GFCC-Based Robust Gender Detection

M. A. Islam

Electrical & Electronic Engineering, International Islamic University Chittagong, Chittagong, Bangladesh
atiq.atrai@gmail.com

Abstract*— Gender classification technique is a part of the signal processing comprises with feature extraction and behavioural gender modelling. Fundamental frequency and pitch are mostly used as feature for gender detection due to their unique characteristics in voice source. In this study, Gammatone Frequency Cepstral Coefficient (GFCC)-based robust gender classification method has been presented. This study was accomplished using speech samples from a text-dependent data set. The prototype gender behavioural modelling was done using Gaussian mixture model (GMM) to obtain better performance and only clean signal was used to train the model. The performance of the proposed method was tested under both clean and contaminated conditions. The clean signal was contaminated using nine different noises at a range of signal-to-noise ratios (SNRs) from 0 dB to 10 dB. The obtained performance showed the proposed method was very robust against noise and the average performance at 0 dB SNR was almost 100% for female and 92% for male irrespective to noises. So, it could be said the proposed method performance was almost noise invariant.

Index Terms*— Gender classification, GFCC, GMM, Modelling, Robustness.

I. INTRODUCTION

Gender classification is an important biometric technique in signal processing in which audio or video or both signals from both male and female are used to identify speaker gender. It is very important to distinguish speaker gender to enhance speaker as well as speech recognition, because separate acoustic behavioural model for male and female provides better performance. In addition, this technique is applied in sorting caller in telephoning, identifying a target speaker's sex among a number of speakers, and intelligibility of man-machine interfacing. Furthermore, the correct gender detection is one of the important criteria for speaker emotion detection.

The voice signal availability makes its application in signal processing very popular compare to fingerprint, face identity, and video signal applications. It is well known, the human voice source information (fundamental frequency) is unique for men and women, which is the basic key to distinguish gender. Moreover, vocal tract-oriented characteristics like formant frequency, amplitude, and bandwidth are also used in gender recognition [1].

The most crucial and important part in any classification technique is to extract hidden feature from the object. Similarly, in gender classification technique feature extraction from speech signal is the most essential stage. A number of studies have been carried on to achieve robust gender classification. Some of them are acoustic feature-based: fundamental frequency and pitch combination [2] is one of the prominent features. It is to be mentioned that the fundamental frequency plays crucial effect on pitch perception and mostly used in identification of speaker gender. The second group contains auditory feature in which feature extraction process is done following auditory peripheral system: auditory nerve (AN) model-based gender classification [4] one of them.

In signal processing, Mel-frequency Cepstral Coefficient (MFCC) is very well known which is designed following auditory peripheral system mechanism. MFCC-based gender classification performance degrades significantly under noisy conditions due to the application of Fast Fourier Transform (FFT) in feature extraction even worse than individual pitch-based performance [3]. To address non-linearity of auditory periphery system and obtain robust performance in gender detection, neurogram-based method has been introduced [4]. In clean, neurogram-based performance was almost 100% irrespective to gender. It was observed in [4], the neurogram-based method's performance was fluctuated quite a lot with the change of types of noise and degraded havoc at 0 dB signal-to-noise ratio (SNR). The combination of pitch and 10^{th} order relative spectral perceptual linear predictive coefficients (RASTA-PLP) [5] provides robust performance in gender classification, but does not reflect non-linearities of the human auditory system. So, a method that reflects auditory periphery non-linearity and provides robust performance is necessary in gender classification technique.

A new method in gender detection including non-linearity of auditory system is presented in this study. An auditory peripheral property-based feature Gamma-tone Frequency Cepstral Coefficient (GFCC) [6] has been used to obtain robust gender classification. Motivated by the robust performance of the GFCC in speaker identification [6, 7] and speech recognition [8], it has been applied in gender classification. Moreover, the application of Gamma-tone filter and cubic operation in GFCC can capture physiological mechanism of the human auditory periphery system.

Once feature extraction is done; most crucial subject is to make gender behavioural model using latent information from extracted feature. In gender classification, gender prototype model is mostly implemented by using either statistical Gaussian mixture model (GMM) or one versus one (OVO) technique-based Support vector machine (SVM) [9]. SVM modelling provides comparable performance to GMM-based modelling in gender classification. The main drawback of SVM-based modelling is that, it requires all training samples at a time to create an object model which needs high configuration machine and also time consuming. Another disadvantage of SVM modelling is the selection of optimal values of cost and gamma variable. The SVM modelling, performance completely relies on the proper selection of these two parameters. On the other hand, GMM can estimate sufficient latent statistical variable from speech signals and needs a small number of mixture components.

In this study, the GMM modelling technique has been applied to create a gender behavioural model to obtain a fast gender detection system. The gender behavioural modelling was done using text-dependent clean speech signals. The performance of the proposed method was evaluated in clean and noisy conditions. Nine different noises were used to validate the proposed method extensively.

978-1-5090-6122-8/16/$31.00 ©2016 IEEE

The rest of the part of this paper has been arranged as follows: the methodological description of feature extraction and gender modelling are given in section II. The obtained performance of the proposed method and robustness issues is described in section III. In section IV, the summarize form of this presented study is given.

II. METHODOLOGY

The methodological procedure of the GFCC feature extraction is shown in Fig. 1. The gender detection process is shown in block diagram in Fig. 2. The assumption machine-based recognition is the features extraction in first step like auditory periphery system and second is to make decision based on stored information regarding testing object. So, the methodology section has two subsections: GFCC feature extraction and GMM-based behavioural modelling for further testing of the presented method's validity.

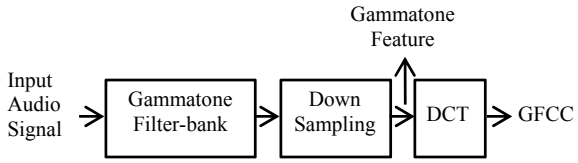

Fig. 1 GFCC feature extraction block diagram.

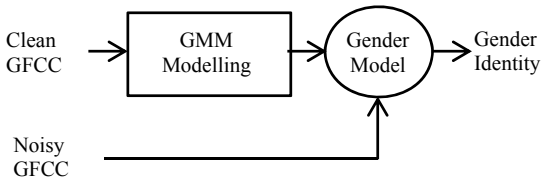

Fig.2 The proposed gender detection method's block diagram using GFCC as a feature.

A. GFCC-Feature Extraction

The Gamma-tone filter is more resemble with the human basilar membrane filter and can reflect psychological and physiological response of auditory cochlea filters. To simulate basilar membrane response, a fourth order Gamma-tone filter has been used. In human auditory system, there have 30000 neurons, which are connected selectively to basilar membrane filter. The basilar membrane is divided into a number of bands and each band is considered as a band pass filter. In this study, only 64 bands have been considered. Filter responses were simulated for centre frequency ranges from 50 Hz to 4 kHz considering the sampling frequency of input signal. So, the Gammatone filter response frequency remains same as it was for input signal.

Once the Gamma-tone time-frequency (T-F) responses were obtained it was down sampled to 100 Hz which is similar to 10 ms windowing technique application on T-F responses. A cubic root operation was applied to introduce cochlea non-linearity in GFCC features. This cubic root application reflects cochlea loudness compression property. The obtain feature is called Gammatone feature (GF). The GF has finer resolution at low frequencies compared to high frequencies. Moreover, the GF has better resolution than FFT-based linear time-frequency spectrogram resolution as shown in Fig. 2 in study [7]. Here is mentioned that, GF has 64 frequency channels which overlaps frequency information and correlates information. It is well known, too many correlation make worse behavioural prototype training model and degrades the system performance. Furthermore, 64-dimension is larger compare to MFCC, neurogram or other conventional features for gender detection.

To reduce the size of Gammatone features discrete cosine transform (DCT) was applied. DCT also de-correlates data and restore spectral feature to instantaneous form.

$$f(i,j) = \sqrt{\frac{2}{M}}\cos\left(\frac{\pi}{2n}.i.(2j-1)\right) \quad i,j = 1,2,\dots M \qquad (1)$$

The content of DCT matrix F is f(i,j), where n is the number of feature dimension. Here, M is 64. To obtain cepstral feature H from Gammatone feature G, DCT matrix F is multiplied by G.

$$H = D \times F \quad F = \{f(i,j)|i,j = 1,2,..M\} \qquad (2)$$

The obtain feature is not exact cepstral feature because a log operation is required between first and second frequency analysis for the de-convolution purpose. The obtained feature name is given Gammatone frequency cepstral coefficient (GFCC). It was observed in this study, the most of energy of Gammatone features retain in 1st to 23rd coefficients after DCT application which is consistent with the finding of the study [7]. Moreover, 1st coefficient was more susceptible to noise. According to suggestion [7], only 2nd to 23rd order GFCC (22-dimensions) was kept for robust gender classification under noisy conditions. The detail GFCC feature extraction procedure can be found in [6].

There are two basic differences between GFCC and MFCC. MFCC uses triangular filter to capture basilar membrane responses, whereas GFCC uses Gammatone filter to do so. Second, GFCC uses cubic operation to reflect cochlea non-linear properties and MFCC uses log operation for the same purpose. It was found in [12], the combination of Gammatone filter and cubic operation on spectral feature in GFCC provides better performance than MFCC.

Once clean GFCC feature was extracted; it was forwarded to create GMM-based behavioural gender model. Noisy GFCC feature was used only to validate the propose method robustness.

B. Gender Modelling

The most important in speech processing is to extract latent variable from the extracted feature from input signal. The performance accuracies of any system mostly depend on accurate modelling. GMM is mostly used in modelling for gender classification [2,4]. GMM follows Gaussian distribution (mean and deviation about the mean) of feature vectors. The main advantage of using GMM as classifier, it can represent each gender phonetic class distribution with few mixture components.

The GMM model is parameterized by mean vectors, covariance matrices and mixture weights from all mixture component densities. In this study, 128 GMM components were used to train the proposed method. It was observed that the application of diagonal covariance instead of full covariance de-correlated the off-diagonal covariance matrices elements and improved gender classification performances. GMM model male and female likelihood was computed as follows

$$P_g(x_t|\mu,\sum\alpha) = \prod_{i=1}^{N} P_g(x_{ti}|\mu,\sum\alpha) \qquad (3)$$

Here, P_g is the gender probability score which indicates gender of testing sample's speaker. N is the total number of speakers and i=1, 2, 3, ……, N are GMM mixture weight. x_{ti}

is the individual speaker's training sample. A is the Gaussian density with mean μ and covariance matrix $\sum_i$.

In testing stage, the testing sample from unknown gender was used to compare with gender trained model. The model which provided maximum likelihood probability score to that testing sample, that model indicated the gender of that testing sample. The probability density function was used to obtain maximum likelihood score of the testing sample's gender using following equation

$$X = \frac{1}{N}\sum_{i=1}^{N} P_g(x_{ti}|\mu, \sum \alpha) \quad (4)$$

If X_{male}(for any x_{ti}) > X_{female}(for any x_{ti}), it indicates the testing x_t is lied in male gender otherwise the gender is detected as female.

It was observed in this study, the application of logarithm in probability density function as it was applied in [4] increased the chance level of testing sample to be male. But, the overall performance was almost similar.

A 2 x 2 confusion matrix was created to provide the percentage accurate classification performance of the proposed method. Once, the target sample's gender was detected, a one was added corresponding gender in confusion matrix. Each instance of male and female gender testing was done independently.

III. Result and Discussion

This section has two parts. Initially, the proposed system's experimental setup has been described and the obtain results and analytical study of the presented method has been described later.

A. Experimental Setup

In this study, a text-dependent dataset named Universiti Malaya (UM) is an asset of University of Malaya has been used. The speech samples were recorded at a sampling frequency of 8 kHz from 39 native Malay speakers. There have 25 males and 14 females. This dataset was mainly used in speaker identification [10] and has also been used in gender detection [4]. Each speaker was asked to utter 'University Malaya' for ten (10) times. 70% samples from each gender were used for GMM-based gender prototype modelling and rest 30% samples were used to test the performance of the system following [4]. So, total 223 samples were used to train the proposed method and 117 samples (75 samples from males and 42 samples from females) were used to test the validation of this study. The behavioural prototype gender model was made using clean GFCC feature. The proposed system was tested in clean and noisy conditions. The input clean signal was contaminated with nine different types of stationary and non-stationary noises. Noises were white Gaussian, pink, street, babble, train, restaurant, exhibition, speech shaped noise (SSN), and factory noise. The noises were added to clean signal for a range of SNRs from 0 dB to 10 dB at a step of 5 dB SNR. The noisy audio signal was used to extract noisy GFCC feature.

B. Results and Discussion

This section presents the performance evaluation of the proposed method using GFCC as a feature. The propose study performance was evaluated for five times for each SNR of each noise and average result was recorded. It is to be recalled, the presented study was tested under both clean and noisy conditions. The gender classification performance of this study using text-dependent speech signal in clean condition is shown in the confusion matrix in TABLE I. It is seen from TABLE I result, the proposed method can detect male and female 100% in clean condition. The proposed method performance as shown in TABLE I indicates that the GFCC-based method can distinguish voice production sources faithfully and segregate male and female accurately. This robust performance could be due to the application of cubic root in GFCC and reflection of basilar membrane filter properties in Gammatone filter.

TABLE II shows the performance of the proposed method under noisy conditions. The results are explicated that the proposed method performances are almost noise invariant and robust to noises.

TABLE I
THE CONFUSION MATRIX OF THE PROPOSED GENDER DETECTION METHOD IN CLEAN.

Gender	Male	Female	Accuracy (%)
Male	75	0	100
Female	0	42	100

TABLE II
THE PROPOSED METHOD ROBUST GENDER DETECTION PERFORMANCE UNDER NOISY CONDITIONS USING NINE DIFFERENT NOISES FOR A RANGE OF SNRS.

Noise type	SNR	Male (%)	Female (%)	Overall (%)
White	10dB	97.33	100	98.67
	5dB	94.67	100	97.34
	0dB	74.67	97.62	86.14
Pink	10dB	100	100	100
	5dB	100	100	100
	0dB	96	100	98
Street	10dB	100	100	100
	5dB	96	100	98
	0dB	88	100	94
Babble	10dB	100	100	100
	5dB	100	100	100
	0dB	98.67	100	99.33
Train	10dB	100	100	100
	5dB	98.67	100	99.33
	0dB	96	100	98
Restaurant	10dB	100	100	100
	5dB	98.67	100	99.33
	0dB	97.33	100	98.67
Exhibition	10dB	100	100	100
	5dB	97.33	100	98.66
	0dB	78.67	100	89.34
SSN	10dB	100	100	100
	5dB	100	100	100
	0dB	100	97.62	98.81
Factory	10dB	100	100	100
	5dB	100	100	100
	0dB	98.67	100	99.34
Clean		100	100	100

The machine-based recognition is always subjected to additive noises and the system performance degrades with the increment of noise level. The performance of the proposed method was significantly dropped at 0 dB SNR for white Gaussian noise (which is considered as high frequency noise)

due to the inclusion of high frequencies information. It has been shown in [11]; the GFCC-based performance could improve under white Gaussian noise considering low frequency information. Besides this, the power spectral density of exhibition noise was used in this study was almost similar to white Gaussian noise and degraded the performance compared to other non-stationary noises. Otherwise, the proposed method performance was on average almost 100% for female and more than 92% for male irrespective of noises at 0 dB SNR. It is known that the fundamental frequencies of males and females are different and the study [5] showed the pitch level of females are higher than males which are distinguishing features. So, the finding of this study is the pitch information of audio signal helps to classify male and female accurately and GFCC feature can extract this from input audio signal.

Generally, the chance level of gender detection technique's performance is 50% i.e. either male or female. The study [4] performance was on average below 70% respective to noises at 0 dB SNR and showed results taking only 10 males and 10 females from UM dataset. This study showed result applying orthogonal polynomial on neurogram. A physiological auditory peripheral-based AN model was employed to compute neurogram. The best performance of the study [4] was shown for restaurant noise and has been compared with the presented study performance for same noise taking same number of speaker samples. The ENV neurogram-based performance has been copied in TABLE III from the study [4]. The comparison results are shown in TABLE III for a range of SNRs from 0 dB to 10 dB.

TABLE III
COMPARISON OF THE PERFORMANCE OF THE PROPOSED METHOD A RELEVANT METHOD UNDER CLEAN AND RESTAURANT NOISE FOR 10 MALES AND 10 FEMALES.

SNR	Male (%) GFCC/ENV	Female (%) GFCC/ENV	Overall (%) GFCC/ENV
Clean	100/93.3	100/96.7	100/95
10 dB	100/96	100/74	100/85
5 dB	100/90	100/64	100/77
0 dB	100/83	100/50	100/67

The GFCC-based presented method performance contrasting to the neurogram-based study was more robust and also very robust to noises. So, it can be said based on this study result; the proposed method can successfully distinguish male and female voice production system as well as identity of gender. The study [12, 13] showed the speaker identification performance improves largely after loudness suppression using cubic operation. This could be said that the inclusion of cubic operation in GFCC provides robust gender detection performance in this study. Moreover, it was observed in this study that the exclusion of cubic operation in GFCC dropped gender detection performance to a remarkable level under noisy conditions.

IV. CONCLUSION

This study is an application of GFCC as a feature in gender detection. GFCC feature has been extracted following auditory peripheral mechanism using Gammatone filter to simulate cochlea filters responses. Twenty-two (22) dimensional GFCC features which contain most energy of 64-channels have been used in this study to capture latent information of gender distinguishing feature. GMM modelling technique has been applied to make gender behavioural prototype model. The proposed method performance accuracy was 100% in quite condition. The proposed method performance was not reduced substantially with the increment of noise level which ensures the robustness of the proposed method. The performance of the proposed gender detection method was almost noise invariant and provided very robust score. The proposed method can also be studied for a large dataset like TIMIT and to dataset which contains different combination of language speeches to extensively validate this method.

ACKNOWLEGEMENT

I would like to thanks anonymous reviewers for their valuable comments those have enhanced this paper quality.

REFERENCES

[1] D. G. Childers, and Wu, K., "Gender recognition from speech. Part II: Fine analysis". The Journal of the Acoustical society of America, 90(4), 1991, pp.1841-1856.

[2] S. Slomka and S. Sridharan, "Automatic gender identification optimized for language independence," in TENCON'97 Brisbane-Australia. Proceedings of IEEE TENCON'97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications. 1997, (Cat. No. 97CH36162).

[3] G. Nisha Meenakshi and P. K. Ghosh, "Automatic gender classification using the mel frequency cepstrum of neutral and whispered speech: A comparative study," in Communications (NCC), 2015 Twenty First National Conference on, 2015, pp. 1-6.

[4] Mamun, Nursadul, Wissam A. Jassim, and Muhammad SA Zilany. "Robust gender classification using neural responses from the model of the auditory system." Functional Electrical Stimulation Society Annual Conference (IFESS), 2014 IEEE 19th International. IEEE.

[5] Y.-M. Zeng, Z.-Y. Wu, T. Falk, and W.-Y. Chan, "Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech," International Conference on Machine Learning and Cybernetics, 2006, pp. 3376-3379.

[6] Shao, Yang, Soundararajan Srinivasan, and DeLiang Wang. "Incorporating auditory feature uncertainties in robust speaker identification." IEEE International Conference on Acoustics, Speech and Signal Processing-2007, ICASSP'07. Vol. 4.

[7] Zhao, Xiaojia, Yang Shao, and DeLiang Wang. "CASA-based robust speaker identification." IEEE Transactions on Audio, Speech, and Language Processing. 2012, pp. 1608-1616.

[8] Shao, Yang, et al. "An auditory-based feature for robust speech recognition." IEEE International Conference on Acoustics, Speech and Signal Processing. 2009.

[9] Burges, Christopher JC. "A tutorial on support vector machines for pattern recognition." Data mining and knowledge discovery. 1998, pp. 121-167.

[10] M. Islam, M. Zilany, and A. Wissam, "Neural-Response-Based Text-Dependent Speaker Identification Under Noisy Conditions," in International Conference for Innovation in Biomedical Engineering and Life Sciences, 2016, pp. 11-14.

[11] Islam, Md Atiqul, Wissam A. Jassim, Ng Siew Cheok, and Muhammad Shamsul Arefeen Zilany. "A Robust Speaker Identification System Using the Responses from a Model of the Auditory Periphery." PloS one 11, no. 7 (2016): e0158520.

[12] Zhao, Xiaojia, and DeLiang Wang. "Analyzing noise robustness of MFCC and GFCC features in speaker identification." IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 7204-7208.

[13] Q. P. Li, "An auditory-based transfrom for audio signal processing," in Applications of Signal Processing to Audio and Acoustics, WASPAA'09. IEEE Workshop on, 2009, pp. 181-184.