Speech based Gender Identification using Feed Forward Neural Networks

Seema Khanum
Department of Computer Science & Engineering
Rajiv Gandhi University
Doimukh, Arunachal Pradesh, India

Marpe Sora, PhD
Department of Computer Science & Engineering
Rajiv Gandhi University
Doimukh, Arunachal Pradesh, India

ABSTRACT

This paper proposes an efficient method of gender identification based on the speaker's voice in a noisy environment. MFCC was used to extract features from the speech sample taken from a noisy speech database; these features are then used to train Artificial Neural Network architecture to classify two different genders (Male and Female). The test result shows that the new proposed ANN architecture can analyze and learn better and faster. The advantage of proposed method is a result of decreasing the number of segments by grouping similar segments in training data using a clustering technique namely fuzzy c means clustering.

General Terms

Gender Identification, Pattern Recognition, Classification

Keywords

Gender Identification,MFCC, ANN, Fuzzy C Mean Clustering

1. INTRODUCTION

Gender identification based on the voice of a speaker consists of detecting if a speech signal is uttered by a male or a female. Automatically detecting the gender of a speaker has several potential applications. Some studies in the literature show that speech recognition and speaker identification would be simpler, if there could be a way to automatically recognize a speaker's gender (sex) at the beginning itself. Gender identification was used primarily as a means to improve recognition performance and to reduce the needed computation. Accurate gender identification has different uses in spoken language systems, where it can permit the synthesis module of a system to respond appropriately to an unknown speaker. The main feature which can distinguish between speaker's genders is the fundamental frequency F0 with typical values of 110 Hz for male speech and 200 Hz for female speech. However, there are Gaussian distributions of these ranges, so that dispersion is wide and thus often could notbe able to categorize the acoustic signal reliably by using this criterion only [1]. Therefore, a research on novel gender identification based on speaker's voice is very essential.

The remainder of this paper is organized as follows. Section 2 describes the methodology of gender identification. Section 3 describes system implementation using the methodologies mentioned in section 2. Section 4 describes the various experiments contacted to evaluate the performance of the gender identification system and their respective results. Finally section 5 concludes the paper.

2. METHODOLOGY OF GENDER IDENTIFICATION

The Gender identification from the speaker's voice can be performed using two basic functions namely, feature extraction and classification. Feature extraction represents the process of transforming a speech signal into a set of parameters essential for speaker gender identification. During the process, only the reliable parts are considered while the unreliable part is discarded. This transformation results in a vector called feature vector.

Classification is the process of classifying the obtained feature vectors into closely related category i.e. either male or female. To accomplish this function well-known neural network architecture named multi-layered perceptron (MLP) with back-propagation training algorithm is used

2.1 Spectral Subtraction Method

In speech communication, the speech signal is always accompanied by some background noise. The noise is the most common causefor degrading the quality of the speech signal in recordings. Therefore, a noise elimination method is used so as to reduce the noise level in the speech signal without affecting its features. This method involves independent spectral subtraction in the frequency bands.

The spectral subtraction method (SSM) is a simple and effective modulefor noise reduction in the noisy speech signal. In this module, signal-to-noise ratio (SNR) is improved by subtracting from each other the estimated average signal spectrum and the average noise spectrum in parts from the original noisy speech signal. It is assumed that the speech signal is corrupted by a wide-band, additive, stationary noise and the estimated noise is the same during analysis and restoration and also the phase is the same in original and restored signal [2].

2.2 Mel-Frequency Cepstrum

MFCC (Mel Frequency Cepstral Coefficient) takes human perception sensitivity with respect to frequencies into consideration, and therefore are best for speech/speaker recognition [3]. An important factor is that human hearing is not equally sensitive to all frequencies. Thus, as compare to the actual frequency which is measured in Hertz, the subjective pitch is measured in mel scale. Mel scale is approximately linear below 1000Hz and logarithmic above 1000Hz. Therefore, the following formula is used for computing the mels for a given frequency f in Hz [4].

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

To create mel-spectrum, filter bank is used wherein one filter corresponds to each desired mel-frequency component. The filter bank has a triangular band pass frequency response which computes the average spectrum around each center frequency with increasing bandwidths. Last step is the conversion of log mel-spectrum back to time domain which can be accomplish by taking Discrete Cosine Transform (DCT) [5].

2.3 Fuzzy C-Means Clustering

Clustering of numerical data forms the basis of many classification and system modeling algorithms. The purpose of clustering is to identify natural grouping of data from a large data set to produce a concise representation of a system’s behavior. In Fuzzy C-Means (FCM) clustering technique, each data point belongs to a cluster to some degree that is specified by a membership grade [6]. All extracted feature vectors are clustered into specific number of clusters using FCM clustering algorithm.

2.4 Feedforward Neural Network

The Artificial Neural Network (ANN) is an efficient pattern recognition mechanism which simulates human brain for neural information processing. ANN is composed of many artificial neurons that are linked together according to some specific network architecture. The objective of the neural network is to transform the inputs into meaningful outputs. Usually, feed forward neural network architecture has three layers. The first layer consists of the input vector and the last layer consists of the output vectors and the intermediate layers or hidden layers learn the data by parsing them through weighted connections. An efficient pattern recognition and classification of data using neural network is effectively performed using the Supervised Learning Neural Network with Back-Propagation Algorithm. For a given set of training input-output pair, this algorithm provides a procedure for changing the weights in a back-propagation network (BPN) to classify the input patterns correctly.

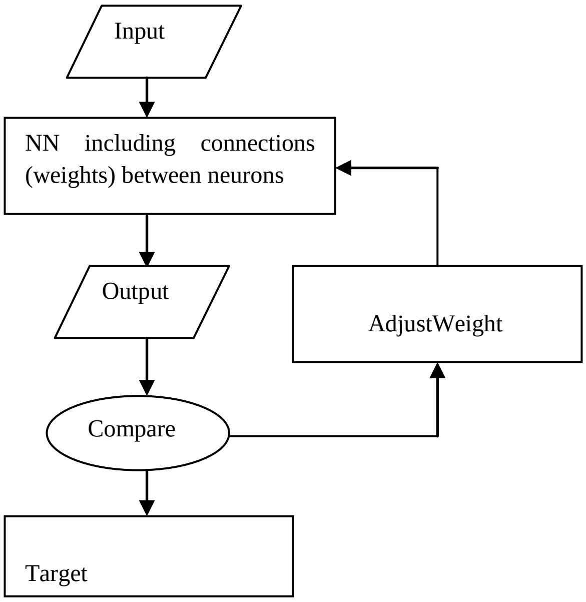


Fig 1: Neural Network (NN)

2.5 Front-End Display using GUI

The final stage of speech based gender identification is its capacity to display the computed results in an efficient and intelligent mode. The graphical user interface (GUI) aids in this aspect to describe the response of the system using figures, text-box and plots. One such example of this technique is shown below. The result generated by the system is displayed in the text box once the user selects the speech signal and pushes either the “Speaker’s Gender” button for output without noise elimination technique or the “Advanced Gender Identification” button for output using noise elimination technique. The GUI also provides an option to hear the selected speech signal by pushing “Play” button.

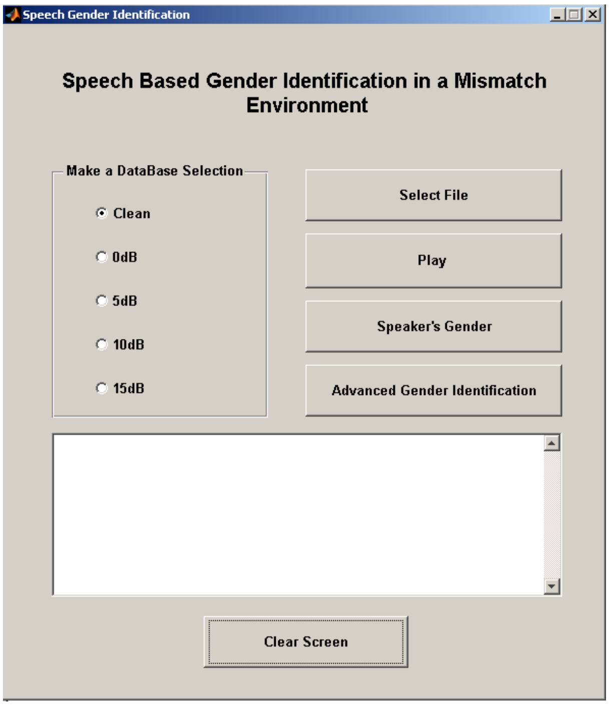
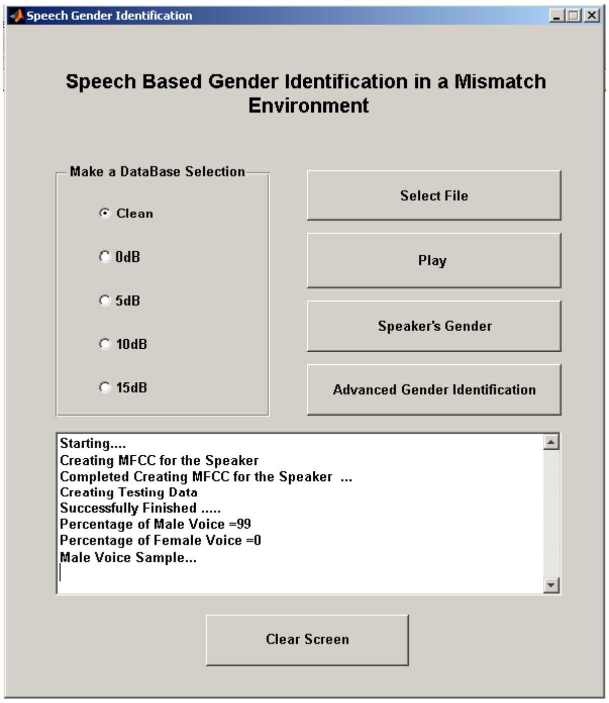

Speech Gender Identification
Speech Based Gender Identification in a Mismatch Environment
Make a DataBase Selection
● Clean
○ 0dB
○ 5dB
○ 10dB
○ 15dB
Select File
Play
Speaker's Gender
Advanced Gender Identification
Clear Screen

Fig 2: Illustration of the Graphical User Interface


Speech Gender Identification
Speech Based Gender Identification in a Mismatch Environment
Make a DataBase Selection
● Clean
○ 0dB
○ 5dB
○ 10dB
○ 15dB
Select File
Play
Speaker's Gender
Advanced Gender Identification
Clear Screen

```
Starting....
Creating MFCC for the Speaker
Completed Creating MFCC for the Speaker  ...
Creating Testing Data
Successfully Finished .....
Percentage of Male Voice =99
Percentage of Female Voice =0
Male Voice Sample...
```

Fig 3: Illustration of the Graphical User Interface Showing Result

3. SYSTEM IMPLEMENTATION

The following stages of speech processing provide a synoptic approach to the speaker's gender identification model.

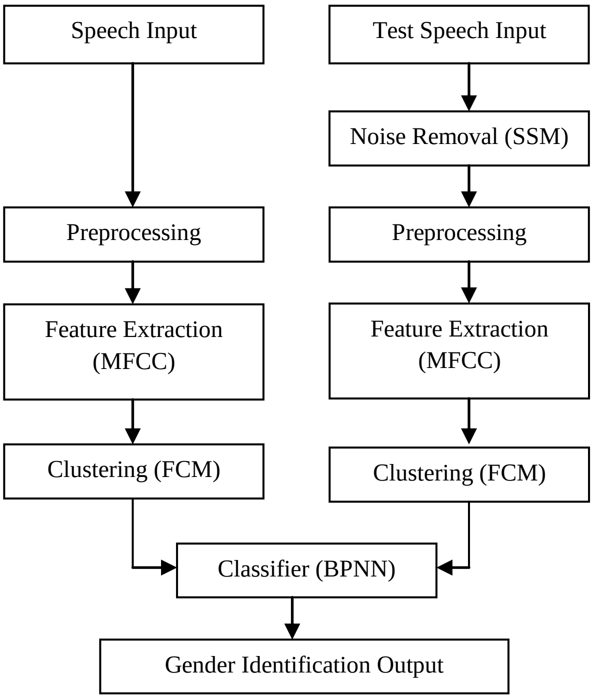


Fig 4: Speech based Gender Identification flow schart

3.1 Input Speech Signal

A noisy speech corpus (NOIZEUS) is used. It includes 30 IEEE sentences which are uttered by 6 speakers (3 male and 3 female). These 30 sentences are actually chosen from 720 sentences available in the IEEE sentence database [7] that contains sentences which are phonetically balanced as well as has relatively low word-context predictability. The basic criteria for choosing these 30 sentences, is that they contain all the phonemes present in the American English language. These sentences are recorded at a sampling frequency of 25 kHz, using the equipment from Tucker Davis Technologies (TDT), in a sound-proof booth. Further, these sentences are down sampled to 8 kHz. The real world noises at different SNR levels being 0db, 5db, 10db and 15db are artificially added to these 30 sentences. For different real world noises, AURORA database is used which contains the recordings from various noisy places such as airport, train station, street, exhibition hall, and so on [8].

3.2 Preprocessing

The speech signals were re-sampled at a sampling frequency of 16 kHz. Then the speech samples were passed through the Low Pass Filter (1 - 0.97z-1) which gives a spectral tilt to the speech samples. MATLAB 8 Software was used for all the computations. The filtered voice samples were segmented into 20ms frames with each frame having 50% overlap with the adjacent frames. Each frame is then multiplied by a Hamming window of length 20ms.

3.3 Feature Extraction

12 MFCC features were calculated from each frame of the speech signal. These feature vectors were then clusters using fuzzy c mean clustering.

3.4 Classifier

3.4.1 Database Generation for Training

In this step MFCC features are extracted from each speech utterance of each speaker. All extracted feature vectors are then clustered into n clusters using fuzzy c-means clustering algorithm.

3.4.2 ANN Training

In the training phase of feedforward neural network, weights are initialized randomly at the beginning and while during learning using the back propagation algorithm they are adjusted to appropriate values. Each speech utterance in the training database is used as a learning sequence to obtain optimal solution.Once the network has reached the desired performance, the learning stage is over and the associated weights are frozen.

3.4.3 ANN Testing

During testing phase, unknown test utterance is given as input to the network to produce an output. The network takes this input as observation sequence and generate output almost as good as the ones produced in the learning stage for similar inputs.

3.5 Display of Output

Input speech utterance is taken from the user and the system then processes and extracts its features using MFCC. The extracted feature vector is given as an input to the trained neural network which will produce output and is displayed at the GUI.

4. EXPERIMENTSAND RESULTS

The use of NOIZEUS database for gender identification, based on speech in a mismatched environment has resulted in the following observations. Different fuzzy cluster sizes have been taken to figure out the optimal solution along with different combination of hidden neurons in the hidden layer of neural network. From the experiments,it is observed that a system with 6 fuzzy clusters and 10 hidden neurons gives an optimal solution. Therefore, it is considered as the model to conduct further experiments on noise elimination techniques.

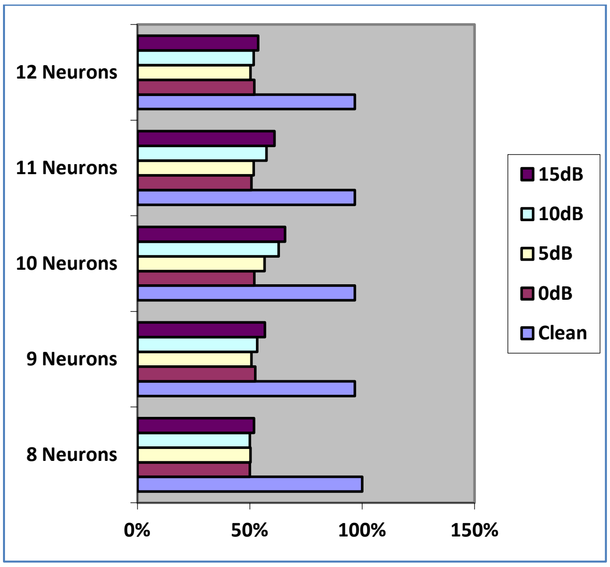


Fig 5: Shows Gender Identification Accuracy for the testing signals at SNRs of 0dB, 5dB, 10dB and 15dB for different number of hidden neurons and No. of Fuzzy clusters = 5

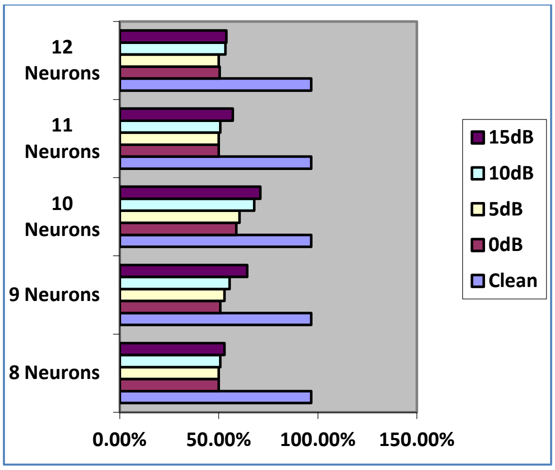


Fig 6: Shows Gender identification Accuracy for the testing signals at SNRs of 0dB, 5dB, 10dB, and 15dB for different numbers of hidden neurons and No. of Fuzzy clusters = 6

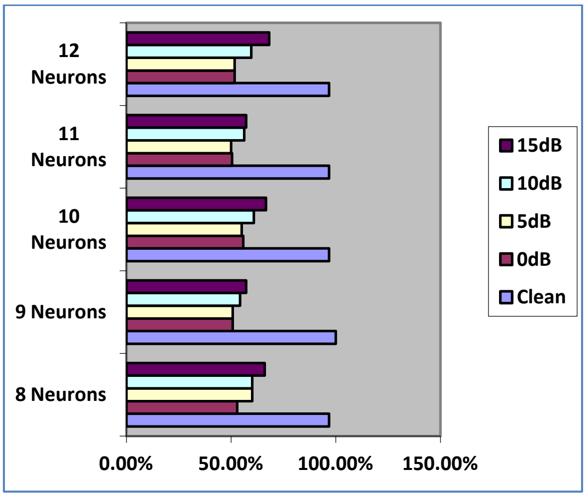


Fig 7: Shows Gender identification Accuracy for the testing signals at SNRs of 0dB, 5dB, 10dB, and 15dB for different numbers of hidden neurons and No. of Fuzzy clusters = 8

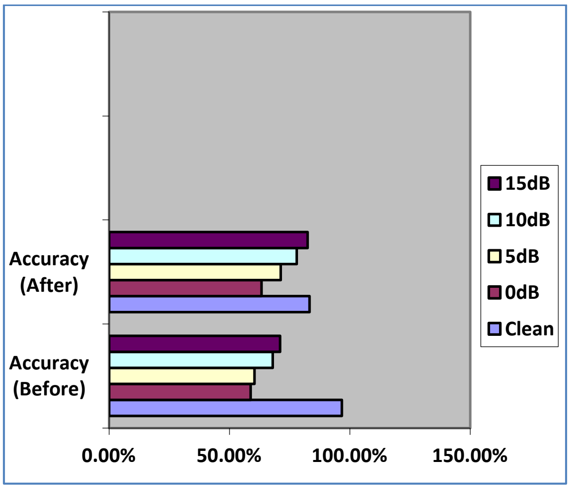


Fig 8: Shows Gender identification Accuracy before and after applying spectral subtraction method for the testing speech signals at SNRs of 0dB, 5dB, 10dB, and 15dB for No. of hidden neurons=10 and No. of Fuzzy clusters = 6

5. CONCLUSION

From experimental results, it can be concluded that Mel Frequency Cepstral Coefficient (MFCC) and Artificial Neural Network (ANN) with noise elimination techniques can identify gender of the speech signal in a mismatched environment better than without using noise elimination techniques. The computations has also been reduced due to the use of fuzzy c mean clusteringthat decreasing the number of segments by grouping similar segments in the training data.The highest identification rate that can be achieved by using the proposed model is 83.3%. This result is achieved by using MFCC with order 12 and ANN with one hidden layer, 10 neurons and tan-sigmoid transfer function for hidden layer and tan-sigmoid transfer function for output layer. An ANN is trained using 30 IEEE sentences (uttered by 3 male and 3 female speakers) in American English language and tested using noise signals at different SNR levels.

6. REFERENCES

[1] Madhavi S. Pednekar, KavitaTiware and SachinBhagwat, “Gender Distinction Using Short Segments of Speech Signal”, IJCSNS International Journal of Computer Science and Network Security, VOL.8, No.10, October 2008.

[2] Czyzewski, Andrzej; Dziubinski, Marek; Kotus, Jozef; Pawlik, Arkadiusz; Rypulak, Andrzej; Szwoch, Grzegorz, “Multitask Noisy Speech Enhancement System”, 26th International Conference: Audio Forensics in the Digital Age (July 2005).

[3] Mukherjee, R, Islam, T.Sankar, R, “Text dependent speaker recognition using shifted MFCC”, Southeastcon, 2013 Proceedings of IEEE, pp. 1 – 4

[4] L. Rabiner, B. H. Juang, “Fundamentals of Speech Recognition”, Pearson Education, 2009.

[5] Madhavi S. Pednekar, KavitaTiware and SachinBhagwat, “Continuous Speech Recognition forMarathi Language Using Statistical Method”, IEEE International Conference on „Computer Vision andInformation Technology, Advances and Applications‟, ACVIT-09, December 2009, pp. 810-816.

[6] Mohammad Reza Homaeinezhad, EhsanTavakkoli, Ali Ghaffari, “Discrete Wavelet based Fuzzy Network Architecture for ECG Rhythm Type Recognition: Feature Extraction and Clustering Oriented Tuning of Fuzzy Inference System” International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 4, No.3, September, 2011.

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

[8] Hu, Y. and Loizou, P. (2007). “Subjective evaluation and comparison of speech enhancement algorithms,” *Speech Communication*, 49, 588-601.