Mel Frequency Cepstral Coefficients (MFCC) has been used to derive features of speech signals that can be used for identifying languages.For classification purposes, Support Vector Machines and Decision Tree classifiers were used and we got accuracies of 76% and 73% respectively.

Keywords: Automatic language identification (LID), Mel frequency cepstral coefficients, Support vector machines, Decision trees.

1. Introduction

Automatic Language Identification is the process of identifying the language used by the speaker from the given digitized speech utterance. Language has played an important role in facilitating communication and in the exchange of ideas among people. When a multitude of languages are spoken in an environment, the first step in communication is the identification of the language used. [1] [2]

The applications of language identification are several in numbers. In countries such as India, where numerous languages are spoken by the people, the existence of an automatic language identification system tends to serve as a means to simplify several existing processes. Consider the customer care center of a telecommunications company that has to answer calls from several customers to address their grievances or to provide information. In order to ensure efficient functioning, it is essential that the customer is able to express their problems through a comfortable means. Generally, when a customer calls a support center, they are first prompted to select the language they would like to use. Instead, if an Automatic Language Identification system was employed, the customer could start explaining their problem or ask for any information right away without having to select an option to choose the language used for the call. In this manner, the average time required to complete each call can be reduced to a certain extent.

Another important application of automatic language identification is in speech to text conversion. [3] When a computer is equipped with an automatic

language identification system, it enables the user to simply speak instead of having to type commands. Automatic Language identification systems are particularly useful when composing messages in different languages. [4] [5] Users need have to manually change the language each time they wish to change the language of the content. The system will be able to identify the language used by the speaker and use the appropriate font to compose the message as it is dictated.

2. Literature

A lot of research has been done towards automatic language identification of Indian languages by combining different feature extraction methods and devising different approaches. Jothilakshmi et. al have proposed a hierarchical system for identifying language of the speech signal. [3] They have proposed a two-step approach in which first, the family of the language is identified and then the specific language among the various candidates in the family is identified. The system has used Mel Frequency Cepstral Coefficients (MFCC) and Shifted delta cepstral coefficients (SDC) as the feature vectors and for classification purposes, Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), and neural networks were used. They obtained 80.56% accuracy when they used GMM for the LID system with MFCC.

Deep Feed forward neural networks were employed for language identification by I. Lopez-Moreno et al. [6] Two approaches using DNNs were proposed. In the first method, the DNN was used as an end-to-end LID classifier taking the speech features as input and computing as output probabilities of various target languages. In the second method, the DNN was used to extract bottleneck features to form inputs for an i-vector system. Their experiments show that DNN based system outperform state-of-art i-vector systems for short duration utterances.

S. Irtza et al. have proposed a hierarchical language identification model that uses language cluster models. [7] The similarities and disparities among the languages was used to transform the problem into a tree of subproblems of language group identification. Muthusamy et al. have compared various language identification methods that were proposed in the previous years. They identified Acoustic phonotactics, Prosodics, Phonotactics and Vocabulary to be the sources from which features could be extracted for LID systems and compared Speaker identification based, Acoustic model per language and Phonotactic based approaches and it was found that phonotactic based systems performed better than broad-category information-based approaches.

3. Acoustic Feature Extraction

3.1 Mel Frequency Cepstral Coefficients

In Automatic Speech Recognition (ASR), MFCC is the most commonly used feature extraction method. It mimics parts of speech perception and production to extract features having details about the linguistic message spoken by the speaker. [3] It tries to eliminate the speaker dependent characteristics by removing their harmonics. Prior to MFCC, Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) along with HMM classifiers were predominantly used. [4]The step involved in extracting MFCCs from a speech signal is shown in the figure and is as follows:

- Frame the signal into short frames.

- For each frame calculate the periodogram estimate of the power spectrum.
- Apply the melfilterbank to the power spectra, sum the energy in each filter.
- Take the logarithm of all filter bank energies.
- Take the DCT of the log filterbank energies.
- Keep DCT coefficients 2-13, discard the rest.

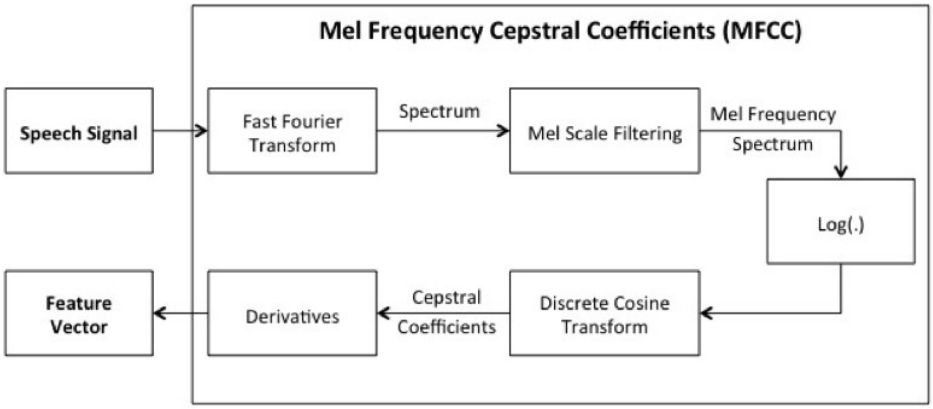

Fig.1 Extraction of MFCC features

4. Acoustic modeling for Language Identification

Each Language can be corresponding to as a class in a classification problem for language recognition. [8] There exist numerous classifier blueprints existing in the literature of machine learning for large-dimension vector classification. [1] This paper consist description of three conventional classifier techniques, namely Support vector machines and Decision trees. The following section consist the description of the mention classifier in detail.

4.1Decision-tree classifier

The decision tree classifier is a supervised learning algorithm that works by putting a set of tests and condition into a tree structure. The source and the inner nodes hold assessment conditions that will be used by the classifier to categorize the different values. All the leaf nodes will be labeled into one of the classes.

Once a tree is constructed using the conditions specified, the class label of a test case can be easily determined. Beginning from the core node, test condition can be applied to the record and go behind the suitable branch depending on the result of the test. Then it guides us to an additional interior node, through which a novel test clause is applied, or to a leaf node. The leaf node is reached then; the label of the class connected with the leaf node is allocated to the record.

For a given problem, several decision trees can be constructed. The biggest problem in constructing an optimal decision tree is finding the questions using which decisions need to be made to arrive at the leaf nodes. A common method to solve this problem is the information gain obtained from each attribute.

4.2 Support Vector Machines

The principle of Structural Risk Minimization (SRM) is used in Support vector machines. Support vector machines are used for pattern classification and non-linear regression, as in Radial Basis Function Neural Net-works (RBFNN). A linear model is constructed to estimate a decision function using non-linear class boundaries that are based on support vectors. [5] SVM, for a linearly separated data, works by training linear machines to obtain optimal hyperplanes that separate data without errors. Training point's closest to optimal separating hyperplanes are called the support vectors. An a priori chosen non-linear mapping is used to map the input patterns to higher dimensional feature space and the high dimensional feature space is constructed using the linear decision surface. A good separation is achieved by the hyper plane that has the largest distance from the nearest training data point of any class for multiclass classification problems.

5. Proposed Work

The proposed automatic language identification process is shown in Fig.2. It has two phases – first, extraction of features from the audio input signals and second, classification is carried out based on the feature extraction. The dataset contains voice recordings of speakers in four languages: English, Hindi, Tamil and Telugu. These are audio files saved in the Waveform Audio file format (WAV). These speech files are passed to a MFCC extraction program that uses each of the audio clippings to generate MFCCs. The generated values are saved in Comma Separated Value file (CSV).

Each row of the CSV file represents a training example. The file contains 39 columns representing the MFCCs extracted from the speech signal corresponding to the training example. The training data is composed of recordings of news broadcasts in four languages – English, Hindi, Tamil and Telugu. This is a multi-class classification problem and it involves supervised training algorithms. The CSV file serves as the input for the classifiers. In the testing phase, each of the speech recordings in the test data, the model computes the MFCC coefficient feature vector and then tests it using the trained model to classify the audio recording into one of the four languages.

6. Experimental results

Experiments conducted in this work were based on database consisting of broadcast news recorded directly from Doordarshan Television Network. The database consisted of 4 languages recorded separately namely Tamil(Ta), Kannada(Ka), Telugu(Te), and Malayalam(Mal). The recording consists of 5 hours of broadcasting data recorded in each language.

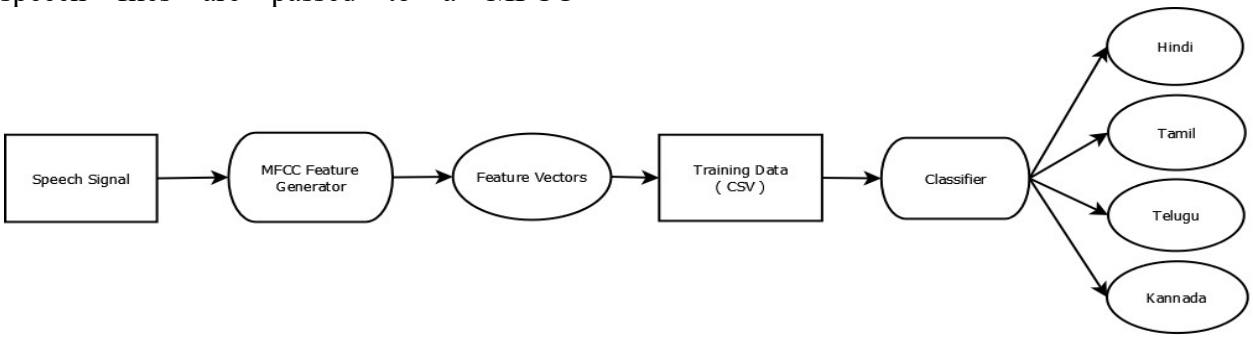

Fig. 2. Automatic Language Identification Model

In Language Identification, amount of training data used plays a major role while training the models. [9] [10]Random selection of training feature vector for majority of experiment processed or conducted are used as the training set for each language to be classified.

The testing and training sets used in the experiments are mutually restricted, and roughly of equal size [11]. The speech files are at first pre-processed by eliminating silences by means of a short-term energy function. The system consists of 3 stages: 1. Feature extraction, 2. Family identification, and 3. Language identification. The features are extracted from all the frames of the speech signal, by means of a frame period of 20ms and frame-shift of 10ms.

The performance can be assessed using detection rate that is defined as

$$\text{Detection rate} = \frac{n_c}{n_c + n_i + n_r}$$

Where n_c is the number of correctly classified, n_i is the number of incorrectly classified keywords, and n_r is the number of rejected keywords. The detection rates for the four languages for two classifiers are recorded below.

Table. 1 Results

Languages	SVM	Decision tree
Tamil	0.4	0.8
Telugu	0.2	0.67
Hindi	0.28	0.2
Kanada	0.33	0.22

7. Conclusion

In this paper, we developed an automatic language identification system that uses Mel Frequency cepstral coefficients extracted from a speech signal to identify the language used by the speaker. The language identification task was modelled as a multi-class classification problem. Feature vectors were extracted from the audio recording to form the training set and then, SVM and Decision Tree classifiers were trained to identify the language spoken by the speaker. We obtained an accuracy of 76% and 73% for SVM and Decision Tree classifiers respectively.

8. References

[1] P. Kumar, A. Biswas, A. .. Mishra and M. Chandra, "Spoken language identification using hybrid feature extraction methods," *Journal of Telecommunication,* vol. I, no. 2, pp. 11-15, March 2010.

[2] C. Madhu, A. George and L. Mary, "Automatic language identification for seven Indian languages using higher level features," in *2017 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, Kollam, India, 2017.

[3] S. Jothilakshmi, V. Ramalingam and S. Palanivel, "A hierarchical language identification system for Indian languages," *Digital Signal Processing,* vol. 22, no. 3, pp. 544-553, 2012.

[4] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing,* vol. 4, no. 1, pp. 31-44, 1996.

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

[6] I. Lopez-Moreno, J. Gonzalez-Dominguez, D. Martinez, O. Plchot, J. Gonzalez-Rodriguez and P. J. Moreno, "On the use of deep feedforward neural networks for automatic language identification," *Computer Speech and Language,* vol. 40, pp. 46-59, 2016.

[7] S. Irtza, V. Sethu, H. Bavattichalil, E. Ambikairajah and H. Li, "A hierarchical framework for language identification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016.

[8] P. P.Shrishrimal, R. R. Deshmukh and V. B. Waghmare, "Indian Language Speech Database: A Review," *International Journal of Computer Applications,* vol. 47, no. 5, pp. 17-21, 2012.

[9] S. Safavi, M. Russell and P. Jančovič, "Automatic speaker, age-group and gender identification from children's speech," *Computer Speech and Language,* vol. 50, pp. 141-156, 2018.

[10] T. Schultz and K. Kirchhoff, "Chapter 8 - Automatic Language Identification," in *Multilingual Speech Processing*, Academic Press, 2006, pp. 233-272.

[11] M. Zissman, "Automatic language identification using Gaussian mixture and hidden Markov models," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* 1993.