Frequency Domain Linear Prediction-Based Robust Text-Dependent Speaker Identification

M. A. Islam

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

Abstract— **Speaker identification is a biometric technique of determining an unknown speaker's identity among a number of speakers using distinguish latent information of uttered speech. Crime investigation, security control, telephone banking and trading, and information reservation are some applications of this technique. Frequency Domain Linear Prediction (FDLP) is a time-frequency-based feature has been derived using 2-D autoregressive model. This feature was constructed from sub-bands short frame energies estimation. FDLP has been used in this study to propose a robust text-dependent speaker identification technique. The clean features were used to obtain speaker behavioural model. Support vector machine has been used to train the proposed method. This presented study was tested in both clean and noisy conditions to validate the method extensively. The proposed method got significant improved performance over all traditional methods performances in noisy conditions. The obtained performance was indicated; the proposed method was very robust to noises and showed consistent performance irrespective to noises.**

Index Terms*— *FDLP, Speaker Identification, Robust, SVM.

I. INTRODUCTION

The fundamental frequency, energy, pitch, and power are the hidden information carried by speech and unique characteristic for individual speaker. These features are varied from man to man due to the variation of physiology of voice production system. The uttered speech not only passes messages to listener but also provides information about gender identity, direction of the sound source, emotions, and talker identity. Ease availability and uniqueness of speech makes its application popular in speech processing.

Speaker identification is a part of speaker recognition. It is a biometric method which uses speaker voice signal to identify a target speaker by matching with a number of prototype speaker models [1]. Speaker identification is two types based on text: text-dependent and text-independent. In a text-dependent speaker identification system, recorded speech samples are aligned to a reference template from registered speakers. On the other side, speech is not limited to a specific length or sentence in a text-independent speaker identification system. This study presents a new approach to develop a robust text-dependent speaker identification system.

A significant number of studies and developments have been done to achieve robust performance on automatic speaker identification for more than four decades. The log operation-based Mel Frequency Cepstral Coefficients (MFCC) [2] and all pole system based Linear Prediction Cepstral Coefficients (LPCCs) [3] are most common feature in automatic speaker identification. LPCC is developed by following human voice production mechanism. On the other hand, MFCC is developed following the auditory peripheral physiology. LPCC features provide better performance in clean condition, but its performance drops significantly with the increment of noise level which causes substantial spectral distortion in the signal [4]. Similarly, Fast Fourier Transform (FFT)-based feature, MFCC, works better in matched condition but in noisy conditions; its performance declines to a remarkably low level [5]. Recently, Gammatone Frequency Cepstral Coefficient (GFCC) has been implemented following auditory peripheral mechanism to obtain robust score in speaker identification system but GFCC-based method's results were not substantially improved under noisy conditions [6]. To obtain robust speaker identification (SID) performance, Auditory Nerve (AN) model-based Neurogram [7] was used. Neurogram is a time-frequency-based synapse response but its extraction process from audio signal is complex and very time consuming. It was also provided comparatively poor performance under non-stationary noises [7].

It is well known, the speaker identification performance of Human listeners irrespective to noises is very robust. So, it is desirable to develop an automatic speaker identification system which performance will be comparable with human listening performance. The machine learning-based speaker identification performances degrade significantly when additive noise or distortion is added to clean signal. Hence, the goal of this study is to present a new method to achieve robust speaker identification performance.

In this study, 2-D autoregressive model-based FDLP [8] has been introduced for robust text-dependent speaker identification. The robust performance of FDLP-based text-independent speaker identification [8] motivates to explore its application in text-dependent speaker identification system. FDLP captures pitch information

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

from acoustic signal which enhance speaker identification performance.

In speaker identification system, Support Vector Machine (SVM) [9], Gaussian Mixture Model (GMM), and Gaussian Mixture Model-Universal Background Model (GMM-UBM) are employed to create speaker behavioural model. GMM and GMM-UBM provides almost similar performance and better in text-independent SID system. Sometimes, SID system's performance varies with the variation of modelling technique. In this study, SVM one versus rest (OVR) speaker modelling technique has been applied to provide robust text-dependent SID performance. The proposed FDLP-based method was validated in both clean and noisy conditions. This study show, SVM speaker modelling provides better performance than the study [7] which was done based on GMM-UBM speaker modelling technique using same dataset.

The residual portion of the presented paper has been arranged in following structure: the methodological description of the proposed method has been given in section II, the evaluation and findings of the presented study is described in section III, and the summarize form of this study is given in conclusion section.

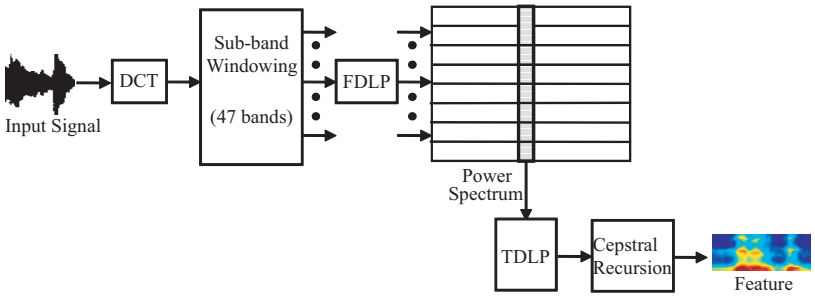

Fig. 1 Illustration of FDLP feature extraction schematic block diagram from an audio signal using 2-D auto regressive model.

II. Methodology

Fig. 1 presents FDLP feature extraction process and Fig. 2 presents the block diagram of the proposed method. Initially, silence period of input signal was removed. In general, each speech utterance contains silence period which is essential to separate words and make sentences intelligible to listeners. In a machine-learning speaker identification system, the removal of silent period is necessary to make a better speaker model, because this part of speech portion does not carry any information and make similarities among speakers in speaker behavioural models. In this study, the silence period of input signals were removed using a standard algorithm called voice-activity-detector (VAD) available in the voice-box toolbox [10]. Silence period free signals were forwarded as input to extract features for robust SID performance.

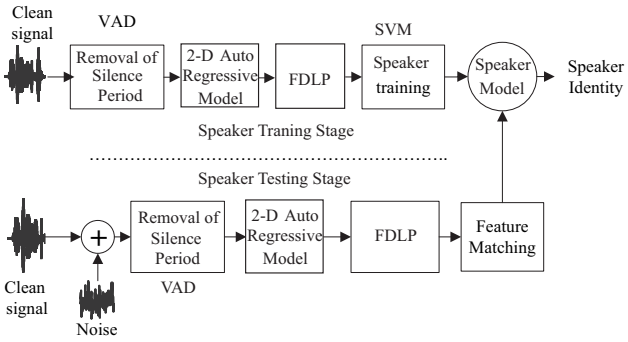

Fig. 2 The methodological block diagram of the proposed FDLP- based method for robust speaker identification.

A. Feature Extraction

A 2-D auto-regressive model-based frequency domain coefficient was developed by Ganapathy et al. [8] applying all poles technique. This feature has been developed based on high-energy peaks estimation of audio signal in the time-frequency (T-F) unit. In this study, the FDLP extraction was done based on the 2-D autoregressive model following [8]. The procedure has several steps. The acoustic time-domain signal was transformed into frequency domain using the DCT. The frequency domain signal was windowed into 47 linear sub-bands with an upper limit of ~ 4 kHz whereas lower limit was set to 0 Hz. Then the linear prediction system was applied to each sub-band to obtain the corresponding Hilbert envelope. A 25-ms length of window with an overlap of 40% between adjacent frames was used to compute 13 cepstral coefficients including the zero-order coefficient (C_0). Delta and acceleration coefficients were also included. So, the total number of FDLP feature dimension was 39 × m. Here, m is the number of frames. The FDLP feature e

B. Existing baseline features

The presented FDLP-based method's results were compared to the results of the existing features such as the GFCC, MFCC, and Neurogram. The feature extraction procedures of the baseline features are given in the following subsections.

i. MFCC

MFCC is a very common feature in automatic speaker recognition and also being used in speech recognition. This feature was computed from input signal using "rastamat toolbox" [11]. 40 bands of triangular filter with centre frequencies from 50 Hz to 4 kHz were used to extract MFCCs. A Hanning window having 25 ms length with 50 % overlap between adjacent windows was applied to extract MFCCs. The dimension of the obtained MFCCs was 13 per frame which is known as static coefficients. The dynamic and acceleration coefficients were also counted to obtain better performance. So, the dimension of MFCC was 39 in this study.

ii. GFCC

GFCC feature extraction process was developed following auditory peripheral system. Gammatone filter-bank was used to reflect basilar membrane filter-bank responses. Because, the Gammatone filter is more similar to basilar membrane band filters according to the physiological observation. The filter centre frequencies were qusi logarithmically spaced from 50 Hz to 4 kHz. Unlike MFCC, Gammatone filter responses were down sampled to 100 Hz. This down sampling was same as 10ms windowing and reduced the dimension of the Gammatone features. A cubic root operation was then applied to compress loudness. GFCC extraction process is different in two ways from MFCC extraction method. Firstly, the application of Gammatone filter-bank instead of triangular filter to simulate basilar membrane response. Secondly, the cubic root operation instead of log operation on power spectrum. It was reported in [6] that most of the speaker-specific information remains in the lowest 23 GFCCs (among 64 bands) due to the compaction property of the DCT. Because the zeroth-order (C_0) was mostly corrupted by noises, only the first 22 GFCCs (excluding C_0) from each frame were used in this study.

iii. Neurogram

Neurogram is the sysnapse responses were derived from an auditory peripheral-based model (AN model) responses. AN-model [12] is a useful simulation-based model for understanding physiological and mechanical mechanism of auditory system. Initially, the input signal was sampled to 100 kHz to adopt with requirement of the AN model and to replicate middle ear response meticulously. All input signals loudness was set at 70 dB. The Neurogram was constructed from the Synapse responses with a ranges of frequencies from 250 Hz to 1 kHz are spaced nonlinearly (using logarithmic space) into 12 bands of frequency called characteristics frequency (CF). The obtained synapse output of AN-model was windowed for 420 points with 50% overlap. This low resolution feature name was envelope (ENV) Neurogram. Only 12 bands i.e. low frequencies information was chosen to provide robust performance in speaker identification. More detail about ENV Neurogram extraction can be found in [7].

C. Speaker Modelling

SVM was used to make speaker behavioural model from the Matlab libsvm toolbox [13] to test the performance of the proposed method. There are two techniques in SVM modelling: one-versus-one (OVO) and one-versus-rest (OVR) classification technique. OVR speaker modelling technique has been applied in this presented work. SVM dimensional space is decided by the input feature vector. Seven (7) speech signals from each speaker were randomly chosen to create individual speaker prototype model and the rest samples were provided to test the identification score of the proposed method. It has been mentioned earlier the FDLP feature dimension was m × 39, where m was the number of frame. On the other side, the MFCC-, GFCC-, and Neurogram- feature sizes were $m \times 39$, m × 22, and $m \times 12$ respectively, where n was the number of frames of the speech signal.

Generally, the input features are normalized to zero (0) mean and one (1) standard deviation. This default normalizing procedure was applied in this study. Radial basis function (RBF) was applied as default to make speaker behavioural model. SVM speaker modelling was done setting cost (c) and gamma (g) function to 4 and 1respectively. The testing samples were then used to obtain unknown speaker identity by comparing with each and every speaker model. The matched speaker model provided maximum probability score to the testing sample and speaker was identified.

III. Results And Evaluations

A text-dependent dataset has been used in this study. A simple text 'University Malaya' was spoken for 10 times by each of the 39 speakers. A comfortable sound-resisted booth was used to record uttered speech. The sampling rate of the recorded speech signals was 8 kHz. This database was designed for the application in text-dependent speaker recognition systems [7].

The proposed FDLP-based performances with other existing methods performances have been described in this section. The performance was tested in both clean and noisy conditions. The clean testing signals were contaminated with white Gaussian (stationary) noise, pink (slow-varying) noise, and street (non-stationary) noise for a range of SNRs -5 dB to 15 dB in steps of 5 dB to proof the robustness of this presented method.

The performance of the proposed method along with alternative methods has been illustrated in Fig. 3. It is well known, any system's performances are reduced when noise level is increased. It is seen from Fig.3, the proposed method-based results are almost similar across different type of noises up to 5 dB SNR but dropped

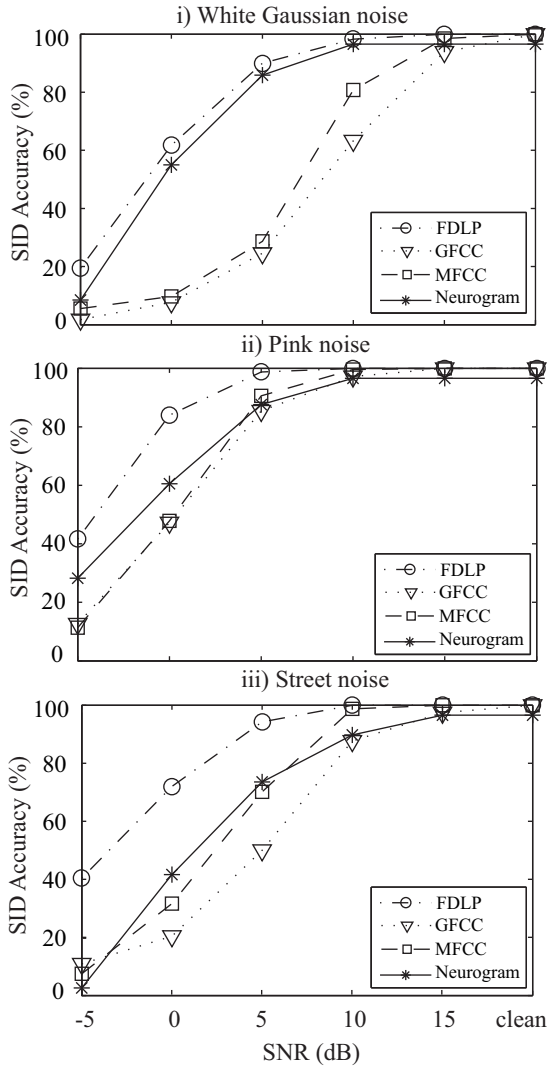

Fig. 3 Comparison of the proposed and substitutive methods performances using SVM speaker modelling. The speaker identification (SID) performance is shown for three different type of noises for SNRs -5 dB to 15 dB.

consecutively below that SNR level. It is also clear from Fig. 3, the FDLP-based method achieved the highest accuracy over existing methods performance irrespective to noises and SNR levels.

Furthermore, the Neurogram-based method's performance outperformed the other two existing methods performance, and was comparable to FDLP-based performance under white Gaussian noise. It is to be mentioned, the Neurogram was obtained considering only low frequencies information. The inclusion of low frequencies information provides better performance under noisy conditions [7]. Again, the proposed method showed a very consistent speaker identification accuracy across different types of noise.

It is to be noted that for SVM-based speaker modelling, MFCC-based method showed better identification results compared to those of GFCC-based system at all SNRs studied (could be due to the less number of support vectors in GFCC-based method). Another important point, the proposed method with low frequencies information (250 Hz to 1 kHz) provides improved performance compared to default system's performance which is studied in the study [7].

The performance shown in Fig. 3 implies that the proposed method was very robust to noises and can separate speakers successfully. The reason of the robustness could be the application of all pole system i.e. linear prediction system in FDLP which can capture speaker distinguishing vocal production system information more accurately.

IV. Conclusion

A robust text-dependent speaker identification method has been presented in this paper based on a novel feature FDLP. This feature was derived from the all pole system using 2-D regressive model. The presented method's performances were tested in clean and noisy conditions. The introduced method's performance outperformed all alternative methods for all SNRs under noisy conditions and performance was comparable in clean. The proposed method also achieved a consistent pattern across different type of noises and provided very robust performance. This study results implicate that the proposed method can model human voice production system very accurately which is the cause of robustness of this presented method. The proposed method can also be applied in speech recognition, gender classification, and speaker recognition. The channel variation study could be an interesting topic for the proposed method.

Acknowledgement

I would like to thanks to all anonymous reviewers for their helpful comments to make this study more acceptable.

References

1. Togneri, R. and D. Pullella, *An overview of speaker identification: Accuracy and robustness issues.* Circuits and Systems Magazine, IEEE, 2011. **11**(2): p. 23-61.
2. Davis, S.B. and P. Mermelstein, *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.* Acoustics, Speech and Signal Processing, IEEE Transactions on, 1980. **28**(4): p. 357-366.
3. Makhoul, J., *Linear prediction: A tutorial review.* Proceedings of the IEEE, 1975. **63**(4): p. 561-580.
4. Li, Q., F.K. Soong, and O. Siohan. *A high-performance auditory feature for robust speech recognition*. in *Interspeech*. 2000.
5. Chi, T.-S., T.-H. Lin, and C.-C. Hsu, *Spectro-temporal modulation energy based mask for robust speaker identification.* The Journal of the Acoustical Society of America, 2012. **131**(5): p. EL368-EL374.
6. Zhao, Xiaojia, Yang Shao, and DeLiang Wang. "CASA-based robust speaker identification." IEEE Transactions on Audio, Speech, and Language Processing 20.5 (2012): 1608-1616.
7. 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.
8. Ganapathy, S., S. Thomas, and H. Hermansky, *Feature extraction using 2-D autoregressive models for speaker recognition.* ISCA Speaker Odyssey, 2012.
9. Burges, Christopher JC. "A tutorial on support vector machines for pattern recognition." Data mining and knowledge discovery 2.2 (1998): 121-167.
10. Brookes M. Voicebox: Speech processing toolbox for matlab. Software, available [Mar 2011] from www.ee.ic.ac uk/hp/staff/dmb/voicebox/voicebox html. 1997.
11. Ellis, D.P., *PLP and RASTA and MFCC and inversion in Matlab.* 2005.
12. Zilany, M.S., I.C. Bruce, and L.H. Carney, *Updated parameters and expanded simulation options for a model of the auditory periphery.* The Journal of the Acoustical Society of America, 2014. **135**(1): p. 283-286.
13. *Chang C.-C. and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology, 2: 27: 1–27: 27, 2011," Software available at http://www. csie. ntu. edu. tw/cjlin/libsvm, 2011.*