IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 27, 2016, accepted November 18, 2016, date of publication January 16, 2017, date of current version March 2, 2017.

*Digital Object Identifier 10.1109/ACCESS.2016.2647229*

# Phoneme Classification Using the Auditory Neurogram

**MD. SHARIFUL ALAM[1], (Student Member, IEEE), MUHAMMAD S. A. ZILANY[1,2], WISSAM A. JASSIM[1,3], AND MOHD YAZED AHMAD[1]**

[1]Department of Biomedical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur 50603, Malaysia
[2]Department of Computer Engineering, College of Computer Science and Engineering, University of Hail, Hail 2440, Saudi Arabia
[3]ADAPT Center, School of Engineering, Trinity College Dublin, the University of Dublin, Dublin 2, Ireland

Corresponding authors: M. S. ALAM (msalam04@yahoo.com) and M. YAZED AHMAD (myaz@um.edu.my)

This work was supported in part by the RP016B-13AET (MSAZ), in part by the Fundamental Research Grant Scheme under Grant FRGS FP036-2014B, in part by the UM Living Lab Grant Programme under Grant LL014-16SUS, and in part by the grant from the University of Hail (MSAZ).

**ABSTRACT** In order to mimic the capability of human listeners identifying speech in noisy environments, this paper proposes a phoneme classification technique using simulated neural responses from a physiologically based computational model of the auditory periphery instead of using features directly from the acoustic signal. The 2-D neurograms were constructed from the simulated responses of the auditory-nerve fibers to speech phonemes. The features of the neurograms were extracted using the Radon transform and used to train the classification system using a deep neural network classifier. Classification performance was evaluated in quiet and under noisy conditions for different types of phonemes extracted from the TIMIT database. Based on simulation results, the proposed method outperformed most of the traditional acoustic-property-based phoneme classification methods for both in quiet and under noisy conditions. The proposed method could easily be extended to develop an automatic speech recognition system.

**INDEX TERMS** Auditory-nerve model, DNN, FDLP, MFCC, neurogram, phoneme classification.

## I. INTRODUCTION

State-of-the-art algorithms for automatic speech recognition (ASR) systems suffer from poorer performance when compared to the ability of human listeners to detect, analyze, and segregate the dynamic acoustic stimuli, especially in complex and under noisy environments [1]–[3]. Performance of ASR systems can be improved by using additional levels of language and context modelling, provided that the input sequence of elementary speech units is sufficiently accurate [4]. To achieve a robust recognition of continuous speech, both sophisticated language-context modeling and accurate predictions of isolated phonemes are required. Indeed, most of the inherent robustness of human speech recognition occurs before and independently of context and language processing [2], [5]. For phoneme recognition, human auditory system's accuracy is already above chance level, at a signal-to-noise ratio (SNR) of −18 dB [2]. Also, several studies have demonstrated the superior performance of human speech recognition compared to machine performance both in quiet and under noisy conditions [6], [7], and thus the ultimate challenge for an ASR is to achieve recognition performance that is close to the performance of human auditory system.

Recently, speech recognition communities have shown a tremendous interests in deep neural networks (DNNs) which were popular during late 80's and early 90's [8], [9]. A neural network becomes a high-quality acoustic model due to some factors: a) they can be made powerful by making the network deeper, b) using a much faster hardware and initializing the weights sensibly, deep neural networks can be trained effectively, and c) the performance can be improved by using a larger number of output units [9].

The perceptual linear prediction (PLP) [10], relative spectra (RASTA) [11], and Mel-frequency cepstral coefficients (MFCCs) [12] are some examples of the preferred traditional features for ASR system. These features are derived by computing the short-term magnitude spectra of speech, and then the nonlinear transformation is applied to model the processing of human auditory system. However, a moderate level of signal distortion due to additive noise or linear filtering may cause a significant departure of feature distribution from the features of the signal in clean condition [13]. As a result, the performance of ASR systems based on these features is far below compared to human performance in adverse conditions [1], [3]. During past years, efforts have

been made to design a robust ASR system motivated by auditory processing. For example, Holmberg *et al.* incorporated a synaptic adaptation into their feature extraction methods and found that the performance of the system improved substantially [14]. Similarly, Strope and Alwan [15] used a model of temporal masking and Perdigao and Sá [16] employed a physiologically-based inner ear model for developing a robust ASR system. However, these models did not include most of the nonlinearities observed at the level of the auditory periphery and thus were not physiologically-accurate.

This study proposes a novel phoneme classification technique in which the features were extracted from the simulated neural responses of a physiologically-accurate model of the auditory system. The responses of this model were successfully used in several speech intelligibility metrics [17], [18], speaker identification system [19] and phoneme classification [20]. Substantial improvements in performances were achieved over conventional systems using the neural-response-based feature instead of employing acoustic-signal-based features. The proposed approach is also expected to improve the robustness of the phoneme classification system by mimicking the response properties observed in the peripheral auditory system in the feature extraction technique. It has been reported in the literature that the auditory-nerve (AN) fiber responds preferentially at a certain phase of the input stimulus, referred to as the phase-locking property, even when the input signal becomes noisy, i.e., neural responses are robust against noise [20]. In the proposed method, the neural responses were simulated using a well-known physiologically-based model of the auditory periphery [21]. This auditory- nerve (AN) model successfully incorporates most of the nonlinear properties observed at the peripheral level of the auditory system such as nonlinear tuning, compression, two-tone suppression, and adaptation in the inner-hair-cell-AN synapse as well as some other nonlinearities observed only at very high sound pressure levels (SPLs) [22], [23]. The model simulates the discharge timings (spike train sequence) of an AN fiber for a given characteristic frequency (CF), and thus a 2-D representation (neurogram) was constructed by simulating the responses of AN fibers over a wide range of CFs. The features for classification (training and testing) were provided by computing the parallel-beam projections (discrete Radon transform, DRT) of the 2-D neurogram image for a wide range of angles. The extracted features were used to train a discriminative classifier, deep neural network (DNN) [8], [9], to classify phonemes for both in quiet and under noisy environments. The performance of the proposed system was compared to the performance of phoneme classification systems based on the widely used features such as the MFCC and the frequency domain linear prediction (FDLP) coefficients [24].

This paper is organized as follows. Section II describes the computational procedure of the proposed phoneme classification technique. The performance of the proposed and traditional methods is provided in section III, and finally the conclusion of this study is presented in section IV.

## II. METHODOLOGY

The block diagram of the proposed neural-response-based phoneme classification method is shown in Fig. 1. The block diagram consists of two stages: training and testing. In the training stage, the clean phoneme signal was applied to the AN model to generate the corresponding neural responses. Neural responses for a range of CFs were simulated to construct the neurogram (2D time-frequency representation). The proposed features were then extracted from the neurogram using the DRT. The features of each phoneme were trained using the framework of the DNN. In the testing stage, either the clean or the noisy/distorted phoneme signal (unknown) was applied to the AN model to generate the neurogram responses, and the features were extracted using the DRT. The extracted features of the testing signal were used as an input to the DNN model to predict the class of the phoneme (label).

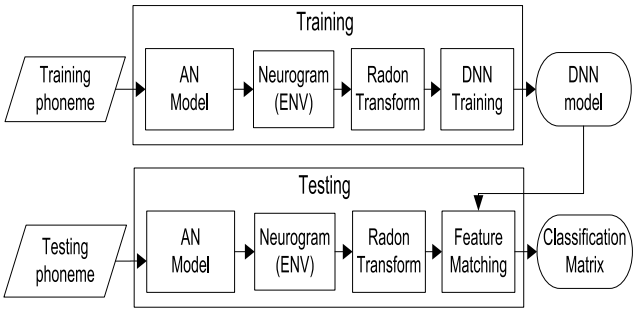


**FIGURE 1.** Block diagram of the proposed phoneme classifier.

### A. AN MODEL AND NEUROGRAM

A physiologically-accurate model of the auditory periphery was employed in this study to simulate the responses at different stages of the peripheral auditory system. The input to the model is an instantaneous pressure waveform of speech signals taken from the TIMIT database [25], and the final output is the spike times from the discharge generator. The schematic diagram and detail description of the employed AN model can be found in [21] and [23]. Each block in the diagram represents a phenomenological description of major components in the auditory periphery from the middle ear to the auditory nerve.

In the AN model, the acoustic signal is passed through a middle-ear filter (first stage), followed by parallel signal-path narrowband (i.e., basilar membrane, BM) and control-path broadband filters (second stage). The control path reflects the active processes in the cochlea. The gain and bandwidth of the nonlinear BM filter are varied according to the output of the control-path filter to account for several level-dependent response properties of the cochlea such as compression, suppression, and nonlinear phase responses in the cochlea. The output of the BM filter is passed through the third stage of the model which simulates inner-hair-cell (IHC) mechanisms with a static nonlinearity followed by a fifth-order low-pass filter. Then the IHC output drives the IHC-AN

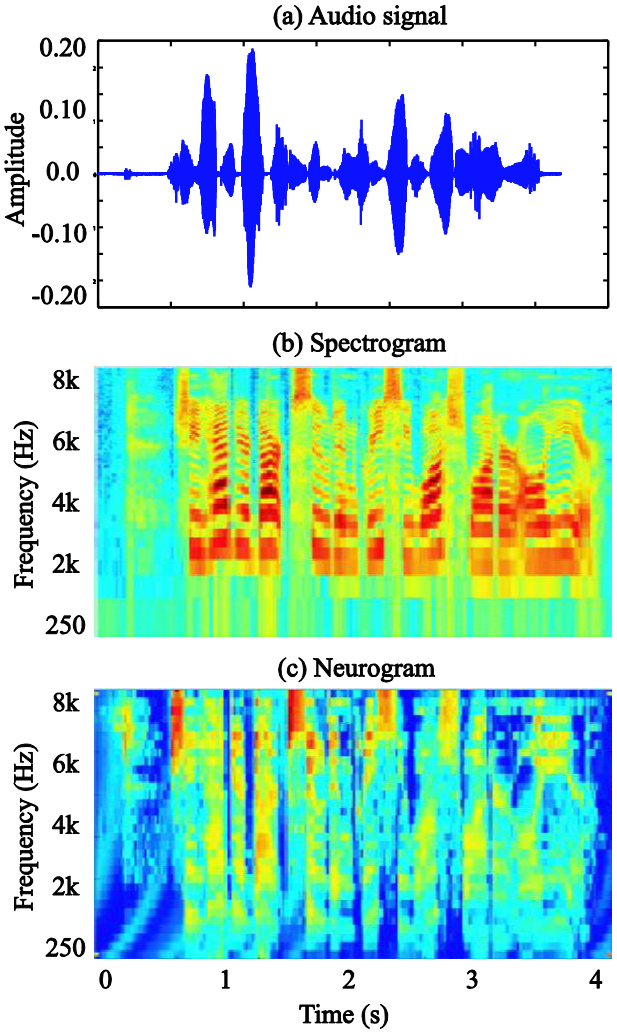


**FIGURE 2.** Illustration of the difference between the spectrogram vs. neurogram: (a) speech signal taken from the TIMIT database ('SA1 sentence'), (b) the corresponding spectrogram and (c) neurogram response.

synapse which provides the instantaneous synaptic release rate as output. Finally, the discharge times are produced by a renewal process that includes both absolute and relative refractory effects. The model responses have been extensively validated against a wide range of physiological recordings from the peripheral auditory system to both simple and complex stimuli [21], [23]. However, to construct the 2-D neurogram, the model IHC-AN synapse output which provides the probability of instantaneous discharge rate of AN fibers as a function of time was used in the present study.

In this study, the IHC-AN synapse outputs for 32 AN fibers with CFs ranging from 150 Hz to 8 kHz were used to construct neurogram. Conceptually, a neurogram is analogous to a spectrogram that gives a pictorial representation of neural responses in the time-frequency domain. Fig. 2 shows the two types of representation, spectrogram and neurogram plots, in response to a typical signal taken from the TIMIT database. The neurogram was initially constructed by averaging the synapse output of each CF (AN fiber) with a bin-width of 100 $\mu$s, and then the resulting response was smoothed by a Hamming window of 128 samples. Thus, the resulting neurogram reflects a relatively slow variation in the amplitude of the input speech signal and is referred to as the envelope (ENV) neurogram [26].

### B. DISCRETE RADON TRANSFORM (DRT)

The multiple parallel-beam projections of the image, $f\ (x, y)$, from different angles are referred to as the discrete Radon transform. The projections are computed by rotating the source around the center of the image [27]. In general, the Radon transform $R_\theta(x^{'})$ of an image is the line integral of $f$ parallel to the $y^{'}$-axis,

$$R_\theta\left(x^{'}\right) = \int_{-\infty}^{\infty} f(x^{'}\cos(\theta) - y^{'}\sin(\theta), x^{'}\sin(\theta) + y^{'}\cos(\theta))dy^{'} \quad (1)$$

where

$$\begin{bmatrix} x^{'} \\ y^{'} \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \quad (2)$$

Fig. 3 illustrates the geometry of the Radon transform. The DRT has been extensively applied in image processing applications such as the computed axial tomography (CAT scan), barcode scanners, and electron microscopy of macromolecular assemblies like viruses and protein complexes. Also, a texture analysis method [28] and a face recognition framework [29] were previously developed based on Radon projections on the input image. Motivated by these applications, in this study, the proposed features were extracted by applying the DRT on the 2-D neurogram image. Quantitatively, the Radon coefficient in this case represents the relative spectro-temporal information across CFs for different angles of projection which would capture the relevant perceptual features of phonemes. The Radon projection coefficients for different angles were then combined to form the feature vector for each neurogram.

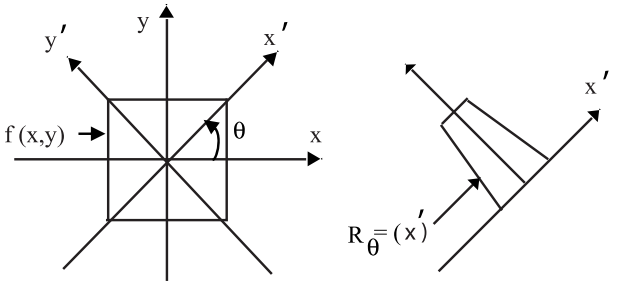


**FIGURE 3.** Geometry of the discrete Radon transform.

### C. DEEP NEURAL NETWORK (DNN)

DNNs are receiving increasing attention for acoustic modeling in speech recognition, especially for large-scale tasks [9], [30]–[32]. In general, the classification problem can be solved using multiple hidden layers. In this study,

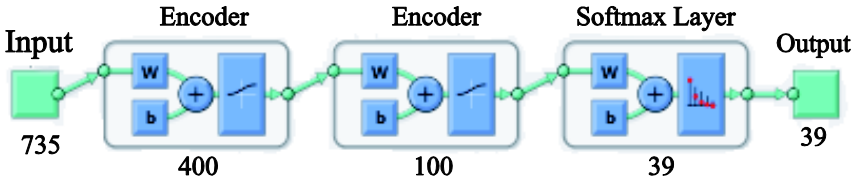


**FIGURE 4.** A diagram of the stacked network that is formed by the encoders from the auto encoders and the softmax layer.

two layers have been used and trained up (each layer individually) using special type of artificial neural networks known as the autoencoder and the sofmax layer. Autoencoder is unsupervised machine learning technique to encode a data set for the purpose of dimensionality reduction. The structure of autoencoder network is similar to that of the traditional feed forward multilayer perceptron (MLP) network but with the output nodes having the same size as the input nodes. In general, the autoencoder technique consists of two stages: encoder and decoder. When the number of nodes in the hidden layer is less than the size of the input layer, a compressed version of the input can be achieved. In other words, the input is mapped to a compressed representation (encoder stage).

Similarly, the autoencoder network can be decoded to map the compressed representation back to the original input (decoder stage). In this study, two autoencoders were employed. The first autoencoder with a hidden layer of 400 neurons was trained to encode the 735-dimensional input features (the Radon projections of the neural responses as explained in subsection E) to a 400-dimensional output. This output was fed to the second autoencoder with a hidden layer of 100 neurons to map the 400-dimensional data to 100-dimensional feature set.

Unlike the autoencoders, the softmax layer is a supervised machine learning technique that uses a multinomial logistic regression to train (classify) the input feature vectors according to the corresponding labels. The input to the softmax layer is the 100-dimensional features set (output of the second autoencoder). The output of the softmax layer consists of 39 neurons which represents the number of phoneme classes.

As a summary, the 735-dimensional proposed feature vectors were compressed to a set of 100-dimensional representation using two networks of autoencoder, and the softmax layer was then employed to classify the compressed form of features into the required classes. The trained network was saved to be used in the testing stage for predicting the classes of the test data-set. Fig. 4 shows the schematic diagram of the neural network.

As a future work, DNNs can be integrated with hidden Markov models (HMMs) for developing a continuous speech recognition system [30] using the same proposed feature.

### D. DATASET AND PHONE CLASSES

Experiments were performed on the complete test set of the TIMIT database [25]. There are two ''sa'' (dialect) sentences in the TIMIT database spoken by all speakers that may result in artificially high classification scores [33]. To avoid any unfair bias, experiments were performed on the ''si'' (diverse) and ''sx'' (compact) sentences. The training data set consists of 3696 utterances from 462 speakers, whereas testing set consists of 1344 utt erances from 168 speakers (not included the training set). The glottal stop /q/ was removed from the class labels, and the 61 TIMIT phoneme labels were collapsed into 39 labels following the standard practice given in [33]. Further, we evaluate the classification performance using a broad phone-class approach proposed by Reynolds and Antoniou [34]. White noise, street noise, station noise and airport noise [35] with different signal-to-noise ratios (SNRs) were added to the clean phoneme signals to evaluate the performance of the proposed and two traditional methods.

### E. PROCEDURE

Each phoneme signal was up-sampled to 100 kHz which was required by the AN model in order to ensure stability of the digital filters implemented for faithful replication of frequency responses of different stages (e.g., middle ear) in the peripheral auditory system. The sound pressure level (SPL) of all phonemes was set to 70 dB which represents the preferred listening level for a monaural listening situation. Because the AN model used in this study is nonlinear, the neural representation would be different at different sound levels.

However, for the purpose of classification, all phonemes were isolated (from TIMIT) and scaled to 70 dB SPL.

In the DNN training phase, the Radon projection coefficients were calculated from the phoneme neurogram using twenty one (21) rotation angles ranging from 0° to 180° in steps of 9°. The vector of each Radon projection was resized to 35 points and then combined together for all angles to form a (1 × 735) feature vector. Thus the total number of features for each phoneme was 735 irrespective of the duration of the phoneme in the time domain. A mapping function was used subsequently to normalize the mean and standard deviation of the feature vector to 0 and 1, respectively. All normalized data from each phoneme were combined together to form an input array for DNN training. The corresponding label vector of phoneme classes was also constructed.

In testing phase, the Radon projection coefficients using the same twenty one rotation angles were calculated from the test (unknown) phoneme neurogram. The label (class) of the test phoneme was identified using the approximated function obtained from the DNN training stage. Fig. 5 shows

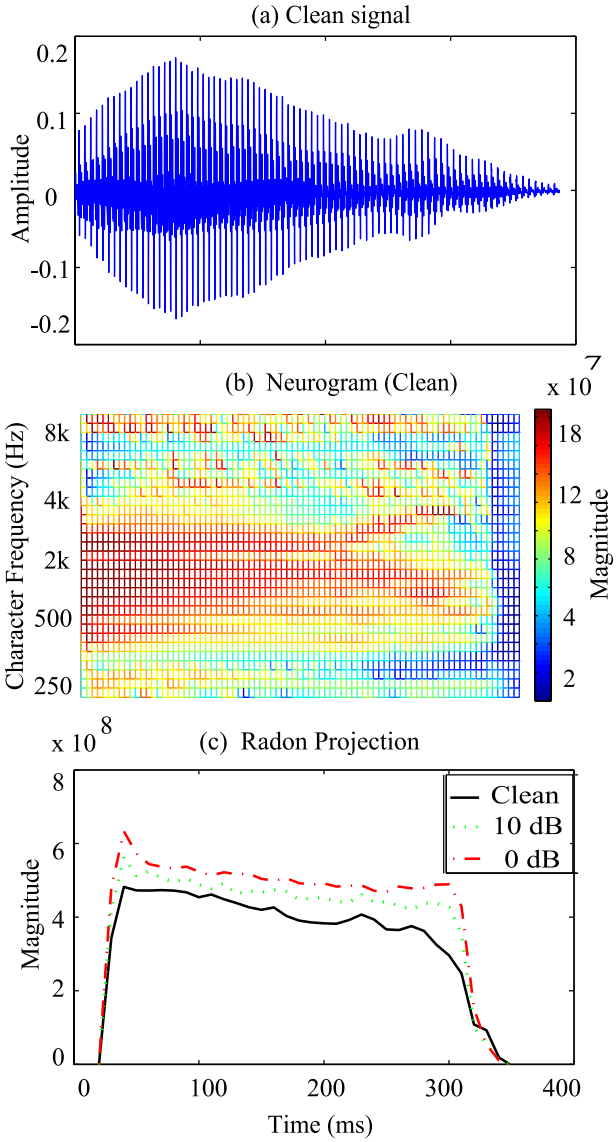


**FIGURE 5.** Neurogram-based feature extraction for the proposed method: (a) a typical phoneme waveform (/aa/), (b) neurogram representation of the phoneme signal, (c) the Radon projection coefficients of the neurogam for an angle of 0° both in quiet and under noisy conditions (0 and 10 dB SNRs).

example features extracted by applying the Radon transform on the neurogram. Fig. 5(a) shows the waveform of a typical phoneme (/aa/) taken from the TIMIT database and the corresponding neurogram representation is shown in Fig. 5(b). Fig. 5(c) shows the Radon projection coefficients of the neurogram for an angle of 0° for the phoneme signal in quiet (solid line) and at SNRs of 0 dB (dashed line) and 10 dB (dotted line).

In order to compare the accuracy of the proposed method to the performance of two traditional methods, the RASTAMAT [36] and FDLP [37], [38] toolboxes were used for extracting MFCC and FDLP features, respectively, from all phonemes. Each phoneme signal was divided into frames using a Hamming window of length 25 ms with an overlap of 10 ms between frames. For each frame, 39 features (dimension of the classifier) consisting of 3 groups such as Ceps (Mel-frequency cepstral coefficients), Del (derivatives of Ceps) and Ddel (derivatives of Del) with 13 features per group were computed.

As a result, for each phoneme the size of the MFCC coefficients array is $l \times 39$, where $l$ is the total number of frames (observations). It is important to mention that the number of features (39) is fixed for all phonemes, whereas $l$ varies depending on signal duration. Similarly, the same size of overlapped frames and window type were used in computing the corresponding 39 FDLP features for each frame. The structure of the FDLP feature array is similar to that of the MFCC method. In the training stage, the DNN network shown in Fig. 4 was employed to train the 39-dimensional MFCC/FDLP features. For both MFCC and FDLP the first autoencoder with a hidden layer of 200 neurons was trained to encode the 39-dimensional input features to a 200-dimensional output. This output was fed to the second autoencoder with a hidden layer of 100 neurons to map the 200-dimensional data to 100-dimensional feature set. The input to the softmax layer is the 100-dimensional features set (output of the second autoencoder). The output of the softmax layer consists of 39 neurons which represents the number of phoneme classes. In the testing stage, the feature vectors of an unknown phoneme with a size of $l \times 39$ were used as an input to the DNN model generated form the training stage. The model predicts $l$ values (the range is from 0 to 39) and the class which has a maximum repetition over the $l$ values determines the final identity (label) of the test phoneme.

## III. RESULTS AND DISCUSSIONS

This section provides simulation results of the proposed, MFCC- and FDLP-based phoneme classification methods. These methods were tested both in quiet and under noisy conditions. For phoneme classification under noisy conditions, the features extracted from original (clean) phoneme samples of the TIMIT train subset were used to train DNN models. In the testing stage, different noise with a particular signal to noise ratio was added to the test phoneme signal from the TIMIT test subset, and proposed features were then extracted.

In this study, different values of SNRs ranging from 0 to 25 dB in steps of 5 dB were considered to evaluate the performance of the classification methods. To generate the confusion matrices of segment classification for the three methods, the full phone-against-phone confusions matrices were first calculated, and all the entries within each broad-class block were then added together to provide one value [34]. The segment classification confusion matrices areshown in Table I for quiet (clean) condition. It is obvious that closures (CLO) were identified more accurately for MFCC and neurogram-based features. For FDLP-based features plosives were less confused with others. For FDLP and neurogram features some of the plosives (PLO) and fricatives (FRI) were confused with other groups, but most of the confusions were observed within these two groups. Similarly, semivowels (SVW) and vowels (VOW) were con-

**TABLE 1. Confusion matrices for segment classification in quiet (clean) condition using three features: MFCC, FDLP, and Auditory Neurogram. Averaged accuracy for MFCC, FDLP and proposed features are 60.55%, 40.94% and 67.71%, respectively.**

| | PLO | FRI | NAS | SVW | VOW | DIP | CLO |
|---|---|---|---|---|---|---|---|
| | | | | MFCC | | | |
| PLO | 38.28 | 8.56 | 7.26 | 14.91 | 16.38 | 0.5 | 14.1 |
| FRI | 10.92 | 39.97 | 8.23 | 7.38 | 8.3 | 1.33 | 23.87 |
| NAS | 10.58 | 2.88 | 57.69 | 7.69 | 9.62 | 0 | 11.54 |
| SVW | 25 | 0 | 0 | 25 | 0 | 25 | 25 |
| VOW | 0.38 | 1.79 | 14.6 | 20.56 | 56.99 | 4.65 | 1.04 |
| DIP | 0 | 0.28 | 5.06 | 22.75 | 47.75 | 24.16 | 0 |
| CLO | 3.75 | 1.42 | 3.08 | 0.29 | 0.47 | 0 | 90.99 |
| | | | | FDLP | | | |
| PLO | 77.59 | 11.2 | 1.68 | 0.28 | 1.96 | 0 | 7.28 |
| FRI | 15.32 | 73.14 | 0.79 | 3.16 | 1.9 | 0 | 5.69 |
| NAS | 3.94 | 3.67 | 58.79 | 9.97 | 14.7 | 0.26 | 8.66 |
| SVW | 4.26 | 2.29 | 7.86 | 67.59 | 13.58 | 1.31 | 3.11 |
| VOW | 3.96 | 1.42 | 6.35 | 11.8 | 68.26 | 3.96 | 4.26 |
| DIP | 1.38 | 0 | 3.21 | 4.13 | 17.89 | 71.56 | 1.83 |
| CLO | 11.93 | 10.17 | 4.59 | 3.1 | 3.67 | 0.15 | 66.39 |
| | | | | Neurogram-Radon | | | |
| PLO | 85.39 | 9.40 | 0.68 | 0.71 | 0.69 | 0.00 | 3.14 |
| FRI | 6.68 | 87.43 | 1.10 | 0.90 | 0.68 | 0.06 | 3.15 |
| NAS | 0.48 | 3.74 | 79.13 | 4.09 | 5.95 | 0.05 | 6.57 |
| SVW | 0.59 | 1.16 | 2.03 | 79.48 | 13.34 | 2.04 | 1.37 |
| VOW | 0.27 | 0.71 | 2.76 | 7.82 | 82.50 | 5.11 | 0.82 |
| DIP | 0.00 | 0.00 | 0.26 | 4.06 | 24.57 | 71.10 | 0.00 |
| CLO | 2.52 | 4.74 | 3.27 | 0.70 | 0.93 | 0.02 | 87.82 |

**TABLE 2. Individual phone class accuracies (%) for different feature extraction techniques using speech with additive noises. The best performance for each condition is indicated in bold.**

| SNR (dB) | 0 | 5 | 10 | 15 | 20 | 25 |
|---|---|---|---|---|---|---|
| | | | White Noise | | | |
| MFCC | 18.1 | **29.94** | **41.3** | **49.74** | **55.49** | **58.46** |
| FDLP | 17.33 | 21.01 | 23.95 | 28.07 | 32.04 | 34.97 |
| Neurogram | **23.53** | 29.56 | 35.77 | 42.61 | 50.88 | 57.99 |
| | | | Street Noise | | | |
| MFCC | 23.66 | 32.55 | 41.04 | 49.13 | 54.84 | 58.29 |
| FDLP | 18.97 | 22.57 | 26.32 | 29.47 | 32 | 38.65 |
| Neurogram | **32.87** | **43.87** | **53.21** | **60.31** | **64.76** | **66.84** |
| | | | Train Station Noise | | | |
| MFCC | 24.79 | 34.24 | 43.55 | 50.84 | 55.22 | 57.12 |
| FDLP | 21.47 | 30.47 | 37.47 | 25.56 | 39.1 | 34.69 |
| Neurogram | **33.61** | **42.3** | **50.89** | **57.6** | **62.24** | **65.51** |
| | | | Airport Noise | | | |
| MFCC | 23.01 | 34.95 | 45.87 | 53.8 | 58.08 | 59.7 |
| FDLP | 21.98 | 26.32 | 29.56 | 31.13 | 37.86 | 39.52 |
| Neurogram | **35.38** | **46.79** | **55.88** | **62.17** | **65.83** | **67.26** |

fused more among these groups compared to other groups for all three methods.

However, the proposed method outperformed the two other traditional methods in terms of accuracy.

Table II shows classification accuracies of individual classes using different features for different noise types at six levels of SNR. For street, train station and airport noises, the performances of the proposed method for all SNRs were better than the results for all other features. MFCC features showed better performance for white noise at all noisy conditions except at 0 dB SNR.

The percentage accuracies of broad phone classes as a function of SNR are shown in Fig. 6. It is clear that the proposed method resulted better accuracy compared to the performance of other two methods at all noisy environments except for white noise. For street, train station and airport noises, the classification accuracy of the proposed method dropped from 82.84% in quiet to ∼58% at 0 dB SNR, whereas for the same condition, the performance of the MFCC- and FDLP-based methods declined from 73.15% to ∼38% and from 51% to ∼30%, respectively. For white noise, MFCC features showed a relatively better performance compared to using other features.

acoustic measurements for phoneme classification, and their broad class accuracy using the complete set (118 speakers) Plenty of research has been done on phoneme classification [13], [32], [39]–[41]. The core test set (7,215 tokens) of the TIMIT database was used in most of the literature for the evaluation of performances of different phoneme classification methods [13], [34], [40], [42]. However, in order to include more variations of phonemes, the proposed method was tested and results are reported here using the complete set of the TIMIT database (50,754 tokens). In quiet, a classification accuracy of 84.1% for the core test set was reported by Reynolds and Antoniou [34], whereas the proposed method achieved an accuracy of 83.48% for the complete test set. Halberstadt *et al*. used heterogeneous was 79.0%, whereas the accuracy using our method was 83.48% (168 speakers) [42]. In 2005, Johnson *et al.* showed the result for the complete set using the HMM as a classifier, and they reported the single phone accuracy of 54.86% and 35.06% using the MFCC and reconstructed phase space (RPS)-based method, respectively [41], whereas the proposed neural-response-based method showed an accuracy of 67.71% for single phones. We also showed the result of 60.55% using the MFCC-based feature with the DNN as a classifier.

Similarly, Saeb *et al.*[40] used the sparse representation for a robust phoneme classification for the core set. At clean condition, their performance was 75.12%, but it dropped to 10% at 20 dB SNR for white noise, whereas the accuracy of the proposed method for the complete set dropped from 67.71 % in quiet to 50.88% for the same SNR condition (20 dB).

To investigate the reason behind robustness of the neural-response-based proposed method, a typical phoneme (/ao/) of length 90 ms was used as an input to the AN model to gen-

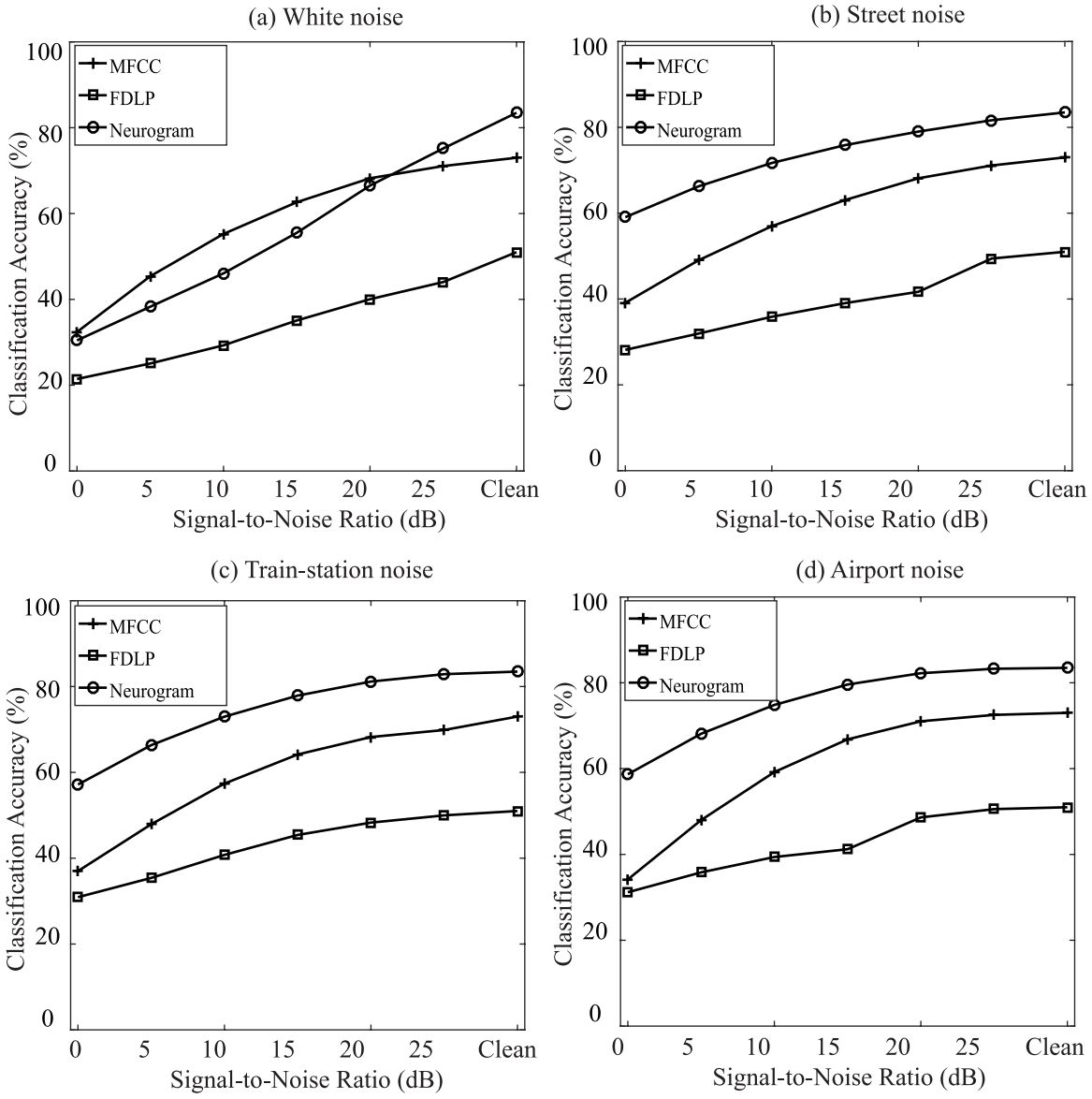


**FIGURE 6.** Broad class accuracies (%) for different feature extraction techniques. Performances were evaluated under four different types of noise (a: white noise, b: street noise, c: train station noise, and d: airport noise). The proposed feature (auditory neurogram) exhibited better performance in comparison to the result for both MFCC and FDLP features in all noisy cases except for white noise.

erate the corresponding neurogram responses. The phoneme waveform in the time domain is shown in Fig. 7(a).

The corresponding MFCC and FDLP coefficients are shown for each frame in Fig. 7(b) and (c), respectively. The size of the generated neurogram image was 32 by 20, where the neural responses were simulated for 32 CFs. All columns of neurogram array were combined together to form a 1-D vector (the new size was $1 \times 640$) which is shown in Fig. 7(d). The responses (features) of the phoneme in quiet are shown by the solid lines, and the responses using the same phoneme distorted by a white noise of 0 dB SNR are shown by the dotted lines in each corresponding plots (b, c, d). The correlation coefficient between these two vectors was com-puted using the following equation and was found to be $\sim$0.54 for the neurogram feature, 0.43 for MFCC, and 0.27 for FDLP coefficients.

$$r = \frac{\sum_m \sum_n \left(A_{mn} - \bar{A}\right)\left(B_{mn} - \bar{B}\right)}{\sqrt{\left(\sum_m \sum_n \left(A_{mn} - \bar{A}\right)^2\right)\left(\sum_m \sum_n \left(B_{mn} - \bar{B}\right)^2\right)}} \quad (3)$$

where $\bar{A}$ and $\bar{B}$ are the mean values of combined vector from the clean (A) and distorted/noisy (B) neurograms or features, respectively.

Based on the similarity index, it is obvious that the proposed neural features were more robust compared to the traditional acoustic-property-based features. This observation is further supported by the findings in study [43] that neural responses are robust to noise due to the phase-locking

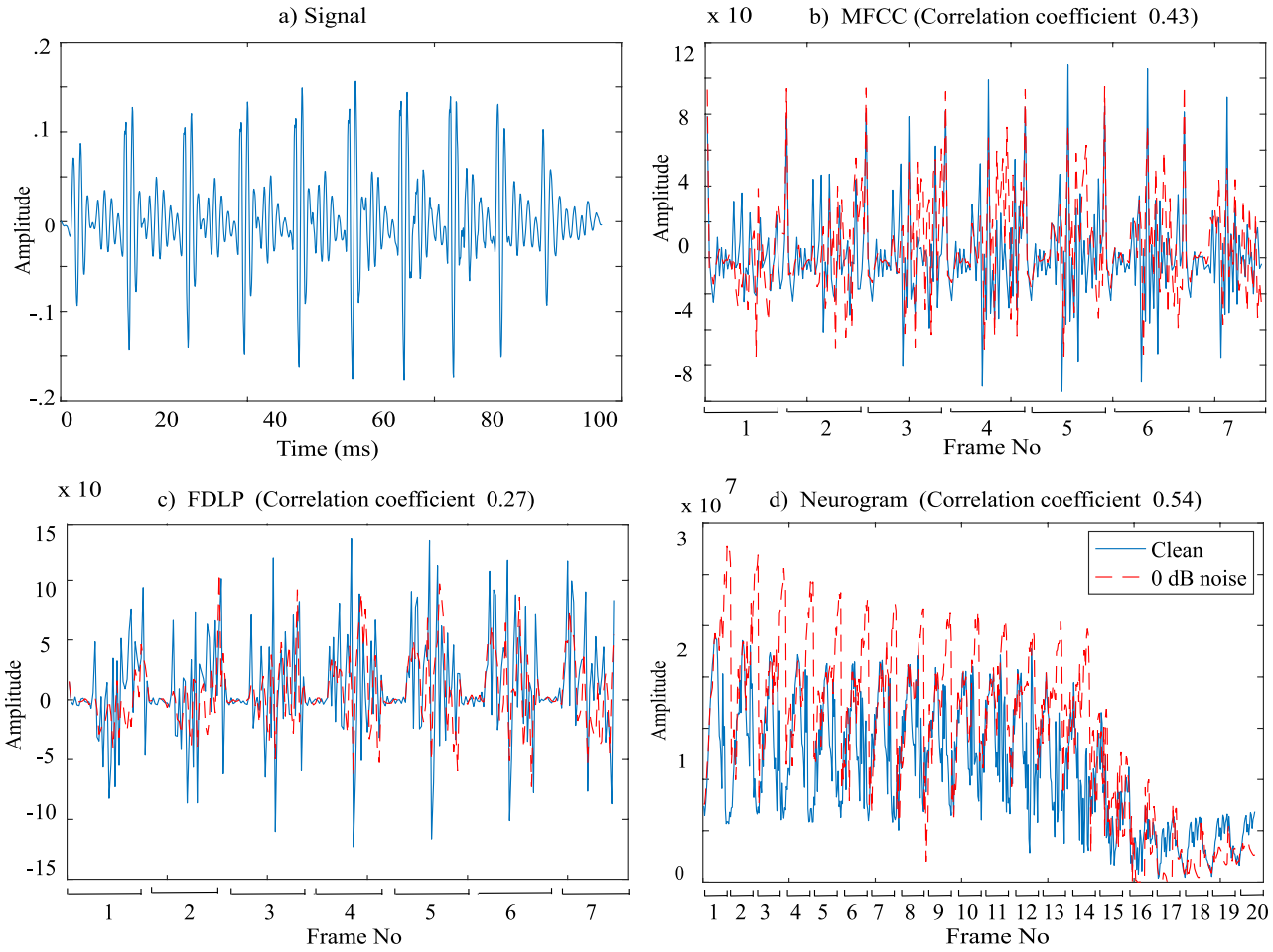


**FIGURE 7.** (a) Signal waveform of the phoneme /ao/, (b) MFCC features extracted from the phoneme /ao/ in quiet (solid line) and at an SNR of 0 dB (dotted line) conditions. The correlation coefficient between the two vectors was 0.43, (c) FDLP features in clean and noisy conditions. The correlation coefficient between the two cases was 0.27, (d) Neurogram responses of the phoneme /ao/ in quiet and noisy conditions. The Correlation coefficient between the two vectors was 0.54. For MFCC and FDLP features, the phoneme was divided into frames and 39 coefficients were computed from each frame.

property of the neuron (especially at lower CFs). Moreover, it was found that in response to a vowel-like stimulus at a conversational speech level, the AN response is phase-locked almost exclusively to the formant frequency closest to the fiber's CF, and this phenomenon is referred to as the synchrony capture [44]. The synchrony capture by the formants makes temporal representations of spectral shape very robust. The auditory-periphery model employed in this study successfully captures all of these properties [45], and thus the proposed neural feature could contribute to the better classification performance similar to the performance of human listeners with normal hearing.

In order to explore the effect of number of Radon angles on classification accuracy, Fig. 8 presents the single classification accuracy for the proposed method as a function of the number of Radon angles. For an angle, the accuracy was almost 30% for the signals in quiet. With the increase of number of angles, the classification accuracy increased substantially up to a number of 21. The number of Radon angles used in this study was chosen as twenty one based on the phoneme classification accuracy for both in quiet and under noisy conditions.

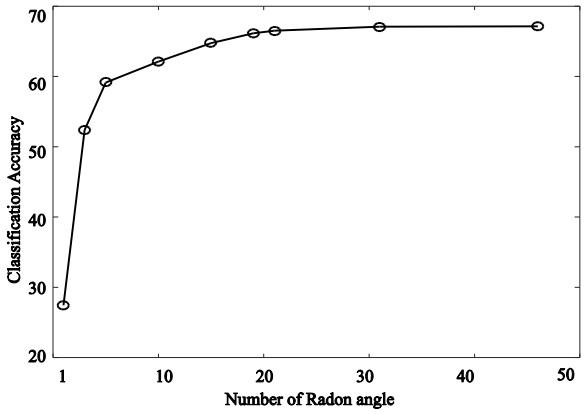


**FIGURE 8.** Phoneme classification accuracy in quiet as a function of the number of Radon angles (between 0° and 180°) used to encode phoneme information.

## IV. CONCLUSIONS

A phoneme classification technique was proposed in this study based on the application of Radon transform on simulated neural responses from a model of the auditory periphery. The performance was evaluated on the complete test set of the TIMIT database and compared to the results using two standard acoustic-property-based methods. In general, the proposed method outperformed MFCC- and FDLP-based classification methods for both in quiet and under most of the noisy conditions. The robustness in the performance of the proposed neural-response-based method could be attributed to the nonlinear properties of the neurons in the peripheral auditory system, and the Radon transform applied on the auditory neurogram successfully captures the relevant perceptual features of the phonemes.

## REFERENCES

[1] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Commun.*, vol. 22, pp. 1–15, Jul. 1997.

[2] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Amer.*, vol. 27, no. 2, pp. 338–352, 1955.

[3] J. J. Sroka and L. D. Braida, "Human and machine consonant recognition," *Speech Commun.*, vol. 45, no. 4, pp. 401–423, 2005.

[4] J. Yousafzai, M. Ager, Z. Cvetkovic, and P. Sollich, "Discriminative and generative machine learning approaches towards robust phoneme classification," in *Proc. Inf. Theory Appl. Workshop*, 2008, pp. 471–475.

[5] G. A. Miller, G. A. Heise, and W. Lichten, "The intelligibility of speech as a function of the context of the test materials," *J. Experim. Psychol.*, vol. 41, no. 5, pp. 329–335, 1951.

[6] J. B. Allen, "How do humans process and recognize speech?" *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 567–577, Oct. 1994.

[7] B. T. Meyer, M. Wächter, T. Brand, and B. Kollmeier, "Phoneme confusions in human and automatic speech recognition," in *Proc. INTERSPEECH*, 2007, pp. 1485–1488.

[8] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[9] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8599–8603.

[10] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[11] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[12] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *J. Comput. Sci. Technol.*, vol. 16, no. 6, pp. 582–589, Nov. 2001.

[13] J. Yousafzai, P. Sollich, Z. Cvetković, and B. Yu, "Combined features and kernel design for noise robust phoneme classification using support vector machines," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 5, pp. 1396–1407, Jul. 2011.

[14] M. Holmberg, D. Gelbart, and W. Hemmert, "Automatic speech recognition with an adaptation model motivated by auditory processing," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 1, pp. 43–49, Jan. 2006.

[15] B. Strope and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 451–464, Sep. 1997.

[16] F. Perdigao and L. Sá, "Auditory models as front-ends for speech recognition," in *Proc. NATO ASI Comput. Hearing*, 1998, pp. 179–184.

[17] A. Hines and N. Harte, "Speech intelligibility prediction using a neurogram similarity index measure," *Speech Commun.*, vol. 54, no. 2, pp. 306–320, 2012.

[18] N. Mamun, W. A. Jassim, and M. S. A. Zilany, "Prediction of speech intelligibility using a neurogram orthogonal polynomial measure (NOPM)," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 4, pp. 760–773, Apr. 2015.

[19] 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, no. 7, p. e0158520, 2016.

[20] M. S. Alam, W. A. Jassim, and M. S. A. Zilany, "Neural response based phoneme classification under noisy condition," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Dec. 2014, pp. 175–179.

[21] M. S. A. 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," *J. Acoust. Soc. Amer.*, vol. 126, no. 5, pp. 2390–2412, 2009.

[22] L. Robles and M. A. Ruggero, "Mechanics of the mammalian cochlea," *Physiol. Rev.*, vol. 81, no. 3, pp. 1305–1352, Jul. 2001.

[23] M. S. A. Zilany and I. C. Bruce, "Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery," *J. Acoust. Soc. Amer.*, vol. 120, no. 3, pp. 1446–1466, 2006.

[24] S. Ganapathy, S. Thomas, and H. Hermansky, "Temporal envelope compensation for robust phoneme recognition using modulation spectrum," *J. Acoust. Soc. Amer.*, vol. 128, pp. 3769–3780, 2010.

[25] J. Garofolo *et al.*, "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, Philadelphia, PA, USA, Tech. Rep. 101, 1993.

[26] B. C. J. Moore, "The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people," *J. Assoc. Res. Otolaryngol.*, vol. 9, no. 4, pp. 399–406, 2008.

[27] R. N. Bracewell, *Two-Dimensional Imaging* (Prentice-Hall Signal Processing Series). Englewood Cliffs, NJ, USA: Prentice-Hall, 1995.

[28] K. Jafari-Khouzani and H. Soltanian-Zadeh, "Rotation-invariant multiresolution texture analysis using Radon and wavelet transforms," *IEEE Trans. Image Process.*, vol. 14, no. 6, pp. 783–795, Jun. 2005.

[29] D. V. Jadhav and R. S. Holambe, "Feature extraction using Radon and wavelet transforms with application to face recognition," *Neurocomputing*, vol. 72, pp. 1951–1959, Mar. 2009.

[30] L. Deng *et al.*, "Recent advances in deep learning for speech research at Microsoft," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8604–8608.

[31] A.-R. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *Proc. NIPS Workshop Deep Learn. Speech Recognit. Rel. Appl.*, 2009, p. 39.

[32] D. Varghese and D. Mathew, "Phoneme classification using Reservoirs with MFCC and Rasta-PLP features," in *Proc. Int. Conf. Comput. Commun. Inform. (ICCCI)*, 2016, pp. 1–6.

[33] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

[34] T. J. Reynolds and C. A. Antoniou, "Experiments in speech recognition using a modular MLP architecture for acoustic modelling," *Inf. Sci.*, vol. 156, pp. 39–54, Nov. 2003.

[35] K. Hopkins and B. C. Moore, "The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise," *J. Acoust. Soc. Amer.*, vol. 125, no. 1, pp. 442–446, 2009.

[36] D. P. Ellis, "PLP and RASTA (and MFCC, and inversion) in MATLAB," Columbia Univ., New York, NY, USA, Tech. Rep. 225, 2005. [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

[37] S. Ganapathy, S. Thomas, and H. Hermansky, "Feature extraction using 2-D autoregressive models for speaker recognition," Johns Hopkins Univ., USA, Tech. Rep. 164, 2012.

[38] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *IEEE Signal Process. Lett.*, vol. 15, pp. 681–684, 2008.

[39] L. D. Vignolo, H. L. Rufiner, D. H. Milone, and J. C. Goddard, "Evolutionary splines for cepstral filterbank optimization in phoneme classification," *EURASIP J. Adv. Signal Process.*, vol. 2011, p. 284791, Jan. 2011.

[40] A. Saeb, F. Razzazi, and M. Babaie-Zadeh, "SR-NBS: A fast sparse representation based N-best class selector for robust phoneme classification," *Eng. Appl. Artif. Intell.*, vol. 28, pp. 155–164, Feb. 2014.

[41] M. T. Johnson, R. J. Povinelli, A. C. Lindgren, J. Ye, X. Liu, and K. M. Indrebo, "Time-domain isolated phoneme classification using reconstructed phase spaces," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 458–466, Jul. 2005.
[42] A. K. Halberstadt and J. R. Glass, "Heterogeneous acoustic measurements for phonetic classification 1," in *Proc. EUROSPEECH*, 1997, pp. 1–4.
[43] M. I. Miller, P. E. Barta, and M. B. Sachs, "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. Amer.*, vol. 81, no. 3, pp. 665–679, 1987.
[44] E. D. Young and M. B. Sachs, "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *J. Acoust. Soc. Amer.*, vol. 66, no. 5, pp. 1381–1403, 1979.
[45] M. S. A. Zilany and I. C. Bruce, "Representation of the vowel/ɛ/in normal and impaired auditory nerve fibers: Model predictions of responses in cats," *J. Acoust. Soc. Amer.*, vol. 122, no. 1, pp. 402–417, 2007.

**MD. SHARIFUL ALAM** was born in Chandpur, Bangladesh. He received the B.Sc. degree (Hons.) in computer science and engineering from the Chittagong University of Engineering and Technology, Bangladesh, in 2006. He is currently pursuing the M.Sc. degree with the Department of Biomedical Engineering, University of Malaya (UM), Kuala Lumpur, Malaysia. From 2014 to 2016, he was a Research Assistant with the Auditory Neuroscience Laboratory, UM. His current research interests include signal, speech, and image Processing.

**MUHAMMAD ZILANY** received the B.Sc. and M.Sc. degrees in electrical and electronic engineering from the Bangladesh University of Engineering and Technology in 1999 and 2002, respectively, and the Ph.D. degree in electrical and computer engineering from McMaster University, ON, Canada, in 2007. From 2008 to 2011, he was a Post-Doctoral Research Associate with the Department of Biomedical Engineering and Nerobiology & Anatomy, University of Rochester, NY, USA. From 2012 to 2016, he was a Senior Lecturer with the Department of Biomedical Engineering, University of Malaya, Kuala Lumpur, Malaysia. Since 2016, he has been an Assistant Professor with the Department of Computer Engineering, College of Computer Science and Engineering, University of Hail, Hail, KSA. He is currently a Charted Engineer with IET, U.K., and a member of the Association for Research in Otolaryngology and the Society for Neuroscience.

**WISSAM A. JASSIM** was born in Baghdad, Iraq, in 1976. He received the B.Sc. and M.Sc. degrees in electrical engineering from Baghdad University, in 1999 and 2002, respectively, and the Ph.D. degree in electrical engineering from the University of Malaya (UM), Malaysia, in 2012. From 2013 to 2015, he was a Visiting Research Fellow with the Department of Biomedical Engineering, UM. From 2015 to 2016, he was a Post-Doctoral Research Associate with the Department of Electrical Engineering, UM. He is currently a Research Fellow with the ADAPT Center, School of Engineering, Trinity College Dublin, the University of Dublin, Dublin 2, Ireland. His current research interests include machine learning, speech, and image processing.

**MOHD YAZED AHMAD** (M'12) received the B.E. degree in electrical engineering from the Department of Electrical Engineering, University of Malaya, Kuala Lumpur, Malaysia, in 2003, the M.Eng.Sc. degree from the Department of Biomedical Engineering, Faculty of Engineering, University of Malaya, in 2006, and the Ph.D. degree from the University of Technology, Sydney, Australia, in 2013. He is currently a Senior Lecturer with the Department of Biomedical Engineering, Faculty of Engineering, University of Malaya. His current research interests include real-time signal processing, instrumentation, and embedded systems.