Speaker Recognition using Random Forest

Khadar Nawas K[1*], *Manish Kumar Barik*[1], *A Nayeemulla Khan*[1]

[1]*School of Computer Sciences and Engineering, Vellore Institute of Technology, Chennai, India*

Abstract. Speaker identification has become a mainstream technology in the field of machine learning that involves determining the identity of a speaker from his/her speech sample. A person's speech note contains many features that can be used to discriminate his/her identity. A model that can identify a speaker has wide applications such as biometric authentication, security, forensics and human-machine interaction. This paper implements a speaker identification system based on Random Forest as a classifier to identify the various speakers using MFCC and RPS as feature extraction techniques. The output obtained from the Random Forest classifier shows promising result. It is observed that the accuracy level is significantly higher in MFCC as compared to the RPS technique on the data taken from the well-known TIMIT corpus dataset.

1 Introduction

A lot of extensive research has been done in the past on the field of speaker recognition in order to achieve a common goal– efficiently identifying a speaker using his/her speech note[1][4]. Speaker Identification systems work on the fact that no two persons in the world have exactly the same voice. Each individual's voice is different as it is attributed to different biological features such as- pitch of the sound, vocal tract's length, sound frequency, the movement of the tongue against the palate, the shaping of the lips, the arrangement of teeth, etc[9].

A state-of-the-art speaker recognition system could really have a significant impact in the field of security. One important application of this would be to identify criminals through their telephonic conversations by listening to some vulnerable keywords[13]. This can eventually save a lot of police work.

Speaker recognition can further be categorized into speaker identification and speaker verification. This paper focuses on the implementation of speaker identification using Random Forest algorithm as a classifier. There are basically three phases in a speaker recognition system: Pre-processing the audio signals, feature extraction and the matching/classification phase. Pre-processing refers to the phase where the inputted audio signal is cleansed, i.e. the signal is made free of the environmental noise in order to efficiently recognize the significant voice signal. The phase of feature extraction involves generating the feature vectors for the given audio signal which constitutes the key characteristics of any individual's voice.

* Corresponding author: khadarnawas.k@vit.ac.in

After the corresponding feature vectors have been obtained for each speaker, these are then modelled to get the specific template for that speaker. Classification algorithms determine distinguishing characteristics among the speakers based on the features obtained. Lastly the testing stage matches the speaker's speech note to the existing classes. The match with the highest score is then assigned to the specified speaker.

The rest of the paper comprises of the below sections. Section II describes the related works and background study for this paper. Section III discusses the dataset used for the model experiment. Section IV illustrates the pre-processing phase of the voice signal, Section V discusses the two feature extraction methods-RPS and MFCC, Section VI discusses the various classification techniques with emphasis on the Random Forest classifier used in our model, Section VII shows the experimental results obtained from the dataset, Section VIII concludes the rest of the paper by citing important inferences obtained from the two methodologies implemented.

2 Related Works and Background

Speech recognition can be mainly divided into three categories: speech recognition (determining the actual content of the spoken utterance), speaker recognition (identifying the speaker's voice) and language recognition (determining the language of the spoken word)[10]. Speaker recognition, a sub-domain of speech recognition can further be categorized into speaker verification and speaker identification[5]. The former deals with confirmation of the affirmed identity by matching

his/her voice characteristics with the one already stored in the system (1:1 relationship)[7]. The latter approach aims to determine the correct voice print of the corresponding speaker from a set of known speakers (1:N relationship). SID system can be implemented as per a text-dependent (where the utterances given by the speaker are already known by the system) and a text-independent system (where the utterances are not stored anywhere and the speaker is identified using only his voice features).

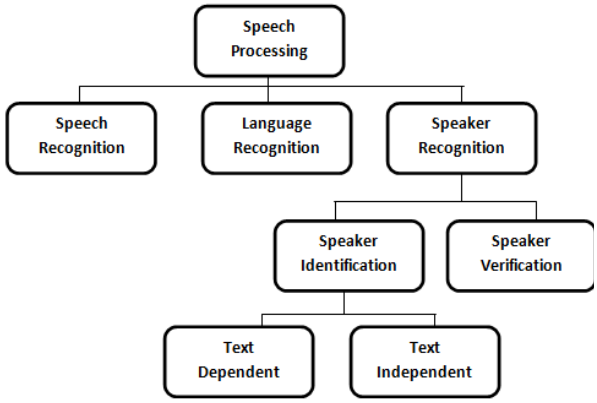


Figure 1. Classification of speaker recognition

3 Dataset

Every machine learning model needs a quality data set in order to operate efficiently. In our implementation of a SID system, we worked with a downloaded dataset, viz. as TIMIT DATASET. TIMIT consists of broadband recordings of 630 speakers of eight major dialects of American English, each further consisting of ten phonetically rich sentences.

We take a set of 38 speakers from the TIMIT dataset. Every speaker's voice notes are contained in a directory named as speaker_dataset that contains speakers labelled as 0 to 37. Each directory contains ten phonetically rich sentences of that particular speaker. So, a total of 380 voice notes will be processed for the extraction of characteristics that are specific to that particular speaker.

4 Pre-processing

Pre-processing is the first phase in the process of recognizing a speaker. It involves the removal of unwanted components such as background noise, unwanted silence in between the uttered words, etc [2]. An efficient speaker identification system must be able to extract the clean speech from the corrupted one. The two most popular methods of pre-processing the speech signal involves spectral subtraction and adaptive noise cancellation[2]. Spectral Subtraction is a method in which non-speech regions in the speech are estimated and are then the average of those regions is subtracted from the speech signal. The efficiency of a pre-processing phase is validated in terms of SNR(Signal-to-Noise Ratio). It refers to the ratio of the power of the correct signal to the noise.

$$SNR = 20 \log_{10}(V_{signal} / V_{noise})$$

where V_{signal} is the voltage of the input signal and V_{noise} is the voltage of the noise.

5 Feature Extraction

Feature extraction is the phase where the pre-processed signal is fed to the model and feature vectors are obtained. Features, in the context of machine learning refers to the numerical characteristics of any given object. In this case the feature vectors can be referred to contain distinct entities of any given speaker such as pitch, amplitude, etc. Feature extraction can be considered as a dimensionality reduction process since the input data fed to the system are too large to be processed and are highly redundant in nature. In our paper, we provide an in depth comparison of two popular feature extraction techniques-MFCC and RPS.

5.1 RPS (Reconstructed Phase Space)

Speaker Identification systems typically uses spectral-based features that reject the state of the phase present in speech signals[18]. Thus, employing some extra features(such as attributing to the turbulence of speech) where the signal phase is not rejected may fill this gap. Moreover, ongoing research on this field have indicated that using nonlinear features may improve the exhibitions of SID systems [16][17]. In one dimensions, the state on any particle can be determined by specifying its position and velocity. In this case, its phase space is a plane. In the case of a voice signal, we have a time series instead of a phase space object. Hence, we need to convert the observations into state vectors. This is phase space reconstruction. RPS method is a multidimensional space where the coordinates are produced by a shift and delay of a sample of a 1-d signal in a time series[6]. In our approach, we embed the speech signal in the reconstructed phase space in order to extract some useful non-linear features. The RPS Distribution plot (scatter graph) in two dimensions is constructed for the vowel /a/ is given in Fig 2.

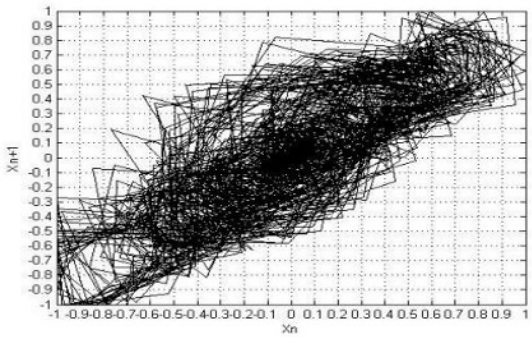


Figure 2. RPS model for the vowel 'a' sound

5.2 MFCC(Mel-Frequency Cepstral Coefficients)

MFCC is one of the most popular and widely used techniques in the field of feature extraction. In our implementation we make use of the librosa library to obtain 40 MFCC features of each speaker's voice sample. This section illustrates the various techniques used to extract features from an audio signal as described in Figure 3.

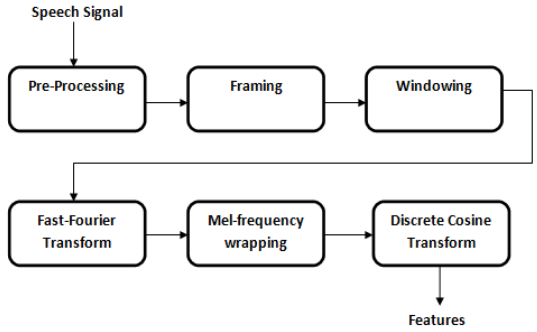


Figure 3. MFCC block diagram

The various phases that are a part of MFCC technique are:

5.2.1 Framing

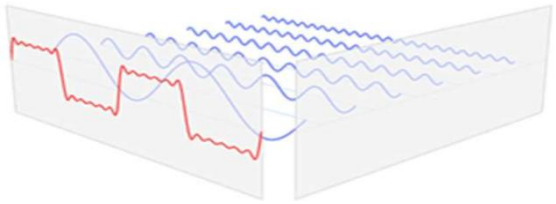
A speech signals uttered by the user are continuously changing. So, in order to process the signal, we segment it at duration of 25ms such that the signal in a frame becomes quasi-stationary (exhibiting stationary behaviour). This segment often corresponds to a particular phoneme such as "duh" in "duck". Further, to avoid discontinuities among consecutive frames, every two frames are mutually overlapped.

5.2.2 Windowing

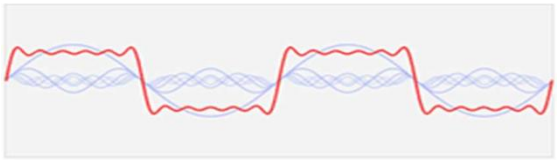
Windowing is done in order to minimize the inconsistent fluctuations at the edges of each frame also referred to as spectral distortion [9]. It is implemented by using a window function that lowers down the value to zero at the edges of each frame.

5.2.3 Fast Fourier Transform

This stage converts the time domain signal(Figure 5) into the corresponding frequency domain signal(Figure 4). This phase basically disintegrates the composite signal into the constituent frequencies that make up the composite signal. The output frequencies obtained are referred to as the spectrum of the given frame. The spectrum of all the frames in the given signal form a spectrogram.



(a)



(b)

Figure 4. Fast-Fourier Transform converts a signal in time domain (a) to a signal in frequency domain (b).

5.2.4 Mel-Frequency wrapping

Several studies have shown that humans perceive frequency not at a linear scale. At low frequencies, the difference in comparing two frequencies as perceived by the human ear is less than the same difference in the frequencies at a higher frequency range [9] as shown in Fig 6. This demands for a different type of scale known as mel scale that sets a threshold frequency below which the mel scale behaves linearly(at low frequencies) and above that logarithmic behaviour(at high frequencies).

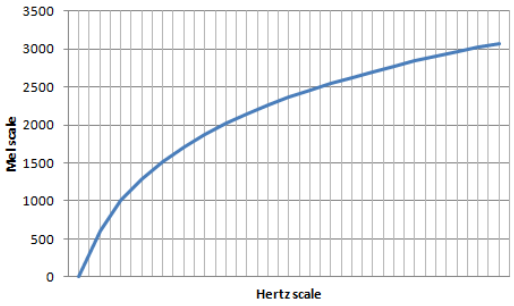


Figure 5. pitch scale vs. mel scale

This is simulated using mel frequency filter bank which has mel frequency scale that converts each derived frequency into its corresponding mel-

frequency [11]. Mel can be calculated using the below formula:

$$\text{Mel}(f) = 2595 \log_{10}(1+f/700)$$

5.2.5 Discrete Cosine Transform.

The log spectrum thus obtained contains a lot of frequencies. So, in order to remove the whole lot of dimensions, we implement dimensionality reduction in which we combine frequencies close to each other. DCT is basically used to convert the obtained log Mel spectrum into time domain known as cepstrum[3]. The output of DCT gives what is referred to as mel frequency cosine coefficients. The set of these coefficients form the feature vectors are ready for modelling.

6 Classification

The extracted MFCC features are then classified into various classes that corresponds to different speakers. Over the years several classification algorithms have been proposed in the task of speaker identification, all with different rates of success. Most of the popular ones include Support Vector Machine(SVM), K-Nearest Neighbours, Random Forest(RF), K-means and the latest among them Deep Neural Networks (DNN)[8]. In our experiment, we split the features extracted from the speaker dataset as per 75% as training data and 25% as testing data that are then fed onto the Random Forest classifier.

Support Vector Machine is one the most widely known ML algorthim that is used in classification purposes. SVM tries to classify the given set of speakers by trying to come up with a hyperplane that seperates the characteristics the constitutes a particular speaker. One drawback of SVM is it is not particularly suitable for large datasets (size > 10,000). Our feature extraction model of 380 voice wav files generate a massive number of rows around 1,99,000 while using MFCC extraction technique and around 1,89,00,000 in case of features obtained from that of the Reconstructed Phase Space(RPS) technique. So, we opt for a different approach that doesn't really hit the performance issues with larger datasets[15]. In this paper we implemented the Random Forest classification model in order to classify the given speakers' feature. Apart from working efficiently with large datasets, fellow researches in their work [12] proposed RF classifiers as the best performing among 179 classifiers in terms of accuracy and precision. Also, in [14] RF classifiers showed better performances compared to KNN and SVM classifiers in terms of accuracy, precision, recall and f-measure.

Random Forest algorithm is a type of supervised classification algorithm that uses a large combination of decision trees and then trains each of them on a different set of observations. The final predictions are done on the random forest based on the calculations done by averaging each prediction of a single tree. A random forest is considered better than a single decision tree because of the fact that this reduces the over-fitting problem which can be produce really inefficient results among various other classification techniques.

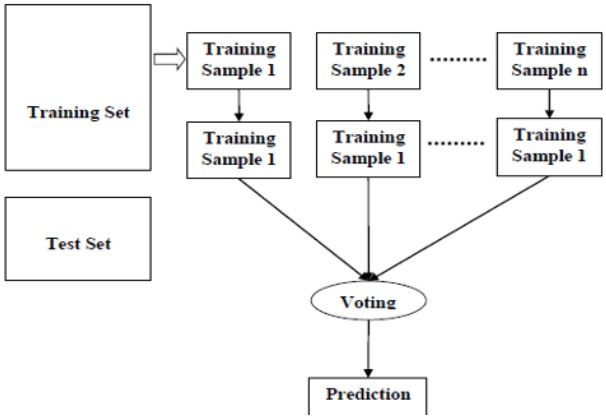


Figure 6. Phases in a Random Forest classifier

7 Experimental results

Table 1. Comparing the two techniques based on accuracy

Speaker Recognition	Training accuracy	Testing accuracy
Random Forest using RPS features	73.6%	71%
Random Forest using MFCC features	100%	97%

Table 1 shows the results obtained from the proposed two methodologies in system identification system using a Random Forest algorithm to classify the features extracted from Mel Frequency Cepstral Coefficient(MFCC) technique and Reconstructed Phase Space(RPS). The MFCC approach shows promising results with an accuracy rate of 100% on training data and 97.4% on testing data as compared to RPS methodology that showed an accuracy of 73.6% on training data and 71% on testing data.

8 Conclusion and Future work

This paper reported the development of a Speaker Identification System using a RF classifier. The paper also briefly described all the stages covered from pre-processing of a speech wav file to eventually identifying the speaker. We infer from the results obtained that spectral features such as MFCC may prove its benefits in showing accurate results and could not be easily replaced by some novel features. But, as a starting point, we propose that future speaker identification software may include a combination of non-linear features(RPS) and the traditional spectral features(MFCC) in order to further improve the performance parameters of speaker identification systems.

References

1. H. S. Jayanna and S. R. Mahadeva Prasanna, May 2009, "Analysis, Feature Extraction, Modeling and Testing Techniques for Speaker Recognition".
2. Yakubu A. Ibrahim, Juliet C. Odiketa, Tunji S. Ibiyemi, 2017, "Preprocessing technique in automatic speech recognition for human computer interaction: An overview", pp186-191
3. Garima Vyas, Barkha Kumari, June 2013, "Speaker Recognition System Based on MFCC and DCT", Vol. 2**,** Issue 5, pp167-169.
4. Jyoti B. Ramgire, Prof. Sumati M.Jagdale, April 2016, "A Survey on Speaker Recognition With Various Feature Extraction And Classification Techniques", Vol. 03, Issue 04, pp709-712.
5. Satyam P. Todkar, Snehal S. Babar, Rudrendra U. Ambike, Prasad B. Suryakar, April 2018, "Speaker Recognition Techniques: A review", pp1-5.
6. Nisha.V.S , M.Jayasheela, February 2013,"Speaker Identification Using Combined MFCC and Phase Information", Vol. 2, Issue 2, pp1149-1152.
7. Shilpa S. Jagtap and D.G.Bhalke, 2015, "Speaker Verification Using Gaussian Mixture Model".
8. Sreenivas Sremath Tirumala, Seyed Reza Shahamiri, November 2016, "A review on Deep Learning approaches in Speaker Identification".
9. https://shodhganga.inflibnet.ac.in/bitstream/10603/183865/9/09-chapter%203.pdf, "From Speech to Feature Vectors", pp83-128.
10. Joseph P. Campbell, September 1997, "Speaker Recognition: A Tutorial", Vol. 85, no. 9.
11. Fang-Yie Leu, Guan-Liang Lin, 2017, "An MFCC-based Speaker Identification System", pp1055-1062.
12. Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, 2014, "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?" .
13. Parashar Dhakal, Praveen Damacharla, Ahmad Y. Javaid, Vijay Devabhaktuni, 2019, "A Near Real-Time Automatic Speaker Recognition Architecture for Voice-Based User Interface".
14. Tumisho Billson Mokgonyane, Tshephisho Joseph Sefara, Mercy Mosibudi Mogale, Madimetja Jonas Manamela, 2019, "Automatic Speaker Recognition System based on Machine Learning Algorithms".
15. Mohammed Zakariah ,September 2014, "Classification of large datasets using Random Forest Algorithm in various applications: Survey".
16. Povinelli RJ, Johnson MT, Lindgren AC, Roberts FM, 2006, "Statistical models of reconstructed phase spaces for signal classification", IEEE Trans Signal Process 2006; 54:2178–86.
17. Yasser Shekofteh and Farshad Almasganj, 2013, "Feature Extraction Based on Speech Attractors in the Reconstructed Phase Space for Automatic Speech Recognition Systems".
18. http://shodhganga.inflibnet.ac.in/bitstream/10603/213852/16/16_chapter7.pdf, "Effective Speaker Spotting based on Nonlinear Properties of Vocal Tract" , pp161-180.