The International Journal of Engineering And Science (IJES)
||Volume|| 2 ||Issue|| 01 ||Pages|| 126-132 ||2013||
ISSN: 2319 – 1813 ISBN: 2319 – 1805

Adult Voice Recognition System using Text Variable Phoneme Model and Coarse Speaking Fundamental Frequency Characteristics

Preeti Sharma
[1]School of Electronics & Electrical Engineering,
Chitkara University, Rajpura, Punjab

Abstract

Speech recognition is a fascinating application of Digital Signal Processing and has many real-world applications. In this paper, a speech recognition system is developed for isolated spoken words using Discrete Wavelet Transforms (DWT) and Artificial Neural Networks (ANN). Speech signals are one-dimensional and are random in nature. This paper investigates Automatic Speech Recognition of gender from speech segments using digital speech processing and pattern recognition techniques. Speaker recognition is an automatic process of recognizing the user on the basis of unique information carried by speech waves. The voice of the speaker is used to verify his or her identity and provide control for access to various services such as, voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers using Speaker recognition technique. Acoustic coefficients were used to form test and reference templates for vowels, voiced and unvoiced fricatives. The effects of different distance measures were comparatively assessed to determine their effectiveness for the task of gender recognition from speech segments.

Daubechies wavelets are and a multi-layer neural network trained with back propagation training algorithm is used for classification purpose

Keywords: *DWT, Speaker identification, ANN*

Date of Submission: 11, December, 2012 Date of Publication: 11, January 2013

I. Introduction

Our voice is the most natural way that used to interact with people and machines, so we can use it to do any job and remote any machine. Speech recognition process is the process in which a computer identifies the spoken words. It means that when you talk to your computer, it will recognize your words. Voice recognition is the technology by which sounds, words or phrases are spoken by humans that are converted into electrical signals, and these signals are transformed into coding patterns to which meaning has been assigned" (Rabiner and Juang, 1993).

Human listeners appear capable of extracting information from the acoustic signal beyond just the linguistic message. Listeners are generally able to identify clues about the speakers personality, emotional state , gender, age , dialect, accent, and the status of his/her health. Current automatic and speech recognition systems are far less capable than human listeners. Factors that limit automated speech and speaker recognition systems include our inability to identify acoustic features sensitive to the task and yet robust enough to accommodate speaker articulation differences, phonemic substitutions, or deletions, prosodic variations, and other factors that influence our recognition ability[5]

The differences between male and female voices depend upon many factors. Many physiological parameters of male and female vocal apparatus have been determined and compared. Reference [1] showed that the ratio of the total length of the female voice tract to that of a male is 0.87 and reference [2] showed that the ratio of the length of the female vocal fold to that of the male is about 0.8. Reference [3] reported that, anatomically, the female larynx also differs from the male larynx in thickness, angle of the thyroid laminae, resting angle of the glottis, vertical convergence angle in the glottis, and in other ways.

II. Wavelets And Speech

Speech signals are one of the most important means of communication in human beings.

The approach in this paper includes the wavelet transform, shown in Figure 1. This figure shows that a one dimensional signal is broken into two signals by low pass and high pass filters. The down samplers eliminate every other sample, so that the two remaining signals are approximately half the size of the original. As this figure shows, the low pass (approximate) signals can be further decomposed giving a second level of resolution (called an octave).

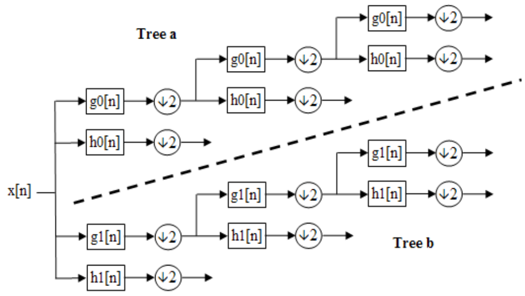
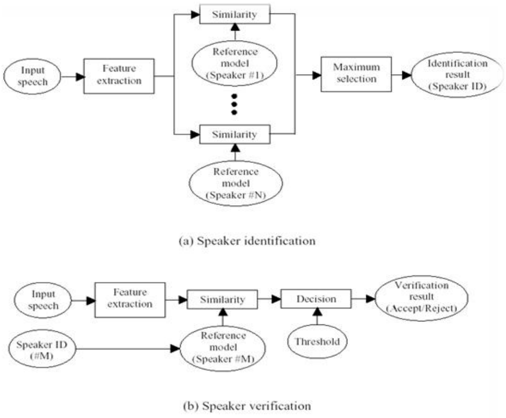


Fig 1; Discrete Wavelet Transform (DWT)

Wavelets express signals as sums of wavelets and their dilations and translations. Wavelets act in a similar way as Fourier analysis but approximate signals, which contain both large and small features, as well as sharp spikes and discontinuities. This is due to the fact that wavelets do not use a fixed time frequency window. The underlying principle of wavelet is to analyze according to scale [6]

Identification and verification can be classified as two stages of Speaker Recognition. The Identification process determines an utterance from registered speaker, whereas verification process accepts or rejects claims of identity of the speaker with the help of an artificial neural network, as shown in Figure 2.


Input speech
Feature extraction
Similarity
Reference model (Speaker #1)
Similarity
Reference model (Speaker #N)
Maximum selection
Identification result (Speaker ID)

(a) Speaker identification

Input speech
Feature extraction
Similarity
Decision
Verification result (Accept/Reject)
Speaker ID (#M)
Reference model (Speaker #M)
Threshold

(b) Speaker verification

Fig 2: Basic Structures of Speaker Recognition systems

Neural networks are an artificial intelligence method for modeling complex non-linear functions. Neural networks can be viewed as massively parallel computing systems consisting of an extremely large number of simple processors called nodes with many interconnections[7]. In the neural network mode, the nodes are artificial neurons and directed edges with weights are connections between neuron outputs and neuron inputs. Inspired by the human brain, neural network models attempt to use some organizational principles such as learning, generalization, adaptivity, fault tolerance etc. [10]. During the learning process, network architecture and connection weights are updated for proper classification. The main advantage of using neural networks is that they have the ability to learn complex nonlinear input-output relationships by using training procedures and adapting themselves to the data. Algorithms based on neural networks are well suitable for addressing speech recognition tasks. If x1, x2, x3, ……xn are the inputs and w1, w2, w3…wn are the corresponding weights, then the total input to the next neuron or the output neuron I is calculated by the summation function

$$I = w_1x_1 + w_2x_2 + \ldots\ldots.+ w_nx_n = {}^{n}\sum_{I=1} n_ix_i \quad (4)$$

The result of the summation function, which is the weighted sum, is transformed to a working output through an algorithmic process called the activation function or the transfer function.

The feed-forward network is the most commonly used type of neural network used in the area of pattern classification, which includes multilayer perceptron. In this work, we use architecture of the Multi Layer Perceptron (MLP) network, which consists of an input layer, one or more hidden layers, and an output layer as shown in Figure 3.

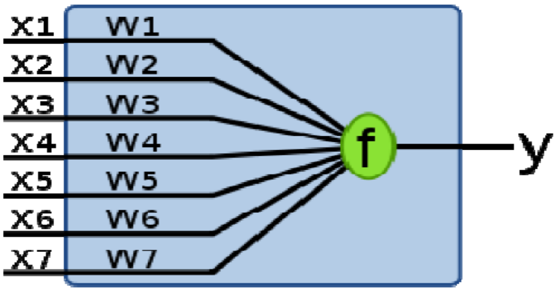


Fig 3: Perceptron Network Architecture [9]

Perceptron is common a pattern recognition machine, based on an analogy to the human nervous system, capable of learning in term of a feedback system which reinforces correct answers and discourages wrong ones. It is a type of artificial neural network and can be seen as the simplest type of feed-forward neural network and a linear classifier.k [9]The algorithm used in this case is the back propagation training algorithm. In this type of network, the input is presented to the network and moves through the weights and nonlinear activation functions towards the output layer, and the error is corrected in a backward direction using the error back propagation correction algorithm. After prolonged training, the network will eventually establish the input-output relationships through the adjusted weights on the network. After training the network, it is tested with the dataset used for testing. The recognition accuracy depends on the feature vectors obtained, training samples selected and the ability of the classifier to learn from these samples.[8] The increasing acceptability of neural network models to solve pattern recognition problems has been mainly due to its low dependence on domain-specific knowledge relative to model-based and rule-based approaches and due to the availability of efficient learning algorithms for users to implement [11].

Speaker recognition methods can also be divided into *text-independent* and *text dependent* methods. In a text-independent system, speaker models capture characteristics of speakers' speech which show up *irrespective of what one is saying*. In a text-dependent system, on the other hand, the recognition of the speaker's identity is based on his or her *speaking one or more specific phrases*, like passwords, card numbers, PIN codes, etc.

All technologies of speaker recognition, identification and verification, text independent and text-dependent, each have its own advantages and disadvantages and may require different treatments and techniques. The choice of which technology to use is application-specific.

Feature Extraction and Feature Matching are the two main modules contained by Speaker Recognition Systems (Refer Figure 2). Extraction process takes a small amount of voice signal which is later used for representing each speaker. The matching process involves the exact procedure to identify the unknown speaker by comparing his extracted features from the set of known speakers. [4].

All speaker recognition systems have to serve two known phases. The first one is called as enrollment or training phase while the second one is called as operation or testing phase. In the *training phase*, each known speaker has to provide samples of their speech so that the system can train a reference model for that speaker. For speaker verification systems, a speaker-specific threshold is also determined from the training samples. During the *testing (operational) phase* (refer to Figure 2), the input speech is matched with stored patterns and recognition decision is made.

Speaker recognition is a difficult task and it is still an active research area. Automatic speaker recognition works on the principle of unique speech characteristics exhibited by any speaker. However this task has been challenged by high variance of input speech signals, the principle source of which comes from the speakers themselves. Signals of speech used in various training and testing sessions are highly different due to many facts such as change of voice with time, health conditions, speaking rates, etc .The other factors which pose a challenge in the technology of speaker recognition are various acoustical noise and variations in recording environments or instruments.

III. Experiment Approach

The algorithm to classify speech into words is done using phoneme matching for a speaker dependant system. It is as follows:
Normalize input signal around x axis,
Normalize the amplitude values,
Use Wavelet Transform (Daubechies 8) to obtain 5 octaves of the same signal,
Compare this signal to template by calculating errors between them, and

Output the best match.

After reading in the signal, the first step is to normalize the incoming signal around the x axis and in amplitude. Normalizing around the x axis occurs by finding the median value of the entire signal, which is the DC component of the signal. This value is then subtracted from the entire signal, which results in moving the signal down around the x axis. Subtracting the minimum value of the signal from the entire signal and then dividing the signal by the maximum value of this signal achieve the normalization of the amplitude. The next step is to eliminate silence by removing any parts of the signal whose amplitude falls under a certain threshold. Then the wavelet transform is used to produce the signals decomposition into five octaves.

Pattern recognition is a broader topic in the field of engineering to which problem of speaker recognition belongs , the goal of which is to classify objects of interest into one of a number of categories or classes of individual speakers. Patterns are the generic objects of interest and are generally sequences of acoustic vectors that are extracted from an input speech since the classification procedure is applied on extracted features in this experiment, it can be also referred to as *feature matching*.

Furthermore, supervised pattern recognition is used where during the training session; each input speech with the ID of the speaker (S1 to S8) is labeled. The training set is comprised of the patterns which are used to derive a classification algorithm. The remaining patterns comprise the test set which are used to test the classification algorithm. The performance of the algorithm can also be evaluated if the correct classes of the individual patterns in the test set are also known.

The recognition process is illustrated in Figure 4 where only two speakers and two dimensions of the acoustic space are shown. The acoustic vectors from the speaker 1are referred as circles while from speaker 2 are referred as triangles. A speaker-specific VQ codebook is generated in the training phase for each known speaker by clustering his/her training acoustic vectors Figure 3 shows the result codewords (centroids) by black circles and black triangles for speaker 1 and 2, respectively. VQ distortion is called as the distance from a vector to the closest codeword of a codebook. An input utterance of an unknown voice is "vector-quantized" in the recognition phase , using each trained codebook and the *total VQ distortion* is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified.

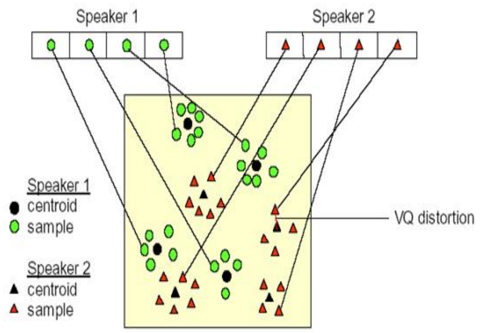


Fig 4: Conceptual Diagram illustrating vector quantization codebook formation. Based on the location of centroids one speaker can be discriminated from another.

IV. Results

Experimentation was conducted on 5 male speakers and 3 female speakers. During the experimentation, the system was trained by making a voice database of these 8 people. The word used for voice input is 'check'. The user could use its own key word also.

Figure 5 shows a plot of signal from database and the User signal.

Figures 6 to 8 show the original signal and the various decomposition level coefficient values.

Each person inputs his/her voice four times. The voice is compared twice with the database of the same person (speaker) while twice it is compared with the other person's voice sample.

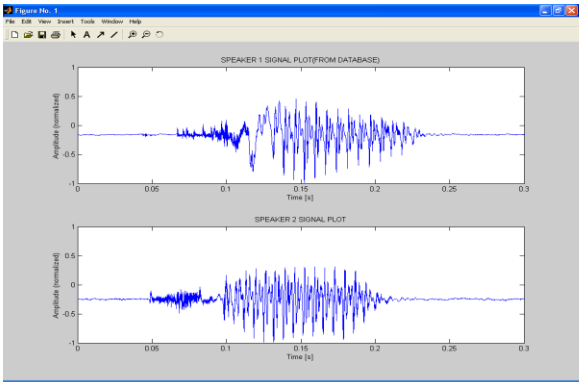


Fig 5: Plot of signal from Database and User Signal

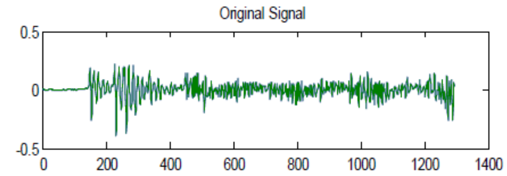


Fig 6: Original Signal (word spoken-'check')

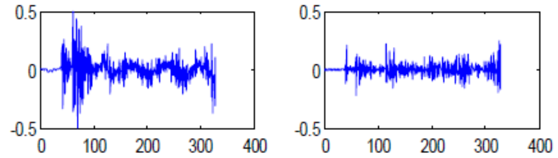


Fig 7: Approximation and detail coefficients (Level 1) of the spoken word 'check'

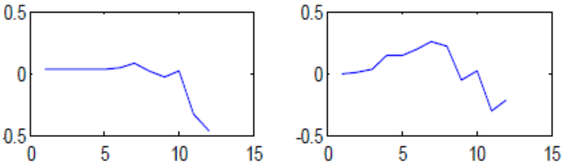


Fig 8: Approximation and Detail coefficients (Level 8) of the spoken word 'check'.

After comparison we get different values of ‘distance’. We have listed the values in the table following. Below the table there is a graphical comparison of speaker’s voice with his/her owns voice sample and also with other voice samples in the database.

Using these statistics, we can set a threshold value for speaker recognition. It can be judged from Figure 4 to Figure 11 that threshold value for the system should lie in between 4 to 5. By taking various threshold values like 4.2, 4.3, 4.4 etc. the accuracy of the system can be seen in table 1 and hence the appropriate threshold can be chosen.

As it is clear from the tableI and Figure 9 to Figure 16, highest accuracy is at threshold value 4.5 and 4.6. One out of these two values can be set for the system.

This system can be affected with many other factors such as background environment, variation in human voice with age and health condition, different speaking rates etc.

Also the whole database is divided into three areas. 70% of the data is used for training, 15% for validation and 15% for testing.

V. Conclusion

The approach taken in this paper is to use the wavelet transform is to extract coefficients from phonemes and to use cross correlation to classify the phoneme.

NAME OF SPEAKER	ENTERED NAME	CALCULATED DISTANCE
Male 1	Male 1	3.566
Male 1	Male 1	4.055
Male 1	Male 2	4.622
Male 1	Male 4	4.655
Male 3	Male 3	4.451
Male 3	Male 3	4.682
Male 3	Male 2	5.561
Male 3	Male 5	4.746
Male 2	Male 2	3.027
Male 2	Male 2	2.802
Male 2	Male 1	4.686
Male 2	Male 3	4.803
Male 4	Male 4	3.113
Male 4	Male 4	3.298
Male 4	Male 3	5.354
Male 4	Male 1	5.622
Male 5	Male 5	3.375
Male 5	Male 5	4.064
Male 5	Male 4	5.126
Male 5	Male 3	4.786
Female 2	Female 2	3.287
Female 2	Female 2	3.109
Female 2	Female 1	5.235
Female 2	Female 3	5.812
Female 1	Female 1	4.148
Female 1	Female 1	3.983
Female 1	Female 3	4.915
Female 1	Male 1	5.523
Female 3	Female 3	4.364
Female 3	Female 3	4.187
Female 3	Female 1	5.182
Female 3	Male 2	5.448

Table I- Experiment Analysis

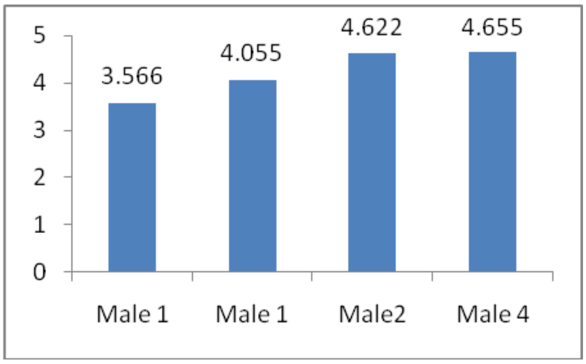


Fig9: Comparison of different speakers with male 1

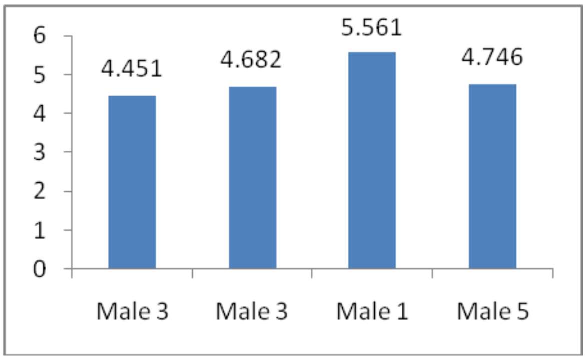


Fig 10: Comparison of different speakers with male 3

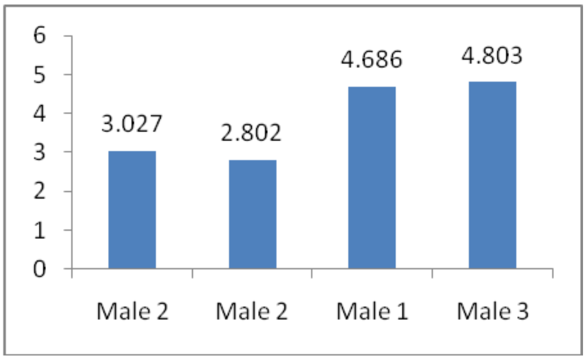


Fig11: Comparison of different speakers with male 2

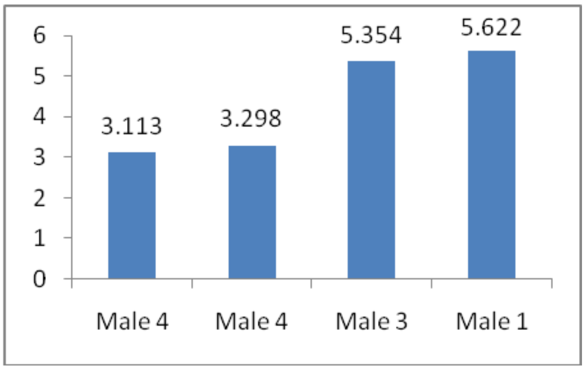


Fig 12: Comparison of different speakers with male 4

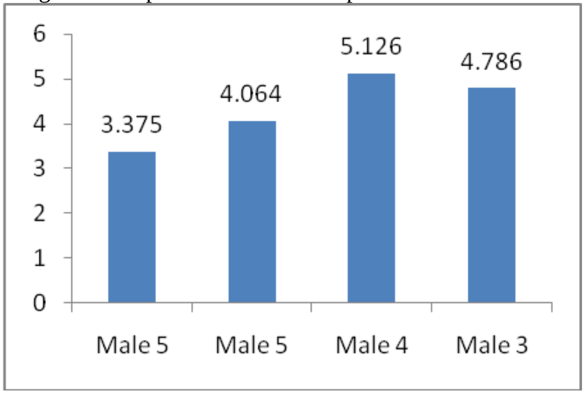


Fig 13: Comparison of different speakers with male 5

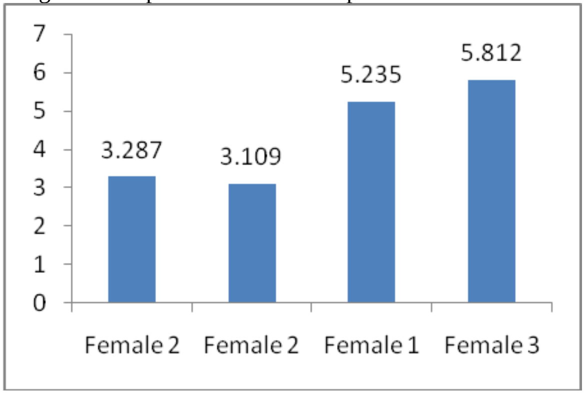


Fig14: Comparison of different speakers with female 2

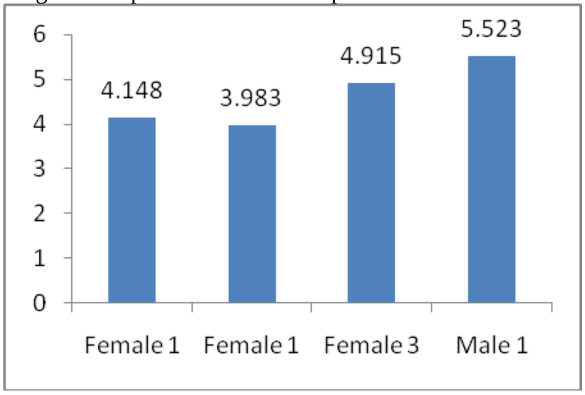


Fig15: Comparison of different speakers with female 1

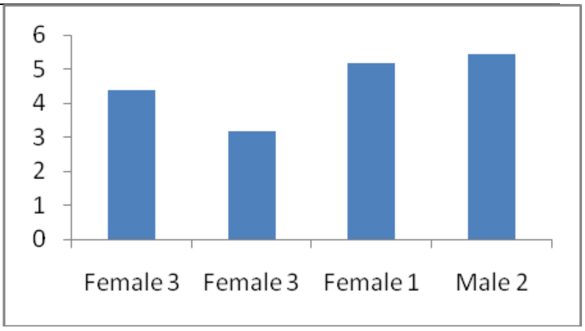


Fig16: Comparison of different speakers with female 3

	% Error
Training	12
Validation	9
Testing	10

Table II: Performance Analysis based on Error Percentage

Cross correlation measures the similarities between the two signals. Normalization of the amplitudes and frequencies are used. The speaker is speaker dependant to make things simpler.

The results show that using the wavelet transform improved the accuracy in correctly identifying the phonemes over not using any method for feature extraction. The results also show that using the approximation coefficients to generate octaves ion the wavelet transform give better efficiency than using the detailed coefficients. The first 3 octaves give the best results, while the accuracy of using the fourth and fifth octaves declines.

The proposed system is an efficient way to ensure security. The using of the ANN model with formants analysis make the recognizing process much faster than using the neural networks method, because it compares numbers to numbers which easier than comparing templates with templates.

Thus, an Automatic Speech Recognition system is designed for isolated spoken words using discrete wavelet transforms and artificial neural networks. A better performance of identification with high recognition accuracy of 90% is obtained from this study. The computational complexity and feature vector size is successfully reduced to a great extent by using discrete wavelet transforms. Thus a wavelet transform is an elegant tool for the analysis of non-stationary signals like speech. The experiment results show that this hybrid architecture using discrete wavelet transforms as the feature extraction tool and artificial neural networks as classifier could effectively extract the features from the speech signal(with 90% efficiency) for automatic speech recognition.(Refer Table II)

References

[1] Fant .G(1976),” Vocal tract energy functions and non uniform scaling,” J Acoust. Soc.Jpn 11,1-18.

[2] Hirano,M,Kurita,J., and Nakahima,T (1983),”Growth, development and ageing of human vocal folds,”in Vocal Fold Physiology, Contemporary Research and Clinical issues”, edited by D.Bless and J.Abbs (College Hill Press, San Diego, CA),pp 22-43.

[3] Titze I.R.(1989),”Physiologic and acoustic differences between male and female voices,” J.Acoust. Soc.Am,85,1699-1707

[4] Linde .Y, Buzo .A & Gray. R(1980), “An algorithm for vector quantizer design”, IEEE Transactions on Communications, Vol. 28, pp.84-95.

[5] Ke Wu, DG Childers(1991) ,” Gender Recognition from Speech. Part I: Coarse Analysis”, J. Acoustical Society of .America pp 1828-1840

[6] CJ Long, S Dutta,” Wavelet Based Feature Extraction for Phoneme Recognition”, Proceedings International Conference on spoken Language Processing, Volume 1, October 1996, pp 264-267

[7] Sonia Sunny, David Peter S,K Poulose Jacob ,”Discrete Wavelet Transforms and Artificial Neural Networks for Recognition of Isolated Spoken Words”, International Journal of Computer Applications (0975 – 8887) Volume 38– No.9, January 2012

[8] Moh’d Rasoul Al-Hadidi,” Speaker Identification System using Autoregressive Model”, Research Journal of Applied Sciences, Engineering and Technology 4(1): 45-50, 2012 ISSN: 2040-7467

[9] Wael Al-Sawalmeh*, Khaled Daqrouq*, Abdel-Rahman Al-Qawasmi*,And Tareq Abu Hilal,” The Use of Wavelets in Speaker Feature Tracking Identification System Using Neural Network”, WSEAS TRANSACTIONS on SIGNAL PROCESSING

[10] Y. Hao, X. Zhu, 2000, A new feature in speech recognition based on wavelet transform, Proc. IEEE 5th Inter. Conf. on Signal Processing, vol 3.

[11] Anil K. Jain, Robert P.W. Duin, Jianchang Mao, 2000, Statistical Pattern Recognition: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22

AUTHOR

PREETI SHARMA, M.TECH (ECE), School of Electronics & Electrical Engineering, Chitkara University, Punjab.