Person Identification through Voice using MFCC and Multi-class SVM

Gaurav Giroti[1], Tejas Nakhate[2], Mahesh Laddha[3], Prof. Manoj Sarve[4]

[123] *Electronics and Communication Engineering, RCOEM, Nagpur*
[4] *Asst. Professor, Dept. of Electronics and Communication Engineering, RCOEM, Nagpur*

Abstract - — *Speaker recognition is basically identification and verification of an authorized personnel who is supposed to access the system. It is used as one of the biometric authentication process available in the world. The biometric verification plays a crucial role in security of the system. Unlike passwords, it cannot be copied from one person to another.*

Speaker recognition can be classified into speaker verification and speaker identification. Speaker verification is the process of accepting or rejecting the identity claim of a speaker.

Speaker identification, on the other hand, is the process of determining which registered speaker provides a given utterance. The process of speaker recognition mainly involves two parts. One is feature extraction and another is feature matching. The feature extraction is used to extract speaker specific features which are further used to do feature matching. These features represent a person in the system. In feature matching these features are matched with the features extracted from the newly recorded voice sample in real time.

Our proposed work is to develop a text-independent speaker identification system, which consists of truncating a recorded voice signal, extracting its features and matching it with stored templates. Cepstral Coefficient Calculation and Mel frequency Cepstral Coefficients (MFCC) are applied for feature extraction purpose. GMM (Gaussian Mixture Modeling) algorithm is used for generating template and feature matching purpose.

Key Words: Mel-Frequency Cepstral Coefficients, Discrete Wavelet Transform, Principle Component Analysis, Support Vector Machines

1. INTRODUCTION

In this project, we have applied the concepts of signal processing to recognize speaker. It involves a four stage system namely speech analysis, feature extraction, modelling and testing. MFCC algorithm is used for the purpose of feature extraction. For the purpose of modelling, we have used the techniques such as Gaussian vector model, support vector machines are used. Although most of the coding in this is done using MATLAB 2017a, use of other languages like python etc. can also be done. The results and the accuracy obtained is usually low in this type of recognition system still we tried to optimize the code as much as we can to obtain the desired results still it needs improvement.

To understand this lets first understand some of the basic properties of the human voice. Human voice consists of two regions, voiced and unvoiced. In order to operate the desired output, it is recommended to do the silence removal of the sample. The voiced speech consists mainly of the vowel sounds which is produced by proper tension on the vocal cord and then pushing air through glottis. The unvoiced portion is formed due to a turbulence usually towards the end of the mouth. There are three parameters that we usually use to detect the voiced and unvoiced part and separate them from each other. They are –

1.1. Zero crossing Rate (ZCR)

Since it is a signal, it crosses the point of minimum amplitude (i.e. zero value) several times during speech. The rate at which this happens can give us an idea whether it is a voiced, unvoiced or a silenced part. Mostly unvoiced signal has higher ZCR than voiced signal due to the presence of higher frequencies as compared to the voiced part. The rate at which the speech signal crosses zero can provide information about the source of its creation. This is because most of the energy in unvoiced speech is found in higher frequencies than in voiced speech, implying a higher ZCR for the former. A possible definition for the ZCR [2] is presented in below equation

$$ZCR = \sum_{m=-\infty}^{\infty} \left| sign[x(m)] - sign[x(m-1)] \right|$$

1.2. Short term Energy (STE)

The energy contained in various frames gives us an idea whether the signal is voiced or unvoiced. Short-term energy of speech signal reflects the amplitude variation. Obviously the amplitude of unvoiced portion is lower than the voiced portion. The amplitude of unvoiced segments is noticeably lower than that of the voiced segments. In order to reflect the amplitude variations in time (for this a short window is necessary), and considering the need for a low pass filter to provide smoothing, h(n) was chosen to be a hamming window powered by 2. It has been shown to give good results in terms of reflecting amplitude variations. [2]

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m) \cdot h(n-m)$$

PGR2026-00033
Krisp Technologies EX1022 Page 1

Below figure shows the plot comparing ZCR and STE of the speech signal.

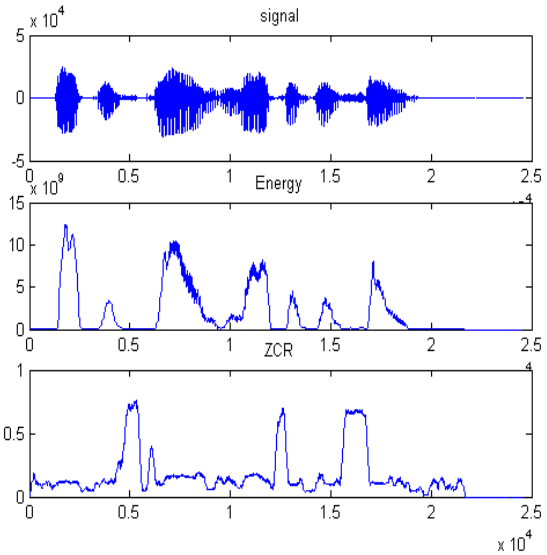

Fig -1: A speech signal (a) with its short-time energy (b) and zero crossing rate (c).

2. FEATURE EXTRACTION

Mel-Frequency Cepstral Coefficients is best from the efficiency point of view [1]. Other techniques like Linear Predictive Coding (LPC) are also used in some cases but we found that MFCC gives better efficiency. MFCCs were first introduced in 1980s [1]. A block diagram of the MFCC algorithm is as shown below.

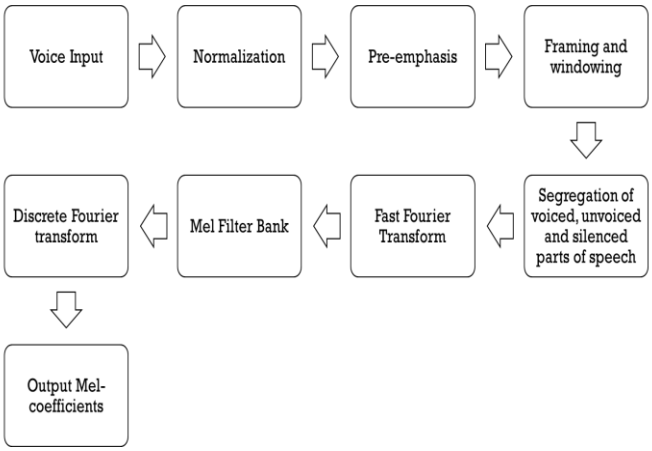

Fig -2: MFCC Extraction

MFCCs are different from typical cepstral analysis methods in a way as they use non-linear frequency scale based on auditory perceptions of human ear called mel scale [22]. These coefficients contain vocal tract information of the speaker [1]. The steps involved in calculating MFCCs are elaborated as follows,

1. The first step involves acquiring voice input in digital form for further processing.
2. In preprocessing the voice signal is normalized so as its amplitude range lies between -1,1. Then the signal is pre-emphasized i.e. the high frequency part of the speech signal is boosted [3].
3. The voice signal is then divided into small frames of 15 to 30 milliseconds
4. The frames obtained are then passed through a hamming window. This is to smoothen the edges of the individual frames to avoid any discontinuities
5. Next the frames containing silenced part are removed using methods involving ZCR or STE calculation
6. Spectral analysis shows that different timbres in speech signals corresponds to different energy distribution over frequencies. Therefore FFT is performed to obtain the magnitude frequency response of each frame [3].
7. Result obtained after performing FFT is then passed through a mel filter bank. Mel filters bank is nothing but triangular filters equally spaced around mel frequency scale [3]. Number of coefficients obtained depends on the number filters present the bank. Each filter can be defined by [22]

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \frac{k-f[m-1]}{f[m]-f[m-1]} & f[m-1] < k \le f[m] \\ \frac{f[m+1]-k}{f[m+1]-f[m]} & f[m] \le k < f[m+1] \\ 0 & k \ge f[m+1] \end{cases}$$

where, M= no. filters

and m={0,1,2,.....M-1}

8. At the end DCT is applied on the out of mel filter bank and obtained MFCC coefficients are stored in database.

3. Modelling and Testing

3.1. Support Vector Machine (SVM)

It is used to detect clusters of data and form boundaries between them. They are also known as Support Vector Networks. They are extensively used in the field of Artificial Intelligence to make decisions. It plays a vital role in the decision making process. What it actually does is it separates samples of data from a cluster of data. It detects similar types of data sets and then forms a boundary around them. The boundaries are formed such that each point on the boundary is equally separated from the boundaries of all the

PGR2026-00033
Krisp Technologies EX1022 Page 2

clusters it is separating. It is a very efficient technique and is pretty simple in case of linear separable data. However, if the data is not linearly separable, then kernels are used to convert it into a linearly separable data first and then GMM is applied. It is also observed that the results of this system gets optimized if we use two dimensional data. The scope of the system can be further increased by its multiclass extension. A multiclass SVM can be used to separate multiple classes from each other. The SVM was initially designed for binary classification but with multiclass SVM, we can extend its applications for classifying more than one classes which is usually the case in most of the practical problems. The multiclass SVM can be applied by decomposing the problem into binary conditions where SVM can be applied directly. Now, to do this there are two possible approaches

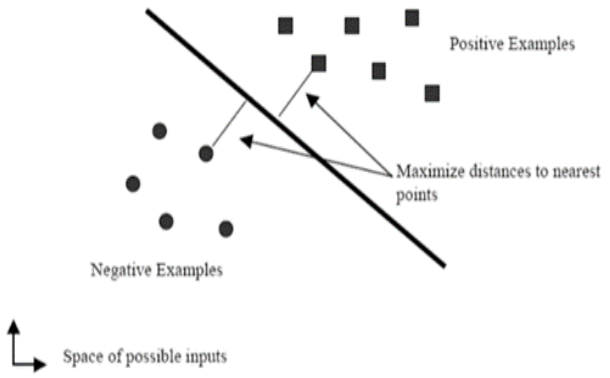

Fig -3: SVM with linear kernel

a) One Versus One (1V1)

In the one versus one approach, k separate binary classifiers are constructed for k class classification. The m-the classifier is trained as positive sample from m-th sample and negative sample from remaining k-1 sample provided no sample is similar to m^{th} one. The drawback is that the ratio of positive to negative classifier is 1: k-1, so the symmetry of the problem is lost. Moreover, if more than one classifiers detected the sample as positive during testing, then it becomes even more difficult to find out the real class.

b) One Versus Rest (1VR)

The one versus one approach is more commonly used. It describes all the possible combinations of samples and forms classifiers for each one of that. The number of classifiers may increase in this case but the accuracy is more. Moreover, in the above type, if the result of any classifier comes to be negative, then it was of no use. But here, since there is direct comparison between two classes, the result is useful weather it comes positive or negative. In fact, there is no positive negative, positive for one class means negative or other and negative for one class means positive for other.

Sometimes the data is not separable linearly, so we need to implement a kernel function in order to convert the data into linearly separable form. There are various types of kernel functions like polynomial kernel, Gaussian kernel, Gaussian radial basis function (RBF), Laplace RBF kernel, hyperbolic tangent kernel, sigmoid kernel etc. [12]. We have used linear kernel in our classification. Use of these kernels is suggested if the data set contains much mixed classes. However, the use of these kernels increases the dimensions of the data which in turn increases the size of the data.

In case of multi-dimensional data, the system becomes complicated. So in order to reduce the dimensions, another technique called Principle Component Analysis is used. An another way to reduce data without losing vital information is to apply Discrete wavelet transform to it both these methods are discussed below

3.2. Principle Component Analysis (PCA)

The Principle Component analysis is used to reduce the dimensions of the data. In a simpler way, it averages the values of a component along which the variation is minimum. Because of which the size of data is reduced and it becomes easy to store and process the data. It also reduces the cost of processing the data since cheap and low power processors could be used. It uses the Eigen vectors and Eigen values to perform this task. The Eigen Vectors and Eigen values are calculated along every dimension.

The use of principle component analysis is necessary since the data is large and most of the data is not of much use which if processed further will produce unnecessary results or will tend to deviate our system from its ideal behavior.

3.3. Discrete Wavelet Transform (DWT)

The PCA can reduce the data only until the dimensions on which the variation of data is large. However, it cannot reduce the data further along that axis. The DWT is used for this purpose. The DWT has a major property which is it reduces the computational power required. There are also other techniques to do this but what makes this technique special is that its ability by virtue of which it reduces the data size without much data loss. Moreover, the wavelet can be reconstructed with the help of scaling function. The scaling function describes all its scaling properties. [9]

Wavelet transforms are in a way similar to the Fourier transform, the similarity lies in the representation of the signal. Both of them represent the signal in terms of basis function or at times with the linear combination of the basis functions. But the main part where wavelet sets itself apart from the Fourier is that the basis functions used in wavelet transform are finite unlike the ones those are used in Fourier transform which are sine and cosine. Due to this feature, the wavelet transform can capture time information as well. [9]

PGR2026-00033
Krisp Technologies EX1022 Page 3

3. CONCLUSION

The Speech Recognition System can be implemented using the MFCC algorithm and the SVM modelling technique. It is suggested to use multiclass SVM if the number of class are more. Also other techniques like the DWT and PCA are useful in improving the quality of the system.

REFERENCES

[1] H. S. Jayanna & S. R. Mahadeva Prasanna "Analysis, Feature, Extraction, Modeling and Testing Techniques for Speaker Recognition", IETE Technical Review, 26:3, 181-190, (2009)

[2] S Upadhya, K Chakraborty & A Talele "Voice Recognition Using MFCC Algorithm", International Journal of Innovative Research in

[3] Advanced Engineering (IJIRAE) ISSN: 2349-2163, Volume 1 Issue 10 (November 2014)

[4] Kawthar Yasmine ZERGAT, Abderrahmane AMROUCHE "Robust Support Vector Machines for Speaker Verification Task", Speech Comm & Signal Proc. Lab.-LCPTS, Faculty of Electronics and Computer Sciences, USTHB, Bab Ezzouar, 16111, Algeria.

[5] Speech and Audio Signal Processing Lab www.jcbrolabs.org/

[6] Joseph Delgadillo "Matlab tutorials" www.josephdelgadillo.com/

[7] http://iitg.vlab.co.in/

[8] youtube.com/

[9] MIT 6.034 Artificial Intelligence, Fall 2010 Lecture 16

[10] https://www.colorado.edu/engineering/CAS/courses.d/SYSID.d/Lectures.d/Discrete.Wavelet.pdf

[11] Chapter 2 Multi-Class Support Vector Machine – by Zhe Wang and Xiangyang Xue

[12] Machine Learning: Multiclass Classification video by Jordan Boyd-Graber (https://www.youtube.com/watch?v=6kzvrq-MIO0)

[13] Kernel Functions-Introduction to SVM Kernel & Examples 12 Aug, 2017 in Machine Learning Tutorials by DF Team (https://data-flair.training/blogs/svm-kernel-functions/)

[14] Gaussian Mixture Models by Douglas Reynolds - MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02140, USA dar@ll.mit.edu

[15] Automatic Speaker Recognition using LPCC and MFCC by P Kumar and S.l. Lahurkar

[16] Robust Support Vector Machines for Speaker Verification Task by Kawthar Yasmine ZERGAT, Abderrahmane AMROUCHE

[17] Performance Analysis of Speaker Identification System Using GMM with VQ M.G. Sumithra A.K. Devika

[18] Sikit learn 2.1 Gaussian Mixture Mode

[19] Zhenhao Ge, Ananth N Iyer, Ram Sundaram, Arvind Ganapathiraju "Neural Network based Speaker Classification and Verification with Enhanced Features" Intelligent system conference, London, 2017

[20] Brain J Love, Jennifer Vining, Xuening Sun "Automatic Speaker Recognitin using Neural Networks"

[21] Songita Mishra, Rabul Laskar, U Baruah, T K Das, P Saha, S P Choudhary "Analysis and Extraction of LP Residual for its application in speaker verification system under Noisy Environment" Multimedia Tools and Applications Volume 76 issue

[22] Bich Ngoc Do "Neural Networks for Automatic Speaker, Language and Sex Identification", Charles University in Prague Faculty of Mathematics and Physics

[23] Atahan Tolunay "Text-Dependent Speaker Verification Implemented in Matlab Using MFCC and DTW" TEKNISKA HÖGSKOLAN LINKÖPINGS UNIVERSITET

PGR2026-00033
Krisp Technologies EX1022 Page 4