TEXT-INDEPENDENT SPEAKER IDENTIFICATION SYSTEM USING AVERAGE PITCH AND FORMANT ANALYSIS

M. A. Bashar[1], Md. Tofael Ahmed[2], Md. Syduzzaman[3], Pritam Jyoti Ray[4] and A. Z. M. Touhidul Islam[5]

[1]Department of Computer Science & Engineering, Comilla University, Bangladesh
[2]Department of Information & Communication Technology, Comilla University, Bangladesh
[3,4]Department of Computer Science and Engineering, SUST, Bangladesh
[5]Department of Information & Communication Engineering, University of Rajshahi, Bangladesh

ABSTRACT

The aim of this paper is to design a closed-set text-independent Speaker Identification system using average pitch and speech features from formant analysis. The speech features represented by the speech signal are potentially characterized by formant analysis (Power Spectral Density). In this paper we have designed two methods: one for average pitch estimation based on Autocorrelation and other for formant analysis. The average pitches of speech signals are calculated and employed with formant analysis. From the performance comparison of the proposed method with some of the existing methods, it is evident that the designed speaker identification system with the proposed method is superior to others.

KEYWORDS

Speaker identification, average pitch, feature extraction, formant analysis

1. INTRODUCTION

Speaker Identification (SI) refers to the process of identifying an individual by extracting and processing information from his/her speech. It is a task of finding the best-matching speaker for unknown speaker from a database of known speakers [1,2]. It is mainly a part of the speech processing, stemmed from digital signal processing and the SI system enables people to have secure information and property access.

Speaker Identification method can be divided into two categories. In Open Set SI, a reference model for the unknown speaker may not exist and, thus, an additional decision alternative, "the unknown does not match any of the models", is required [3]. On the other hand, in Closed Set SI, a set of N distinct speaker models may be stored in the identification system by extracting abstract parameters from the speech samples of N speakers. In speaker identification task, similar parameters from new speech input are extracted first and then decide which one of the N known speakers mostly matches with the input speech parameters [3-6].
One can divide Speaker Identification methods into two: Text-dependent and Text-independent methods. Although text-dependent method requires speaker to provide utterances of the key words or sentences which have the same text for both the training and identification trials, the

DOI : 10.5121/ijit.2014.3303

text-independent method does not rely on a specific text being spoken.

The aim of this work is to design a closed-set and text-independent Speaker Identification System (SIS). The SIS system has been developed using Matlab programming language [7-8].

2. RELATED WORKS

A brief review of relevant work of this paper is stated as follows. Authors in Ref. [9] studied the performance of text-independent, multilingual speaker identification system using MFCC feature, pitch based DMFCC feature and the combination of these two features. They shown that combination of features modeled on the human vocal tract and auditory system provides better performance than individual component model. Their study also revealed that Gaussian Mixture Model (GMM) is efficient for language and text-independent speaker identification. Reynolds et al. [10] shown that GMM provide a robust speaker representation for the text-independent speaker identification using corrupted, unconstrained speech.

The authors in Ref. [11] implemented a robust and secure text-independent voice recognition system using three levels of encryption for data security and autocorrelation based approach to find the pitch of the sample. Their proposed algorithm outperforms the conventional algorithms in actual identification tasks even under noisy environments.

3. SPEAKER IDENTIFICATION CONCEPT

The overall architecture of Speaker Identification System is illustrated in Fig. 1.

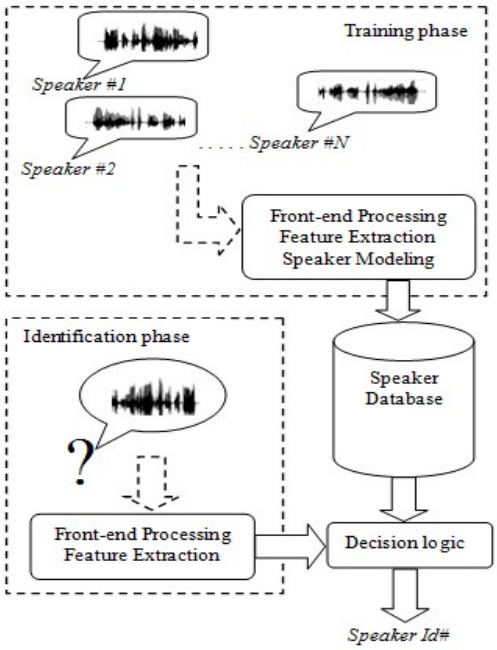

Figure 1. System architecture of closed-set and text-independent SIS.

From the above figure we can see that a Speaker Identification system is composed of the following modules:

a) Front-end processing: It is the "signal processing" part, which converts the sampled speech signal into set of feature vectors, which characterize the properties of speech that can separate different speakers. Front-end processing is performed both in training and identification phases.
b) Speaker modeling: It performs a reduction of feature data by modeling the distributions of the feature vectors.
c) Speaker database: The speaker models are stored here.
d) Decision logic: It makes the final decision about the identity of the speaker by comparing unknown speaker to all models in the database and selecting the best matching model.

Among several speech parameterization methods, we focus on average pitch estimation based on auto-correlation method. There are many classification approaches, but all have some limitations at some particular field. At present the state-of-art classification engine in the Speaker Identification technology are the Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Vector Quantization (VQ), Artificial Neural Network (ANN) and Formant [12]. In this paper the formant analysis is based on power spectral density (PSD).

4. AVERAGE PITCH ESTIMATION

Pitch represents the perceived fundamental frequency (F0) of a sound and is one of the major auditory attributes of sounds along with loudness and quality [13-14]. Here we are interested to find out the average pitch of a speech signal. A method is designed for estimating average pitch. We named this method Avgpitch. The flowchart of Avgpitch is shown in Fig. 2.

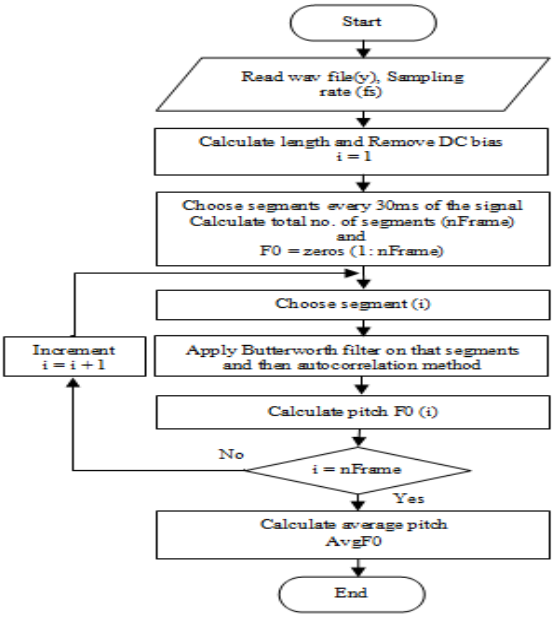

Figure 2. Flowchart of average Pitch estimation (Avgpitch).

Average pitch was used to reduce the comparison task in formant analysis. We calculated average pitch for “speaker.wav” (the unknown speaker in identification phase) file as well as for all trained files in speaker database. Pitch contour and average pitch (158.6062Hz) of “speaker.wav” file is shown in Fig. 3.

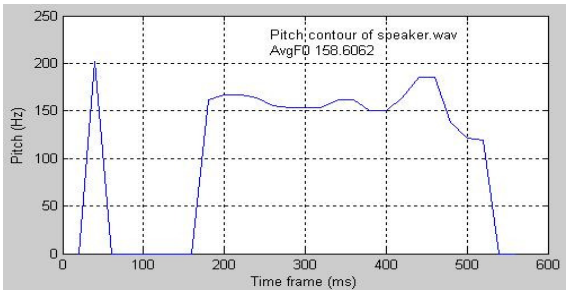

Figure 3. Pitch outline of "speaker.wav" file.

Then we calculated average pitch differences between the "speaker.wav" file and all the trained speech files. To illustrate this with figure we used 40 trained files in database. Fig. 4 shows average pitch differences between the unknown speaker and 40 trained speakers.

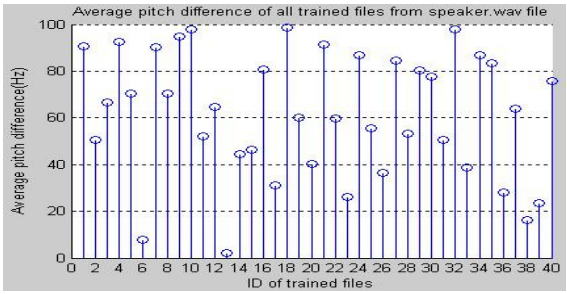

Figure 4. Plot of average pitch differences of 40 trained files from "speaker.wav" file.

Fig. 4 gives us a closer look in identification task. We can see that some of the differences are small enough while others are so high. As the average pitch differences could potentially characterize a speaker so we can prune out some of trained files with high average pitch differences from our consideration. Actually in our proposed system we discard a significant number of trained files based on a certain difference limit (roughly above 40Hz). And rest of the trained files are used in next consideration, that is, for formant analysis. From Fig. 4 we can see 10 speakers are with ID (in orderly) 13, 6, 38, 39, 21, 36, 17, 26, 31 and 20 whose average pitch differences are not more than 40 Hz. So we will do formant analysis on these ten selected trained files to identify the best match speaker ID for the unknown speaker (speaker.wav file).

5. FORMANT ANALYSIS

Formants are the meaningful frequency components of human speech [3]. The information that humans require to distinguish between vowels can be represented by the frequency content of the vowel sounds. In speech, these are the characteristic part that identifies vowels to the listener. We designed an algorithm for formants analysis. The flowchart of formant analysis algorithm is presented in Fig. 5.

Applying this algorithm we get the PSD of speech signal. The vector position of the peaks in the power spectral density is also calculated that can be used to characterize a particular voice file. Fig. 6 shows first four peaks in power spectral density of "speaker.wav" file.

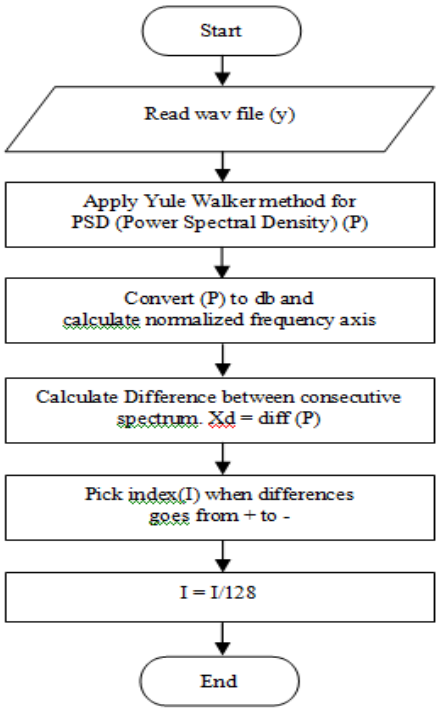

Figure 5: Flowchart of Formant Analysis.

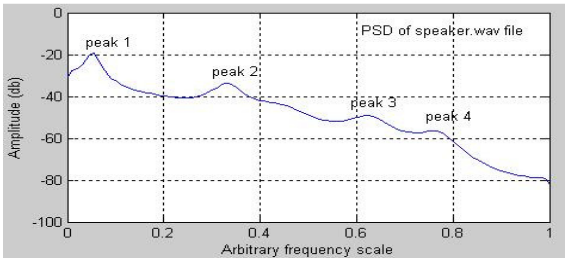

Figure 6. Plot of the first four peaks in power spectral density of "speaker.wav" file.

Formant analysis was also done on ten selected trained speaker files getting from the previous section. Fig. 7 shows the PSD of ten trained speaker files with ID 13, 6, 38, 39, 21, 36, 17, 26, 31, and 20 respectively. We calculated formant vector (vector positions of peaks) of "speaker.wav" file as well as of ten selected trained files. The purpose of these formant vectors is to find out the difference of peaks between the "speaker.wav" file and all other trained files. Then the root mean square (rms) value of the differences is calculated each time to get the single value of formant peak difference. Fig. 8 shows the formant peak differences of ten selected trained files from "speaker.wav" file.

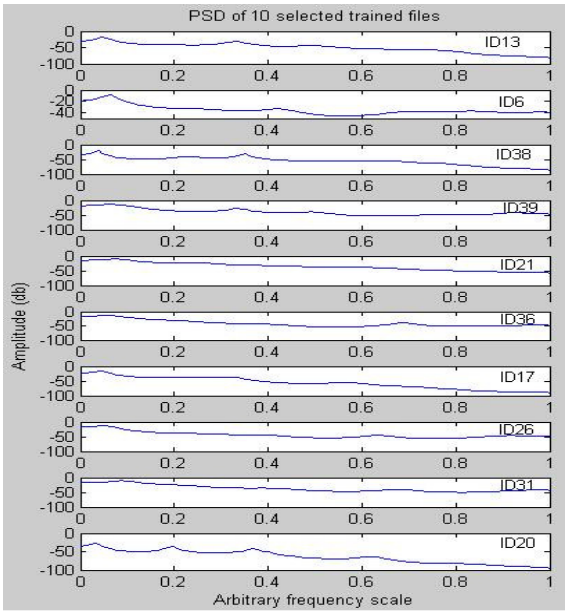

Figure 7. PSD of ten selected trained files (ID 13, 6, 38, 39, 21, 36, 17, 26, 31 and 20).

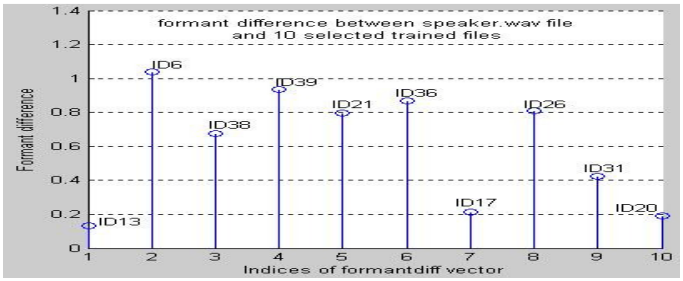

Figure 8. Plot of formant peak differences between “speaker.wav” file and ten selected trained files.

6. RESULTS AND DISCUSSION

Using the information obtained from Fig 8, the result of this system could easily be found. The ID of speaker that has the minimum formant difference should be the best matched speaker for the unknown speaker (speaker.wav). From Fig. 8 we can see that the lowest formant difference is for speaker ID13. The next best matching speakers are found easily from the sorted formant difference vector between “speaker.wav” file and ten selected trained files. This is shown in Fig. 9. From Fig. 9 we get the best matching speakers with ID 13, 20, 17, 31, 38, 21, 26, 36, 39 and 6 respectively. We checked out the trained file with ID 13 and the unknown speaker (speaker.wav) and found that two voices are of the same speaker.

The Speaker Identification code has been written using the MATLAB. It was found that comparison based on average pitch helped us to reduce the number of trained file to be compared in formant analysis. And comparison based on formant analysis produced results with most accuracy.

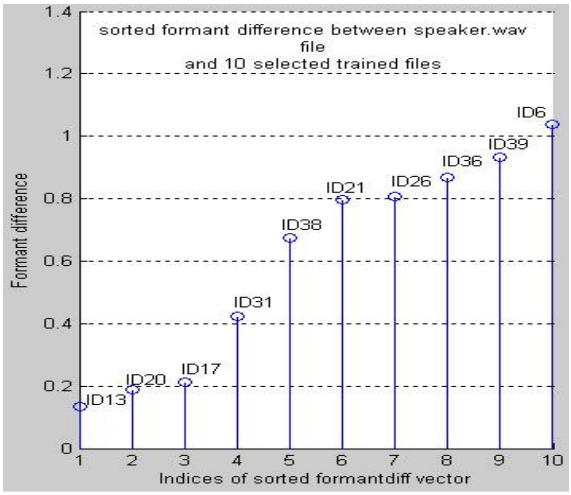

Figure 9. Plot of formant peak differences between "speaker.wav" file and ten selected trained files.

To verify the performance of the proposed Speaker Identification system, the speech signals of 80 speakers are recorded in the laboratory environment. For identification phase some speech signals also recorded in laboratory and in noisy environment as well. We got about 90% accuracy for normal voices (in laboratory environment). We got about 75% accuracy for the twisted (change the form of speaking style) voice in identification phase and about 70% when the testing signal is noisy.

7. CONCLUSIONS

In this paper a closed-set text-independent Speaker Identification system has been proposed using average pitch and formant analysis. The highest Speaker Identification accuracy is 91.75%, which satisfies the practical demands. All experiments were done in a laboratory environment which was not fully noise proof. The accuracy of this system will increase considerably in a fully noise proof environment. We successfully extracted feature parameters of each speech signal with the MATLAB implementation of feature extraction. For characterizing the signal, it was broken down into discrete parameters because it can significantly reduce memory required for storing the signal data. It can also shorten computation time because only a small, finite set of numbers are used for parallel comparison of speakers' identities. We hope that may be one day, we will expand this work and make an even better version of Speaker Identification system.

REFERENCES

[1] K. Shikano, "Text-Independent Speaker Recognition Experiments using Codebooks in Vector Quantization", CMU Dept. of Computer Science, April 9, 1985.
[2] S. Furui, "An overview of Speaker Recognition Technology", ESCA workshop on Au tomatic Speaker Recognition, Identification and Verification, 1994.
[3] Wikipedia. http://en.wikipedia.org/wiki/.
[4] Lincoln Mike, "Characterization of Speakers for Improved Automatic Speech Recogni tion", Thesis paper, University of East Anglia, 1999.
[5] B. Atal "Automatic Recognition of Speakers from Their Voices", Proceedings of the IEEE, vol. 64, April 1976, pp. 460-475.
[6] H. Poor, "An Introduction to Signal Detection and Estimation", New York: Springer-Verlag, 1985.
[7] Royce Chan and Michael Ko, Speaker Identification by MATLAB, June 14, 2000.

[8] Vinay K. Ingle, John G. Proakis, "Digital Signal Processing Using Matlab V4", PWS Publishing Company, 1997.
[9] Todor Dim itrov Ganchev, "Speaker Recognition", PhD Thesis, Wire Communication Laboratory, Dept. of Computer Science and Engineering, University of Patras, Greece, November 2005.
[10] S. S. Nidhyananthan and R. S. Kumari, "Language and Text-Independent Speaker Identification System using GMM", WSEAS Transactions on Signal Processing, Vol.9, pp. 185-194, 2013.
[11] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models", IEEE Transactions on Speech and Audio Processing, Vol. 3, pp. 72-83, 1995.
[12] A. Chadha, D. Jyoti, and M. M. Roja, "Text-Independent Speaker Recognition for Low SNR Environments with Encryption", International Journal of Computer Applications, Vol. 31, pp. 43-50, 2011.
[13] D. Gerhard. "Pitch Extraction and Fundamental Frequency: History and Current Tech-niques", technical report, Dept. of Computer Science, University of Regina, 2003.
[14] Dmitry Terez, "Fundamental frequency estimation using signal embedding in state space". Journal of the Acoustical Society of America, 112(5):2279, November 2002.