ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.4

A Study of various Techniques and Algorithms for Speaker Recognititon

Gaurav Giroti[1], Mahesh Laddha[2], Tejas Nakhate[3]
Department of Electronics and Communication
RCOEM, Nagpur, India

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 authentication 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 identification and Speaker verification. Speaker verification is the process of accepting or rejecting the identity claimed by the speaker. Speaker identification, on the other hand, is the process of determining the individual on basis of voice parameters. Further, depending on the mode of operation in text-dependent and text independent Speaker recognition system. The text-dependent system only works with predefined words or phrase to be used in training and testing, while text independent system works for any utterances made by the speaker. The process of speaker recognition mainly involves feature extraction and creating a classification model. The feature extraction is used to extract speaker specific features like pitch, vocal tract information, excitation source information,etc. which are then used to create a classifier. 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.

Keywords: Mel-Frequency Cepstral Coefficients; Discrete Wavelet Transform; Principle Component Analysis; Support Vector Machines; Gaussian Mixture Model; Neural Networks

I. INTRODUCTION

In this document, we have tried to list down all the possible techniques and algorithms, that we came across, during the course of our project Person Identification through Voice. The whole Speaker recognition system can be divided into four subsystems viz. Analysis, Feature Extraction, Modeling and Training and Testing. Task of each of these subsystems is described and various techniques used in them are explained briefly.

II. SPEECH ANALYSIS AND FEATURE EXTRACTION

A) Linear predictive Coding

LPC is seldom used by itself for speaker recognition in modern day Speaker recognition systems, but here it will only serve as a basis for comparison for the other methods [21].Linear predictive coding (LPC) offers a powerful, yet simple method to provide exactly this type of information. Basically, the LPC algorithm produces a vector of coefficients that represent a smooth spectral envelope of the DFT magnitude of a temporal input signal. These coefficients are found by modeling each temporal sample as a linear combination of the previous p samples as shown below [20]:

$$x(n) = \sum_{k=1}^{p} a_k\, x(n-k) + e(n) = \hat{x}(n) + e(n)$$

In above formula, the estimated value of the n$\hat{x}$th sample, and e(n) is the difference between the estimate and the true value. The p coefficients, a_k that minimize the total error between the signals and x are known as the p$\hat{x}$th order LPC coefficients for the signal x. It is observed that LPC using 8 coefficients has a better recognition rate than other LPC coefficients for codebook of size 32. The results however are not unexpected. The two most significant factors that affect the recognition results are the quality of the speech signal together with the size of the codebook. Increasing the size of the codebook and LPC coefficients increases the effect of noise on the signal, as the signal will contain more information where noise can be present. The results obtained for LPC using codebook size of 64 are pretty much similar to those using codebook sizes of 32. The recognition rate decreases from 66.67%, to hovering around 40% to 50%, as the number of coefficients used increases.

B) Mel-frequency Cepstral Coefficients

Mel-frequency Cepstral coefficients algorithm is a technique which takes voice sample as inputs. MFCC takes human perception sensitivity with respect to frequencies into consideration, and therefore are best for speaker recognition. MFCC represent the human perception of sound. Before implementing MFCC algorithm the input speech signal has to be pre-emphasized i.e. boost up the high frequency content in the speech signal. The segmental analysis of the signal is done by framing signal in frames of length about 15-25 milli-seconds [1]. All the frames are then multiplied by a Hamming window of the same size as that of the frame. The coefficients of MFCC are obtained by taking Fourier transform of each of these frames. The Fourier transform is then passed through the Mel-filter bank which consists of logarithmic triangular filters. This filter bank resembles the (Mel scale). In the end, DCT is applied on logarithmic power of the frame which is nothing but output from Mel-filter bank. Although MFCCs are susceptible to the ambience noise, the simplicity of the procedure for implementation of MFCC makes it most preferred technique for voice recognition **Error! Reference source not found.**.

C) LP Residual Extraction

Speech signal is produced by the convolution of excitation source and time varying vocaltract system components. This excitation source is separated from the vocal tract information by de-convolving the speech. Deconvolution can be done in both frequency domain as well as time domain [21]. However, the disadvantage with the cepstral analysis is that, the deconvolution becomes computationally complex and also some of the information is lost when transverses through frequency domain. This does not happen in case of the Linear Prediction analysis which is processed in time domain. A n order LP-analysis can be done by sampling for a frame size of 320 samples (20ms) with 160 samples (10ms) shift. The sampling frequency used is 16 KHz with a resolution of 16 bit per sample. n LPCs can be computed for every 160 samples. Second order autocorrelation among the samples are used for calculation of LPCs. LP-residual is obtained by using an all-zero filter which filters out the LPCs from the speech thus removing the second order autocorrelation among the samples. These LPCs of the speech signal is passed through all-zero filter to extract LP-residual. The LPresidual obtained is one dimensional and thus need to be reshaped into vectors for ease in processing. Parametric vectors are hard to obtain for such LP-residual signals because they are less correlated. The use of conventional spectral processing tools may lead to the loss of information from such features. Thus, vectors in such cases are obtained by sub-segmenting the LP-residual features into smaller frame sizes and are known as non-parametric vectors. Sub-segmentation is done with frame size of 5ms, 4ms, and 3ms with a single sample shift. To form a 5ms frame size vector, the obtained raw LP-residual is reshaped into a nonparametric vector where the first vector is the sample starting from 1st to 40th sample and the 2nd vector starts from the 2nd sample of the LP-residual and ends with the 41th samplewith a single sample shift between two vectors. This continues till the last sample of the LPresidualis reached. The final non-parametric vector will be a matrix of size M ∗40, where M the length of the vector and 40 is the dimension of the vector. Similar non-parametric vectors are formed for 4ms and 3ms frame size whose matrix sizes are N ∗ 32 and P ∗ 24 where N and P are the respective vector size.

III. MODELLING AND TESTING

After extraction of features, various processes are used to model the data so that it becomes useful for testing. Most of the widely used techniques are mentioned here.

1) Support Vector Machines (SVM)

Support Vector Machine (SVM) is a binary linear classifier in its basic form. It has been recently adopted in speaker recognition task. Given a set of linearly separable two-class training data, there are many possible solutions for a discriminative classifier. An SVM seeks to find the Optimal Separating Hyperplane (OSH) between classes by focusing on the training cases that lie at the edge of the class distributions, the support vectors, with the other training cases effectively discarded [12]. The Support vector machines work great when the data is linearly separable. Also, its implementation for linearly separable data is not much complicated. But, when the data is not linearly separable, we need to use kernels. These kernels when operated on the data it converts it into linearly separable form and then SVM can easily form its classes. Thus the implementation during non-linear separable data becomes quiet complicated and here, we have used the same non-linear SVM [16].

2) Gaussian Mixture Model

Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. One can think of mixture models as generalizing k-means clustering to incorporate information about the covariance structure of the data as well as the centers of the latent Gaussians. They are used as a parametric model of the probability distribution of continuous measurements or features. **Error! Reference source not found. Error! Reference source not found. Error! Reference source not found.** A Gaussian mixture model is a weighted sum of M component Gaussian densities as given by the equation,

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i \, g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

The Gaussian Mixture object implements the expectation-maximization (EM) algorithm for fitting mixture-of-Gaussian models. It can also draw confidence ellipsoids for multivariate models, and compute the Bayesian Information Criterion Speech signal is produced by the convolution of excitation source and time varying vocal tract system components. This excitation source is separated from the vocal tract information by de-convolving the speech. Deconvolution can be done in both frequency domain as well as time domain. However, the disadvantage with the cepstral analysis is that, the deconvolutionbecomes computationally complex and also some of the information is lost when to assess the number of clusters in the data. A Gaussian Mixture. fit method is provided that learns a Gaussian Mixture Model from train data. Given test data, it can assign to each sample the Gaussian it mostly probably belongs to using the Gaussian Mixture. predict method. The Gaussian Mixture comes with different options to constrain the covariance of the difference classes estimated: spherical, diagonal, tied or full covariance.

3) Vector Quantization

A speaker identification system must able to estimate probability distributions of the computed feature vectors. Storing every single vector that generate from the training mode is impossible, since these distributions are defined over a high-dimensional space. It is often easier to start by quantizing each feature vector to one of a relatively small number of template vectors, with a process called vector quantization. It is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a code word. **Error! Reference source not found.** The collection of all code words is called a codebook. The training material is used to estimate the code book. Here a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors hence, a vector quantizer Q of dimension k and size N is a mapping from a vector in the k-dimensional space into one of N centroids in the space. The next important step is to build a speaker-specific VQ codebook for this speaker using those training vectors. There is a well-known algorithm, namely the LBG algorithm [10], for clustering a set of L training vectors into a set of M codebook vectors.**Error! Reference source not found.**

4) ANN: Use of Artificial Neural Networks in Speaker recognition is one of the latest topic of study in Speaker Recognition. The feed forward Multi-Layer Perceptron (MLP) is an artificial neural network that can model non-linear functions by using non-linear sigmoid functions in its hidden layer. In fact, it has been proven that a MLP can model any arbitrary function using only three layers, given that it has enough inputs and hidden units [20]. This property makes the MLP a universal classifier/identifier and a perfect candidate for our speaker identification purposes. The MLP networks each contained only one hidden layer with hidden units ranging from 8 to 15, depending on the types of input data and the number of outputs. With each new training set, initially trained with only 8 hidden units and are then adjusted depending on validation error values. The number of outputs of each network was the number of speakers that the network is designed to distinguish between. The number of epochs that the network trained on ranged from 100 to 200 epochs[19]. In general, using more hidden units did not always give the optimal result. We also learned that changing the number of hidden units had much more impact than changing the number of epochs. Generally, as the number of outputs (i.e. the number of speakers to identify) increased, the network required more hidden units (12-15) as well as more training epochs. Cepstral coefficients also required more hidden units (12) to produce the best results.

IV. FUTURE SCOPE

Looking into the future, this kind of technology has a lot of scope since it can make the devices completely hands-free. But the only drawback of this is its efficiency. Looking at all the research done until date in the industry, the efficiency of this system is only found to be 40-50% for text-independent speaker identification systems, which is not feasible. Although text-dependent systems have higher accuracy reaching almost 90%, but they don’t have many applications other the for security systems and are not quite user-friendly. The modelling techniques Support Vector Machines (SVM) along with Gaussian Mixture Models (GMM) provide yield good performance for text-independent systems whereas Vector Quantization (VQ) is mostly applied in text-dependent speaker verification systems.

V. 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 Environmnet” Multimedia Tools and Applications Volume 76 issue