Text Independent Gender Identification in NoisyEnvironmental Conditions

Seema Khanum
Department of Computer Science & Engineering,
Rajiv Gandhi University
Doimukh, Arunachal Pradesh, India
seema.khanum@rgu.ac.in

Firos A
Department of Computer Science & Engineering,
Rajiv Gandhi University
Doimukh, Arunachal Pradesh, India
firos.a@rgu.ac.in

***Abstract*—Thispaper proposes a competent system that is not only text independent in identifying gender of a speaker but can also work efficiently in noisy environmental conditions in real time. The noisy environmental conditions are the places where noise signals are generated at different SNRs (Signal to Noise Ratios) such as train station, restaurant, exhibition hall, airport, and so on. The algorithms used in the proposed system are MFCC (Mel-Frequency Cepstral Coefficients) for feature extraction from the speech and ANN (Artificial Neural Network) for classification between the genders (Male and Female).**

Keywords—Gender Identification; Noisy Environment; MFCC; ANN

I. INTRODUCTION

Literature survey shows that the gender identification is a crucial task for carrying out efficiently other speech processing tasks such as speaker recognition, language identification, speech compression, and so on. There already exist many gender identification systems few of them worked considering different parameters of speech such as pitch, ZCR (Zero Crossing Rate), STE (Short Time Energy), and so on. Few other gender identification systems have used different combinations of algorithms such as Gaussian Mixture Model (GMM), Multilayer Perceptrons (MLP), Vector Quantization (VQ) and Learning Vector Quantization (LVQ) along with Mel-Frequency Cepstral Coefficients (MFCCs) [1]. Most of these existing systems are either text dependent or works with a clean speech (i.e. without background noises), which is not acceptable in real world scenario and hence a failure. Therefore, a system for automatic gender identification of a speaker which is text independent and also works perfectly in different noisy environmental conditions is very essential in the field of speech processing.

The remainder of this paper is organized as follows. Section II describes the methodology used in the proposed text independent gender identification system. Section III describes the implementation of the proposed system using the methodologies mentioned earlier in Section II. Section IV represents the experiments conducted on the final implemented system to evaluate its performance in term of accuracy and their respective results. Finally, Section V concludes the paper.

II. METHODOLOGY

The basic methodology involved in speech based gender identification is broadly divided into two procedures. The first procedure is to extract features from the speech signal and the second procedure involves classification of the extracted features properly into two gender groups (male and female). There exist different procedures for feature extraction from speech signal but the proposed system employs MFCC. Similarly, for classification process the method used is feed forward neural network with back-propagation training algorithm.

A. Mel Frequency Cepstral Coefficient

The following diagram depicts various steps involved in calculating the MFCCs.

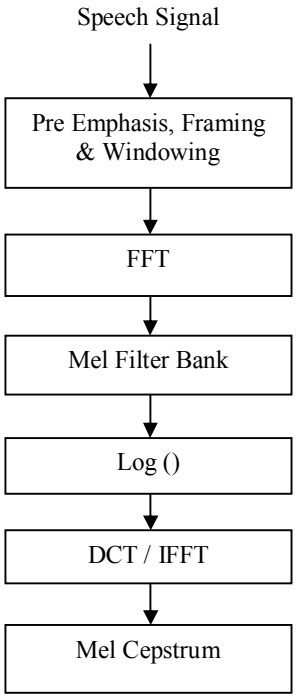

Fig.1. A flowchart of MFCC calculations

ISBN:978-1-5090-6471-7/17/$31.00 ©2017 IEEE

MFCC is a well-known feature extraction method, used in the field of speech processing, which employs logarithmic Mel Scale that works based on human hearing scale. To be precise, the Mel Scale is linear up to 1000 Hz and logarithmic at frequencies higher than 1000Hz. Thus, to calculate Mels for a particular frequency f in Hz, we use the following formula [2].

$$mel(f) = 2595 \times \log_{10}(1+f/700)$$

B. Fuzzy C-Mean Clustering

FCM clustering is also called soft clustering technique where data elements can belong to more than one cluster to some degree that is specified based on a set of membership levels. The purpose of clustering at this stage is to identify natural grouping of data from a large data set to produce a concise representation of a system's behavior.

In FCM, the center of a cluster c_k is the mean of all points, weighted by their degree of belonging to the cluster. Any point x has a set of coefficients giving the degree of being in the k^{th} cluster $w_k(x)$. It also depends on a parameter m, which controls how much weight is given to the closest center.

$$c_k = \frac{\sum w_k(x)^m x}{\sum_x w_k(x)^m}$$

C. Back Propagation Neural Network

Back Propagation Neural Network (BPNN) is a simple feed forward neural network with back propagation learning algorithm. The back propagation learning algorithm used here is the *trainscg* (scaled conjugate gradient), whose advantage over other training algorithms is that, it requires less memory and much faster than standard gradient descent algorithms. The function of BPN is to learn the mapping of a set of input MFCC patterns to a set of output patterns. As the network is trained with different MFCC patterns, it develops the ability to generalize over similar features in the different patterns.

D. Spectral Subtraction Method

As the purposed system works in the noisy environmental conditions, an algorithm must be required for noise removal from the input speech signal. There exist many noise removal techniques but our system utilizes Spectral Subtraction Method (SSM) as it considerably reduces the noise level keeping other important features of the original speech signal intact.

The working principle of the SSM is quiet straight forward, it involves estimating of an average signal spectrum and noise spectrum in parts of the recoding and then subtracted from each other so that the average SNR (Signal to Noise Ratio) is improved. If y(m) is the noisy signal and n(m) is the estimated noise then the desired signal x(m) without noise is given by,

$$x(m)=y(m)- n(m)$$

III. System Implementation

The following Fig.2 shows the different stages involved in purposed competent system for gender identification [3], in noisy environmental conditions.

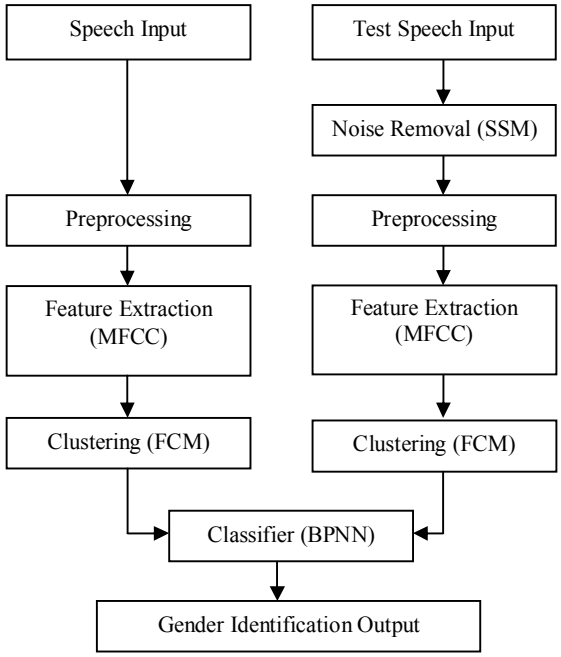

Fig.2. Stages of text independent gender identification system

A. Input Speech Signal

Input speech signals are taken from noisy speech corpus (NOIZEUS). This corpus consists of thirty IEEE sentences which are originally taken from IEEE database [4] spoken by male and female speakers. These thirty IEEE sentences are phonetically balanced with relatively low word context predictability and also include all phonemes in the American English language [5]. These sentences were recorded using Tucker Davis Technologies (TDT) recording equipment in a sound proof booth. These 30 sentences are then corrupted by various real world noises (taken from AURORA database) at different SNRs of 0dB, 5dB, 10dB and 15dB. The real world noises include noises at the airport, exhibition hall, train station, street, restaurant, and so on [6].

B. Noise Removal From Speech Signal

The purposed system works on the speech signals at a sampling rate of 16 kHz, thus it requires the sentences in NOIZEUS to be re-sampled which are at the sampling rate of 8 kHz. The noise is removed from the input speech signal using the spectral subtraction method as briefed in the Section II.D.

C. Preprocessing

After the background noise is removed, the input speech signal is divided into frames of 20ms. Each time frame has an overlapping of 50% with the next frame. Overlapping is necessary during the segmentation for smooth transition from one frame to another. Frames are then windowed with Hamming window to remove any discontinuities at the edges. For the current sample n, the Hamming window W(n) is calculated as follows:

$$W(n) = 0.54 - 0.46 \cos\left[\frac{2\pi n}{N-1}\right]; \quad 0 \le n \le N-1$$
$$= 0; \quad \text{otherwise}$$

Where, N is the total number of samples.

D. Feature Extraction

Feature vectors are extracted from the speech signal using MFCC algorithm as described in Section II A. These MFCC coefficients are then processed by fuzzy clustering method so as to group large amount of data generated into some specific number of clusters and hence helps in reducing computational cost and time for real time identification

E. Classifier

After the clusters are created by the fuzzy c-mean clustering method, they are arranged in proper format to feed into the artificial neural network for recognition. In the training stage the weights of the feed forward neural network was assigned by some random value and is then adjusted for optimal during the learning process by using back propagation algorithm.

In the testing stage, the neural network is tested against various test samples of speech, to check whether the obtained system properly classifies the speech into male voice and female voice.

F. GUI for Implemented System

The below Fig.3 represents the snapshot of GUI (Graphical User Interface) created for the implemented system in order to display the response of the system to the end users in more professional and intellectual way. MATLAB 8 Software was used as a platform to create the front end display. As shown in the GUI, "Select File" button allows user to either select the clean speech or the noisy speech at SNRs of 0dB, 5dB, 10dB and 15dB listed under "Make a DataBase Selection". Once the speech is selected, user can hear the speech by pressing "Play" button. User is provided with the option to identify the speaker's gender without using the noise elimination technique by pressing "Speaker's Gender" button and at the same time with using noise elimination technique by pressing "Advanced Gender Identification" button, thus can compare the results from the two methodologies effortlessly. "Advanced Gender Identification" button demonstrates the implementation of the proposed system for text independent gender identification in noisy environmental conditions.

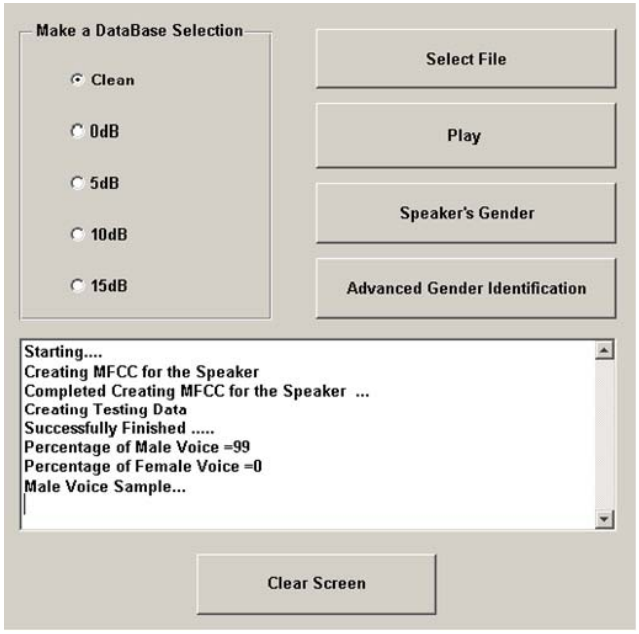

Fig.3. Snapshot of the GUI displaying results

IV. EXPERIMENTS & RESULTS

A system with 6 fuzzy clusters and 10 hidden neurons gives an optimal solution for speech based gender identification [3]. Therefore, further experiments are conducted based on this model. MATLAB 8 Software was used as a platform to develop the proposed system.

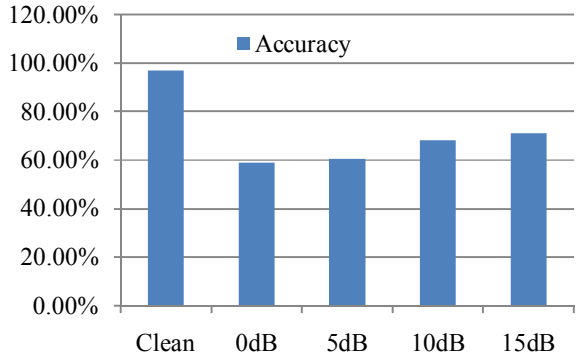

Fig.4. System Accuracy before applying SSM on input speech signal at different SNRs

The Fig 4, above shows the accuracy resulted from the system when noise is not removed from the input speech using the noise elimination technique. Observe that in case of clean speech, the accuracy is higher compare to the noisy speech where the noise is added at different level of SNRs. Similarly, the Fig.5 below shows system accuracy after applying spectral subtraction method for input speech signals.

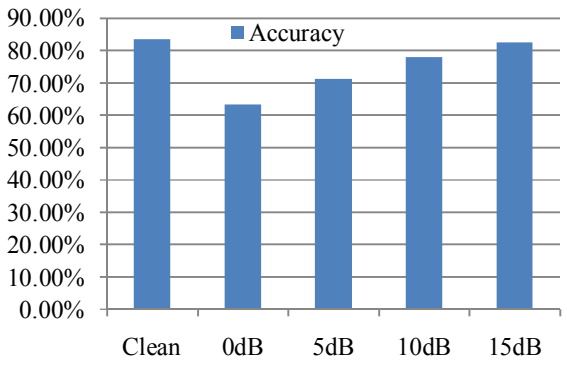

Fig.5. System Accuracy after applying SSM on input speech signal at different SNRs.

V. CONCLUSION

The purposed system for text independent gender identification in noisy environmental conditions is implemented and experiments have been conducted to check its efficiency in terms of accuracy. From the experimental result, it can be seen that the accuracy of the system, after applying the spectral subtraction method, has increased compare to the accuracy of system before applying the noise elimination technique. The highest accuracy achieved for the noisy speech signal in identification of gender is 83.3%. This outcome is also a consequence of applying the most robust algorithms i.e. MFCC for feature extraction and feed forward neural network with back propagation logarithm in learning the input patterns for classification. Fuzzy c means clustering reduces and fixes the input data set to neural network, which also plays an important role in system outcome.

REFERENCES

[1] Diemili R, Bourouba, H. and Korba," A speech signal based gender identification system using four classifiers", Multimedia Computing and Systems (ICMCS), 2012 International Conference , IEEE

[2] L. Rabiner, B. H. Juang, "Fundamentals of Speech Recognition", Pearson Education, 2009.

[3] Seema Khanum and Marpe Sora, "Speech Based Gender Identification Using Feed Forward Neural Networks", International Journal of Computer Applications (0975 –8887) :National Conference on Recent Trends in Information Technology (NCIT 2015)

[4] Ma, J., Hu, Y. and Loizou, P. (2009). "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions", Journal of the Acoustical Society of America, 125(5), 3387-3405

[5] IEEE Subcommittee (1969). IEEE Recommended Practice for Speech Quality Measurements. IEEE Trans. Audio and Electroacoustics, AU-17(3), 225-246

[6] H. Hirsch, and D. Pearce (2000). "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions.", ISCA ITRW ASR2000, Paris, France, September 18-20.