EASR

Engineering and Applied Science Research

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

Bangla dataset and MMFCC in text-dependent speaker identification

Md Atiqul Islam* and An-Nazmus Sakib

Electrical and Electronic Engineering, International Islamic University Chittagong, Chittagong - 3414, Bangladesh

Received 3 August 2018
Revised 28 December 2018
Accepted 2 January 2019

Abstract

Automatic Speaker Identification (SID) is a challenging research topic that is mostly done based on either text-dependent or text-independent speech materials. Generally, an automatic SID system is designed based on English speech. The main goal of this study is to present a text-dependent dataset based on Bangla speech. We explored three different feature extractors as a front-end processor: the Mel-frequency Cepstral Coefficient (MFCC), the Gammatone Frequency Cepstral Coefficient (GFCC), and a newly developed feature – a Modified MFCC (MMFCC) to simulate SID accuracy. The SID accuracies were simulated under clean and noisy conditions. Four types of noises were added to clean signals to generate noisy signals for a range of signal to noise ratios (SNRs) from -5 dB to 15 dB. A standard dataset based on English speech is also presented to compare the SID accuracies with the presented Bangla dataset SID accuracies. The second goal of this study is to examine MMFCC and introduce its novelty in a text-dependent SID system. It is seen from the results of this study, the MMFCC-based method results significantly outperform the MFCC and GFCC-based methods under noisy conditions and produce comparable results in a clean environment.

Keywords*:* Bangla dataset, UM dataset, SID system, MMFCC, GFCC, MFCC, Robust performance

1. Introduction

An automatic SID system is a biometric system that uses a speaker voice signal to build speaker behavioral statistical models and identify a target speaker by matching the test sample against those of statistical models. A biometric SID system employs several processes: collection of speech samples, pre-processing, feature extraction from audio signals, extraction of statistical latent information from front-end features to generate speaker models, and feature matching with speaker models. So, the most primitive part in an automatic SID system is to build a speech dataset.

The applications of an SID system are increasing day by day with the advancement of technology. Progress is difficult because of the uniqueness of each individual's speech. Currently, Intel uses face detection technology in laptops and Siri in i-phones uses biometric authentication to access these devices. The most recent smartphones use finger print scanners and alpha-numeric passwords as a means of access to these devices and individual accounts. A biometric SID system can be explored as an alternative to existing means of security access because of speech signal availability, feasibility, and separation ability. Moreover, a biometric SID system has vast application in many fields including crime investigation, banking over a telephone network, and to search for a person in a large corpora of speech data.

A biometric SID system has two systems based on speech signals: text-dependent and text-independent. A SID system is said to be text-dependent when a fixed utterance is spoken by same speaker several times and among those samples, a few are used for training and the remainder are used for testing the system performance. In contrast, a system is said to be text-independent when speech is not limited to a specific length of time or sentences while training and testing samples are different. In this paper, a text-dependent SID system is presented.

The most useful datasets have been built using English utterances (TIMIT, YOHO, TIDIGIT, and GRID, for example) [1-2] and the availability of Bangla dataset is very rare in SID studies. More precisely, we have found only one online full-text SID publication using Bangla words from 10 speakers containing 80 samples in total [3]. So, it is expected that a freely available online-based Bangla dataset will significantly contribute to biometric SID system studies and this is the novelty of the current study.

The features are used in an SID system are categorized in a broad sense into two groups, auditory peripheral-based features and voice production system-based features. Mel-frequency Cepstral Coefficients (MFCC) [4], Gammatone Frequency Cepstral Coefficient (GFCC) [5], and Neurograms [1], among others, are examples of auditory peripheral-based features. In contrast, Linear Prediction Cepstral Coefficients (LPCCs) [6] based on an all pole system, Perceptual Linear Prediction (PLP) coefficients [6] and Frequency Domain Linear Prediction (FDLP) systems

*Corresponding author.
Email address: atiq.atrai@gmail.com
doi: 10.14456/easr.2019.7

[7] are in the second category. We did this study using auditory features.

The MFCC is considered a standard feature in SID systems. This feature is used as a baseline feature to establish feature performance in a newly developed SID system. The MFCC-based method provides better SID accuracy for a clean signal, but its accuracy degrades with an enhanced SNR level [6]. In contrast, the GFCC-based method [5] also provides similar performance to MFCC for a clean signal, but has improved performance at higher noise levels. It was found [8] that application of a cube root in GFCC is the main factor that improves its SID score under noisy conditions. Very recently, a Modified MFCC (MMFCC) [9] was introduced in a text-dependent SID system that produced 100% accuracy for clean signals. However, it obtains significantly improved SID accuracy over a conventional MFCC under noisy conditions. The objectives of this study are to introduce a text-dependent Bangla dataset, explore MFCC, MMFCC, and GFCC to simulate SID accuracies for this dataset, and to compare the Bangla dataset performance with that of an English speech-based dataset.

The remainder of this paper is organized as follows. Data collection and feature extraction procedures with their block diagrams are presented in the Methodology section. The simulation results for the Bangla and English datasets using various features is shown and an analytical study of the results are discussed in the Results section. Finally, the study is summarized in the Conclusions section.

2. Materials and methods

In this section, a methodological description of the current study is presentated including data collection for a new Bangla speech dataset. This is followed by a description of the experimental setup, front-end feature extraction, and speaker modeling. The block diagram shown in Figure 1 depicts the SID system.

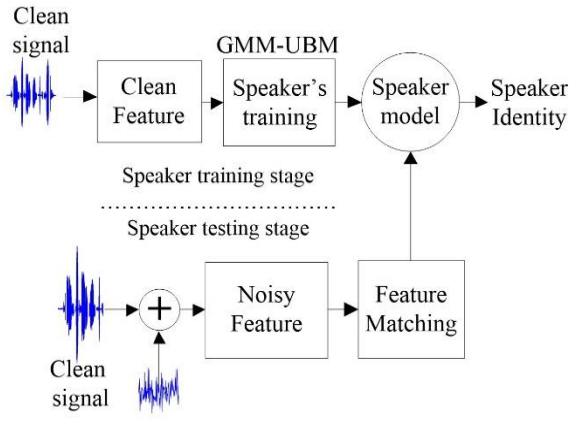

Figure 1 A methodological block diagram for the presented SID system. The dotted line separates the training and testing stages of this study. There is no overlap between training and testing samples.

2.1 Dataset description

In this section, the description of Bangla and English accent-based datasets is given. The most important, primitive, and time-consuming task is to record speech utterances from speakers for use in an automatic SID system.

2.1.1 Bangla dataset

Speakers were recruited to provide speech samples for this dataset. Verbal consent was obtained from each speaker after the purposes of this data collection were clarified. Forty (40) native speakers with a high variation in age (8-29 years old) were chosen for this dataset. There were 36 male and 4 female volunteer speakers. We attempted to recruit an equal number of female speakers, but religious sensibilities prevented their cooperation. Each speaker was asked to utter "Ami vhat khai", meaning, "I eat rice". Each speaker did this in three modes: fast, normal, and slow. Each mode contained ten (10) samples. There were 30 samples per speaker and in total 1,200 sentence utterances.

All speeches samples were recorded using a mobile handset employing a 16-bit pulse code modulation (PCM) encoding technique. The collected data were saved in ".wav" audio file format to support them in MATLAB, in both old and new versions. Generally, recordings were made with a sampling frequency of 16 kHz, therefore, this dataset's sampling frequency was kept at 16 kHz. Each utterance required, on average, 2 seconds. However, all data for the Bangla dataset were collected in a natural environment with no soundproof booth. The entire dataset is freely available at: https://drive.google.com/drive/folders/1TNcp6t7_ZJ1E3TDhqZTW_DKhrkXMHbs1?usp=sharing

2.1.2 UM dataset

A known English accent-based text-dependent dataset, the UM dataset [10], was used in this study as a standard to assess the quality of the Bangla dataset. This dataset contains speech samples of thirty-nine (39) Malaysian native speakers. Each person was asked to utter "University Malaya" ten (10) times. So, the number of speech samples was 390. Each speaker was instructed to deliver speech in normal mode and data was recorded in a soundproof booth. The speech materials were collected at a sampling frequency of 8 kHz by researchers at the University of Malaya.

The sound pressure levels (SPL) of these two datasets are depicted in Figure 2. The SPL of a volunteer's speech (x) with N samples was calculated as follows:

$$P = 20 \times log_{10}\left(\frac{\sqrt{\sum_{i=1}^{n}\frac{x_i^2}{N}}}{20 \times 10^{-6}}\right) \quad (1)$$

Here, 20×10^{-6} Pa is the smallest sound pressure that we can hear. It is seen from Figure 2 that each speaker's SPL had a unique fundamental frequency and the SPL varied with the environment. Bangla and UM dataset speakers have an average SPL 78.33 dB and 66.53 dB, respectively. The Bangla dataset has comparatively less standard deviation in SPL than the UM dataset. Individual SPL uniqueness is such that the SPL can be used as a speaker and accent distinguishing feature.

The sample Pearson correlation coefficient (r) for five different speakers was computed to evaluate the degree to

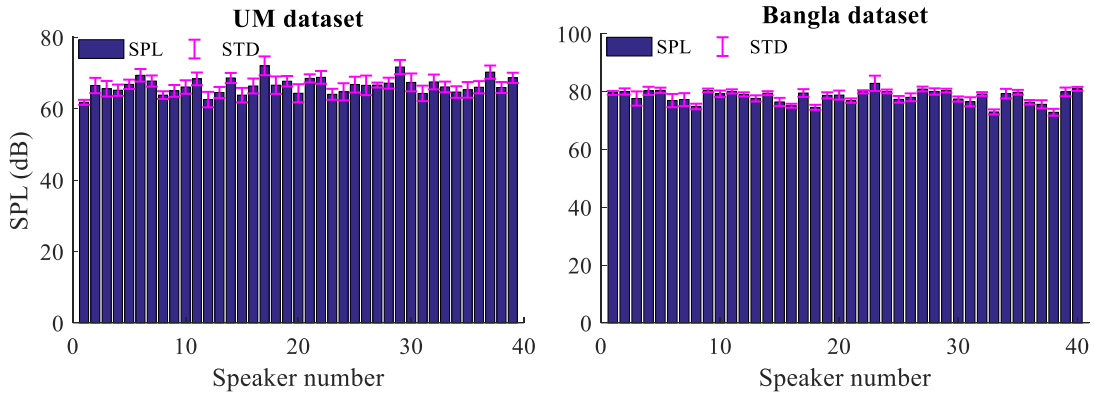

Figure 2 The average sound pressure level (SPL) with standard deviations (STD) of the UM dataset (left) and Bangla dataset (right) speakers.

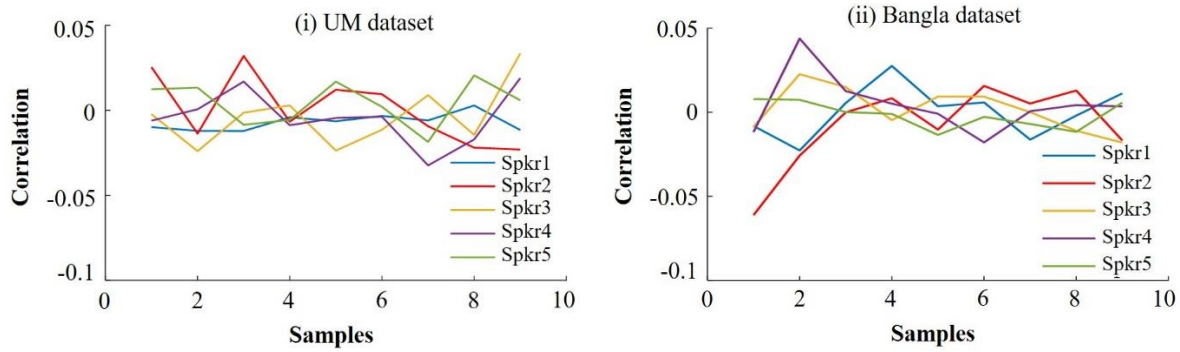

Figure 3 The correlation coefficient among samples of individual speakers from the UM (left) and Bangla datasets (right).

which individual speakers are distinguishable from each other. We calculated the correlation score (r) using following equation 2:

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{\left[N\sum x^2 - (\sum x)^2\right]\left[N\sum y^2 - (\sum y)^2\right]}} \quad (2)$$

N is the number of elements in a speech sample. Each sample (x) is compared with next sample (y) to calculate a correlation score between samples of a speaker. Since, there were ten samples for each speaker, the number of cross correlation scores was nine, excluding the reference sample's correlation score. The resulting correlation coefficient scores for five different speakers are shown in Figure 3. It can be seen from Figure 3, both dataset's speakers are linearly separable and have almost similar patterns of sample correlation.

2.2 Experimental setup

Figure 1 illustrates the experimental procedure used in this study. Thirty-nine out of 40 speakers were from the Bangla dataset to maintain similarity with number of speakers in the UM dataset [10]. The SID accuracies were evaluated in both clean and adverse environments. Only the normal mode of speech from Bangla dataset was used in this study. Seven samples from each speaker were taken randomly to obtain speaker models and the remaining three samples were kept for further testing of system performance. The noisy signals were obtained by adding four different noises to clean signals: white (stationary) noise, pink (slow-varying) noise, babble (speech-shaped) noise, and street (non-stationary) noise with SNRs ranging from -5 dB to 15 dB in steps of 5 dB.

2.3. Feature extraction

In this stage, the MMFCC, MFCC, and GFCC feature extraction procedures from input speech signals are described.

2.3.1 Modified Mel-frequency cepstral coefficient (MMFCC)

MMFCC is a similar type feature unlike MFCC. The MMFCC extraction procedure is most similar to MFCC except that the use of a cube root operation instead of a log operation and computation of frame energy, which may also be called an envelope. The envelope is computed using equation (3):

$$E(i,k) = C(i, 1+j : j+L) \times C(i, 1+j : j+L)' \quad (3)$$

where, $j = (k-1) \times overlap$, is the starting point of each window, i is the number of channels, L is the window length and k is the number of frames. In this study, a 25ms window with a 50% overlap between adjacent frames was used. The MMFCC computational block diagram is given in Figure 4.

Initially, a raw signal is windowed with 25ms and a discrete Fourier transform (DFT) is applied to obtain the frequency spectrum of that signal. This spectrum is forwarded to 25 bands of a rectangular filter in which the

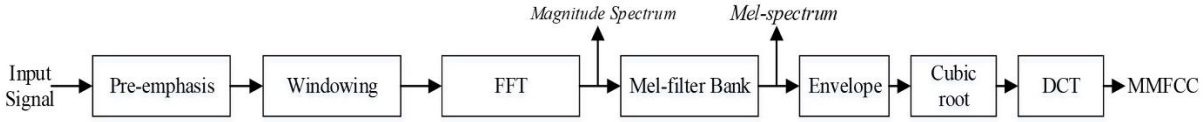

Figure 4 A block diagram of MMFCC extraction from an input speech signal.

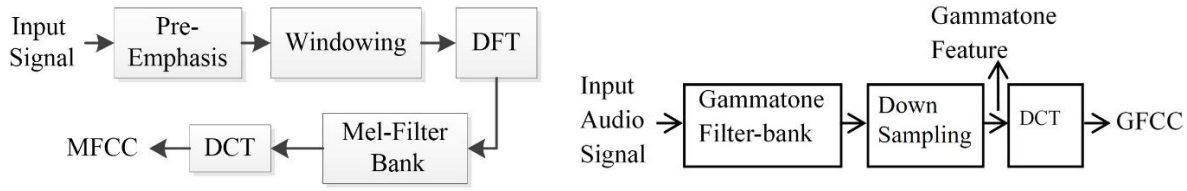

Figure 5 The MFCC (left) and GFCC (right) feature extraction process from an audio signal.

frequency range is 50 Hz to 3.8 kHz to obtain best performance. The energy in each band is calculated using equation (3). A scale of loudness suppression (cube root) is applied to the envelope to emulate cochlear suppression and the warping effect using equation (4):

$$y(i,j) = E(i,j)^{1/3} \quad (4)$$

A logarithm operation was not applied because there was wide fluctuation of variance at low energy. Finally, a discrete cosine transform (DCT) was applied to revert the energy spectrum to a time domain signal. This feature is called MMFCC. The simulated MMFCC feature average dimension was 25 ×180.

2.3.2 Mel-frequency cepstral coefficient (MFCC)

A methodological block diagram for MFCC feature extraction from an input speech signal is shown in Figure 5. MFCC and MMFCC extraction from an input signal are exactly the same until the mel-scale filter-bank in Figure 4. A frequency range of 50 Hz to 3.8 kHz with 25 bands is also applied to compute the MFCC feature as in the MMFCC and GFCC features. A DCT is applied on the filter output to obtain a cepstral coefficient. In this study only static information was considered and dynamic coefficients (del and ddel) were excluded. It was observed in this study that the dymaic coefficients are most affected by noises. So, the average MFCC feature dimension was 175 x 13. Here, there were 13 cepstral coefficients and 175 frames.

2.3.3 Gammatone Filter Cepstral Coefficient (GFCC)

GFCC is different than MFCC in two ways, in terms of the filter is used (Gammatone filter instead of Traingular filter) and addition of cochlea non-linearity (cubic root application instead of log operation).

The GFCC feature extraction proces is illustrated step by step in Figure 5 (right side). This feature extraction has three steps: (i) generation of gammatone filter impulse response, (ii) obtaining the Gammatone filter (GF) feature, and (iii) application of DCT matrix. In this study, a 64 bands of a fourth order Gammatone filter is used which centered frequencies are scaled in an equivalent rectangular bandwidth (ERB) scale with a frequency range used for MFCC and MMFCC for comaprative study. Fast Fourier Transform (FFT) is applied on both Gammatone impulse response and signal of each band and then, the scaler multiplication is done between them. At the end of this stage, an Inverse FFT (IFFT) is applied to reverse the frequency domain signal into the time domain. This obtained feature is then downsampled to 100 Hz to reduce the size of the feature and a cube root operation is applied producing what is known as a Gammatone feature (GF). Finally, a DCT matrix was created and multiplied by the GF features to obtain the GFCC. The dimension of resulting feature was 64 x 220 on average. It was found [11] that the coefficients above the 23rd band were almost zero and 1st band information amplitude was too large. So, only the 2nd to 23rd band coefficients were considered in this study. Inclusion of higher band information makes the system slow and more succeptible to noise. So, the GFCC size was on average 22 x 220.

2.4 Speaker Modelling

The classifier GMM usually uses a Gaussian probability distribution function (PDF) to model a speaker, accent, or gender [12]. The Gaussian PDF is parameterized by mean, weight and covariance matrices for individual mixture components (M). So, the behavioural model is presented as the weighted sum of individual PDFs. The PDF of a random vector feature $x_n$for a covariance matrix $\sum_g$, where g =1, 2, 3, ….. M is calculated as:

$$f(x_n|\lambda) = \sum_{g=1}^{M} \pi_g \mathcal{N}(x_n|\mu_g, \sum_g) \quad (5)$$

where π_g and μ_g are the weight and mean of the gth mixture components. So the GMM model can be presented as:

$$\lambda = \{\pi_g, \mu_g, \sum_g | g = 1,2, \dots M\} \quad (6)$$

Finally, the matching probability is computed for a given GMM model as:

$$p(\boldsymbol{X}|\lambda) = \prod_{n=1}^{T} p(x_n|\lambda) \quad (7)$$

where $X = \{x_n | \, n \in 1 \dots . T\}$ is a sequence of feature vectors.

The GMM can be an effective model without prior knowledge of the speech content and it has become a popular classifier in identification tasks. The GMM modelling technique uses an expectation maximization (EM) algorithm [13] to estimate latent feature parameters to iteratively

increase the accuracy of the data used in the model. This model also able to adapt to new data using maximum likelihood linear regression (MLLR) or maximum a-posteriori (MAP) adaptation [14]. It provides better identification accuracy when adapted with UBM [14]. We used GMM-UBM as a back-end processor because of its characteristics and prior successful applications in SID systems [1-2, 9, 12]. We did not explore use of a deep neural network (DNN) [15] as a back-end processor in this study due to our limitations. This would be a worthy topic of study in the future.

3. Result and Analytical Study

In this section, Bangla and UM dataset-based SID accuracies are illustrated using MMFCC, MFCC, and GFCC. All of the aforementioned feature-based speaker behavioral models were created using the same GMM-UBM algorithm so that a fair comparison can be made. This section is divided into two sub-sections, Results and Analytical study.

3.1 Results

The Bangla and UM dataset-based SID accuracies are shown in Figure 6. Since, the speakers from both dataset are linearly separable as shown in Figure 3, it is expected that all automatic SID methods should provide an almost 100% SID rate and the obtained results support this hypothesis. It was also observed that the use of a cube root operation in both GFCC and MMFCC provides better SID scores that log and operation-based MFCC. So, it can be postualted that the cube

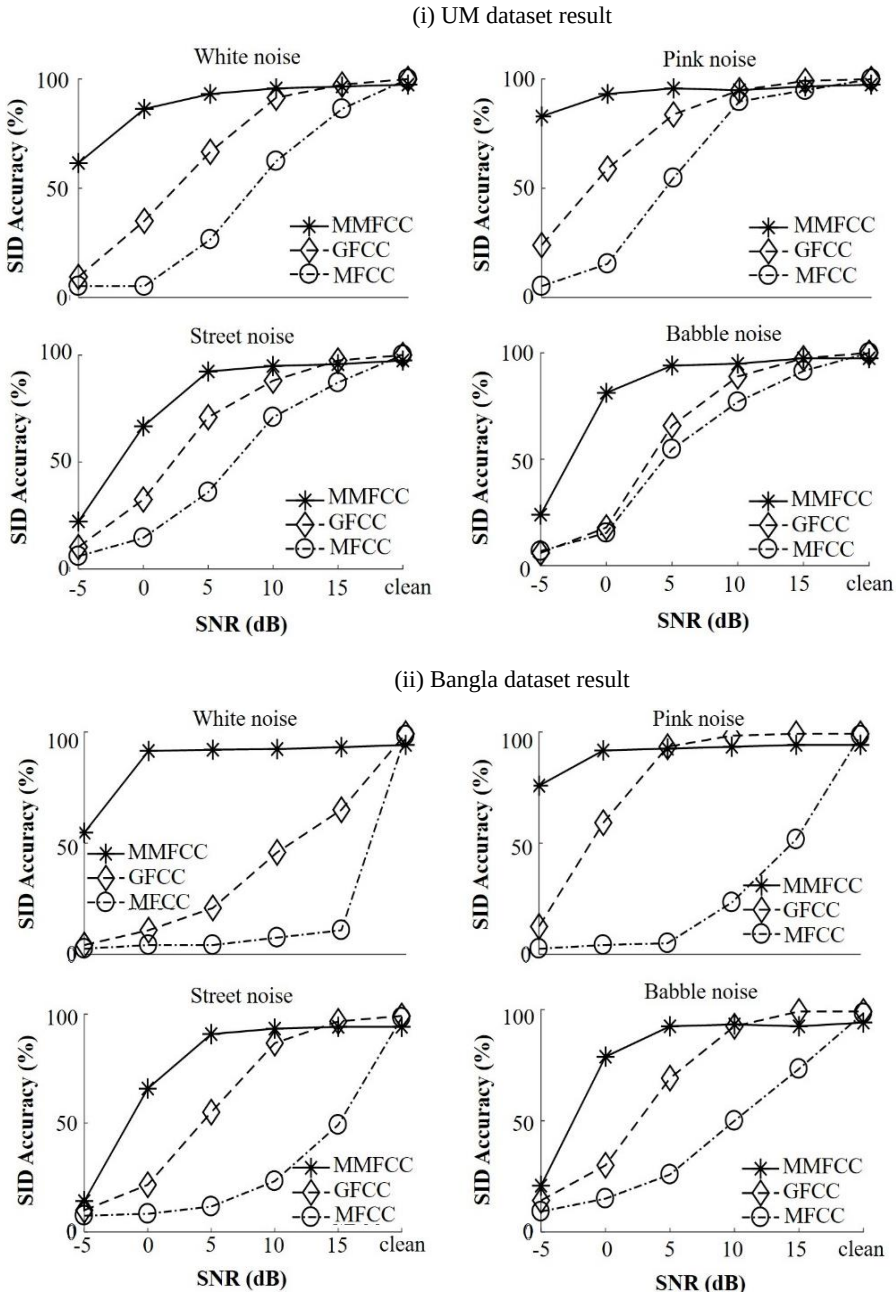

Figure 6 Bangla dataset (top) and UM dataset (buttom) speech materials are used to simulate this result using MMFCC, GFCC, and MFCC to show the SID performance and compare their perfromance.

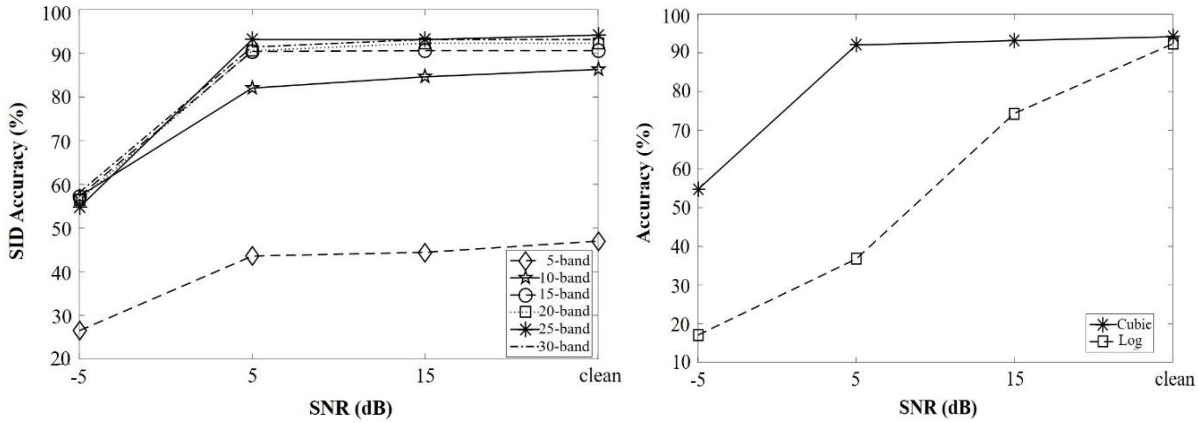

Figure 7 The graphic representation of channel number (left) and cochlear non-linearity (right) effect on SID in clean and noisy conditions using a MMFCC-based method.

root application emulates human cochlear non-linearity and adds better loudness suppression than a log operation in front-end features. The SID accuracies of MMFCC-based method were very robust irrespective of noise and the dataset used below 5 dB SNR. In contrast, MFCC and GFCC-based SID accuracies fluctuated with the types of noise and datasets. The SID accuracies of MFCC and GFCC-basedGFCC-based methods decreased almost lineraly with incremental increases in SNR levels. However, the GFCC-based method provided the best perfromance for pink noise while the MFCC-based method produced almost similar SID scores irrespective of the dataset under babble noise. The results indicate that the MMFCC model outperformed other methods under low SNR levels (SNR < 5 dB). The obtained MMFCC performance in this study was quite similar to that of a previous study [10].

In the current study, the Bangla and UM datasets had almost identical SID performance and MFCC and GFCC-based model performance were also similar to those of previous studies [2, 10-11]. Moreover, similar SID accuracies of both datasets indicate that the Bangla and UM datsets were of similar quality as is supported by Figures 2 and 3. MMFCC produced significanlty better SID scores that the MFCC and GFCC methods irrespective of the datset used and noise under low SNR conditions. Additionally, the SID accuracy of the MMFCC was significantly improved over a recent Auditory Nerve (AN) model [16] synapse response-based study of SID accuracies with the UM dataset irrespective of noise [17].

3.2 Analytical study

We explored MMFCC to analyze the effects of various parameters in supporting robust SID performance. The Bangla dataset speech materials were used to investigate these effects on SID performance. White noise was used to obtain noisy signals. The effect of channel number, cochlear non-linearity, frequency selectivity, and speech utterance mode on SID accuracy will be presented below.

3.2.1 Channel number effect

Figure 7 presents the impact of channel number on the SID system. To investigate this effect, speech signals with a frequency range from 50 Hz to 3.8 kHz were simulated using 5, 10, 15, 20, 25, and 30 filter channels. It can be seen from Figure 7 (left) that the MMFCC with 25 filter channels produced the best SID score with a clean signal and its perfromance is comparable to a 30 channel-based method. In contrast, 15 filter channels generated the best SID score in adverse environments. The performance decreased when decrementing the number of channels for a clean signal. It improved under noisy conditions except when 5 filter bands were used. Based on the results presented in Figure 7, the narrowband filter response produced comparatively better SID results compared to the wideband filter response. This is consistent with a previous study [1].

3.2.2 Non-linearity Effect

In an SID system, a log or cube root operation [18-19] is used to mimic cochlear non-linearity. Here, we investigated the difference in performance resulting from log and cube root operations, to model cochlea non-linearity in the SID system. This was done with MMFCC extraction for each of these operations. The results are depicted in Figure 7 (left). The cube root-based MMFCC clearly outperformed log-based MMFCC under clean and noisy conditions. This is consistent with the findings of pervious studies [8, 20]. Additionally, Figure 6 shows that the cube operation-based MMFCC and GFCC provided much better performance than the log operation-based MFCC method under both clean and noisy conditions.

3.2.3 Frequency slectivity effect

The frequency of an acoustic signal in a human cochlea is analyzed. Selective neurons fire according to the frequency to which they are exposed and this frequency information is transmitted to the brain. So, it is important to learn which frequency range is most important in identifying speakers under both clean and noisy conditions.

To study the effect of frequency resolution on SID system, four different frequency ranges, 50 Hz to 1 kHz, 2 kHz, 3 kHz, and 8 kHz respectively were selected with a fixed number of channels (25). The results of this experiment are shown in Figure 8. It can be seen that SID accuracy for all ranges of information for clean signals is quite similar. Low frequency information provides very robust SID performance irrespective of SNR as shown in Figure 6. This has been reported by other researchers [1]. It can be clearly seen from Figure 8 that the SID accuracy decreases with the incremental increases in frequency under noisy conditions. The worst performance is in the 8 kHz range. This is because white noise has a high frequency spectral density. However, the SID performances was also evaluated for street (low

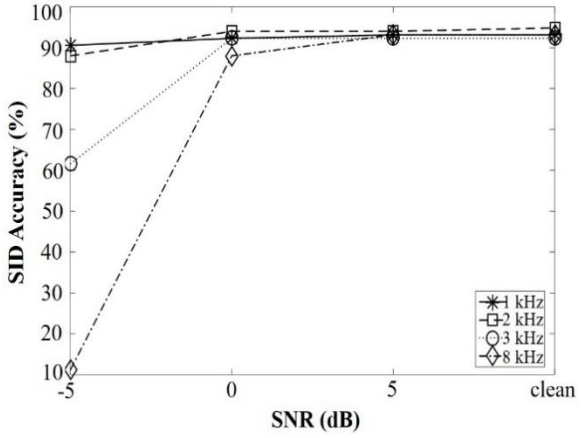

Figure 8 The frequency resolution effect on SID accuracy using MMFCC as feature.

frequency) noise and babble (speech shaped) noise for SNR -5 dB. The SID accuracies were 4.17%, 9.07%, 11.11%, and 17.48% for street noise and 9.4%, 13.68%, 17.98%, and 21.37% for babble noise for a frequency limit 1 kHz, 2 kHz, 3 kHz, and 8 kHz respectively Based on these results, the SID system achieved high accuracy under stationary noise with the low frequency, but a high frequency range is required for non-stationary noise. Moreover, this study implies that a narrowband filter is better in improving SID performance than a wideband filter under noisy conditions.

3.2.4 Effect of speech utterance mode

The Figure 9 shows the influence of three different modes of specch utterance from the Bangla dataset in an SID system. Three modes of speech (fast, normal, and slow) from the Bangla dataset were used to investigate the utterance effect mode on SID system performance. Each utterance mode was simulated with a frequency range from 50 Hz to 3.8 kHz with 25 channels. Figure 9 shows that the SID accuracies varied with the speech utterance mode under noisy conditions, but the accuracies are almost similar for clean signals. It can be seen from Figure 9 that the medium (normal) mode of speech provides the best SID performance. Slow utterance mode performance is comparatively lower

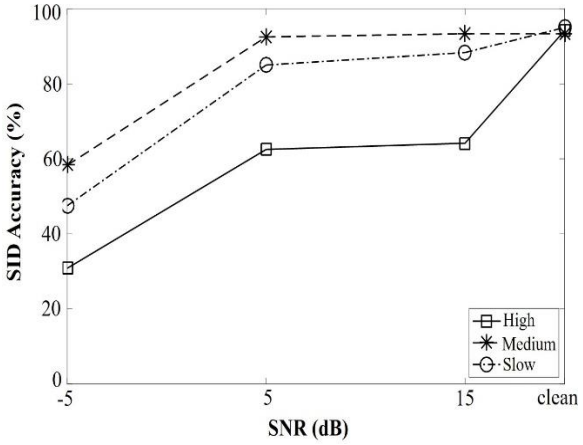

Figure 9 The utterance mode effect on SID system performance under clean and noisy conditions. Three modes of speech utternaces from the Bangla dataset were used to simulate this result.

than that of the medium utterance mode. The high (fast) utterance mode of speech shows poor performance since it originates from contaminated conditions that affect the human listener.

4. Conclusions

This study presents a Bangla dataset for a text-dependent SID system. To asses the quality of this dataset, another text-dependent dataset, the UM dataset, was also used in this study. These two datasets were compared in terms of SPL, sample Pearson correlation coefficients, and SID accuracies. MMFCC, MFCC, and GFCC were used to simulate SID accuracies under both clean and noisy conditions. Noisy speech signals were generated by distorting clean signal with white, pink, street, and babble noise. A range of SNRs from -5 dB to 15 dB were applied to test these methods using the Bangla and UM datasets under noisy coniditions. The simulated results show that both dataset speakers are linearly separable and have almost 100% SID accuracy with clean signals. The GFCC and MFCC-based methods SID score varies under noisy conditions depending on dataset. In comparison, the MMFCC-based method SID accuracies are very robust to noises, datasets, and mode of utterance. Additionally, the MMFCC-based method significantly outperformed the MFCC and GFCC-based methods under noisy conditions below 5 dB SNRs irrespective of noise pattern. The simulation results also indicate that inclusion of a cube root operation instead of a log operation in the front-end processor produces significantly improved SID performance. Additionally, a narrowband filter-bank generates better SID perfromance than a wideband filter. In the future, this dataset can be used to study gender detection, speech utterance mode separation, cues to changes in voice mode. MMFCC can be used in text-independent speaker and speech identification.

5. Acknowledgment

We would like to thank all subjects for their selfless effort to build this Bangla dataset to augment research on audio signal processing. We would also like to thank the reviewers for their valuable comments and suggestions to improve this work.

6. References

[1] Islam MA, Jassim WA, Cheok NS, Zilany MSA. A robust speaker identification system using the responses from a model of the auditory periphery. PloS one. 2016;11(7):e0158520. doi: 10.1371/journal.pone.0158520.

[2] Reynolds DA, Rose RC. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE trans speech audio process. 1995;3(1):72-83.

[3] Das D. Utterance based speaker identification using ANN. Int J Comput Sci Eng Appl. 2014;4(4): 15-28.

[4] Davis SB, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process. 1980;28(4):357-66.

[5] Shao Y, Srinivasan S, Wang D. Incorporating auditory feature uncertainties in robust speaker identification. 2007 IEEE International Conference on Acoustics,

Speech and Signal Processing; 2007 Apr 15-20; Honolulu, USA. USA: IEEE; 2007. p. 277-80.
[6] Makhoul J. Linear prediction: a tutorial review. Proceedings of the IEEE. 1975;63(4):561-80.
[7] Ganapathy S, Thomas S, Hermansky H. Feature extraction using 2-D autoregressive models for speaker recognition. Odyssey 2012: The Speaker and Language Recognition Workshop; 2012 Jun 25-28; Singapore.
[8] Xiaojia Z, Wang De. Analyzing noise robustness of MFCC and GFCC features in speaker identification. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; 2013 May 26-31; Vancouver, Canada. USA: IEEE; 2013. p. 7204-8.
[9] Islam MA. Modified mel-frequency cepstral coefficients (MMFCC) in robust text-dependent speaker identification. 2017 4th International Conference on Advances in Electrical Engineering (ICAEE); 2017 Sep 28-30; Dhaka, Bangladesh. USA: IEEE; 2017. p. 505-9.
[10] Islam MA, Zilany MSA, Wissam AJ. Neural-response-based text-dependent speaker identification under noisy conditions. In: Ibrahim F, Usman J, Mohktar M, Ahmad M, editors. International Conference for Innovation in Biomedical Engineering and Life Sciences, ICIBEL 2015; 2015 Dec 6-8; Putrajaya, Malaysia. Singapore: Springer; 2015. p. 11-4.
[11] Zhao X, Shao Y, Wang D. CASA-based robust speaker identification. IEEE Trans Audio Speech Lang Process. 2012;20(5):1608-16.
[12] Hansen JH, Hasan T. Speaker recognition by machines and humans: a tutorial review. IEEE Signal Process Mag. 2015;32(6):74-99.
[13] Bilmes JA. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute. 1998;4(510):126.
[14] Togneri R, Pullella D. An overview of speaker identification: accuracy and robustness issues. IEEE Circ Syst Mag. 2011;11(2):23-61.
[15] Ghahabi O, Hernando J. I-vector modeling with deep belief networks for multi-session speaker recognition. Odyssey 2014: The Speaker and Language Recognition Workshop; 2014 Jun 16-19; Joensuu, Finland. p. 305-10.
[16] Zilany MS, Bruce IC. Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. J Acoust Soc Am. 2006;120(3):1446-66.
[17] Zilany MS. A novel neural feature for a text-dependent speaker identification system. Eng Appl Sci Res. 2018;45(2):112-9.
[18] Stevens SS. On the psychophysical law. Psychol Rev. 1957;64(3):153-81.
[19] Stevens S. Perceived level of noise by Mark VII and decibels (E). J Acoust Soc Am. 1972;51(2B):575-601.
[20] Li Q, Huang Y. An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. IEEE Trans Audio Speech Lang Process. 2011;19(6):1791-801.