MFCC Based Text-Dependent Speaker Identification Using BPNN

S. S. Wali and S. M. Hatture
Dept. Computer Science and Engineering, Basaveshwar Engineering College, Bagalkot, India
Email: swathiwali@gmail.com

S. Nandyal
Dept. Computer Science and Engineering, Poojya Doddappa Appa College of Engineering, Gulbarga, India

***Abstract*—Speech processing has emerged as one of the important application area of digital signal processing. Various fields for research in speech processing are speech recognition, speaker recognition, speech synthesis, speech coding etc. Speaker recognition is one of the most useful and popular biometric recognition techniques in the world especially related to areas in which security is a major concern. This paper presents an automatic Speaker Recognition model which is Text-Dependent where the speaker is allowed to speak only fixed text. Automatic speaker recognition is the use of a machine to recognize a person from a spoken phrase, based on individual information (characteristics of voice) included in speech waves. Recognizer block employs MFCC (Mel Frequency Cepstrum Co-efficients) technique to get hybrid features for speaker identification/verification system. These features are used to train the ANN classifier in the training phase. Later in the testing phase the speaker is recognised based on the ANN classifier. The accuracy of 92% is achieved.**

***Index Terms*—speaker recognition, MFCC, ANN classifier**

I. Introduction

Most of the existing traditional authentication systems used for human computer interface uses the password, user name for authentication. Today's world is internet based, hence all use e-business applications thus improving the robustness of the system and provide an obstruction for theft is an important chore. Since the traditional system used password, Personal Identification Number they are in brink to evaporate.

This made the way for Biometrics. They are many kinds of biometrics some of them are Fingerprint, Palm-print, Iris, Face Recognition, hand geometry, speech recognition etc. the robustness provided by these biometrics is not vigorous because of the spoofing attacks. Various fields from research in speech processing are speech recognition, speaker recognition, speech synthesis, speech coding etc. Speaker recognition is one of the most useful and popular biometric recognition techniques in the world especially related to areas in which security is a major concern. The ability of recognizing voices of those familiar to us is a vital part of oral communication between humans. Research has considered automatic computer based speaker recognition since the early 1970's taking advantage of advances in the related field of speech recognition. Speaker recognition is one such technology that creates new services which parallely helps to lead more and more secure life.

Speech processing, the study of speech signals, can be regarded as a special case of digital signal processing, applied to speech signal. In this domain, Speaker recognition is an interesting and challenging problem. Speaker recognition is the process of recognizing automatically who is speaking on the basis of individual information included in speech waves. This technique uses the speaker's voice to verify their identity. The objective of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity. Speaker identification entails the classification of an unknown speaker using the database.

The human speech contains numerous discriminative features that can be used to identify speakers. Speech contains significant energy from zero frequency up to around 5kHz. The property of speech signal changes markedly as a function of time. To study the spectral properties of speech signal the concept of time varying Fourier representation is used. However, the temporal properties of speech signal such, as energy, zero crossing, correlation etc. are assumed constant over a short period. That is its characteristics are short-time stationary. Therefore, using hamming window, Speech signal is divided into a number of blocks of short duration so that normal Fourier transform can be used.

There are two types of speaker recognition systems: Text-dependent meaning the text must be the same for enrolment and recognition which improve performance especially with cooperative users. Text independent system has no advance knowledge of the presenter's phrasing and is much more flexible in situations where speaker is not cooperative.

All speaker recognition systems have to serve two distinguished phases. The first one is referred to the enrolment or training phase, while the second one is referred to as the operational or testing phase. In the

Manuscript received June 27, 2014; revised September 22, 2014.

doi: 10.12720/ijsps.3.1.30-34

training phase, each registered speaker has to provide samples of their speech so that the system can build or train a reference model for that speaker. In the testing phase, the input speech is matched with stored reference model(s) and a recognition decision is made.

In this paper we have presented a robust approach for text dependent voice recognition. The recognition system involves MFCC (Mel Frequency Cepstrum Co-efficients) technique for extracting features. These features are utilized for training the classifier in the training stage. In the testing stage the database serves as an input for ANN classifier which recognises the speaker based on his or her voice.

II. Proposed Methodology

At the highest level, all speaker recognition systems contain two main modules: feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers. The block diagram of our work is as shown in Fig. 1.

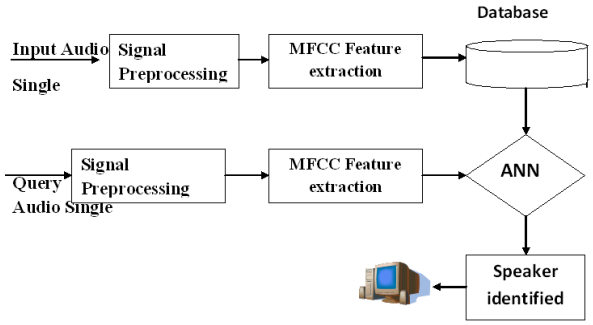

Figure 1 Generic speaker recognition system

A. Signal Pre-Processing

To capture the speech signal, sampling frequency of 11025Hz, sampling solution of 16-bits, mono recording channel and recorded file format, *.wav, have been considered. The speech preprocessing part has a vital role for the efficiency of learning. After acquisition of speech utterances, wiener filter has been used to remove the background noise from the original speech utterances [1]-[3]. Speech end points detection and silence part removal algorithm has been used to detect the presence of speech and to remove pulse and silences in a background noise [4]-[8]. To detect word boundary, the frame energy is computed using the short-term log energy equation,

$$E_{(i)}=10\log\sum_{t=u1}^{n1+N-1}S2(t) \quad (1)$$

where….$E_{(t)}$ is energy, $S(t)$ is voice sample.

Further, from different types of windowing techniques, Hamming window has been used for this system. The purpose of using windowing is to reduce the effect of the spectral artifacts that results from the framing process [9]-[11].

The concept applied here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame.

If we define the window as w(n), $0 \le n \le N-1$, where N is the frame length, then the result of windowing is the signal

$$Y(n)=x(n)\,w(n),\ 0\le n\le N\text{-}1 \quad (2)$$

The hamming window can be defined as follows:

$$w(n)=0.54-0.46\cos\left(2\pi\frac{n}{N}\right),\ 0\le n\le N \quad (3)$$

B. Feature Extraction

The general methodology of audio classification involves extracting discriminatory features from the audio data and feeding them to a pattern classifier. The purpose of this module is to convert the speech waveform into a set of features or rather feature vectors (at a considerably lower information rate) for further analysis. This is often referred to as the signal-processing front end.

The attributes of an ideal feature extraction strategy include:

a) The features should be resistant to an environmental noise and channel distortion
b) Variations in voice caused by speaker's health or aging should not degrade the performance of feature extraction methodology.
c) Feature extractor should maintain high inter-speaker discrimination and as little as possible of intra-speaker variability.
d) The speaker-characteristic features extracted from speech should be relatively easy to calculate.
e) The feature extraction method should be difficult to imitate or mimic using speech of imposters.

The above attributes are difficult to achieve in a single feature extraction procedure. This is because some of the attributes listed above have an inverse relationship; if one is improved the other deteriorates.

MFCC Algorithm: Human perception of frequency contents of sounds for speech signal does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The mel frequency scale is a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. As a reference point, the pitch of a 1KHz tone, 40dB above the perceptual hearing threshold, is defined as 1000mels. Therefore we can use the following approximate formula to compute the mels for a given frequency f in Hz.

$$\text{Mel(f)}=2595*\log 10\left(1+\frac{\text{f}}{700}\right) \quad (4)$$

Ours approach to simulate the subjective spectrum is to use a filter bank, one filter for each desired mel-frequency component. That filter bank has a triangular band pass frequency response and the spacing as well as the bandwidth is determined by a constant mel-frequency interval. The mel scale filter bank is a series of l triangular band pass filters that have been designed to simulate the band pass filtering believed to occur in the auditory system. This corresponds to series of band pass

filters with constant bandwidth and spacing on a mel frequency scale.

Cepstrum: In this final step, we convert the log mel spectrum back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the discrete cosine transform (DCT). In this final step log mel spectrum is converted back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC). The discrete cosine transform is done for transforming the mel coefficients back to time domain.

$$C_n = \sum(\log S_k)\cos\left\{n(k-0.5)*\frac{\pi}{k}\right\},\ n = 1,2,3\ldots k \quad (5)$$

whereas S_k, $K = 1, 2, \ldots K$ are the outputs of last step.

The MFCC feature extracted from fixed length signal frames effectively capture the characteristics of the speakers. It was also reported that the MFCC performs well for the task of speaker verification if the frame size ranging from 20ms to 50ms, and the frame step ranging from 1/6 to 1/3 of the frame size is used to analyze the speech. Thus keeping in view these recommendations, the MFCC based feature extraction method was implemented on short view these recommendations, the MFCC based feature extraction method was implemented on short-time signal (frame by frame basis) using frames of length 20ms with 10ms of overlap between adjacent frames.

C. ANN

Artificial neural networks in general are machine learning models which are genuinely inspired by the animals' central nervous systems (in particular the brain) that are capable of machine learning and pattern recognition. They are usually presented as systems of interconnected "neurons" that can compute values from inputs by feeding information through the network. The input to this neuron = (1, 2, ...,) is a feature vector in an *n*-dimensional feature space. The weight vector = (1, 2, ...,) may represent the template of a certain target.

The learning procedure tries is to find a set of connections w that gives a mapping that fits the training set well. Furthermore, neural networks can be viewed as highly nonlinear functions with the basic form $F(X, W) = Y$, where x is the input vector presented to the network, w are the weights of the network, and y is the corresponding output vector approximated or predicted by the network. The weight vector w is commonly ordered first by layer, then by neurons, and finally by the weights of each neuron plus its bias.

The extracted sixteen MFCC features from 50 users are used to train the ANN classifier with Back Propagation Neural Network (BPNN) individual's voice characteristics. The ANN classifier is trained for 50 classes (users), with 10000 iterations (epoch).

III. Results and Discussion

In the proposed work, the user is authenticated using voice biometric modality. The developed method contains four stages like voice signal acquisition, preprocessing, feature extraction, vector quantization and template generation and classification/categorization. The voice sample for a text-dependent speaker identification, the "Basaveshwar Engineering College" phrase is collected from 50 users with 10 phrases from each users. For every user 7 voice samples are used for training the ANN using BPNN algorithm. Sixteen MFCC features are extracted for every user through VQ method. Hence, for 50 users total of 5600 (350 samples × 16 features) feature values are used for training. The training of ANN Classifier using BPNN is shown in Fig. 2.

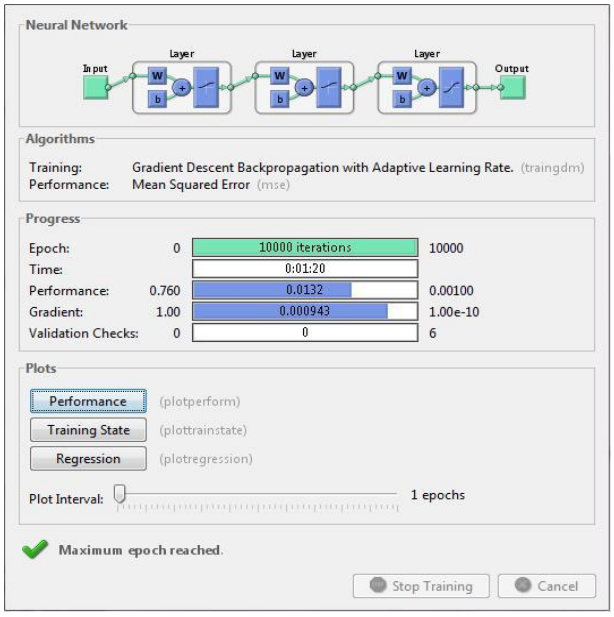

Figure 2. ANN training session

The developed system is tested for 10, 20, 30, 40 and 50 users, separately in order to test the effectiveness. The accuracy of 92% is achieved for 10 users, 82% is achieved for 20 users, 76% is achieved for 30 users, 72% is achieved for 40 users and last of all 70% is achieved for 50 users.

IV. Performance Analysis

The use of neural network was introduced in this work to improve the robustness of the biometric system that provide authentication. The user voice signal is recorded is from the Sony audio recorder ICD PX720. It PC compatible with Microsoft Windows and has a 288 hours of recording time (LP mode). The storage capacity of 1 GB Flash Memory and Stereo recording with external microphone with Correct dictation in playback. The recorded voice sample is converted to 11.1 kHz using Digital Voice Editor Software. Voice samples are was stored in different folder for training and testing. The voice sample for the phrase "Basaveshwar Engineering College" is shown in Fig. 3. Since the project is based on text-dependent phrase with a phrase of "Basaveshwar Engineering College".

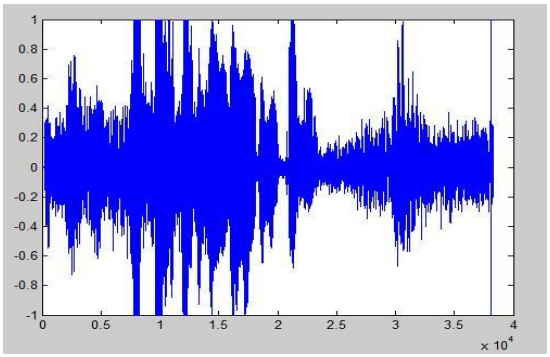

Figure 3. Voice sample of "basaveshwar engineering college" phrase

The performance of biometrics system is measured using correct identification rate (CIR) or correct recognition rate (CRR), false acceptance rate (FAR) [12] and false rejection rate (FRR) [12]. The CIR is the ratio of the number of authorized users accepted by the biometric system to the total number of identification attempts made. It is stated as follows:

$$CIR = \frac{Numberof\ correctly\ identified\ claims}{Totalnumberof\ claims} \times 100\% \quad (6)$$

The FAR or 'type 2 error' is the ratio of the number of unauthorized users accepted by the biometric system to the total number of identification attempts made. It is stated as follows:

$$FAR = \frac{Numberof\ wrong\ identified\ claims}{Totalnumberof\ claims} \times 100\% \quad (7)$$

The FRR or 'type 1 error' is the ratio of the number of authorized users rejected by the biometric system to the total number of attempts made. It is stated as follows:

$$FRR = \frac{Numberof\ wrongly\ rejected\ claims}{Totalnumberof\ claims} \times 100\% \quad (8)$$

The performance of the proposed system is tabulated in Table I.

TABLE I. RECOGNITION PERFORMANCE

Sl. No	Number Of Users	Number of tested voice samples	CIR in %	FAR in %	FRR in %
1	10	100	92	5	3
2	20	200	82	11	7
3	30	300	76	14	10
4	40	400	72	17	11
5	50	500	70	18	12

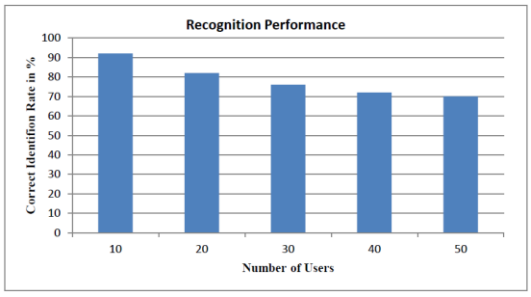

Figure 4. Recognition performance of different set of users

The Fig. 4 shows the recognition performance for a set of 10, 20, 30, 40 and 50 users.

Hence, by using these parameters the identification accuracy of around 92% to 70% is achieved for 10 to 50 users and false rejection rate of 4%vto 12% is obtained for the 10 to 50 users.

V. CONCLUSION

This paper has presented a Neural Network based Personal Authentication using Voice Modality. Using text-dependent speaker identification/verification method the voice samples using voice recorder; with a fixed phrase "Basaveshwar Engineering College", the Mel Frequency Cepstrum Coefficients features are extracted and represented using vector quantization. Further, the multilayer feed-forward back propagation neural network classifier using back propagation is used and the recognition accuracy of around 92% to 70% is achieved for 10 to 50 users and false rejection rate of 4%vto 12% is obtained for the 10 to 50 users.

In future, the words in the voice sample phrase are to be increased and different soft computing classifier can be adopted to accomplish better recognition results. Since the proposed work is on text -dependent also future scope might be for text-independent phrase.

REFERENCES

[1] S. Doclo and M. Moonen, "On the output SNR of the speech-distortion weighted multichannel wiener filter," *IEEE Signal Processing Letters* vol. 12, no. 12, 2005.

[2] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*, New York: Wiely, 1964.

[3] N. Wiener and R. E. A. C. Paley, *Fourier Transforms in the Complex Domains*, Providence, RI: American Mathematical Society, 1934.

[4] K. Kitayama, M. Goto, K. Itou, and T. Kobayashi, "Speech starter: Noise-Robust endpoint detection by using filled pauses," in *Proc. Eurospeech 2003*, Geneva, 2003, pp. 1237-1240.

[5] S. E. B.-Ghazale and K. Assaleh, "A robust endpoint detection of speech for noisy environments with application to automatic speech recognition," in *Proc. ICASSP2002*, 2002, pp. 3808-3811.

[6] A. Martin, D. Charlet, and L. Mauuary, "Robust speech/non-speech detection using LDA applied to MFCC," in *Proc. ASSP2001*, 2001, pp. 237-240.

[7] R. O. Duda, P. E. Hart, and D. G. Strok, *Pattern Classification*, 2nd ed. A Wiley-Interscience Publication, John Wiley& Sons, Inc, 2001.

[8] V. Sarma and D. Venugopal, "Studies on pattern recognition approach to voiced-unvoiced-silence classification," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP'78*, 1978, pp. 1-4.

[9] F. Owens, *Signal Processing of Speech (Macmillan New Electronics)*, Macmillan, 1993.

[10] F. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51-84, 1978.

[11] J. Proakis and D. Manolakis, *Digital Signal Processing, Principles, Algorithms and Applications*, 2nd ed. New York: Macmillan Publishing Company, 1992.

[12] S. A. Angadi and S. M. Hatture, "A novel spectral graph theoretic approach to user identification using hand geometry," *International Journal of Machine Intelligence*, vol. 3, no. 4, pp. 282-288, 2011.

Miss. Swathi S. Wali was born on 12-10-1990, Bijapur. She is currently perceiving PG degree in Computer science and Engineering at Basaveshwar Engineering College Bagalkot-587102, Karnataka, India. Area interest is image processing and signal processing.

Sanjeevakumar M. Hatture was born on 20-12-1973, Bijapur, India Karnataka. He received the Bachelor's Degree in Electronics and Communication Engineering from Karnataka University, Dharwad, Karnataka State, India, and the Master Degree in Computer Science and Engineering from the Visvesvaraya Technological University, Belgaum, Karnataka, India, and currently pursuing PhD Degree in the Research Centre, Department of Computer Science and Engineering at Basaveshwar Engineering College, Bagalkot under Visvesvaraya Technological University, Belgaum, Karnataka, India. His research interests include biometrics, image processing, pattern recognition, and Soft computing and network security. He is life member of professional bodies like IEI and ISTE.

Dr. Suvarna Nandyal was born on 01-01-1972, Gulabarga India. She is presently working as Professor in Department of Computer Science and Engineering, Poojya Doddappa Appa College of Engg, Gulbarga, Karnataka, India. She received her BE degree in Computer Science & Engineering from Gulbarga University Gulbarga in 1993, M Tech degree in Computer Science & engineering from VTU Belgaum in 2003 and Ph.D. degree in Computer Science & Engg from Jawaharlal Technological University, Hyderabad, Andrapradesh India in 2013. She has teaching experience of 20 Years. She has extensive research interests in Image processing, Video Processing, Speech Recognition, Cloud Computing and information retrieval. She is member of ISTE and IETE. She has published number of research papers in International/National journals and conferences. She is involved in various academic activities such as curriculum development and professional society activities. She has guided post graduate students and two Ph.D scholars are working under her guidance.