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

A novel neural feature for a text-dependent speaker identification system

Muhammad S. A. Zilany*

Department of Computer Engineering, Faculty of Computer Science and Engineering, University of Hail, Hail 2440, Saudi Arabia

Received 31 March 2017
Accepted 18 August 2017

Abstract

A novel feature based on the simulated neural response of the auditory periphery was proposed in this study for a speaker identification system. A well-known computational model of the auditory-nerve (AN) fiber by Zilany and colleagues, which incorporates most of the stages and the relevant nonlinearities observed in the peripheral auditory system, was employed to simulate neural responses to speech signals from different speakers. Neurograms were constructed from responses of inner-hair-cell (IHC)-AN synapses with characteristic frequencies spanning the dynamic range of hearing. The synapse responses were subjected to an analytical function to incorporate the effects of absolute and relative refractory periods. The proposed IHC-AN neurogram feature was then used to train and test the text-dependent speaker identification system using standard classifiers. The performance of the proposed method was compared to the results from existing baseline methods for both quiet and noisy conditions. While the performance using the proposed feature was comparable to the results of existing methods in quiet environments, the neural feature exhibited a substantially better classification accuracy in noisy conditions, especially with white Gaussian and street noises. Also, the performance of the proposed system was relatively independent of various types of distortions in the acoustic signals and classifiers. The proposed feature can be employed to design a robust speech recognition system.

Keywords: Speaker identification, Auditory-nerve model, MFCC, GMM, SVM

1. Introduction

The dynamic variation in acoustic features conveys important speech and speaker-specific information to the listener. However, these features degrade substantially when the speech signal is subjected to environmental noises or distortions, and, thus, an acoustic feature based automatic speech and speaker identification system suffers from poorer performance under noisy conditions. Alternatively, human performance on both of these tasks is robust to noise or distortions in speech signal [1]. In this study, a robust neural feature is presented in the context of developing a text-dependent speaker identification (SID) system.

Efforts have been made over the last few decades to develop speaker identification systems that work well both under quiet and noisy conditions. Most of the traditional methods derive features by mimicking the response properties of the cochlea such as the mel-frequency cepstral coefficients (MFCCs) [2] or by modeling the vocal tract of the human auditory system, such as the linear prediction cepstral coefficients (LPCCs) [3]. In general, MFCC, LPCC and variant methods provide a reasonable performance in quiet and matched conditions [4-5], but in mismatched conditions (training on clean data, but testing under noisy conditions), the performance declines to a very low level due to the substantial spectral or temporal distortion in the speech signal [6-8].

To achieve better performance under noisy conditions, Shao et al. [9] proposed a Gammatone frequency cepstral coefficient (GFCC) feature based on the responses of an auditory system based Gammatone filter-bank. Recently, a frequency domain linear prediction (FDLP) feature was also developed based on a speech integrated short-frame energy estimation [10] to propose a robust speaker identification system. Under quiet conditions, the performance of the FDLP and GFCC based methods was comparable to that of a MFCC based method. However, performance was not substantially improved under noisy conditions [11]. Based on the cortical representation and tensor factorization of a speech signal, Wu et al. [7] proposed a feature that provided slightly better results under noisy conditions. However, the auditory model employed in this method was linear, although it is well established that the auditory system is highly nonlinear, and thus this method cannot handle the effects of sound presentation level. Therefore, this method was not tested for text-dependent databases.

In this paper, the identification of a speaker is proposed using a feature based on the neural responses of a physiologically based computational model of auditory nerve (AN) fibers [12]. This AN model reflects most of the observed nonlinear properties of the peripheral auditory

*Corresponding author. Tel.: +9665 5012 8991
Email address: ms.zilany@uoh.edu.sa
doi: 10.14456/easr.2018.11

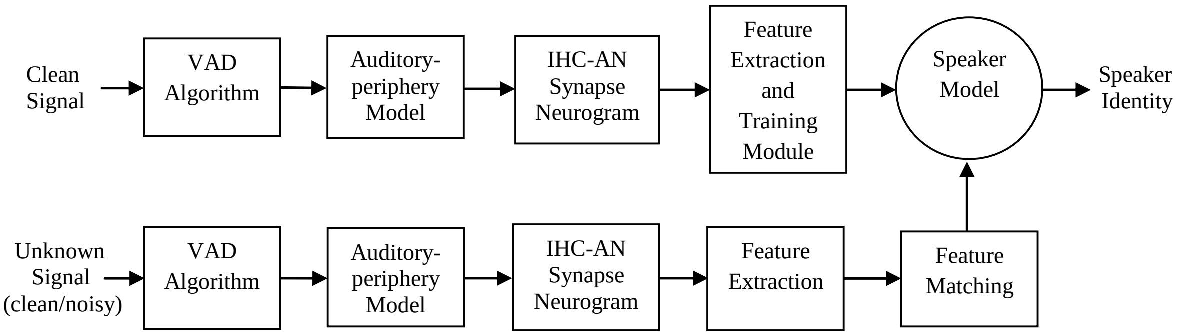

Figure 1 Schematic diagram of the proposed neural-feature based speaker identification (SID) system

system, such as nonlinear tuning, compression, two-tone suppression, and adaptation in the inner hair cell AN synapse. This approach was motivated by the observation that neural responses from physiological recordings are robust against different types of noises [13], and the model employed in this study successfully captured this phenomenon.

Based on simulated responses using the same auditory-periphery model [12], a number of speech intelligibility metrics such as the neurogram similarity index measure (NSIM) [14] and the neurogram orthogonal polynomial measure (NOPM) [15] were previously proposed. Both of these metrics reasonably predicted the subjective intelligibility scores for listeners with normal hearing and hearing loss under mismatched conditions. Results of phoneme classification using the simulated neural responses from the same AN model also showed a substantial improvement over existing baseline methods, especially under noisy conditions [16]. These observations motivated the development of speaker identification and verification systems using features extracted from simulated neural responses at different levels (e.g., discharge generator, inner hair cell AN synapses) of the auditory system [17-18].

Figure 1 presents a detailed block diagram of the proposed neural response based SID system. Clean and noisy speech signals were subjected to a voice activity detection (VAD) algorithm before simulating the neural responses by the AN model. A synapse neurogram is a 2D time-frequency representation (similar to spectrogram) which is constructed by simulating the inner hair cell AN synapse responses to speech signals over a wide range of characteristic frequencies (CFs). An analytical function was used to accurately estimate the instantaneous discharge rate of each AN fiber. The resulting neurogram contained important underlying information about the identity of the speaker that was used to train and test the speaker model. A feature-ranked algorithm was employed to reduce the number of features necessary to capture the underlying speaker specific information. In this study, three commonly used classifiers, the support vector machine (SVM), Gaussian mixture model (GMM), and Gaussian mixture model-universal background model (GMM-UBM) were employed for training (using clean signals) and testing (unknown clean/noisy signal) of the SID system. These standard classifiers were used to evaluate the performance of the SID system in quiet and under mismatched conditions. We hypothesized that the simulated neural response would capture adequate information about the speaker to reflect the pattern of human speaker identification performance in quiet and adverse conditions.

2. Methodology

In this article, a text-dependent speech database was employed in which each of 39 speakers uttered ‘Universiti Malaya’ ten times. The speech stimuli were recorded in a sound-proof booth, and the sampling frequency was 8 kHz. This database was recorded for use in the text-dependent speaker recognition systems [17]. Since silent periods do not contain any useful information (but could cause similarities among speakers), a voice-activity-detector (VAD) algorithm [19] was used in the present study to remove silent periods from the input speech signal.

2.1 AN model and feature extraction

The computational AN model developed by Zilany et al. [12] was employed in this study to simulate neural responses to speech signals from different speakers. The AN model has several stages. Each stage represents a phenomenological description of a major functional component in the auditory periphery from the middle ear to the auditory nerve [12]. The input to the AN model was the speech signal, and the output was the probability of instantaneous discharge rates from the inner hair cell AN synapse section of the model (instead of the output of the model discharge generator). The final stage of the model (discharge generator) provides discharge times of the AN fiber responses (post-stimulus time histogram) by a renewal process which includes the effects of refractoriness. However, it requires simulation of multiple repetitions of the same stimulus. In order to make the system computationally more efficient, an analytical function was applied, in this study, on the synapse responses requiring no repetition of the stimulus. This was done to accurately estimate the mean instantaneous discharge rate that takes into account the effects of absolute and relative refractory periods of the AN fiber responses [20-21].

The synapse responses were simulated for a range of characteristic frequencies (CFs) of AN fibers. The CF is the most sensitive frequency for an AN fiber that corresponds to a given place (tuned to that frequency) on the basilar membrane (BM) of the cochlea. A schematic block diagram of the AN model can be found in Zilany and Bruce [21]. The AN model requires the original speech signal to be resampled to 100 kHz, which is required to resemble the physiological frequency response properties of different parts of the AN model [21]. The sound presentation level of each speech signal was set to 70 dB SPL (~ conversational speech level) [22-23]. The estimated mean instantaneous discharge rates of each CF were divided into frames using a Hamming window with a length of 25 ms and a 60% overlap among adjacent frames. Then neural responses were averaged for synchronization to frequencies up to ~67 Hz [1/(100×10^{-6}×250×0.6)]. It has been reported in the literature that the neurons in the midbrain (inferior colliculus) show synchronized responses to the envelop of the signal (~20-100 Hz) [24]. In this study, the resulting neurogram does not include information about the temporal fine structure (which might go up to ~4 kHz).

The output of a population of model IHC-AN synapses displays the time-varying discharge rates as a function of time and is referred to as the synapse neurogram. Alternatively, a spectrogram is a FFT based representation of an acoustic signal. In this study, a synapse neurogram was constructed from the responses of 26 model IHC-AN synapses with CFs ranging from 250 to 4000 Hz (logarithmically spaced). To be consistent with physiological observations, the responses of high, medium and low spontaneous rate (SR) fibers were simulated for each CF and weighted by 0.6, 0.2 and 0.2, respectively [25].

It was reported in [18] that the responses of model AN fibers above 850 Hz (index of CF 12 and above) were not quite correlated (between clean and noisy responses) for the same speech samples with varying amount of noise. Thus, we considered only the responses of IHC-AN synapses with CFs up to 850 Hz (index 12) to develop a robust speaker identification system. The neural features were then ranked using a Laplacian feature selection technique to reduce redundant information and also to improve the learning process [26]. In the current study, the feature dimension was reduced to 10 (instead of 12), which provided a comparable performance (± 4%) when all data were employed.

2.2 Existing baseline features

In this section, a brief description of three existing baseline feature based methods (MFCC, GFCC, and FDLP) is provided. These methods were employed in this study to compare the performance of the proposed system using the neural feature.

2.2.1 MFCC

The mel-frequency cepstral coefficient was computed based on the linear cosine transform of the log power spectrum of an incoming sound signal. In this study, the VAD algorithm was also applied to remove the silent periods from the speech signal. The “Rastamat toolbox” developed by Ellis [27] was then used to compute MFCC coefficients from the processed signal. The signal was first divided into frames using a Hanning window of 25 ms with an overlap of 50% among consecutive frames, and MFCCs were calculated. Each frame consisted of 39 MFCC features: Ceps (13 cepstral coefficients), Del (13 derivatives of ceps) and delta del (13 derivatives of Del).

2.2.2 FDLP

The frequency domain linear prediction (FDLP) feature was developed in [28] based on high-energy peaks in the time-frequency (T-F) domain. In this study, a 2-D autoregressive model proposed by Zhao et al. [11] was employed to extract FDLP features. In this method, the time-domain signal was converted to a frequency domain signal, which was divided into 47 linear sub-bands. Then, the Hilbert envelop for each sub-band was computed. The same windowing technique employed in the MFCC was applied to compute 39 coefficients for each frame.

2.2.3 GFCC

The GFCC feature was derived from the responses of a Gammatone filter-bank, which physiologically more closely resembles the cochlear filter-bank. The same window size and overlap between frames, as employed in MFCC, were used to compute the GFCCs. A nonlinear cubic root operation was applied in estimating the GFCC feature [11]. For each frame, 64 coefficients were computed, and only the first 22 coefficients (the first was excluded due to noise) were considered for the SID task [11].

2.3 Speaker modeling

In this work, three standard classifiers, SVM, GMM, and GMM-UBM, were used to estimate the speaker model. The performance of these classifiers for this particular text dependent SID task was evaluated. Undistorted (i.e., clean) signals were always used to develop the speaker model (in the training phase), whereas both clean and distorted signals were used in the testing phase.

2.3.1 Support Vector Machine (SVM)

The Matlab Libsvm toolbox [29] was used in this study to generate the speaker model using the SVM classifier. Seven speech samples from each speaker were used in the training stage. The remaining three samples were used to evaluate the SID performance of the system. The proposed method used m × 10 features for each speech sample, where m was the number of temporal envelop points in the processed neurogram. There were n × 22 features for the GFCC based method, whereas for the MFCC and GFCC based methods, n × 39 features were used for each speech sample, where n was the number of speech frames.

In this study, the proposed neural feature was normalized in the training stage to have a mean of zero and a standard deviation of one, and the default kernel function of radial basis function (RBF) was used. The best parameters associated with the RBF kernel function were chosen to ensure the best accuracy using a cross validation algorithm. In the present study, the magnitude of cost function (c), gamma (g), SVM type (s), and shrinking parameter (h) were set to 4, 1, 0, and 0, respectively. However, it was observed that MFCC and FDLP based systems provided a better SID accuracy with a different set of parameters, $c = 1$ and $g = 0.0125$. These were employed to estimate the performance of the two methods. Finally, the speaker model with the maximum decision value was used to determine the identity (speaker) of the unknown test speech signal.

2.3.2 Gaussian mixture model (GMM)

With the application of the expectation maximization algorithm [30], GMM served as a common standard classifier for SID tasks. For all methods, 39 GMM speaker models were generated using training samples (UM database). In the proposed method, 10-dimensional feature vectors were obtained using 16 mixture components of the GMM. The performance was found to be degraded slightly with a higher number of mixture components. Alternatively, MFCC, FDLP, and GFCC based GMM speaker modeling was done using 32 mixture components that exhibited a maximum SID performance for these methods. Finally, features extracted from the test signals were compared to all GMM speaker models to identify the speaker.

2.3.3 GMM-UBM

In GMM-UBM, the universal background model optimized the range of speaker-specific features, and then GMM speaker models were generated using those features. In this study, training data were adapted using the maximum a-posteriori (MAP) adaptation technique [31]. To achieve the maximum SID performance using the proposed feature, 64 mixture components were used. However, it was observed that MFCC and FDLP based methods provided a better SID performance with 128 mixture components. In GMM-UBM speaker modeling, features were arranged in n × n (dimension × frames), and means, co-variances, and weights of the UBM were calculated from the training data to adapt to the GMM classifier.

3. Evaluation results

This section presents the performance of the SID system using the proposed neural feature under quiet and noisy conditions. To examine the robustness of the proposed feature under distorted or environmentally noisy conditions, the unknown test speech signals were added to white Gaussian (stationary) noise, pink (slow-varying) noise, and street (non-stationary) noise at different SNRs.

The estimated SID accuracy using the proposed feature was also compared with the accuracy of the MFCC, FDLP, and GFCC based SID systems. To ensure the reliability of the proposed method, its performance was independently evaluated four times at each SNR. The average results were reported using GMM, SVM, and GMM-UBM classifiers. The standard deviation of the SID scores at each SNR was found to be negligibly small and, thus, is not reported here. For a fair comparison, the performance of the baseline feature-based systems was also evaluated four times, and the mean SID scores were reported.

Figure 2 shows the SID performance of the proposed system (solid line) along with the identification performance of the baseline-feature-based methods using the GMM (first column), SVM (second column) and GMM-UBM (third column) as the speaker model. The performance is shown as a function of SNR. The performance of all methods in quiet conditions was very comparable (~100%). As an unknown speech signal was distorted by noise, the SID performance of all systems declined according to the level of noise added to the clean signal.

Using the GMM as a classifier (left panels in Figure 2), the proposed neural feature based system outperformed all existing baseline systems in white Gaussian noise at all SNRs studied. Also, the GFCC based system exhibited better performance compared to the SID accuracy of the MFCC based system in noisy conditions, which is also consistent with the observations in [9]. It is also obvious for the GMM in Figure 2 that the performance of all methods under pink noise was better compared to the results with the other two types of noise conditions. The performance using the proposed feature decreased gradually as a function of the noise level (SNR). The SID accuracy of the proposed method was relatively similar across different types of noise. Alternatively, the performance of the baseline feature-based systems was strongly dependent on the type of distortion (i.e., noise) in the acoustic signal.

The middle panels (second column) in Figure 2 show the performance of the SID system using the SVM as a classifier. In general, the results are similar across different types of noise for each SNR. It is evident that the FDLP based method showed the highest accuracy irrespective of noise and SNR levels. However, the proposed neural feature based method exhibited better results than the other two methods. The accuracy of the proposed method was comparable to that of the FDLP based SID system, except under conditions with distortions due to street noise. In general, the SVM speaker modeling produced better accuracy compared to those

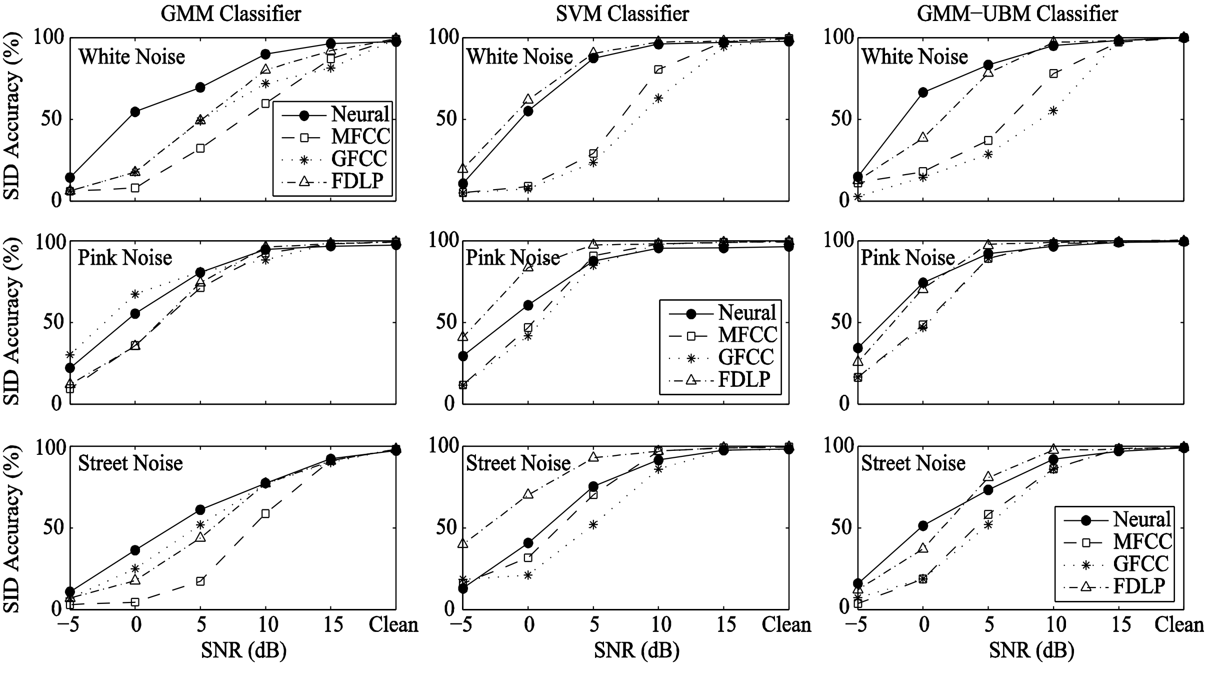

Figure 2 Performance of the SID task using the proposed neural and existing (MFCC, FDLP, and GFCC) features with the GMM (column one: left panels), SVM (column two: middle panels), and GMM-UBM (column three: right panels) as a classifier. The performance is shown for three different types of noise (top: white, middle: pink, and bottom: street) at SNRs ranging from -5 to 15 dB in steps of 5 dB.

applying the GMM classifier for the task (left vs. middle panels). Again, the proposed feature exhibited a very consistent SID accuracy across different types of noise. However, the FDLP based method also achieved a very consistent result across different types of distortions (including under white Gaussian noise) using the SVM as a classifier. It is notable that for this classifier, the MFCC based method provided better SID accuracy compared to those of the GFCC based system. This could be attributed to the use of a low number of support vectors in the GFCC based SID system.

The performance of the speaker identification systems using the GMM-UBM as a classifier has been shown in the right panels of Figure 2. Again, it is obvious that the proposed method outperformed the other methods under white Gaussian noise and also in mismatched conditions in general. Unlike baseline feature-based methods, the performance of the proposed method was very consistent across different types of noise. For this classifier, the GFCC based method provided poorer accuracy under white Gaussian noise, whereas the FDLP based method exhibited a better accuracy under street noise conditions.

4. Discussion

Humans can easily distinguish two speakers irrespective of the text dependent or text independent speech signal both in quiet and under noisy environments. A physiologically based computational model of the auditory system was employed in this study to capture the necessary and useful features. The goal was to propose a simple feature from the neural responses that can mimic the pattern of speaker identification accuracies observed behaviorally. In this regard, the proposed neural response based system was very effective and robust against different types of noise, as illustrated in Figure 2.

The most important finding of this study was that the performance (pattern of accuracy as a function of SNR) of the proposed neural-response-based method was very consistent across different types of noise, whereas other baseline feature-based methods produced identification accuracies strongly dependent on the types of distortion in the acoustic signal. Thus, the proposed neural features successfully incorporated the distinguishing aspects of different speakers, which were degraded with a consistent (gradual) pattern across different types of distortion (white Gaussian, pink, and street noises).

4.1 Consistent performance

Neural responses were simulated to three speech samples (same text) from three different speakers to investigate the consistency of performance of the proposed feature. The energy at each CF (12) was calculated for each speech signal and is presented in Figure 3. It is clear that the energies for different speech samples (same text) from the same speaker were very consistent (close in magnitude) and well-separated from those of other speakers. Thus it can be inferred that the proposed method can capture underlying speaker distinguishing information very accurately in the neural responses.

4.2 Performance is classifier-invariant

Figure 2 shows the performance of the various methods using three different classifiers. It is obvious that the proposed method provided the most consistent performance across different types of classifiers. This means that the pattern and magnitude of identification accuracies were comparable among different classifiers. Alternatively, the performance of existing baseline feature based methods was highly dependent on the classifier. For example, the FDLP based method showed a substantially improved performance when the SVM was employed as a classifier.

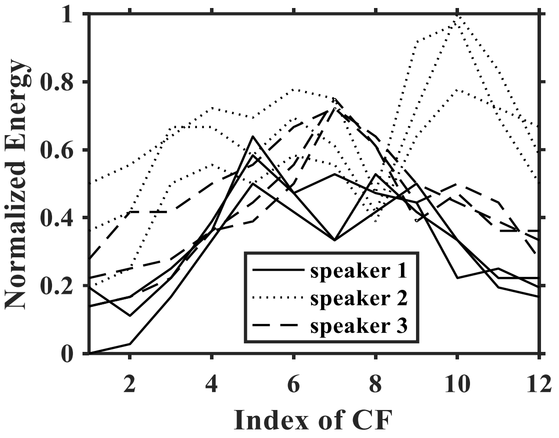

Figure 3 Illustration of speech energies as a function of CF. Neural responses were simulated and energy was measured for each CF to speech signals from three different speakers. Three speech samples (same text) were taken for each speaker.

4.3 Robustness of the proposed feature

Five speech samples (same text) were chosen from the same speaker and then distorted by adding white Gaussian noise at SNRs of 0 and 10 dB to examine the robustness of the proposed neural feature underlying the SID task. The neural responses were simulated for 26 CFs ranging from 250 Hz to 4 kHz (logarithmically), and the correlation coefficient between the neural responses to clean and noisy signals was measured for each CF. The correlation measure was averaged across five samples for each CF, and the means and standard deviations are shown in Figure 4.

It is clear that the lower CF responses (<850 Hz) showed higher correlation between the neural responses to clean and noisy signals, whereas the correlation at higher CFs was less reliable (higher standard deviation). In general, the neural responses were robust to different types of noise (produced a similar pattern and magnitude of correlation as a function of CF, results not shown), and even at a SNR of 0 dB, the correlation between the neural responses to clean and noisy signals was ~0.7 for lower CFs. It can be observed in Figure 5 that the responses at lower frequencies (below 850 Hz) were relatively less distorted by noises than the responses at higher frequencies, consistent with the range of phase-locking properties of the auditory neurons at lower CFs (<~4 kHz).

The auditory system has several nonlinearities such as nonlinear tuning, compression, two-tone rate suppression, phase-locking, synchrony capture, and adaptation in the inner hair cell AN synapse. The AN model employed in this study to simulate neural responses faithfully captured all of these nonlinearities in the peripheral auditory system [21]. Although it would be difficult to attribute the contribution of individual nonlinear phenomenon to the speaker

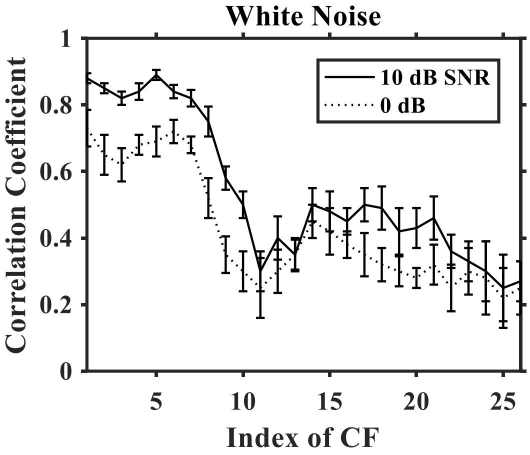

Figure 4 Robustness of the proposed neural feature. Neural responses were simulated using five independent speech samples (same text) from the same speaker. The correlation coefficient was measured between neural responses to each respective clean and noisy signal, and the mean and standard deviation of correlation measure across five samples are shown. White Gaussian noise at SNRs of 0 and 10 dB was considered.

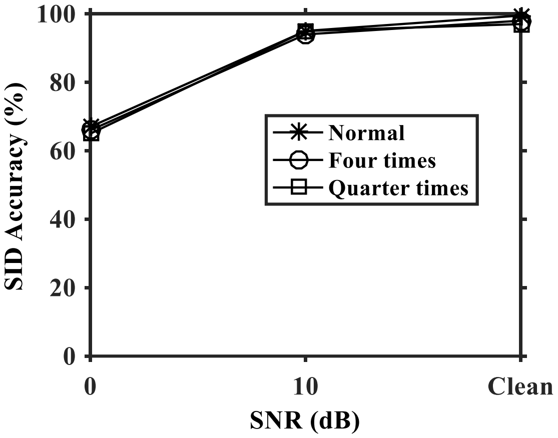

Figure 5 Illustration of the effect basilar-membrane (BM) filter bandwidth on the SID accuracy of the proposed neural feature based method.

identification tasks, the nonlinear phenomena incorporated in the AN model certainly played a significant role. Most of the existing methods derive features from an auditory system based linear filter bank and cannot account for the nonlinearities observed physiologically. Thus, the performance exhibited was substantially poorer compared to the scores from the neural response based method.

4.4 Effect of bandwidth of the auditory filters

Frequency selectivity in the cochlea plays an important role in resolving and segregating different sounds perceptually. Thus, the filter bandwidth of the cochlea might have an implication on the SID performance. To explore the effect of channel bandwidth on the speaker identification performance, neural responses were simulated for three different bandwidths (normal, 0.25, and 4 times broader) by varying Q10 values of the cochlear filter of the AN model.

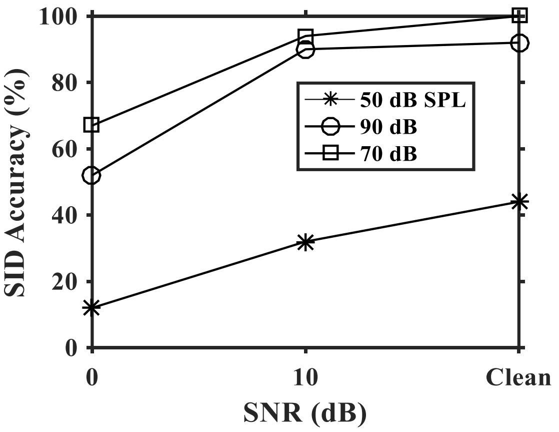

Figure 6 Illustration of the effect of sound presentation level on the performance of the proposed SID system.

Figure 5 shows the SID accuracy of the proposed system for three different bandwidths of the cochlear filter. It is obvious that the filter bandwidth had no substantive effect on text dependent speaker identification tasks. The effect of various nonlinearities on the SID task could be pursued in a future study.

4.5 Effect of sound presentation level

In general, speech intelligibility decreases with the increase of sound pressure level (SPL) above the conversational speech level (65~70 dB) [22-23]. It is thus expected that the change of SPL above or below the conversational speech level might affect the performance of the speaker identification system. To investigate the effect of SPL on the performance of the proposed system, neural responses were simulated at three different SPLs (50, 70, and 90 dB). Each instance was run for three SNR levels (0 and 10 dB, and clean conditions). The clean signal was distorted by introducing white Gaussian background noise. The effect of SPL on the identification accuracy of the proposed system is illustrated in Figure 6. It is clear that the highest accuracy was achieved at a conversational speech level of 70 dB SPL for all SNRs, and the performance degraded at higher or lower sound presentation levels. At lower SPLs, the neural responses at certain CFs might not be strong enough to represent the acoustic signal adequately, and thus the performance became poorer. It is to be noted that the acoustic cepstral-based features (MFCC, FDLP, and GFCC) are independent of the sound presentation level since these methods do account for the effects of nonlinearities observed in the auditory system.

5. Conclusions

In this paper, a novel neural-response-based feature is proposed for a text-dependent speaker identification system that was derived from the responses of a well-known physiologically based computational model of the auditory periphery. The performance of the proposed system was compared to the performance of three acoustic feature based SID systems (GFCC, FDLP, and MFCC) both in quiet and noisy conditions. In quiet, the performance was close to 100% for all methods. However, the proposed system outperformed most of the baseline feature based methods iunder noisy conditions. The neural response based system showed substantially better performance in low and negative

SNRs, especially in white Gaussian and street noises for the GMM and GMM-UBM classifiers. Additionally, the accuracy of the proposed neural response based method was classifier independent. The proposed neural feature also exhibited a similar pattern across different types of noise, reflecting the fact that the neural responses are more robust to noise. The effects of some supra-threshold nonlinearities on the identification tasks were also successfully captured by the proposed method.

6. Acknowledgements

This study was supported by the grant from the University of Hail. The author would like to thank Md. Atiqul Islam for helping with the code and simulations for the earlier version of the manuscript.

7. References

[1] Wenndt SJ, Mitchell RL. Machine recognition vs human recognition of voices. 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing; 2012 Mar 25-30; Kyoto, Japan. Kyoto: IEEE; 2012. p. 4245-8.

[2] 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.

[3] Makhoul J. Linear prediction: a tutorial review. Proceedings of the IEEE. 1975;63(4):561-80.

[4] Nakagawa S, Wang L, Ohtsuka S. Speaker identification and verification by combining MFCC and phase information. IEEE Trans Audio Speech Lang Process. 2012;20(4):1085-95.

[5] Wu Q, Zhang L. Auditory sparse representation for robust speaker recognition based on tensor structure. J Audio Speech Music Proc. 2008;2008:1-9.

[6] Chi TS, Lin TH, Hsu CC. Spectro-temporal modulation energy based mask for robust speaker identification. J Acoust Soc Am. 2012;131(5):EL368-EL74.

[7] Wu Q, Zhang L, Shi G. Robust feature extraction for speaker recognition based on constrained nonnegative tensor factorization. J Comput Sci Tech. 2010;25(4): 745-54.

8] Wang JC, Wang CY, Chin YH, Liu YT, Chen ET, Chang PC. Spectral-temporal receptive fields and MFCC balanced feature extraction for robust speaker recognition. Multimed Tool Appl. 2017;76:4055-68.

[9] Shao Y, Srinivasan S, Wang D. Incorporating auditory feature uncertainties in robust speaker identification. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2007 Apr 15-20; Honolulu, USA. USA: IEEE; 2007. p. IV277-IV80.

[10] Ganapathy S, Thomas S, Hermansky H. Feature extraction using 2-D autoregressive models for speaker recognition. Singapore: Odyssey 2012-The Speaker and Language Recognition Workshop; 2012.

[11] Zhao X, Shao Y, Wang D. CASA-based robust speaker identification. IEEE Trans Audio Speech Lang Process. 2012;20(5):1608-16.

[12] Zilany MS, Bruce IC, Carney LH. Updated parameters and expanded simulation options for a model of the auditory periphery. J Acoust Soc Am. 2014;135(1): 283-6.

[13] Miller MI, Barta PE, Sachs MB. Strategies for the representation of a tone in background noise in the temporal aspects of the discharge patterns of auditory-nerve fibers. J Acoust Soc Am. 1987;81:665-79.

[14] Hines A, Harte N. Speech intelligibility prediction using a neurogram similarity index measure. Speech Comm. 2012;54(2):306-20.

[15] Mamun N, Jassim WA, Zilany MS. Prediction of speech intelligibility using a neurogram orthogonal polynomial measure (NOPM). IEEE Trans Audio Speech Lang Process. 2015;23(4):760-73.

[16] Alam MS, Zilany MS, Jassim WA, Ahmad MY. Phoneme classification using the auditory neurogram. IEEE Access. 2017;5:633-42.

[17] Islam MA, Zilany MS, 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; 2015 Dec 6-8; Putrajaya, Malaysia. Singapore: Springer; 2016. p. 11-4.

[18] Islam MA, Jassim WA, Cheok NS, Zilany MS. A robust speaker identification system using the responses from a model of the auditory periphery. Plos One. 2016;11(7):1-21.

[19] Brookes M. Voicebox: speech processing toolbox for matlab [Internet]. UK: Software; 1997 [cited 2017]. Available from: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[20] Edwards BW, Wakefield GH. On the statistics of binned neural point processes: The Bernoulli approximation and AR representation of the PST histogram. Biol Cybern. 1990;64:145-53.

[21] 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.

[22] Studebaker GA, Sherbecoe RL, McDaniel DM, Gwaltney CA. Monosyllabic word recognition at higher-than-normal speech and noise levels. J Acoust Soc Am. 1999;105(4):2431-44.

[23] Dubno JR, Horwitz AR, Ahlstrom JB. Word recognition in noise at higher-than-normal levels: Decreases in scores and increases in masking. J Acoust Soc Am. 2005;118(2):914-22.

[24] Krishna BL, Semple MN. Auditory temporal processing: Responses to sinusoidally amplitude-modulated tones in the inferior colliculus. J Neurophysiol. 2000;1978;84:255-73.

[25] Liberman MC. Auditory-nerve response from cats raised in a low-noise chamber. J Acoust Soc Am. 1978;63(2):442-55.

[26] Roffo G. Report: feature selection techniques for classification [Internet]. USA: Computing Research Repository; 2016. Available from: http://arxiv.org/abs/1607.01327

[27] Ellis DP. PLP and RASTA (and MFCC, and inversion) in Matlab [Internet]. USA: 2005 [cited 2017]. Available from: http://www.ee.columbia.edu/ln/rosa/matlab/rastamat/

[28] Ganapathy S, Thomas S, Hermansky H. Front-end for far-field speech recognition based on frequency domain linear prediction. Switzerland: IDIAP Research Institute; 2008. Report No.: IDIAP-RR 08-17.

[29] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol (TIST). 2011;2(3):1-27.

[30] 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):1-15.

[31] Reynolds DA, Quatieri TF, Dunn RB. Speaker verification using adapted Gaussian mixture models. Digit Signal Process. 2000;10(1):19-41.