ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Audio Signal Recognition System Based On Vocal Features.

A Jose Albin *, and NM Nandhitha.

*Research Scholar, Faculty of Computing,Sathyabama University, Tamil Nadu, India.
Professor, Faculty of Electrical & Electronics, Sathyabama University, Tamil Nadu, India.

ABSTRACT

In biometrics, identification of an individual human being is based on the speech signal generated by their speech production system. In this paper, a novel text independent speaker recognition system using the physiological characteristics of human speech production system is developed. Pitch, fundamental period and the number of peaks are the features considered for the development of speaker recognition system. Feature extraction based on Auto-correlation, Discrete Wavelet Transform and Cepstrum techniques are performed. Evaluation of the system performance for the various feature extraction techniques is done in terms ofsensitivity.

Keywords: Speaker Recognition System, Pitch, Fundamental Period, Number Of Peaks, Auto-Correlation, Discrete Wavelet Transform, Cepstrum, Sensitivity.

**Corresponding author*

INTRODUCTION

Speaker recognition system is a biometric process for the recognition of an individual's voice. The biometric process identifies the speaker based on their speech waveform produced by the human speech production system and the structure of their vocal tract [1-4].The voluntary movement of the anatomical structures present in our human speech production system generates an acoustic sound pressure wave called the speech waveform. Vocal folds, soft palate, tongue, teeth and lips are the finer anatomical components involved in the human speech production [5]. These finer anatomical components are also called as articulators. Based on the movements and positionof the articulators, the modulation occurs in the speech waveform. Larynx plays an important role in speech production as it provides the periodic excitation to the system for speech sound called as voice. Since the human system varies with time, the spectral characteristics of the speech waveform possess non stationary property.The prosodic features involved in the study of voicing are the fundamental period, fundamental frequency and pitch [6-8].

Feature extraction and classification are the two main components present in any speaker recognition system. Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), Linear Predictive Cepstral Coefficients (LPCC), Bark Frequency Cepstral Coefficient (BFCC), Revised Perceptual Linear Prediction (RPLP), Cepstral Coefficients were the feature extraction techniques. Vector quantization, Hidden Markov Model, Gaussian Mixture Models, Adaptive Resonance Theory, Back Propagation Network and State Vector Machine were the classification techniques [6, 7, 9-15].The robustness of any speaker recognition system depends on the appropriate selection of acoustic features, feature extraction technique and the classifier involved in the system. In this paper, an approach towards a novel text independent speaker recognition system is made by considering the features like pitch, fundamental period and number of peaks obtained by the anatomical structures present in our human speech production system. Auto correlation, cepstrum and Discrete Wavelet Transform are the feature extraction techniques used. Adaptive Resonance Theory (ART) is the classifier used for pattern matching. Performance is measured in terms of sensitivity.

This paper is organized as follows: In section 2, an overview of statistical metrics is presented. In section 3, literature survey is presented. The proposed system is presented in section 4. Results and discussions are dealt in section 5. Section 6 concludes this paper with future work.

OVERVIEW OF STATISTICAL METRICS TO REPRESENT VOICE FEATURES

Speech signal provides glottal information of the speaker, which by proper feature extraction gives the identity of the speaker. The features namely fundamental period, fundamental frequency, pitch and number of peaksfor a smaller co articulation of the speakers along with a suitable feature extraction technique is responsible to explore the speaker identity [7]**.**Fundamental period T, is the time taken between two vocal fold openings and depends on the size and tension of the speaker's vocal folds. The rate of vibration is said to be the fundamental frequency and is obtained by finding the reciprocal of T. The perceived fundamental frequency of a sound is termed as pitch. Every speaker has a pitch range depending on their larynx [4]. So by obtaining the voicing parameters of an unknown human or speaker through feature extraction techniques like autocorrelation, cepstrum and Discrete Wavelet Transform, an effective speaker recognition system can identify the speaker.

The measure of similarity between two waveforms is correlation. For a frame of speech signal { x(n), n = 0,1,..,N-1}, the autocorrelation function is given by equation 1 [16].

$$X(k) = \frac{1}{N} \sum_{n=0}^{N-1-k} x(n)x(n+k), \quad k=0,1,..,N-1 \quad (1)$$

The cepstrum of a signal is defined as the inverse Discrete Fourier Transform of the log magnitude of the DFT of the signal. For a frame of speech {x(n), n = 0,1,..,N-1}, the cepstrum function is defined by equation 2 [6].

$$C(n) = \sum_{n=0}^{N-1} \log\left(\sum_{n=0}^{N-1} \left| x(n)e^{-j\frac{2\pi}{N}kn} \right| \right) e^{j\frac{2\pi}{N}kn} \qquad (2)$$

DWT is used to obtain a time frequency spectrum of a speech signal. For a frame of speech signal {x(n), n = 0,1,..,N-1}, the wavelet transform is defined by equation 3.

$$\tilde{x}_{\psi}(a,b) = \frac{1}{\sqrt{a}} \int_{\infty}^{-\infty} x(t)\psi(\frac{t - b}{a})dt \qquad (3)$$

Wavelet spectrum of a signal x (t) is $\tilde{x}_{\psi}(a,b)$, scale is represented by 'a' and time translation by 'b'. $\psi(t)$ is the considered wavelet of interest [17].

RELATED WORKS

Li et al (2014) proposed a new architecture that incorporates cost sensitive learning technology for speaker verification system. The system measures the discontinuities in the pitch envelope for estimating speaker's fundamental frequency at various emotions [18].

Bhandavle et al (2014) presented emotion based speaker recognition system using gender based modified Mel Frequency Cepstral Coefficients (MFCC). The emotional features namely pitch, intensity and frequency were extracted for speaker recognition usingVector Quantization with K-means algorithm as classifier. Higher accuracy was achieved only for small databases [19].

Rathore and Tripathi (2014) proposed a multilingual person identification system considering the features namely pitch and formant frequency. For classification the authors used neural network. Speaker identification along with their gender and language was identified by their proposed system [20].

Sharma et al (2014) developed speaker and gender recognition system. MFCC and delta MFCC were used for feature extraction. Radial basis function network and back propagation algorithm is used of classification. Radial basis function network outperformed back propagation algorithm as a classifier for multilingual speaker and gender recognition system [21].

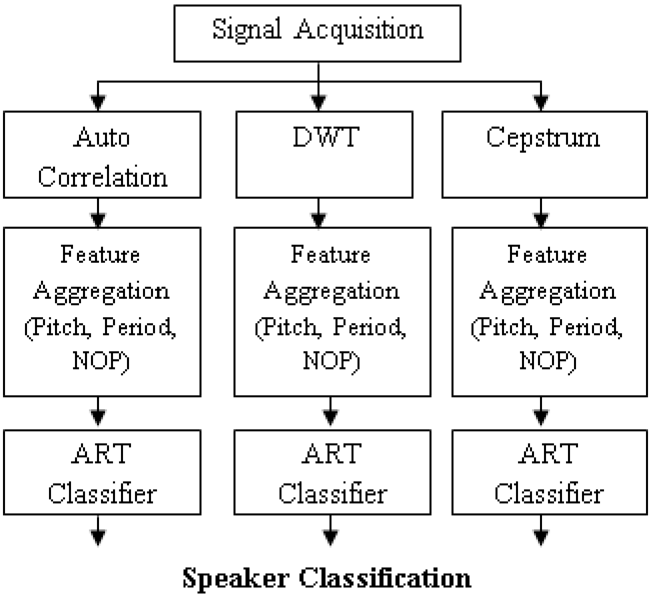
Dixit and Mulge (2014) proposed a noise effective approach for a speaker recognition system. Spectral subtraction method was used for high level noise filtering and LPC for feature extraction process. Hidden Markov Model (HMM) and Neural network were used for classification [22].

Kovoor, Supriya and Jacob (2014) implemented amultimodal behavioral biometric authentication system that considers human speech and handwritten signature.MFCCs, Spectral Flux, Spectral Centroid and Spectral Roll off were used for feature extraction. VQ was the classifier incorporated in the system.Multimodal authentication system provided better results thanunimodel system [23].

PROPOSED WORK

Ten speech signals from 10 different speakers are acquired and a research database is maintained. The prosodic features namely Pitch, Fundamental Period and the Number Of Peaks (NOP) occurring in the signal for a particular frame are extracted from the acquired speech signals. Auto correlation, cepstrum and Discrete Wavelet Transform (DWT) with Discrete Meyer wavelet (Dmey) are used for feature extraction process. The extracted features are given to an ART classifier for an effective speaker recognition.ART classifier is trained with different datasets (autocorrelation features, DWT features, cepstrum features, auto correlation & DWT features, DWT &cepstrum features, autocorrelation &cepstrum features, autocorrelation & DWT &cepstrum features). The effectiveness of the extracted features is evaluated using sensitivity.

Figure 1: System Architecture for the proposed system


Signal Acquisition
Auto Correlation
DWT
Cepstrum
Feature Aggregation (Pitch, Period, NOP)
Feature Aggregation (Pitch, Period, NOP)
Feature Aggregation (Pitch, Period, NOP)
ART Classifier
ART Classifier
ART Classifier
Speaker Classification

Case 1: Feature Extraction using Autocorrelation features

Auto correlation functionis an invulnerable solution for pitch estimation in noisy environment. Since speech is a non stationary signal, instead of long frame length autocorrelation measurement, a short frame length autocorrelation measurement gives better results [24]. A frame length of 1024 sec at t=10000sec is considered for all the speech signals, and autocorrelation function is applied for feature extraction over that region of interest. Pitch, fundamental period and number of peaks were extracted from the obtained correlation coefficients. The extracted features were normalized and given to an ART classifier.

Case 2: Feature Extraction using DWT features

DWT decomposes the speech signal into approximation and detailed coefficients. Discrete Meyer (Dmey) is the chosen wavelet since it provides maximum number of lesser intraclass variance and maximum number of higher interclass variance [25]. Pitch, fundamental period and number of peaks were extracted from wavelet coefficients. Normalized extracted features were given to an ART network for speaker classification.

Case 3: Feature Extraction using cepstrum features

Cepstrum is an effective method for finding out the exact position and shape of human vocal tract from a speech signal [6]. Cepstrum function is applied over all the speech signals. Pitch, fundamental period and number of peaks were extracted from the cepstrum coefficients. The extracted features were given to an ART classifier after normalization.

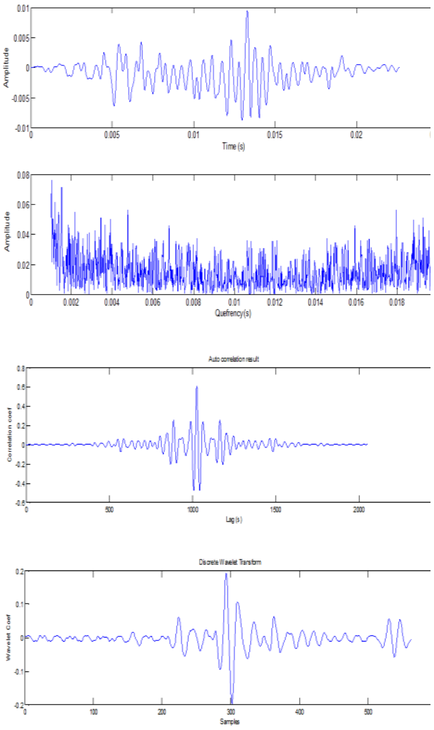
Case 4: Feature Extraction using combined features(autocorrelation, cepstrum and DWT)

The features obtained from autocorrelation function and DWT (6 features), autocorrelation and cepstrum (6 features), DWT and cepstrum (6 features), autocorrelation & DWT &cepstrum (9 features) were given separately to an ART classifier. For all the cases, the performance is studied and evaluated using sensitivity.

RESULTS AND DISCUSSIONS

All the speech signals are stored in '.wav' format. Ten signals of each speaker are considered for training purpose.For testing, five signals out of ten signals of each speaker are considered. A total of 100 speech signals are maintained for training and 50 speech signals for the testing dataset. Feature extraction techniques namely autocorrelation, DWT and cepstrum are applied over the training dataset. The waveforms obtained for the first sample speech signal of speaker 1 using the above mentioned techniques is shown in Figure2. Pitch, period and the number of peaks are aggregated from the feature extraction coefficients.

Figure 2: Waveforms for sample 1 of speaker 1 a) Frame of original speech signal b) Cepstrum waveform c) Auto correlation waveform d) DWT waveform



Feature extraction and classification of speaker using ART network is performed for all the four cases. The efficiency of the proposed work for all the cases is evaluated using Sensitivity [25].

$$\text{Sensitivity(\%)} = \text{TP} / (\text{TP} + \text{FN}) \qquad (4)$$

Where TPisTrue Positivei.e.identifying the correct speaker.FNisFalseNegative i.e. not identifying the correct speaker. The sensitivity obtained for the proposed speaker recognition system using the various feature extraction techniques (case 1- case 3) are listed in Table 1. For the combined features from various features extraction techniques (case 4) the sensitivity calculated is listed in Table 2. Sensitivity comparison for various feature extraction techniques (case 1- case 3) is shown in Figure 3. Sensitivity comparison for the combination of various feature extraction techniques (case 4) is shown in Figure 4.

Table1: Sensitivity for various feature extraction techniques

Speaker	Sensitivity (%)		
	Auto-Corr	DWT	Cepstrum
1	80	100	100
2	100	100	100
3	100	80	80
4	100	100	80
5	100	100	100
6	60	100	80
7	100	100	80
8	100	100	100
9	100	100	100
10	100	60	100

Table 2: Sensitivity for combined feature extraction techniques

Speaker	Sensitivity (%)			
	DWT &Cepstrum	Auto-Corr&Cepstrum	DWT & Auto-Corr	DWT &Cepstrum&Auto-Corr
1	100	100	100	100
2	100	100	100	100
3	100	100	100	100
4	100	100	100	100
5	100	100	100	100
6	100	100	100	100
7	100	100	100	100
8	100	100	100	100
9	100	100	100	100
10	100	100	100	100

Figure 3: Sensitivity comparison for various feature extraction techniques

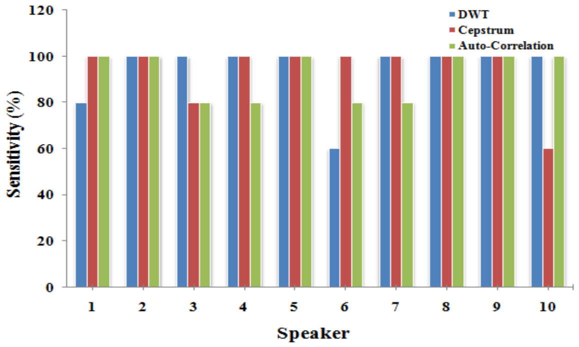
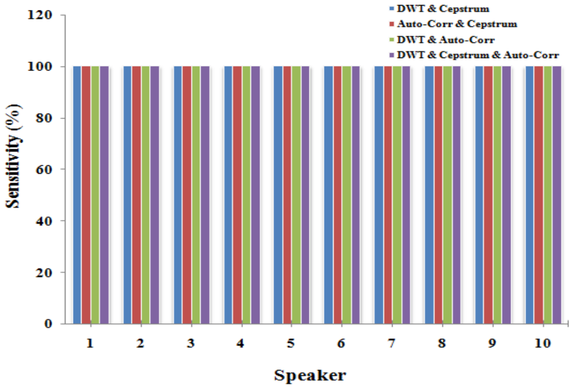


Figure 4: Sensitivity comparison for the combination of various feature extraction techniques



CONCLUSION

Identification of speaker based on the position and shape of their vocal tract is an efficient way of speaker recognition. Since each human have unique speech production system and vocal tract structure, speech waveform generated possess different Pitch, fundamental period and number of peaks. In this paper, feature extraction based on auto-correlation, Discrete Wavelet Transform and cepstrum techniques were performedconsidering the prosodic features namely pitch, fundamental period and number of peaks. Sensitivity obtained using DWT and auto-correlation is higher compared to cepstrum. Combined features provide 100 % sensitivity.The proposed work is well suited for speaker recognition for audio search engines. For other real time applications, a separate set of training and testing datasets is needed.

REFERENCES

[1] http://www.fbi.gov/about-us/cjis/fingerprints_biometrics/biometric-center-of-excellence / files / speaker - recognition.pdf

[2] MN Nachappa, AM Bojamma, CN Prasad, Nithya M. Int J Res Stud Comp Sci Eng 2014; 1(3):26-32.

[3] Sanghpal Humane, Abhilash Ukande, Sampada Adekar, Nitin Sakhare, Omprakash Yerne, Purva Gogte. Int J Adv Res Compu Sci Software Eng 2014;4(3):487-489.

[4] Rashmi CR. Int J Comp SciInform Technol 2014;5(4):5258-5262

[5] John R Deller, John H.L. Hansen, John G Proakis. Discrete-Time Processing of Speech Signals", An IEEE Press Classic Reissue, John Wiley & Sons

[6] Sirisha Devi, Srinivas Yarramalle, Siva Prasad Nandyala. Int J Compr Network Inform Security 2014;7:61-77

[7] Joaquin Gonzalez--Rodriguez. Evaluating Automatic Speaker Recognition systems: An overview of the NIST Speaker Recognition Evaluations (1996--2014), Loquens 1(1) January 2014.

[8] Kashyap Patel, RK Prasad. Int J Emerg Sci Eng 2013;1(7):33-37.

[9] B Rama Subbaiah, M Suleman Basha, B Ravi Teja. Int J Eng Res 2014;3(2S):22-25.

[10] Ashutosh Parab, Joyeb Mulla, Pankaj Bhadoria, and Vikram Bangar. J Res Electr Electron Eng 2014;3(4).

[11] Aman Chanda, Dhivya Jyothi, Mani Roja. Int J Comp Appl 2011;31(10).

[12] Jose Albin, NM Nandhitha, S Emalda Roslin. Adv Intell Syst Comput 2015;327:473-480.

[13] N Selvarasu, Alamelu Nachiappan and NM Nandhitha. European J Sci Res 2012;80(1):10-19.

[14] Selvarasu N, Alamelu Nachiappan and Nandhitha NM. National J Adv Comp Managl 2012;3(1):68-74.

[15] Emalda Roslin S, and Gomathy C. European J Sci Res 79(4):541 – 550.

[16] Brett R Wildermoth and Kuldip K Paliwal. Use of Voicing and Pitch Information for Speaker Recognition, SST-2000: 8th Australian International conference Speech Science and Technology, pp 324-328.

[17] Mariusz Ziółko, Rafał Samborski, Jakub Gałka, Bartosz Ziółko. Wavelet-Fourier Analysis For Speaker Recognition, Zakopane–Ko´scielisko, September 2011.

[18] Dongdong Li, Yingchun Yang, Weihui Dai. Sci World J 2014: 1-9.

[19] Shraddha Bhandavle, Rasika Inamdar, Aarti Bakshi. Emotion based Speaker Recognition with Vector Quantization, International Conference on Electronics & Computing Technologies, ICONECT-2014, pp 9-12.

[20] Praveen Singh Rathore, Neeta Tripathi. Int J Eng Trends Technol 2010;10(1).

[21] Samiksha Sharma, Anupam Shukla and Pankaj Mishra. Int J Innovative Sci Eng Technol 2010;1(4).

[22] Sunita Dixit, MD Yusuf Mulge. Int J Comp Sci Mobile Comp 2014;3(8):275 – 284.

[23] Binsu C Kovoor, Supriya MH, K Poulose Jacob. Int J Adv Res Compr Sci Software Eng 2014;4(9).

[24] Albin AJ, Nandhitha NM, Roslin SE. Text Independent Speaker Recognition System using Back Propagation Network with Wavelet Features. In: IEEE International Conference on Communication and Signal Processing, pp. 942–946 (2014).

[25] http://en.wikipedia.org/wiki/Sensitivity_and_specificity

[26] N Selvarasu, Alamelu Nachiappan, NM Nandhitha. Extraction and Quantification Techniques for Abnormality Detection from Medical Thermographs in Human, CD proceedings of IEEE International Conference on Computing, Communication and Networking (ICCCN - 2010), ID 56

[27] Chakravarthi R, Nandhitha NM, Emalda Roslin S, Sangeetha MS. Int J App Eng Res 2014;9(21):9927-9940.