2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019

A Robust Text Dependent Speaker Identification Using Neural Responses from the Model of the Auditory System

Md. Ibrahim Khalil, Nursadul Mamun, Khadija Akter
Department of Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology, Chittagong, Bangladesh
ibrahimete13@gmai.com, nursad49@gmail.com, khadija007008@gmail.com

***Abstract*—Speaker recognition is considered as a behavioral biometric to identify speaker's identity based on their voice features. In this study, a new speaker identification system is proposed using the neural responses at the level of the auditory nerve (AN). For this, a very well-developed physiological based computational model of auditory periphery is used to simulate the neural responses for a given speech. The output, in the form of synapse responses, is then analyzed for the feature extraction. Neurograms are constructed for a range of characteristic frequencies from the output responses. Features are then calculated from the neurogram to train the system. The same extracted features for a given speaker are then used to identify the speaker in the testing phase. To test the reliability of the proposed system, the model has been tested both in quiet and noisy conditions. The results show that, neural response-based speaker identification system can substitute the existing technology and thus improve the performance for application of remote authentication and security system.**

***Keywords*—speech identification, speech recognition, neural network, auditory nerve model, TFS, Envelope, synapse responses.**

I. Introduction

Speaker identification (SID) is the process of identifying a person characterized by the vocal voice. Voice or speech is a biometric property of a human which can be used for uniquely characterizing a person. The basic features of speech vary from person to person uniquely with a proper distinguishable range.

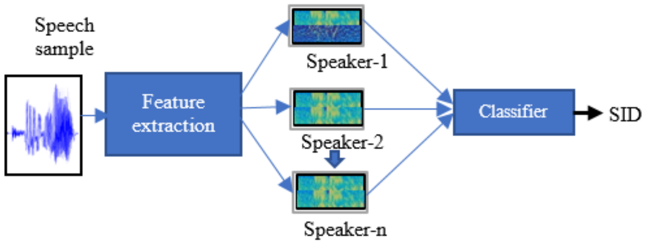


Fig. 1: Illustration of speaker identification system.

Speaker identification refers to identify the speaker from a set of pre-trained dataset available in a particular database as shown in the Fig. 1. On the other hand, speaker verification is done by testing one particular data against specific one's trained data among a dataset. Unlike to the traditional biometrics such as fingerprint, face, iris etc., speech or voice biometric is a combination of behavioral and physiological property of a person. The physiological properties depend on the different parts of our body parts such as mouth, nasal cavity, weight, throat, larynx, tongue and so on. On the other hand, behavioral properties depend on geographic area, pronunciation or manner of articulation, language, fluency, accent, dialect etc. Our eyes, skin and all other sensory organs help us to interact with environment. This interaction is done with the help of some process and for this, stimulations are converted into a sequence of signals. Then these signals of ear are represented into digital form and can be used for speaker identification. The conversion within human auditory system is one of the most sophisticated and complex system. However, different environmental noise degrades the acoustic signal and make then ineffective to use for speaker identification. The sound which impinges on outer ear goes through the physiological mechanism across middle and inner ear which is connected to auditory nerve fiber. Auditory nerve fiber response according to the perceived signal by our hearing system. This gives the nonlinearity of the human auditory sound processing, which is the main features used in this study using computational model of Auditory periphery by Zilany [1, 2].

Several studies have been done for speaker identification based on acoustic signal features. Traditional Speaker Identification (SID) system such as Mel-frequency cepstral (MFCC), Linear predictive Coding (LPC), Relative Spectral Transform Perceptual Linear Prediction (RASTA-PLP), Gammatone Frequency Cepstral Coefficients (GFCC) etc. [3] perform well in quite conditions. However, there performance significantly decreases under noisy environment [4]. Human hearing system along with the brain is able to identify a speaker even with the coefficients background noise or in noisy environment. This study proposes a speaker identification system using a computational neural response based auditory model which is robust to noise [5]. As neural response shows the phase-locking property to a periodic input up for a frequency range, this means that neural response-based model is very robust to noise. This is the reason behind choosing auditory system-based speaker identification method rather than acoustic signal-based method.

II. Methods

The purpose of a speaker identification system is to make sure that the identity of a speaker does belong to the speaker model.

The proposed system is divided into two parts, training phase and testing phase as shown in Fig. 2. The unvoiced speech

978-1-5386-9111-3/19/$31.00 ©2019 IEEE

signals are removed and passed through the AN model. Then the obtained neurograms are processed for features. The features are used for speaker identification for different classifiers. In training phase, speaker model is created and save to a database. In testing phase, the speech sample of tested speaker will be tested against the model database.

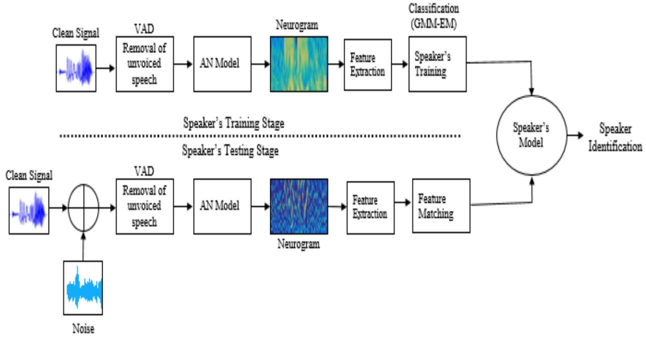


Fig. 2. Proposed speaker identification method.

A. *Pre-processing:*

In this stage, Voice Activity Detector (VAD) [6] algorithm was applied to detect and remove the silence period of speech. The unvoiced or very low energy signals are also detected in this algorithm. Then the Dynamic Time Wrapping (DTW) [7] is used for the alignment of all the speech samples of a speaker at the same timing and amplitude to get the wrapped version.

B. *Computational model of the Auditory Nerve (AN):*

The computational AN model developed by Zilany and colleagues considered as a useful tool for understanding the physiological and behavioral process of Human auditory peripheral. The phenomenological description of activity of every part can be represented by AN model from middle ear to auditory nerve. The model consists of 3 blocks and 4 filters, total 7 parts; middle ear, feed forward control path, C1 filter, C2 filter, IHC, OHC Synapse Model and Discharge generator. An instantaneous pressure waveform (sound) is the input of the AN model where spike times are the output The instantaneous discharge rate of auditory nerve fibers as a function of time is the output of synapse response which is used to construct neurogram [8, 9].

C. *Neurogram:*

The pictorial representation of the output of AN model in time and frequency domain is called as neurogram. The output of AN model is typically visualized through electrical recording of auditory nerve. In this study, neural responses are simulated for a range of 20 characteristic frequency (CFs) to analyze the acoustic waveform. The neurogram is basically two types which represent speech contents. They are called Envelope (ENV) and Temporal Fine Structure (TFS). The ENV usually carries the voicing manner of speaker, vowels identity and how speech is articulated. The TFS carries the formant information of speech and contains fine timing structure of auditory nerve spike. Considering speech stimulus as an input of AN model, the output of AN model in forms of ENV and TFS for a range of CFs.

D. *Feature Extraction:*

As speech is considered to be a one-dimensional signal, the mean value technique is applied to further process the neurogram. The response of AN model is a two-dimensional array which is changed to one dimensional array by taking the mean across time. The feature extractor returns a column vector containing the mean of the elements in each row. The mean value is taken with respect to 20 CFs.

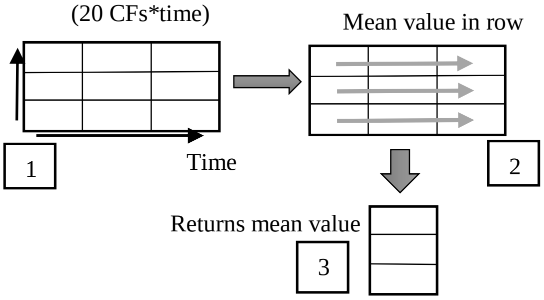


Fig. 3: ENV or TFS neurogram blocks transformation column mean value according to the CFs.

Figure 3 illustrates on how features of the original ENV or TFS neurogram is extracted. In Fig. 3. (1) shows the process of applying mean in row wise. Each two-dimensional CFs is averaged at first. Then it has been converted into one dimensional array in Fig. 3. (2). The block have values of rows with 20 CFs(Column) which contains necessary information to distinguish one speaker from the other. Finally, this feature is used in artificial neural network.

All speech samples are given as an input to the AN model and subsequently model responses are to produce ENV and TFS neurograms using synapse output. The whole ENV and TFS in the range of 20 CFs is used as a feature in GMM. Where 70% of the data is used for training and remaining 30% is used as the testing data.

E. *Classification Techiques:*

<u>Gaussian Mixture Model (GMM):</u>

GMM classification techniques firstly generate the model of 39 speakers from the output of AN model (ENV & TFS) using Expectation Maximization (EM) algorithm. All the speech samples of each speakers are concatenated to the speaker model in training phase. Assuming that a GMM model for the speaker j represented by lambda (λ_j), is defined as the sum of all K components of the Gaussian densities for the feature vectors (x_t) of that particular speaker. Defining the probability of x_t based on the GMM model or it's weighting probability function as[10]:

$$p(x_t\,|\lambda_j\,) = \sum_{i=1}^{K} g_i\,\mathcal{N}(x_t;\mu_{i,\sum i}) \tag{1}$$

Here $\sum_I$ = covariance matrix & μ_i = mean for feature vectors $\mathcal{N}$ = individual component densities parameterized by the feature vector, mean vector and covariance matrix for a D-variate Gaussian function. 20 variables (CFs) with different (4, 8, 16, 32, 64, 128, 256) GMM components is used for training purpose.

In testing phase, speech samples along with the GMM model is given as input to a PDF and a vector is generated as output. The maximum value of vectors gives the identification of speaker [11].

Neural Network:

At first, the neurograms are applied for feature extraction and generated dataset for training and testing separately. The training dataset of each speaker are concatenated. Using back propagation algorithm, the neural network learns from the input data. By using this process, the neural network can map between inputs and desired outputs by adjusting weighted value of connections of the network. The goal of back propagation training is to converge to a near-optimal solution based on the total squared error calculated in equation 2 [12].

$$E_C = \frac{1}{2}\sum_{c=1}^{C}\left(D_c - O_c\right)^2 \qquad (2)$$

Where C represents the number of units in the output layer, D_c is the desired network output (from the output vector) corresponding to the current output layer unit, and O_c is the actual network output corresponding to the current output layer unit.

I-vectors

The neural features were used as an input of I-vector. At training phase, I-vector for a given utterance can be extracted as follow[13],

$$w = (I + T^T\Sigma^{-1}NT)^{-1}.T^T\Sigma^{-1}F \qquad (3)$$

Here 'I' is an identity matrix of CF×CF dimension, N is a diagonal matrix with F ×F blocks, c (c = 1, 2, ...C) where C is the mixture number, and the super-vector, F of dimension (CF×1), is obtained through the concatenation of the centralized first-order Baum Welch Statistics (BWS) statistics, (.) t denotes transpose. The covariance matrix, Σ, represents the residual variability not captured by T. An efficient procedure of estimating the total-variability subspace, T, is described in. The training algorithm of the total-variability space (T) is similar to JFA eigen training only for voice except for single difference. In JFA Eigen voice training, all the sessions of given speaker are considered to be the same person [14].

F. Speech database:

This study used 'University of Malaya (UM)' database to identify the speaker. It is a text dependent database consisting of 390 signals collected from 39 speakers (10 samples from each speaker). Among them, 25 are male and 14 are female with aged within 22 to 24 years old. The audio signal was recorded in a noiseless room and processed with a sampling rate of 8 kHz. Each speaker uttered 'University Malaya' 10 times in a quiet room in different session. In this study, 70% of recorded random data from each speaker is used for training purpose and remaining 30% are used in order to test the performance of the proposed system. Clean (unchanged) signals were used for training the speaker model. Both clean and noisy signals were used for testing purpose. Noise is a greater consideration in case of practical world having most effect in speech. Clean signal is used for training and different noise is added to clean signal for testing. Several noise such as white Gaussian noise, speech shape noise and pink noise is examined in this study.

III. Results

The performance of the proposed system is described in this section. The performance of the proposed system using ENV and TFS features along with neural network as a classifier is shown in Fig. 4.

To examine the robustness of the system, the system performance was evaluated under different types of noise. Noise are combined with clean speech signal after sampling. The results show that, the proposed system is more robust for pink noise than any other types of noise because of Power Spectral Density (PSD) which is inversely proportional to the frequency of the signal. On the other hand, the system is sensitive to Gaussian noise. Gaussian noise is additive because it is added to any noise that might be intrinsic to the information system. And for this reason, it gives lowest performance among these noisy conditions. The speech shape noise is present on the speech signals based on existing *speech* Corpus (Multiple speech files). Table I and Table II represent the performance of the proposed system for different types of noise using three different classifiers using ENV and TFS response, respectively. The comparison of different classifier described in Table I gives different results under noisy condition. But the results for all of the three classifiers are same in clean condition. There don't have any mismatching in this condition. The I-vector classifiers show the less accuracy then the other two methods whereas the GMM shows the best result. GMM can be bootstrapped with flat data where neural network needs to be trained with more accurate data. Neural network can't be guaranteed to converge to an optimal point where GMM can be granted. In noisy condition NN gives better result comparatively to I-vector. This is because NN is less affected by noise than I-vector.

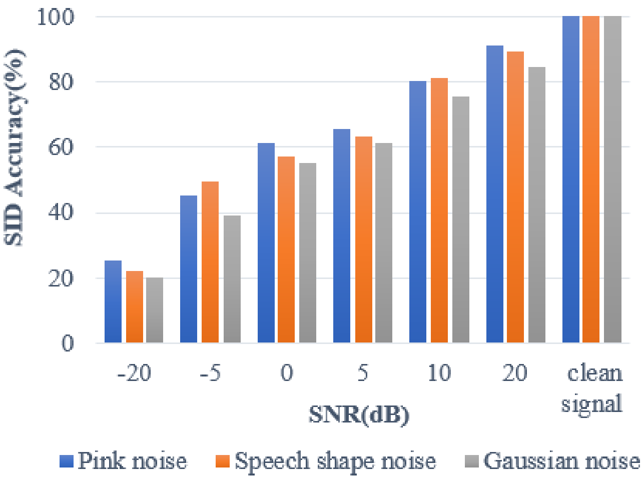
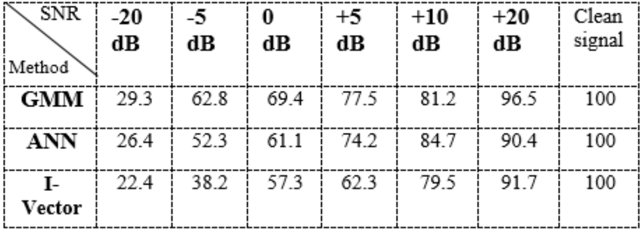


Fig. 4: Speaker Identification (SID) accuracy with different noise types in different SNR using neural network as a classifier.

Table I: The Speaker Identification performance under noisy conditions using ENV response

SNR / Method	-20 dB	-5 dB	0 dB	+5 dB	+10 dB	+20 dB	Clean signal
GMM	27.8	58.2	67.3	72.4	89.7	97.3	100
ANN	22.2	49.9	57.1	63.3	81.1	89.2	100
I-Vector	19.1	38.9	54	56.6	79.1	91.1	100

Table II: The Speaker Identification performance under noisy conditions using TFS response



SNR Method	-20 dB	-5 dB	0 dB	+5 dB	+10 dB	+20 dB	Clean signal
GMM	29.3	62.8	69.4	77.5	81.2	96.5	100
ANN	26.4	52.3	61.1	74.2	84.7	90.4	100
I-Vector	22.4	38.2	57.3	62.3	79.5	91.7	100

TFS information depends on phase locking to individual cycles of the stimulus waveform. In general, the performance of the proposed system declines as more and more noise is added to the speech signal, consistent with the results from the behavioural studies. Under quiet condition, speaker identification performance is 100%, whereas it drops to~21% when a background noise of SNR -20 dB is used. Although the performance using TFS and ENV is comparable at very high and low SNRs, TFS information gives better speaker identification performance in the intermediate SNR levels (-5 to 10 dB). This finding suggests that phase-locking information to the individual stimulus frequency is important for speaker identification, which is supported by the well-known fact that the difference in fundamental frequency plays a big role in speaker identification. TFS is important for pitch perception and sound localization. And in general pitch is used to distinguish different words.

IV. CONCLUSION

This study approaches a neural response-based speaker identification method using a physiologically-based model of the auditory system which simulated the auditory nerve responses for a wide range of characteristic frequencies. ENV and TFS are the two extracted features that are taken as an output from this physiological based auditory nerve model. The proposed speaker identification system employed three types of classifiers such as: GMM, ANN and I-vector. Two extracted features of ENV and TFS are given as an input to the classifiers. The obtained results from this work revealed that the performance of the proposed neural response-based system was far better than the performance of the traditional acoustic feature-based speaker identification system especially under noisy conditions. To quantify the robustness of this proposed speaker identification system different noise effect on this system were calculated. Among the three classifiers, GMM classifier gives the better result than other two even in noisy condition. The performance of the proposed system also showed that TFS response-based identification system perform better than the ENV response-based system, meaning that the TFS contains more information related to the identity of speakers than in the ENV information.

REFERENCES

[1] M. S. Zilany, I. C. Bruce, P. C. Nelson, and L. H. Carney, "A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics," *The Journal of the Acoustical Society of America*, vol. 126, pp. 2390-2412, 2009.

[2] S. B. Davis and P. Mermelstein, "Readings in speech recognition. chapter Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," ed: Volume, 1990.

[3] N. A. Bansal, "Speaker Recognition using MFCC front end analysis and VQ Modeling Technique for Hindi words using MATLAB," *International Journal of Computer Applications,* vol. 45, pp. 48-52, 2012.

[4] E. B. Tazi, A. Benabbou, and M. Harti, "Efficient text independent speaker identification based on GFCC and CMN methods," in *Multimedia Computing and Systems (ICMCS), 2012 International Conference on*, 2012, pp. 90-95.

[5] M. A. Islam, W. A. Jassim, N. S. Cheok, and M. S. A. Zilany, "A robust speaker identification system using the responses from a model of the auditory periphery," *PloS one,* vol. 11, p. e0158520, 2016.

[6] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE signal processing letters,* vol. 6, pp. 1-3, 1999.

[7] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *KDD workshop*, 1994, pp. 359-370.

[8] N. Mamun, W. A. Jassim, and M. S. Zilany, "Robust gender classification using neural responses from the model of the auditory system," in *Functional Electrical Stimulation Society Annual Conference (IFESS), 2014 IEEE 19th International*, 2014, pp. 1-4.

[9] N. Mamun, W. A. Jassim, and M. S. Zilany, "Prediction of speech intelligibility using a neurogram orthogonal polynomial measure (NOPM)," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 23, pp. 760-773, 2015.

[10] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE transactions on speech and audio processing,* vol. 3, pp. 72-83, 1995.

[11] D. A. Reynolds, "A Gaussian mixture modeling approach to text-independent speaker identification," 1993.

[12] M. Minsky and S. A. Papert, *Perceptrons: An introduction to computational geometry*: MIT press, 2017.

[13] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, pp. 788-798, 2011.

[14] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 16, pp. 980-988, 2008.