International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

LANGUAGE DETECTION IN SPEECH

Mr. M. Dhana Raju[1], Dr.K.Kranthi Kumar[2], R.Shiva[3], M.Sai Kamal[4], B.Dhruv Kumar[5], T.Akanksha[6]

[1]Assistant Professor, [2]Associate Professor, Department of Information Technology, Sreenidhi Institute of Science and Technology, Telangana, India
[3,4,,5,6]B. Tech Student, Department of Information Technology, Sreenidhi Institute of Science and Technology, Telangana, India

---

Abstract - *Language Detection in Speech is an approach to detect the language from audio. It is the process of detecting the language of an utterance by an anonymous speaker, irrespective of gender pronunciations. The major task is to identify those features or parameters which could be used to clearly distinguish between languages. The system uses Support Vector Machine (SVM) to the handle the problem of multi class classification. In existing system speech is converted to text. Here insights are obtained on certain constraints. System for recognizing a language in a speech is not found.* Our proposed system is a one stop solution for the existing issues. Our system aims at recognizing languages as accurately as possible. The major task is to identify features or parameters which could be used to clearly distinguish between languages and produce output.

Key Words: MFCC, SVM, Speech Recognition, Audio file, Language Classification.

1. INTRODUCTION

Identification of languages is the method of defining the language spoken in an utterance. Automatic language recognition is the question of determining the language a person is speaking from a speaking sample. As with speech recognition, today human beings are the world's most effective language processing systems. People may decide, within seconds of hearing speech, if it is a language they know. When it is a language they are unfamiliar with, they will also make moral assumptions about its resemblance to a language they know.

1.1 MFCC

Mel-frequency cepstral coefficients (MFCCs) are a parametric representation of the speech signal, which is widely used in automated speech recognition, but have proven to be effective for other purposes as well, including speaker identification and emotion recognition.

A mel is a unit of perceived tonal pitch or frequency measurement. MFCCs allow a signal representation that is closer to human perception by mapping onto the Melscale, which is an adaptation of the Hertz-scale for frequency to the human sense of hearing. They are calculated by applying a Mel-scale filter bank to the Fourier transform of a windowed signal. A DCT (discrete cosine transformation) subsequently transforms the logarithms spectrum within cestrum.

Mel filter banks consist of alternating triangular filters with cut-off frequencies defined by the centre frequencies of the two neighbouring filters. The filters have linearly spaced centre frequencies and fixed bandwidth on the mel scale. The logarithm transforms multiplication into addition. It transforms the multiplication of the magnitude in the FT into additional.

1.2 SVM

Extremely challenging task is to identify emotion-related speech features. Support Vector Machine is used as a classifier to classify different emotional states such as anger, sadness, fear, happiness, boredom. SVM is a simple and efficient algorithm which has very good performance in classification compared to other classifiers. SVM is the common classification, regression and other learning tasks learning system. On a small amount of training samples SVM has better classification performance. But guidelines on choosing a better kernel with optimized SVM parameters are lacking. With its parameters and kernel function with its parameters, there is no uniform pattern used to choose SVM. The paper proposed methods for selecting optimized parameters and SVM kernel function.

SVM's main principle is to establish a hyperplane as the decision surface to maximize the margin of separation between negative samples and positive ones. Thus SVM is designed for classification of patterns in two class. A combination of binary support vector machines can solve multiple pattern classification problems.

2. EXPERMENTAL ANALYSIS

Output screens are shown here along with the overview of the various functionalities in our application.

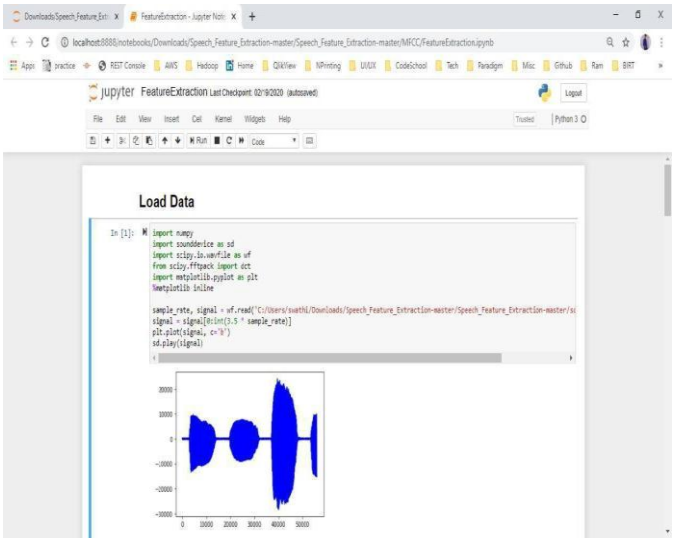

Fig -1: Loading Audio file

The packages and libraries such as numpy, sound device, scipy.io. Wav file, fftpack are imported and the audio file path is set in system. The audio input file is read first as the starting step. The audio signal is played using the method PLAY method. The audio file is divided into different signals. These signals are plotted then we get the graph plot as shown in the figure.

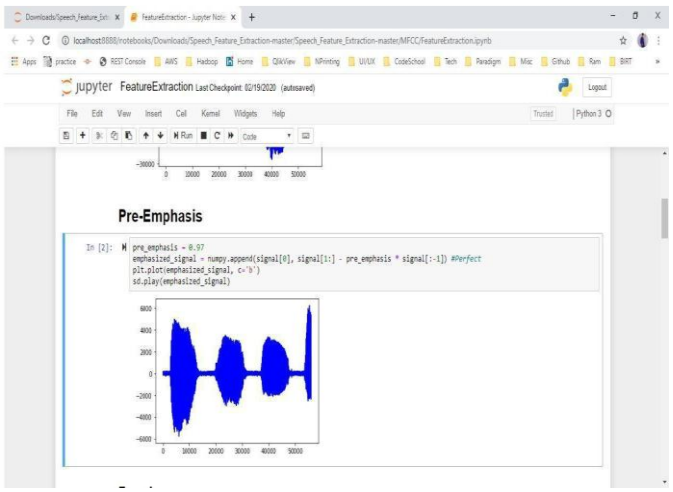

Fig -2: Pre Emphasis

Pre-emphasis works since speech signal is bandlimited and relatively low frequency pre-emphasis boosts the high frequency component.so, high frequencies are distorted more. To avoid this, SNR at high frequency has to be boosted, while lower frequencies SNR can be lower. Pre-emphasis is used in high speed digital transmission to improve the signal quality at the output of a data transmission. The transmission medium can introduce distortions when transmitting signals at high data rates so pre-emphasis is used to distort the transmitted signal to correct this distortion.

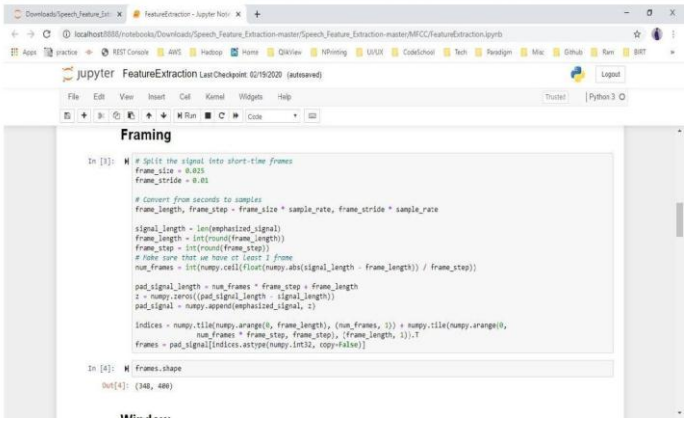

Fig -3: Framing

Here we capture the variability in the waveform, every state of the art system firstly segments the signal into frames. At a given, very short time frame (10-50ms), the speech segment is close to stationary. In this interval, speech signal remain unchanged. The features are extracted from these frames. The original Fourier transform operates on a signal of theoretically infinite length, so the STFT requires that each frame somehow explained to infinite length. For extracting the A short-term analysis is applied to the spectral characteristics of a speech signal .Once the frame blocking procedure is completed, to every frame a windowing function is applied to suppress the effect of discontinuities at frames edges.

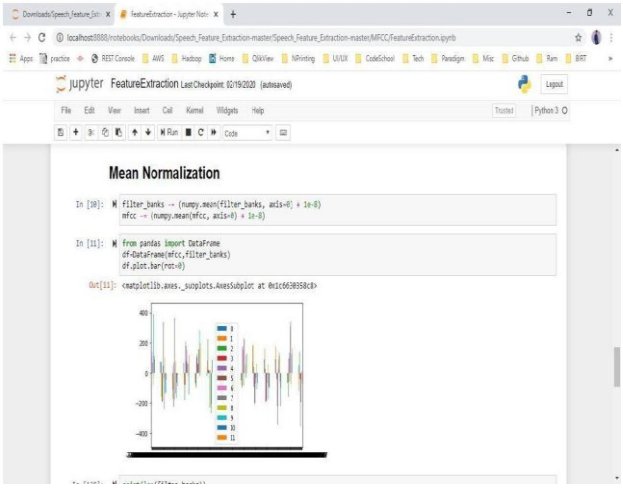

Fig -4: Mean Normalization

In the process of normalization each sample value of the speech signal is separated by the highest value of the

sample amplitude. The mean value of the speech signal is subtracted from each sample to eliminate the DC offset and any of the perturbations caused by recording instruments. Due to possible inconsistencies between training and test conditions, it is considered good practice to reduce as much variation as possible in the data that does not carry important speech information. For example, loudness differences between the recordings are negligible for recognition. Normalisation transforms are applied to the insignificant sources of variance.

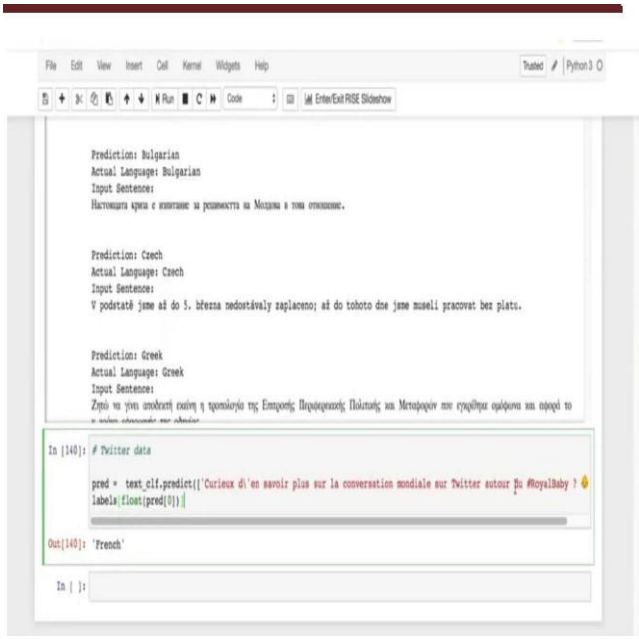

Fig -5: Language Classification

They consider applying SVMs to speech recognition and language recognition. A key element of our approach is the use of a kernel which compares sequences of feature vectors and produces a similarity measurement. Support vector machines (SVMs) proved a effective pattern classification technique.

3. CONCLUSION

The current system is capable of identifying different languages with an appreciable accuracy.. The major barrier with any System research is the availability of standard multi lingual speech corpus for training. This project has not made use of any standard dataset, but still competes for a good accuracy. Experiments were conducted by forming a speech corpus using speech samples obtained from online podcasts and audio books. This corpus comprises of utterances, each of them spanning over a uniform duration of 10 seconds. The entire corpus is split into two sets, larger unit as the training dataset and a smaller set as the test set. Preliminary results indicate an overall accuracy of 96%. A more comprehensive and rigorous test indicates an overall accuracy of 80%. Thus, the acoustic model employing mean values of MFCC proves to be a viable approach for Language Identification.

REFERENCES

[1] K. M. Berkling, T. Arai and E. Barnard, "Analysis of phoneme-based features for language identification", in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 94, Adelaide, Australia, April 1994.

[2] J. Hieronymous and S. Kadambe, "Spoken Language Identification Using Large Vocabulary Speech Recognition", in Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP 96), Philadelphia, USA, 1996.

[3] K. M. Berkling and E. Barnard, "Language Identification of Six Languages Based on a Common Set of Broad Phonemes", in Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP 94), Yokohama, Japan, September 1994.

[4] K. M. Berkling and E. Barnard, "Theoretical Error Prediction for a Language Identification System using Optimal Phoneme Clustering", in Proceedings 4rd European Conference on Speech Communication and Technology (Eurospeech 95), Madrid, Spain, September 1995.

[5] Y. K. Muthusamy, "A Segmental Approach to Automatic Language Identification", Ph.D thesis, Oregon Graduate Institute of Science & Technology, July 1993.

[6] M. A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", in IEEE Trans. Speech and Audio Proc., SAP-4(1), January 1996.

[7] Chi-Yueh Lin, Hsiao-Chuan Wang, “Language identification using pitch contour information”, from Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan