Noise-robust text-dependent speaker identification using cochlear models

Md. Atiqul Islam, Ying Xu, Travis Monk, et al.

Citation: The Journal of the Acoustical Society of America **151**, 500 (2022); doi: 10.1121/10.0009314

View online: https://doi.org/10.1121/10.0009314

View Table of Contents: https://asa.scitation.org/toc/jas/151/1

Published by the Acoustical Society of America

JASA ARTICLE

Noise-robust text-dependent speaker identification using cochlear models

Md. Atiqul Islam,[a)] Ying Xu, Travis Monk, Saeed Afshar, and André van Schaik
International Centre for Neuromorphic Systems in the MARCS Institute for Brain, Behaviour, and Development,
Western Sydney University, Penrith, New South Wales, 2751, Australia

ABSTRACT:
One challenging issue in speaker identification (SID) is to achieve noise-robust performance. Humans can accurately identify speakers, even in noisy environments. We can leverage our knowledge of the function and anatomy of the human auditory pathway to design SID systems that achieve better noise-robust performance than conventional approaches. We propose a text-dependent SID system based on a real-time cochlear model called cascade of asymmetric resonators with fast-acting compression (CARFAC). We investigate the SID performance of CARFAC on signals corrupted by noise of various types and levels. We compare its performance with conventional auditory feature generators including mel-frequency cepstrum coefficients, frequency domain linear predictions, as well as another biologically inspired model called the auditory nerve model. We show that CARFAC outperforms other approaches when signals are corrupted by noise. Our results are consistent across datasets, types and levels of noise, different speaking speeds, and back-end classifiers. We show that the noise-robust SID performance of CARFAC is largely due to its nonlinear processing of auditory input signals. Presumably, the human auditory system achieves noise-robust performance via inherent nonlinearities as well. https://doi.org/10.1121/10.0009314

(Received 30 March 2021; revised 16 November 2021; accepted 27 December 2021; published online 28 January 2022)

[Editor: Michael I Mandel] Pages: 500–516

I. INTRODUCTION

Biometric authentication has a wide range of applications including human-machine interfaces, online banking, shopping, forensic testing, and crime investigation. Nowadays, iPhone's Siri, Google Assistant, Samsung's Bixby, and other smartphone assistants use audio biometric authentication. Recently, biometric authentication has been implemented on several neuromorphic systems such as TrueNorth (Modha, 2014), Loihi-Intel (Davies *et al.*, 2018), and BrainChip's Akida (Turchin, 2019). These hardware implementations should expand applications of biometric authentication in mobile devices, cars, computers, and beyond.

Speaker identification (SID) is a biometric authentication system that uses speaker utterances to identify a target speaker. Each speaker utterance is unique because of vocal fold size, larynx length, vocal tract physiology, and articulation habits (Ghazanfar and Rendall, 2008). A rigorous SID system should have the ability to extract characterizing features from speech. These features should enable a better speaker model at the back-end and result in higher SID accuracy regardless of background noise, variances in speech, or our choice of back-end classifier.

In general, most SID systems use fast Fourier transform (FFT)-based front-ends to generate spectral features, such as mel-frequency cepstral coefficient (MFCC) (Davis and Mermelstein, 1990), gammatone frequency cepstral coefficient (GFCC) (Shao *et al.*, 2007), and power normalized cepstral coefficient (PNCC) (Nayana *et al.*, 2017). These methods achieve almost 100% SID accuracy when speech is noiseless. However, the accuracies of these systems fall rapidly in the presence of background noise. This accuracy reduction occurs because the FFT distributes frequency bands linearly and their spectral distortion under noisy conditions affects the system accuracy (Li and Huang, 2011). Other front-ends use voice production mechanisms, such as perceptual linear prediction (PLP) (Makhoul, 1975) and a 2D autoregressive model-based frequency domain linear prediction (FDLP) (Ganapathy *et al.*, 2012). These systems use all-pole models to provide close to 100% SID accuracy given clean speech. However, in the presence of noise, these front-ends poorly classify speakers as they also use FFT-based power spectra to extract speech features.

In contrast, the human auditory system is not only capable of performing a variety of speech processing tasks but is also highly robust to background noise (Hansen and Hasan, 2015). At the input stage of the human auditory system, the cochlea decomposes, converts, and amplifies sound waves nonlinearly into electrical signals that are input to the nervous system. Helmholtz, who is considered as the first to hypothesize that the ear performs spectral analysis (Von Helmholtz, 1863), proposed that the function of the basilar membrane (BM) in the cochlea can be emulated with a series of resonators with different frequencies covering the audible range. Rhode (1971, 1978) was first to observe

[a)]Electronic mail: Atiq.Islam@westernsydney.edu.au, ORCID: 0000-0001-5105-9262.

extensive cochlear nonlinearities in a squirrel monkey and a guinea pig. Another study (Allen, 2001) showed that cochlear nonlinearities are critical in determining the range of auditory processing of hearing sub-systems. For example, nonlinear compression is necessary to detect and recognize high-frequency sounds, two-tone suppression produces a sharp formant for spoken vowels, and level-dependent nonlinearity handles a wide dynamic range of sound levels in hearing.

More recently, several cochlear models have been proposed that emulate the physiological and psychoacoustic characteristics of the human auditory periphery system (Lyon, 2011b; Saremi and Stenfelt, 2013; Verhulst *et al.*, 2012; Zilany and Bruce, 2006). In a previous study (Saremi *et al.*, 2016), seven recent cochlear models were compared in terms of compatibility with the evaluation of cochlear excitation patterns, frequency selectivity, nonlinear response growth, amplitude modulation processing, computational cost, level-dependent tuning, and input-output functions. The evaluation was executed for frequencies of 0.5, 1, 2, 4, and 8 kHz which are used for clinical hearing assessment. The evaluated results were compared with the available physical experimental results. The study concluded that the cascade of asymmetric resonator with fast acting compression (CARFAC) and auditory nerve (AN) cochlear models (Zilany and Bruce, 2006) best fit those physical experimental results among seven cochlear models (Saremi *et al.*, 2016). The CARFAC closely fit experimental data in 12 out of 13 experiments. It could have fit all 13 experiments by changing the value of one parameter (V_{offset}) in the model (Saremi and Lyon, 2018). The study also showed that the computational time of the CARFAC is substantially lower than the other nonlinear cochlear models for a specific task (Saremi *et al.*, 2016). Despite the best fit to human auditory physiological and psychoacoustic data, nobody has applied the CARFAC or the AN models to a SID task.

We explore the CARFAC model as a front-end in a SID system and compare its performance to that of the AN model and other conventional FFT feature generators. The CARFAC model simulates the cochlea as cascaded asymmetric resonators with an automatic gain control (AGC) feedback loop. It models instantaneous and dynamic nonlinearities that capture the full range of nonlinear processing in the human auditory pathway. We hypothesize that these nonlinearities will yield accurate SID, even when input signals are corrupted with noise.

We pair our front-ends with simple and transparent back-end classifiers such as linear support vector machines (SVM) (Chang and Lin, 2011; Cristianini and Shawe-Taylor, 2000) and Gaussian mixture models (GMM) with the universal background model (UBM). Recent works demonstrate remarkable performances of deep neural networks on SID tasks (Nassif *et al.*, 2021; Snyder *et al.*, 2018; Sztahó *et al.*, 2019). Our purpose is not to maximize SID accuracies with state of the art back-ends, but rather to investigate whether and how nonlinearities in biologically inspired front-ends might help back-ends learn better speaker models given noisy speech. We also want to investigate how various elements of the auditory pathway model contribute to this task, and better understand how the human auditory system achieves noise robustness. While deep neural networks represent state of the art back-ends across a range of tasks, the reason for their superior performance is often obscure, and the system is usually treated as a black box. Their large number of hidden activation units obfuscate the effects of cochlear nonlinearities on SID accuracy, which is what we want to investigate. Moreover, the training of a deep neural network often requires significant training data which is not always available, e.g., for the datasets we consider here. However, for completeness, we will also use the extreme learning machine (ELM) back-end (Huang *et al.*, 2006) to indicate the possible SID performance of CARFAC when paired a neural network. The ELM requires less data for training without sacrificing much performance capability (Al-Kaltakchi *et al.*, 2021).

We compare the SID performance of the CARFAC model with three other SID pre-processing front-ends. The first is MFCC, which is a standard FFT-based front-end often used as a baseline for comparison (Alam and Zilany, 2019; Li and Huang, 2011; Zhang *et al.*, 2018). The second is FDLP, which emphasizes acoustic cues related to the human voice production system, and generally produces a higher SID accuracy than MFCC front-ends (Ganapathy *et al.*, 2012). The third is the AN model, which has been shown to outperform MFCCs, GFCCs, and FDLP in SID tasks under noisy conditions (Ganapathy *et al.*, 2012). We compare the SID performances of these four front-ends under a wide range of conditions. We corrupt input signals with white, pink, and a variety of non-stationary types of noise. We also vary the signal-to-noise ratios (SNRs) of those noise types. We use two text-dependent datasets in two different languages as input. We then explore how different nonlinearities in the CARFAC model impact its SID performance. Our results show that the CARFAC model consistently outperforms the other three feature generation front-ends, particularly when input signals are noisy. This outperformance is consistent across datasets, types of background noise, and back-end classifiers. More broadly, our results show that the inherent nonlinear processing capabilities of biological auditory systems are at least partially responsible for their robust SID performance in noisy environments. The necessity of cochlear nonlinearities changes with the noise types, and becomes more essential under non-stationary noise conditions.

II. METHOD AND MATERIALS

We briefly describe our SID front-ends, back-end classifiers, and datasets.

A. Front-end feature extraction

In this section, we describe the CARFAC, AN, MFCC, FDLP, and GFCC extraction process from an input audio signal.

1. The CARFAC front-end

Figure 1 presents a block diagram of the CARFAC front-end. In the training stage (top of Fig. 1), we present clean speech to the CARFAC model. We calculate the energy (energy calculation block, Fig. 1) of its output, i.e., the BM and the inner hair cell (IHCs) responses. We then train the back-end classifiers, either a GMM-UBM or an SVM, with the CARFAC output energy. The classifier then learns a map between a speaker's BM features and their identity.

The testing stage (bottom of Fig. 1) is similar to the training stage. The key difference is that the input signal is corrupted by some types of noise at some SNR values. The noisy signal is added to the original input utterance, and we calculate the CARFAC output energy as we did in the training stage. We present those front-end features to the trained back-end classifier to guess the speaker's identity. We next discuss some of the blocks in Fig. 1 in more detail.

a. The CARFAC block. The CARFAC model is described in Lyon (2017). It uses a cascade of second-order asymmetric resonators (the CAR section) to model the BM response to a transduced traveling wave. The transfer function of the CAR is

$$H = \frac{Y}{X} = g\frac{z^2 - (2a_0 - hc_0)rz + r^2}{z^2 - 2a_0 rz + r^2}, \tag{1}$$

where $g = \left[1 - 2a_0 r + r^2/1 - (2a_0 - hc_0)r + r^2\right]$, $a_0 = \cos(2\pi f_c/f_s)$; $c_0 = \sin(2\pi f_c/f_s)$; and $h < 2(1+a_o)/c_o$. Here, f_s and f_c are the sampling frequency and the cut-off frequency. The parameter r is determined by the FAC section. Without the FAC section, the value of r is set to 1. The values of h control the pole-zero distance. Consequently, h also controls the gain and bandwidth of the CAR filter. The CARFAC also models the IHC function that represents the sound as transduced on the auditory nerve. We will soon compare the effect of training the back-end on either BM or IHC output on SID performance.

The FAC section in the CARFAC includes the outer hair cells and automatic gain control with smoothing filters (Lyon, 2017). The resonator pole and zero locations control the damping factor, which in turn changes the gain and bandwidth of the BM filter. It emulates the level-dependent compressive nonlinearity in the model (Lyon, 2017). The impact of the h values on the BM response is shown in Fig. SuppPub1A (Islam, 2022). In this study, we use h=0.35$*C_0$, because empirical results suggest that value achieves high SID accuracy. In our simulations, we set the values of f_s to 16 kHz and f_c is determined by the Greenwood function (Greenwood, 1961) to map 25 channels from 125 Hz to 3 kHz. We set the upper-frequency limit at 3 kHz because most SID cues, such as the speaker's fundamental frequency, pitch, and formants (f_1 and f_2), are below this frequency (Stemple *et al.*, 2018). The CARFAC damping factor is a parameter that controls the compression of BM responses. In human hearing research, typical values of the damping factor range from 0.1 to 0.4 (Lyon, 2017). We set the damping factor to 0.15 to compute the minimum pole-zero radius that achieves maximum damping, which in turn facilitates high SID accuracy as shown in Fig. SuppPub1B (Islam, 2022).

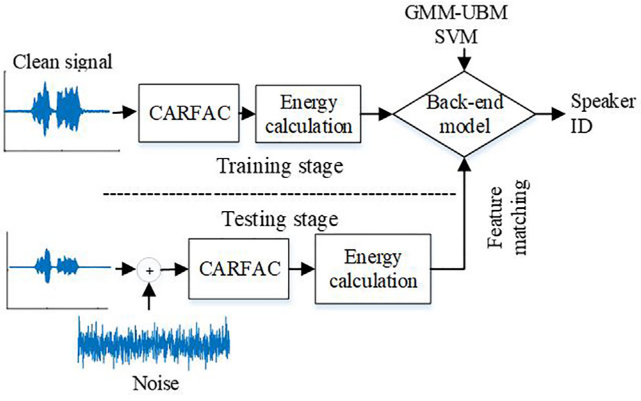

FIG. 1. (Color online) The block diagram of the CARFAC-based speaker identification system.

b. The CARFAC energy calculation block. The energy of the CARFAC output is the feature that we use to identify speakers. The CARFAC response was discretized over 25 separate channels. We calculated the energy of each channel by constructing time windows that overlapped each other. We then calculated the BM energy E in channel i using

$$E(i) = \sum_{j=1}^{L} C(i,\ 1+j : j+L)^2, \tag{2}$$

where j is the starting index for each time window, L is the window duration, and $C(i)$ contains the output samples of channel i. Empirically, we found that a 50 ms window duration with 50% overlap provided high SID accuracy for both clean and noisy input speech. We also observed that the two lowest frequency channels contained the most energy, and eliminating them revealed richer speaker-defining features in the other channels. So the size of the output BM energy was $23 \times F$, where F is the number of frames in the input signal.

Figure 2(A) shows examples of the CARFAC energy, i.e., the BM energy (right column), as well as the input signals that generated them (left column). The left column shows waveforms of a speech utterance corrupted by different types of noise as indicated. The right column shows the BM energies of those waveforms, which are the inputs for our back-end classifier. The x axis is time, and the y axis is the channel number. Comparing the top three rows of the right column of FIG. 2(A), we see that CARFAC filters out the superimposed noise well. The outputs for speech corrupted by white noise and pink noise strongly resemble those of noiseless speech (top row). Figure 2(B) shows how the output BM energy can be used as features to identify speakers. The left three panels of Fig. 2(B) show waveforms for the same utterance spoken by three different people.

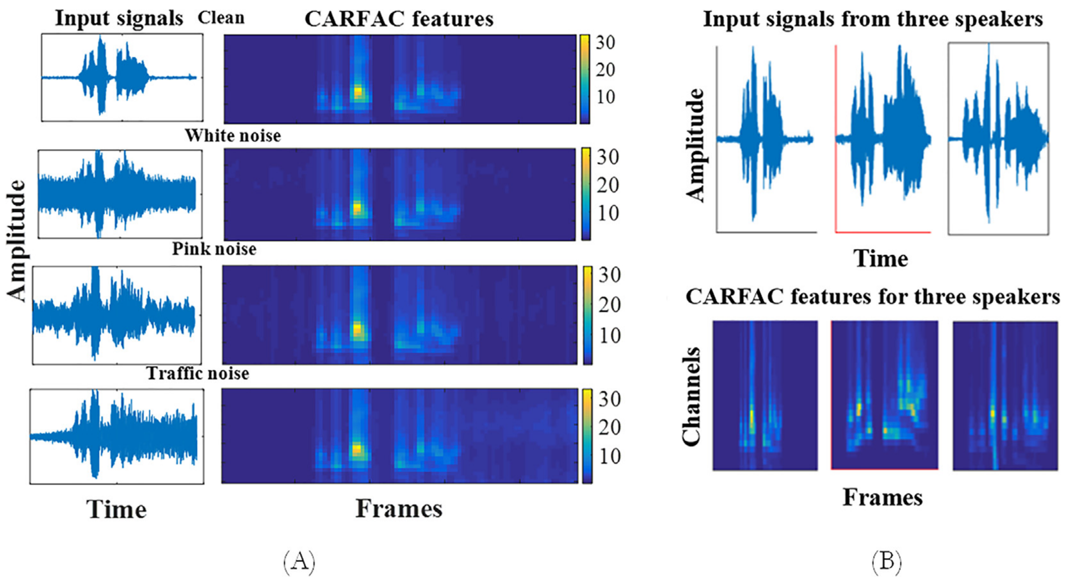

FIG. 2. (Color online) The CARFAC features in response to given input signals are shown in (A). Responses are shown for clean and noisy signals (0 dB signal-to-noise ratio) to show the noise-robustness of CARFAC. (B) Utterances from three speakers and their BM responses are shown to illustrate the speaker distinguishing capability of CARFAC.

The right three panels of Fig. 2(B) show the output BM energies for those three waveforms. The three examples are clearly different, illustrating that the CARFAC generates speaker-distinguishing features from their input signals.

2. The AN front-end

The details of the AN model and related equations can be found in Zilany and Bruce (2006). Updated versions of this model are presented in Zilany *et al.* (2014) and Bruce *et al.* (2018). Both the neurogram (Alam and Zilany, 2019; Islam *et al.*, 2016) and the synapse response (Zilany, 2018) from the AN model have been used in SID systems. The extraction of these responses from an input signal is very time-consuming in software simulations. They require a very high sampling rate for the AN model to emulate the cochlear response faithfully. The neurogram-based SID result is comparable to those obtained with MFCCs (Islam *et al.*, 2016).

To address these limitations, we only use the BM response of the AN model in this work. This feature extraction is illustrated in Fig. 3. It uses only the signal path filter (chirping C1 block, Fig. 3) along with feedback nonlinearity from the AN model (lower blocks, Fig. 3). Our version of the AN model does not require the linear filter (C2) or the IHC sections of its antecedent [see Fig. 1 in Ref. Zilany and Bruce (2006)]. These simplifications halve the computation time compared to computing neurograms (Islam *et al.*, 2016) while improving its resultant SID accuracies.

The control path (control path filter block, Fig. 3) controls the gain and bandwidth of the C1 filter and is responsible for the cochlea's level-dependent nonlinearity. The C1

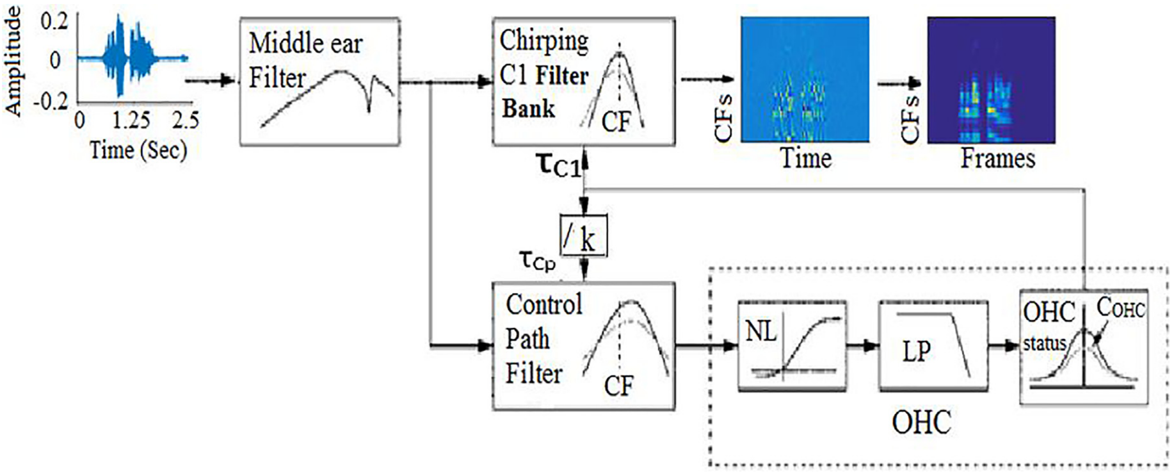

FIG. 3. (Color online) Block diagram for BM feature extraction from the AN model. The signal path filter and the control path from the AN model are shown in the block diagram. This figure has been modified from Zilany and Bruce (2006).

filter is a 10th order Chirp filter and the control path filter is a third-order Gammatone filter. The control path filter has a lower time constant than the C1 filter, which results in a two-tone suppression nonlinearity in the signal path. The Boltzman nonlinear function (NL block, Fig. 3) followed by a low pass filter (LP block, Fig. 3) with an 800 Hz cut-off frequency compresses a wide range of sound levels in the C1 filter. The OHC nonlinear function (OHC block, Fig. 3) converts the low-pass filter output to a time-varying time-constant for the C1 filter. In this way, the control path controls the gain and bandwidth of the C1 filter based on the input signal and emulates two-tone suppression and compression nonlinearity in the model.

Like previous SID applications of the AN model (Islam *et al.*, 2016; Zilany, 2018), we used 25 channels with a frequency ranging from 125 Hz to 3 kHz in a logarithmic scale to simulate the BM response. To facilitate a fair comparison with the CARFAC model, we compute the BM energy via Eq. (1) with the same window length and overlap as used in the CARFAC model.

3. The FDLP front-end

Figure 4 presents a block diagram of the FDLP feature extraction process. A full description of the FDLP and the equations it implements can be found in Ganapathy *et al.* (2012). FDLP has been used in SID (Alam and Zilany, 2019; Islam *et al.*, 2016) and gender detection (Islam, 2016) applications. The FDLP estimates high-energy peaks of a spectrogram. Initially, high frequencies in the input signal are boosted in a pre-emphasis stage (top row, Fig. 4). A DCT converts the input into the frequency domain. The full-band DCT signal is split into successive sub-bands using a windowing technique to create a power spectrum. Next, an IFFT is applied to the power spectrum to generate autocorrelation coefficients for a recursion process. This recursion process uses the Levinson-Durbin algorithm (Franke, 1985) to produce the auto-regressive (AR) model coefficients according to the model order for a prescribed autocorrelation sequence. Here, the model order is 40 following the study of Ganapathy *et al.* (2012) to fit the temporal envelope to pitch pulses.

The generated AR model coefficients are transformed into a power spectrum by applying a FFT, and then the resultant power spectrum matrix is inverted. This inverted power spectrum for a full-band signal is called an FDLP envelope (bottom row, Fig. 4). Each band of the envelope is buffered using a 50 ms frame size with 50% overlap between frames. A Hamming window then estimates the short-term energy in each band. A log operation generates a log-energy spectrum, and another DCT is applied to convert it into 13 cepstral coefficients. We calculate the differences across those coefficients (del, bottom row, Fig. 4), and the difference of those differences (ddel, bottom row, Fig. 4). Together with the cepstral coefficients in the extracted feature vector, our FDLP feature dimensionality is $39 \times F$, where F is the number of frames in the signal. Empirical results found that including both del and ddel in the FDLP feature provides a better SID result. The bottom-left plot in Figure 4 shows an example of the FDLP feature output. This FDLP feature is then forwarded to the back-end to evaluate the FDLP method's SID accuracy.

4. The MFCC front-end

MFCCs (Davis and Mermelstein, 1990) are often used as a benchmark in SID applications. We use the same frequency ranges for all algorithms, with 25 channels of triangular filters, and a 50 ms frame with 50% overlap between frames. We exclude the delta and delta-delta features that we included in the FDLP front-end because they cause a reduction of MFCC performance as we have empirically

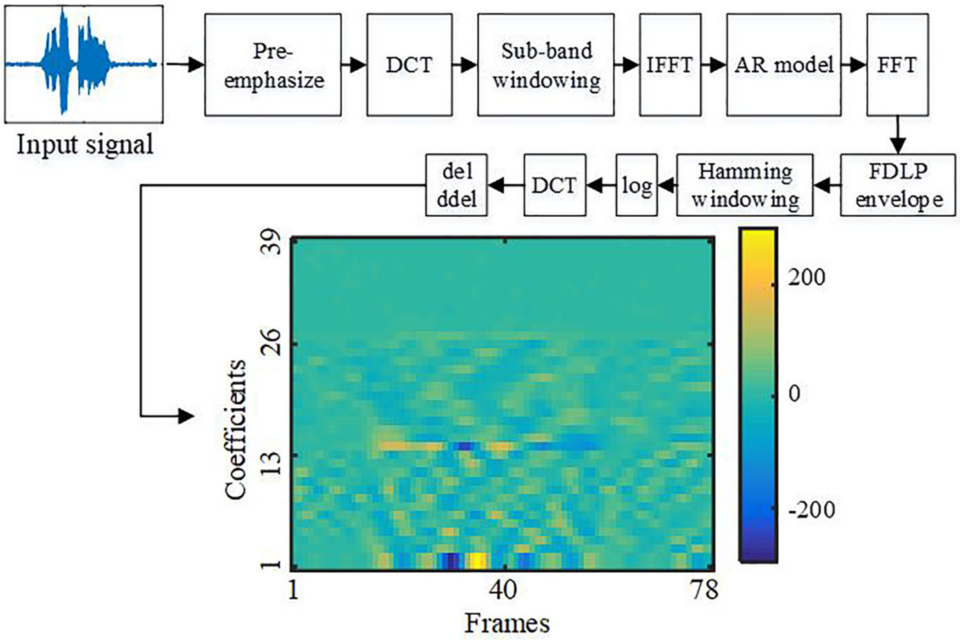

FIG. 4. (Color online) Block diagram of the FDLP feature extraction from an input speech. Abbreviations: discrete cosine transform (DCT), fast Fourier transform (FFT), inverse FFT (IFFT), and frequency domain linear prediction (FDLP).

found in this study. We used the RASTAMAT toolbox of MATLAB (Ellis, 2005) to extract MFCCs from input speech.

All extracted front-end features were normalized such that each channel mean is zero and each channel standard deviation is one. This normalization reduces the variation between clean and noisy speech features and enhances SID performance. These normalized speech features are input for the back-end classifier to identify the speaker.

5. The GFCC front-end

Gammatone frequency cepstral coefficients (GFCC) have been applied to many SID tasks (Shao and Wang, 2008; Zhao *et al.*, 2012; Zhao and Wang, 2013). Our implementation of the GFCC front-end is similar to Shao and Wang (2008). We used a Gammatone filter with 25 channels in a frequency range from 125 Hz to 3 kHz, in the same manner as the other front-ends. We applied a cube root and DCT to the resultant spectral features. We then omitted the lowest frequency channel, which contained the highest energy and was prone to noise [see also Zhao *et al.* (2012)]. The output GFCC feature size was then $24 \times n$, where n is the number of samples per channel. We have utilized this GFCC for a comparison with cochlear methods applying cube root and DCT (described later).

B. Back-end speaker models

In our experiment, we used either the GMM-UBM or the SVM as a back-end classifier. We briefly discuss each in turn. The GMM-UBM has enjoyed previous success in SID systems (Islam, 2017; Islam *et al.*, 2016; Reynolds *et al.*, 2000; Togneri and Pullella, 2011; Zilany, 2018). The general GMM, denoted λ_{UBM}, is generated in the UBM step. It is parameterized by the mean vector μ_i, weights W_i, and a covariance matrix $\sum_i$ obtained from all mixture components M as $\lambda_{UBM}=\{W_k, \mu_k, \sum_k\}$, where $k=1,2,\ldots, M$. We used $M=128$ mixture components to build the λ_{UBM} models. The expectation-maximization (EM) algorithm (Dempster *et al.*, 1977) trains λ_{UBM} iteratively [33]. On each EM iteration, the model's likelihood is increased for each training utterance sample $(x_t|\ t=1,2,3,\ldots, T)$ in the full dataset by the following formulas:

$$W_k = \frac{1}{T}\sum_{t=1}^{T} p(k|x_t, \lambda), \tag{3}$$

$$\mu_k = \frac{\sum_{t=1}^{T} p(k|x_t, \lambda)x_t}{\sum_{t=1}^{T} p(k|x_t, \lambda)}, \tag{4}$$

$$\sigma_k^2 = \frac{\sum_{t=1}^{T} p(k|x_t, \lambda)x_t^2}{\sum_{t=1}^{T} p(k|x_t, \lambda)} - \mu_k^2. \tag{5}$$

Here, T is the total number of samples in the training dataset pooling from all speakers and σ_k^2 is the variance matrix, i.e., the diagonal elements of the covariance matrix for all mixture components. The initial GMM in the UBM step consists of the training dataset's mean vector, variance matrix, and weight vector.

Each individual GMM step starts by calculating the posterior probability $p(k|x_t)$ of the training vectors $(x_t|t=1,2,3,\ldots,T_t)$ for the UBM distribution components. Here, T_t is the total number of training samples for each speaker. $p(k|x_t)$ for each component k in the UBM is

$$p(k|x_t) = \frac{W_k p_k(x_t)}{\sum_{k=1}^{M} W_k p_k(x_t)}, \tag{6}$$

where $p_k(x_t)$ is the Gaussian distribution of mixture k. The probability counts L_k, mean μ_k^{New} and variance $\sigma_k^{New^2}$ for all training samples from each speaker for the mixture k are computed using $p(k|x_t)$,

$$L_k = \sum_{t=1}^{T_t} p(k|x_t), \tag{7}$$

$$\mu_k^{New} = \frac{1}{L_k}\sum_{t=1}^{T_t} p(k|x_t)x_t, \tag{8}$$

$$\sigma_k^{New^2} = \frac{1}{L_k}\sum_{t=1}^{T_t} p(k|x_t)x_t^2. \tag{9}$$

This step is similar to the expectation step in UBM development (Reynolds *et al.*, 2000). Next, the estimated new parameters for the GMM and the old parameters of the UBM are used to tune the new GMM parameters for a speaker,

$$W_k^G = \left[\frac{\alpha L_k}{T} + (1-\alpha)W_k\right]\gamma, \tag{10}$$

$$\mu_k^G = \alpha\mu_k^{New} + (1-\alpha)\mu_k, \tag{11}$$

$$\sigma_k^{G^2} = \alpha\sigma_k^{New^2} + (1-\alpha)\left(\sigma_k^2 + \mu_k^2\right) - \mu_k^{G2}, \tag{12}$$

where α is the adaptation coefficients for the weights, means, and variances, respectively. This coefficient balances the UBM parameters and new estimates. In Eq. (10), γ is a scaling factor that ensures $\sum_{k=1}^{M} W_k^G = 1$. The adaptation coefficient $\alpha = L_k/(L_k+v)$, where v is the relevance or adaptation factor, which we empirically set to $v=10$. Finally, the hypothesized GMM speaker model is $\lambda_{GMM} = \{W_k^G, \mu_k^G, \sigma_k^{G^2}\}$, $k=1,2,\ldots,M$. Then the log-likelihood for a test sequence of feature vectors X is computed as

$$\wedge(X) = \log p(X|\lambda_{GMM}), \tag{13}$$

where $X = \{x_t|t \in 1,\ 2,\ 3,\ldots,\ T1\}$ and $T1$ is the number of testing samples. Each testing sample has a log-likelihood score against each speaker model. The maximum testing score against a speaker model and index of that model yields the most likely identity of the target speaker. A confusion matrix of speakers counts the diagonal indices of the highest matched speaker against all speaker models. The SID score is computed using speakers S and testing samples from each speaker, R,

$$\text{SID accuracy} = \frac{\sum_{m=1}^{n}\sum_{l=1}^{n} D_{ml}}{R \times S}, \tag{14}$$

where m and l are the rows and columns indices of the confusion matrix (D), respectively.

The SVM (Cortes and Vapnik, 1995) is a supervised classifier and widely used for object classification. In the training stage, a nonlinear SVM kernel maps input training data into a higher dimensional feature space to make them linearly separable by hyperplanes. These hyperplanes help to classify data points depending on their location relative to the hyperplanes. The data points closest to the hyperplane are called support vectors (SVs) and control the orientation and location of hyperplanes. The SVM determines SVs to maximize the margin between SVs and the hyperplane. In the testing stage, the SVM returns predicted labels and probabilities for the testing set. We used the predicted labels to compute SID accuracies. The predicted label is matched with the given label for the testing sample and the maximum matching label indicates the identity of the target speaker.

We used the C-support vector classifier (C-SVC), proposed in Cortes and Vapnik (1995) and available in the LIBSVM library (Chang and Lin, 2011) to classify speakers. We used the radial basis function (RBF) kernel. To train the SVM we must tune two parameters C and γ. The parameter γ is inversely proportional to the span of the kernel and C is inversely proportional to the margin between SVs. We used a cross-validation algorithm (Cortes and Vapnik, 1995) to find the values of C and γ that give the best result for each of our SID systems. Empirically, we found that $C = 2$ and $\gamma = 0.09$ yield a better SID accuracy for the cochlear front-ends, and $C = 2$ and $\gamma = 0.05$ for the MFCC and FDLP front-ends.

C. Datasets

We use two datasets as input to our SID systems. The first is the University of Malaya (UM) speech dataset (Islam *et al.*, 2015), which contains 39 native Malaysian speakers. The second is the Bangla dataset (Islam and Sakib, 2019), which contains 40 Bangladeshi speakers. Both datasets are publicly available (Islam *et al.*, 2022).

In both datasets, each speaker produces 10 samples of a short phrase. The utterances from both datasets have a wide dynamic range from 20 to 90 dB. The spoken phrase in the Bangla dataset is "Ami vat khai (I eat rice)" and it is "University Malaya" in the UM dataset. Their average durations are 3 and 2.5 s, respectively. Phrases from the UM dataset were recorded in a soundproof booth in Kuala Lumpur, Malaysia. Phrases from the Bangla dataset were recorded with a mobile phone in a quiet environment in Noakhali, Bangladesh.

Both datasets are text-dependent and comprise brief samples. Each of these characteristics is desirable for our experiments because they simplify the SID task, so we can focus on the feature generation process (i.e., the front-end). Text-independent SID tasks are generally more challenging than text-dependent ones, requiring more data and longer utterances to train the back-end classifier (Poddar *et al.*, 2015). Having more training data and longer utterances improves the noise-robustness of the classifier. So, the noise-robustness of the front-end feature extractor and its impact on SID accuracy is obscured if we train our system on large datasets. Our smaller datasets with shorter utterances reduce the ability of the classifier to compensate for signal noise. While our SID task is text-dependent with small datasets, it still has applications in areas such as voice activation of smart devices, or identifying a suspect from a voice lineup.

The Bangla dataset has slow, normal, and fast modes of utterances from each speaker. This speaking speed variation allows us to investigate their effects on SID performance. Each mode of utterance contains 10 samples from 40 speakers. Our results typically refer to the normal mode of utterance. In one subsection, we will investigate how SID performance depends on speaking speed.

Our SID systems were trained on clean speech, i.e., speech uncorrupted by noise. We randomly chose seven of the ten samples from each speaker to train the back-ends. We used the remaining three samples for testing. We added various types of background noise at various signal-to-noise ratios (SNRs) to that testing data. Our noise types were white (Gaussian), pink (1/*f* spectrum), or traffic (nonstationary) noise. Our SNRs ranged from –5 to 15 dB in increments of 5 dB. We also evaluated performances on clean testing data.

III. RESULTS

We apply cross-validation in all experiments with 6 independent trials. In the following figures, the solid bars display the average SID accuracy over those trials. The error bars display the maximum and minimum values of our 6 trials.

A. Comparing SID performances on noisy speech

Figure 5 presents the performances of the CARFAC model when we use BM energy (green bars) or IHC output (blue bars) to train the back-end. We use the same channel number and frequency information to fairly compare their performances. The results were generated using the GMM-UBM back-end. Figure 5 shows that the GMM-UBM learns a more accurate speaker model when trained on BM energy than IHC output. The IHC output requires similar computation time as BM energy for the CARFAC front-end, but

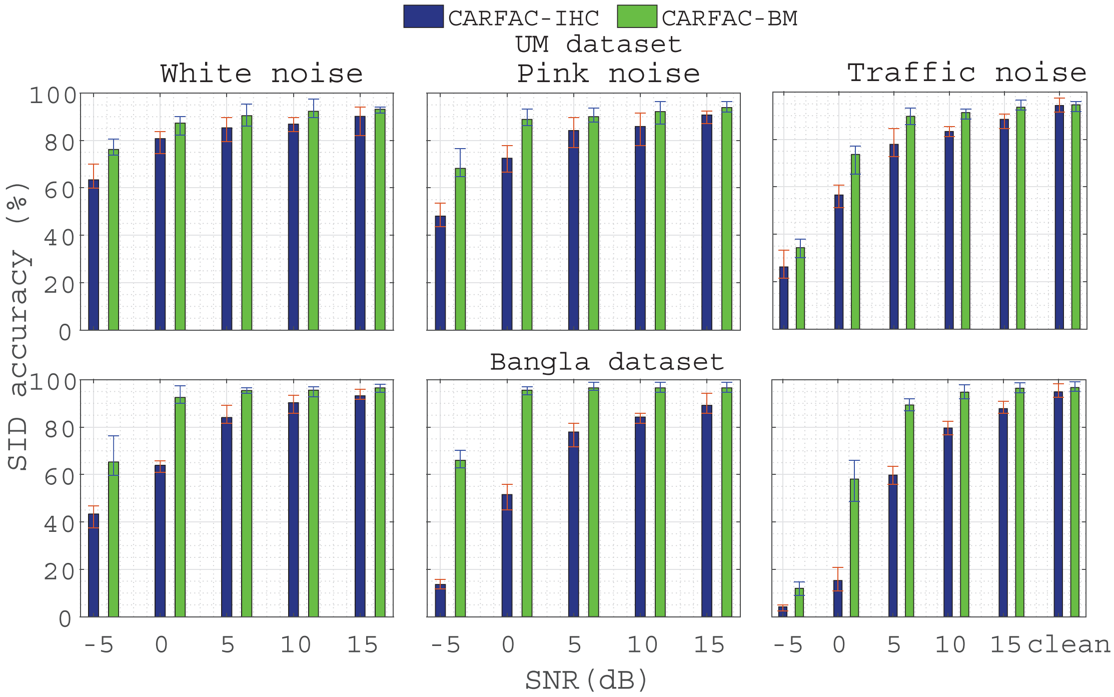

FIG. 5. (Color online) Results are shown for the IHC, and BM responses from the CARFAC model for the clean and noisy conditions using the GMM-UBM as a back-end model.

significantly more time for the AN model. Also, FFT-based approaches simulate the BM response from an audio signal. Therefore, we use CARFAC's BM energy to train the back-ends for all subsequent experiments.

Figure 6 compares the SID performances of the CARFAC, AN, FDLP, and MFCC front-ends (Fig. 6 legend) on the UM dataset. We used the GMM-UBM (top row) and SVM (bottom row) as back-end classifiers. The columns of Fig. 6 specify the type of background noise added to the testing dataset, and the x axes of the panels indicate the SNR. Figure 6 shows that all four front-ends have similar performances when the testing dataset had no added background noise (clean, far-right bars in Fig. 6). But their performances on noisy data vary. For example, the MFCC front-end noticeably drops in SID accuracy, even for relatively high SNRs. That drop is consistent across noise types

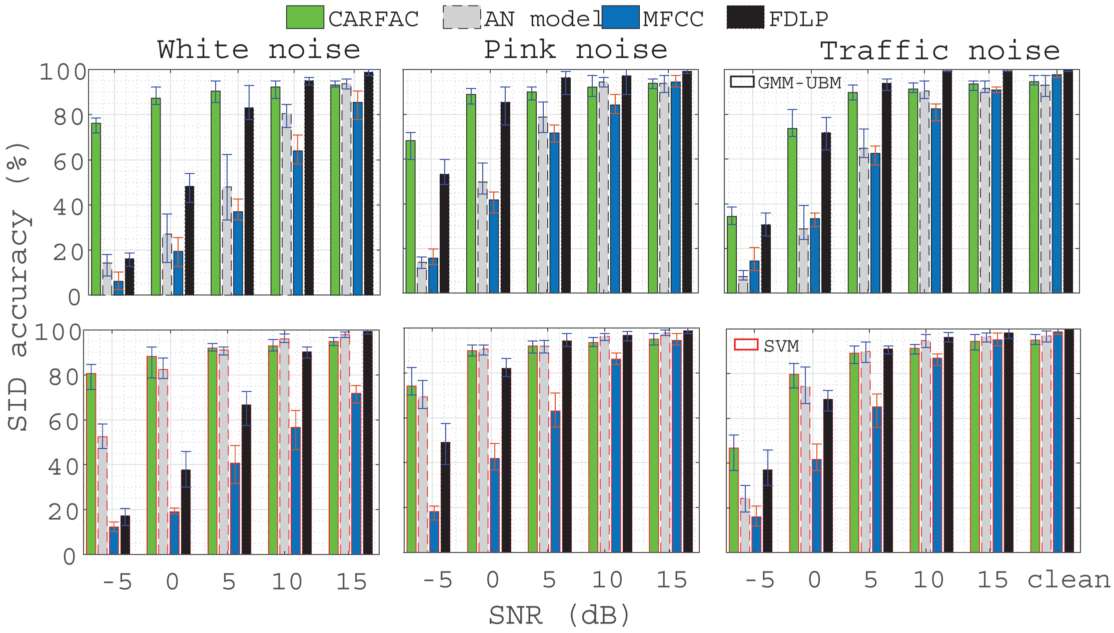

FIG. 6. (Color online) The SID accuracies for the CARFAC and alternative methods using the UM dataset. Results are shown for white noise (left column), pink noise (middle column), and traffic noise (right column) with a range of SNRs (x axes) using a GMM-UBM (top row) or SVM (bottom row) classifier.

and our choice of back-end classifier. The FDLP front-end maintains a high SID accuracy if the SNR is high. For pink and traffic noise types, the FDLP front-end has the highest SID accuracy when the SNR is 15 dB, as previously reported (Islam *et al.*, 2016). But their performances dramatically decrease as the SNR decreases. In particular, the SID accuracy at –5 dB SNR is on average below 36% for all noise types and both back-end classifiers, and often much lower than that.

The AN front-end yields higher SID accuracies than the MFCC and FDLP front-ends at low SNRs (except for traffic noise), but only if we use the SVM as a classifier. The CARFAC front-end also yields high SID accuracies at low SNRs, but its performance is less sensitive to our choice of classifier. However, with the GMM-UBM back-end, CARFAC outperforms all others when data is noisy (i.e., low SNR), for all noise types. All SID front-ends suffer low performances with traffic noise at –5 dB SNR. Traffic noise is a fluctuating noise and affects the whole speech spectrum. There is a significant difference between clean and noisy speech features, particularly at very low SNRs. It is difficult to identify speakers accurately because training and testing data strongly differ. Later we will investigate how nonlinearities in cochlear front-ends can improve SID accuracy on noisy, nonstationary speech.

Figure 7 presents analogous results to Fig. 6, but on the Bangla dataset. We again observe the same principal results from Fig. 6. The MFCC front-end (blue bars, Fig. 7) classifies speakers accurately only for clean testing data. The FDLP front-end (black bars) classifies accurately for pink and traffic noise at high SNRs. The AN front-end (gray bars) outperforms MFCCs at lower SNRs irrespective of back-end classifiers. When the SVM is used as the back-end classifier, the CARFAC front-end (green bars) significantly outperforms all others at low SNRs for all noise types. All front-ends struggle to correctly classify speakers with traffic noise at low SNRs (leftmost bars of right panel). However, the CARFAC front-end can achieve more than 50% correct SID accuracy under traffic noise with the SVM classifier.

Collectively, Figs. 6 and 7 show that CARFAC classifies noisy speech better than alternative front-ends, particularly when the noise is stationary. Figures 6 and 7 also show that CARFAC is robust to noise types up to 5 dB SNR, which is the threshold for a good conservational SNR level (Rindel, 2019). These results of CARFAC are invariant to our choice of dataset or back-end classifier. Moreover, CARFAC's classification accuracy generally varied less over six independent trials. We observe this trend for both datasets and both classifiers. In contrast, the performance of alternative methods varied depending on the input dataset.

B. Noisy speech at different speaking speeds

SID systems usually use input speech at normal conversational speeds. We investigated the impact of speaking speed on the SID performance of our four front-ends using the SVM as a back-end. We used the Bangla dataset for this investigation because it contains samples spoken at three different speeds.Figure 8 displays spectrograms of input speech from the Bangla dataset for a sample spoken quickly (left panel), normally (middle panel), and slowly (right panel). Figure 8 illustrates three reasons why our front-ends might classify speakers for slow and normal speech more accurately than for fast speech.

First, the spectrogram of fast speech contains less spectral energy for the last word of a phrase (at time = 1 s, left panel, Fig. 8) than the spectrograms of normal and slow speeds (at times 1.5 and 2 s, middle and right panels, Fig. 8).

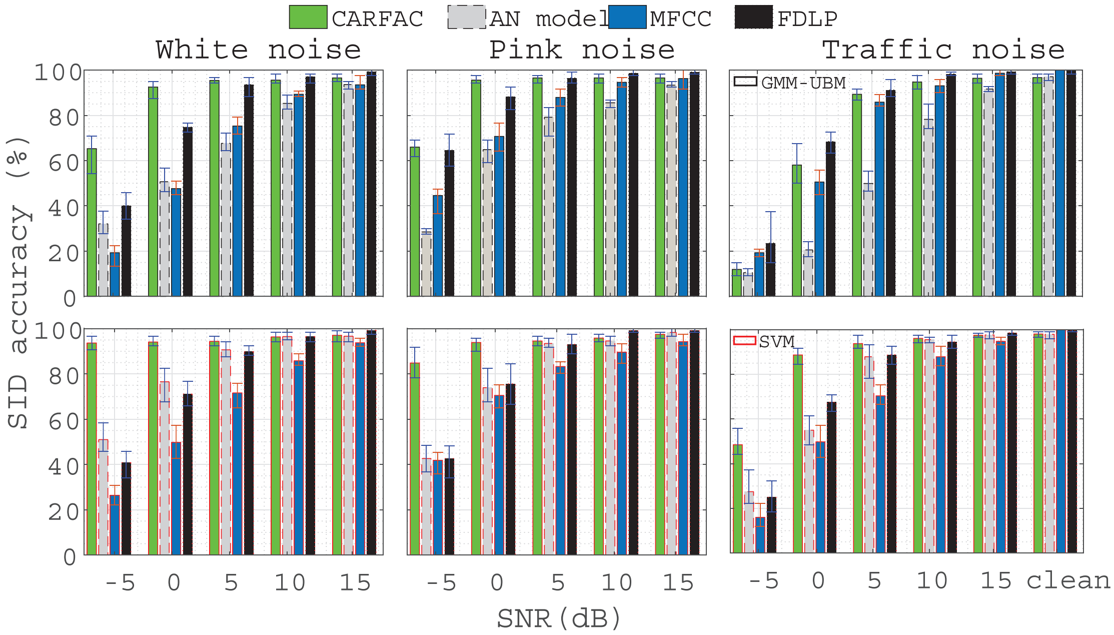

FIG. 7. (Color online) The SID accuracies for the CARFAC-based and alternative methods using the Bangla dataset. The layout is analogous to Fig. 6.

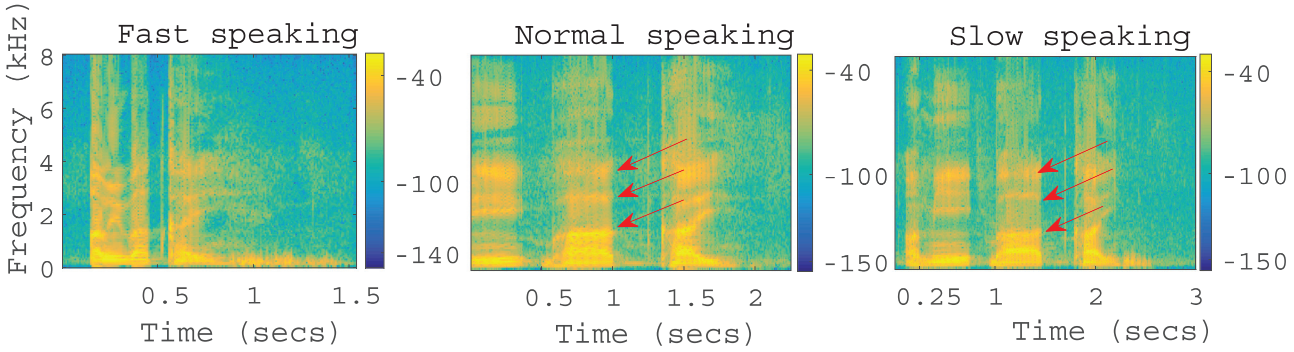

FIG. 8. (Color online) Spectrogram of three speaking speeds of the same utterance to illustrate the effect of speaking speed on energy distribution and formant patterns.

The last word of fast speech often provides less information about the identity of a speaker than other speaking speeds. Second, the final word in an utterance is sometimes slurred, so it provides less spectral information about the speaker's identity. For example, the right panel of Fig. 8 shows clear boundaries marking the final phoneme in the sample (at time = 2 s), whereas the left panel of Fig. 8 shows that the final boundary is smeared. Third, the second and third formants are less distinguishable in the fast utterance than they are in the normal and slow utterance (red arrows, Fig. 8).

Figure 9 presents the performances of our four SID front-ends on slow (left panel), normal (middle panel), or fast (right panel) utterance speeds. We used the SVM as the back-end. Figure 9 shows that speaking speed affects the SID performance for all front-ends, but some front-ends are affected more than others. The SID accuracy of MFCCs (blue bars) barely increases or decreases as speaking speed increases or decreases for all noise levels. Curiously, the FDLP (black bars) classifies less accurately for slow speech than fast speech. Cochlea-inspired front-ends (green and gray bars) yield higher classification accuracies for normal and slow utterance speeds than they do for fast speed. This result is particularly true at –5 dB SNR. The AN model has also a similar pattern of results to CARFAC for different speeds of utterances. However, the CARFAC front-end significantly outperforms the other three given very noisy input data (i.e.,–5 dB SNR), regardless of speaking speed.

C. Varying the number of channels

Figure 10 compares the SID accuracies of our front-ends while varying their number of channels. Specifically, we set the channels of each front-end to 15, 25, 35, and 45 channels (left to right panels). All front-ends used the frequency information ranging from 125 Hz to 3 kHz. We used the SVM as a back-end classifier for each front-end, and we used the Bangla dataset as input in all experiments. The noise in our testing dataset was pink noise with SNR as indicated. Figure 10 shows that, as we increase the number of channels above 15, the SID performance of the CARFAC (green bars) becomes better. In contrast, the AN model (gray bars) provides poor SID accuracy while the number of channels is varied from 25, particularly at –5 dB SNR (left gray bars, first, third, fourth panels). These results suggest that the CARFAC front-end needs a higher number of channels to produce a better result. In contrast, the variation of channel numbers affects the performance of the AN model significantly, particularly at –5 dB SNR. For example, the CARFAC produces a better SID performance than the AN model using only 25 channels. This variation of channel numbers causes a change of spectral information. Thus it seems the bio-inspired front-ends are more sensitive to changes in channel numbers. In contrast to cochlear methods, the MFCC (blue bars) and FDLP (black bars) front-ends have less variation of performance with the change of the number of channels, regardless of SNRs. Their similar

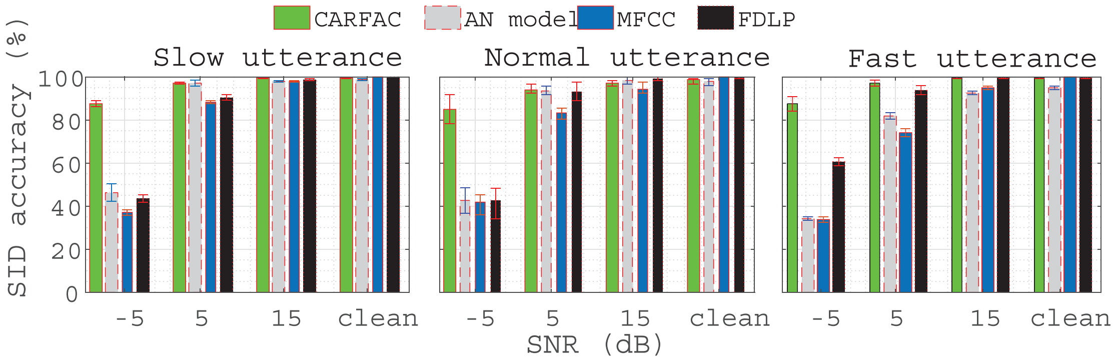

FIG. 9. (Color online) The effect of speaking speed on SID accuracy for our front-ends. Each method results are simulated for pink noise (SNRs: –5 dB, 5 dB, 15 dB), and clean condition.

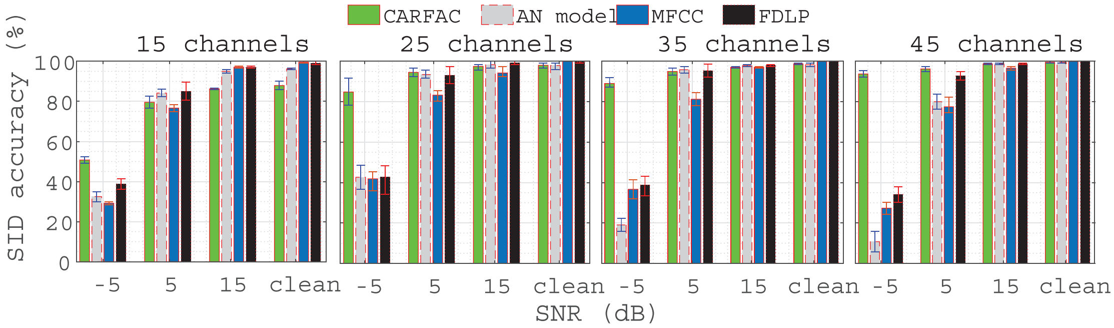

FIG. 10. (Color online) The effect of channel numbers on SID performance given clean and noisy speech. We added pink noise to testing data (at SNRs –5 dB, 5 dB, and 15 dB), and also tested performances on noiseless testing data (clean, all panels).

responses to variations in channel numbers indicates that the performance of the MFCC and FDLP is robust to channel numbers.

D. CARFAC nonlinearities and their effect on performance

The CARFAC front-end implements nonlinear computations in two ways. First, it performs level-dependent multi-rate nonlinearities through the AGC operation that models cochlear functions (Lyon, 2017). Second, an instantaneous nonlinear function (NLF) interacts with the input waveforms and produces a combination tone, like the cubic distortion tone (CDT) in the cochlea.

To investigate the effect of these nonlinearities on the SID task, we compared the performances of four variants of the CARFAC model. The first is the linear CAR section of CARFAC, i.e., the front-end does not implement nonlinear operations. The second and third are the linear CAR section combined with an AGC and NLF components, respectively. The fourth is the full CARFAC front-end which includes both nonlinearities functions (Lyon, 2017).

Figure 11 compares the performances of these four CARFAC variants on the Bangla dataset with the SVM as a back-end classifier. We generated a separate SVM speaker model for each CARFAC variant using their training samples. Figure 11 shows that the full CARFAC front-end (green bars) identifies speakers most accurately across all noise types and SNRs, particularly when compared to the linear CAR (red bars) under clean and noisy conditions. This result suggests that the necessity of cochlear nonlinearities is essential to identify a speaker more accurately.

The variants of CARFAC produce similar performances above 5 dB SNR irrespective of types of noise. The CAR with AGC (gray bars) produces a similar or better result than the CAR with NLF (purple bars) at –5 dB SNR, particularly under pink (middle panel) and traffic (right panel) noise. This result indicates that the compressive nonlinearity (AGC) might be more useful than the instantaneous NLF to classify speaker accurately under noisy conditions. This observation is particularly true given low SNRs with time-varying noise signals (cf. leftmost gray and purple bars of the right panel). The CAR with NLF outperforms the linear CAR at –5 dB SNR, particularly for pink and white noise. The NLF function produces distortion tones that decrease the similarity between clean and noisy speech features and cause a reduction of SID accuracy of the CAR with the NLF method. However, both of those nonlinearities are less effective at classifying noisy speech if they operate in isolation as shown in FIG. 11. The full CARFAC front-end adds a two-tone suppression effect via the AGC, which suppresses the instantaneous distortion. Figure 11 suggests that the two cochlear nonlinearities working in tandem can boost SID performance, particularly in the presence of noise.

E. Additional nonlinearities applied to cochlear features

The CARFAC and the AN models are not the only front-ends to utilize nonlinear computations. The FFT-based

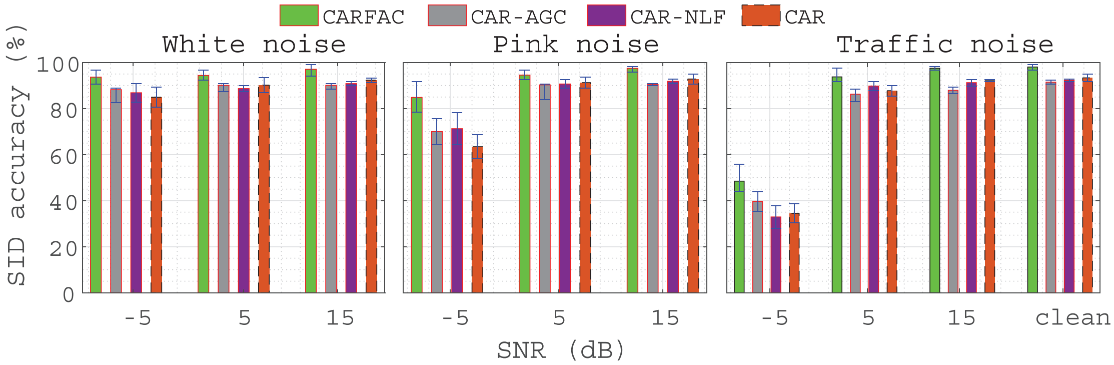

FIG. 11. (Color online) An evaluation of the contribution of each stage from the CARFAC in the SID system performance.

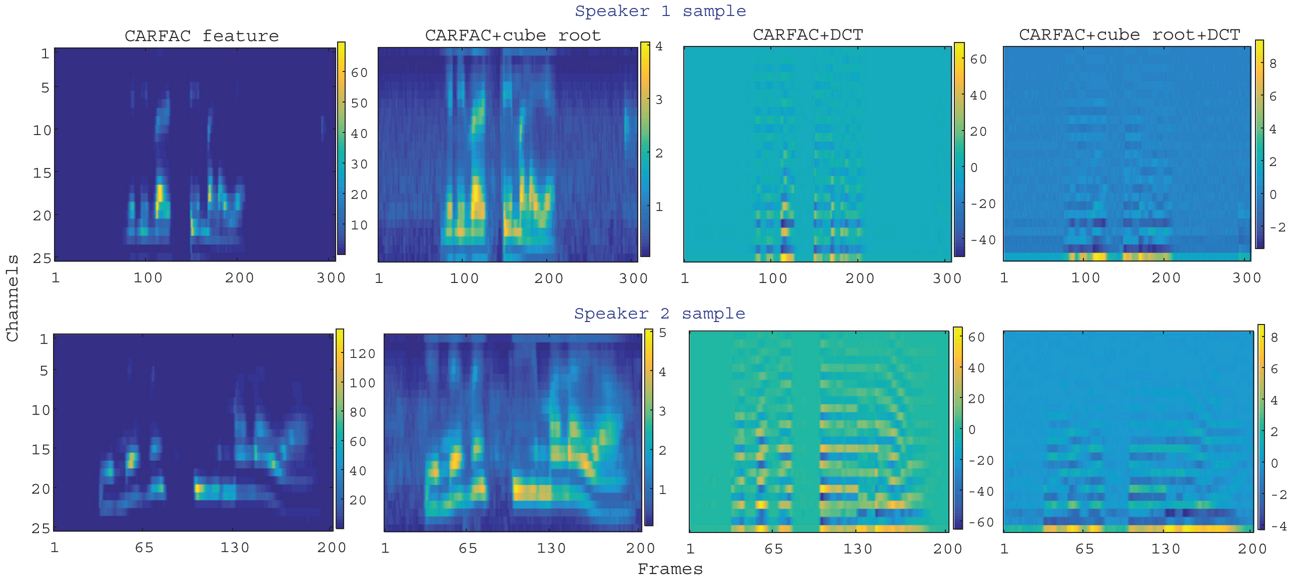

FIG. 12. (Color online) The effect of the cube root (second column) and DCT (third column) on CARFAC's BM features (left). The cube root and DCT effect is shown in the right column. All features are shown for two speakers (top and bottom rows).

GFCC is analogous to the MFCC in operation, for example, applying a cube root exponent (as an instantaneous nonlinearity) and a DCT on the Gammatone spectral features to extract nonlinear input features. However, the cube root is an instantaneous nonlinearity, and not dynamic like the FAC section of the CARFAC model. We applied a cube root to mimic Stevens's power-law (Stevens, 1972; Stevens, 1957) to the front-end features. We wondered whether applying analogous nonlinearities to the AN model and CARFAC's output features would boost SID performance.

Figure 12 illustrates the effect of applying a cube root exponent and DCT to the CARFAC's output features. The left panels in Fig. 12 display a typical CARFAC energy feature. The middle panels separately apply a cube root exponent and DCT to CARFAC's output features. The cube root exponent dynamically adapts signal intensity. It amplifies unvoiced speech and suppresses the intensity of loud parts in the input (left-middle panel). These effects increase the variation between speakers, which we expect will yield higher SID accuracy. The standalone DCT emphasizes only voiced speech (right-middle panels) which helps to achieve a higher SID accuracy, particularly under noisy conditions. The DCT also compresses energies toward lower-frequency channels and decorrelates input features. The DCT causes cochlear energy features to be symmetrically distributed, as the Gaussian distribution is. Therefore, we expect that the GMM can more accurately model speakers with compressed high-frequency channel information, particularly in noisy environments. The right panels in Fig. 12 illustrate the combined effect of the cube root and DCT. The cube root nonlinearly amplifies the input signal and boosts the unvoiced portion, as shown in Fig. 12 (third column) and Fig. 13 (second column). In contrast, the DCT emphasizes the voiced signal, which increases the similarity between clean and noisy signals.

Figure 13 shows that the DCT transforms data to be approximately symmetric about the origin (third and fourth panels) compared to CARFAC's BM energy (first panel) or cube root output (second panel). Thus, the DCT should distribute input data in a way that helps symmetric generative functions (e.g., GMMs) to learn more accurate speaker models. However, the application of the DCT followed by the cube root amplifies unvoiced input (i.e., noise) via the cube root. We expect SID performance would suffer as a result.

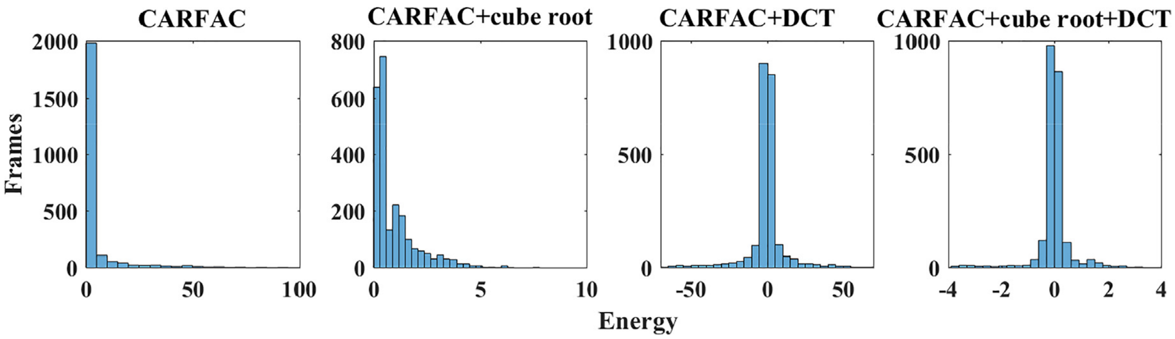

FIG. 13. (Color online) The effect of cube root and DCT on the data distribution.

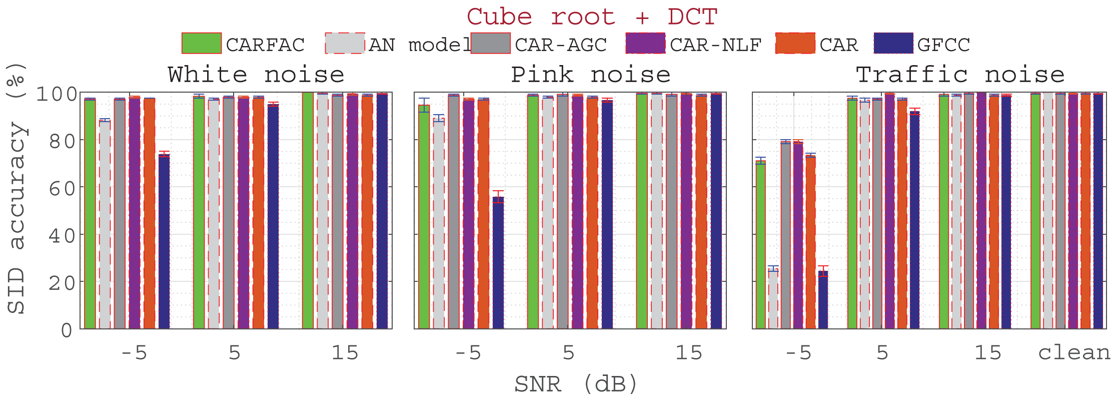

FIG. 14. (Color online) The nonlinearity effect on SID system performance is shown for CARFAC-based and alternative methods. The results are simulated for clean and noisy signals using the SVM classifier.

Figures 12 and 13 show that the combined nonlinearities provide better compression and decorrelation than isolated nonlinearities (right panel in Fig. 12).

Figure 14 compares the SID performances of our four CARFAC variants (from FIG. 11) and the AN front-end (light gray bars). We applied the cube root exponent and a DCT to the extracted features.

We then trained the SVM classifier with those transformed features from the Bangla dataset. Figure 14 also compares our CARFAC variants with GFCCs (blue bars), which employs the cube root and DCT nonlinearities on spectral features (Shao and Wang, 2008).

Figure 14 shows that the inclusion of the cube root and DCT nonlinearities significantly improves the SID performance of all CARFAC variants (compared to Fig. 11). All cochlear models, including the full CARFAC and AN front-ends, achieve significantly higher SID performance on traffic noise at –5 dB than we observed previously (cf. Figs. 6 and 7). The CARFAC variants, the original CARFAC model, and the AN outperform the GFCC at −5 dB when the cube root and DCT are applied under all noise types. All CARFAC variant front-ends achieve a much higher SID accuracy than the AN model at low SNR, regardless of noise types. Comparing Figs. 11 and 14, we see that that the SID performance of cochlear models is sensitive to changes in the additional nonlinear computations that they implement. For example, adding the cube root followed by the DCT to the output CARFAC features improves SID performance on non-stationary noisy data (compare the right panels of Figs. 11 and 14). Comparing Figs. 11 and 15, the CAR section followed by the cube root and DCT outperforms the CARFAC model without these nonlinearities. Hence, applying a static nonlinearity (e.g., a cube root) followed by DCT can improve performance beyond the dynamic/adaptive nonlinearity of CARFAC.

Figure 15 displays results of analogous experiments to Fig. 14, but with a GMM-UBM back-end classifier. We used clean data to train the back-end. Both Figs. 14 and 15 display similar results. Cochlear front-ends may be well-suited for physiological and psychoacoustic data, but they require additional and specific nonlinearities to optimize performance in SID tasks.

Figures 14 and 15 show that the cochlear front-ends can classify speakers more accurately when we apply the cube root and DCT to their output. We also investigate how accurately cochlear front-ends classify speech given other types of nonstationary noise such as airport, factory, restaurant, train, cocktail party, street, and exhibition noise as a background. We applied the cube root and DCT to the CARFAC, AN model, and CAR output features to compare

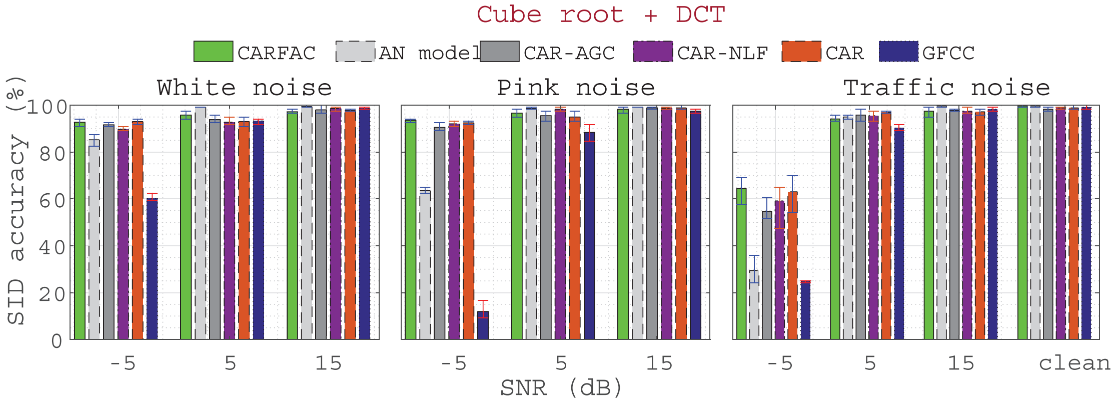

FIG. 15. (Color online) The nonlinearity effect on SID system performance is shown for CARFAC-based and alternative methods. The results are simulated for clean and noisy signals using the GMM-UBM classifier.

their performance under other non-stationary noise. We obtained similar results to traffic noise conditions [results are shown in Fig. SuppPub2 in Ref. Islam (2022)]. Therefore, the cochlear models need to incorporate conventional compressive nonlinearities such as the cube root followed by DCT to produce a higher SID accuracy under all types of noise.

F. Performance with ELM

Finally, we use the ELM as a back-end to present preliminary SID results for a biologically inspired cochlear front-end coupled with a state of the art back-end. We chose the ELM as our back-end instead of a deep neural network because the former requires significantly less training data than the latter. Figure 16 presents our results. To facilitate a fair comparison, we used the same input features for the ELM as we did for the GMM-UBM and SVM, but we resized the input features to be 22 × 22. Empirically we observed that other resizing, such as 28 × 28 or 64 × 64, yields lower SID accuracy than 22 × 22. For the ELM, we used the root mean square propagation training technique and the initial learning rate was 0.01. The regularization rate was 0.00005. The maximum epoch was 30 with a batch size of 22. We used these settings for both the UM (top row) and Bangla (bottom row) datasets. We also used the same settings for the GMM-UBM and SVM back-ends on both datasets.

Figure 16 shows that the ELM (blue bars) produces similar results to the GMM-UBM (dark green bars), particularly under noisy conditions, while the SVM back-end (light green bars) generally outperforms both. Presumably the ELM requires more training data to achieve similar SID accuracies as simpler back-ends. Figure 16 shows that ELM performance remains consistent irrespective of SNR, except for the traffic noise (blue bars, right panel). Applying the cube root and DCT to CARFAC's output BM features (red bars) can further boost SID accuracy. This improvement indicates that the nonlinearities in the front-end can play an important role in identifying speakers from noisy data. Training the ELM or DNN on more data should further enhance SID accuracies. But our goal is to show how nonlinear processing in the front-end can help the back-end learn more noise-robust speaker models, whether or not we have sufficient data to train a state-of-the-art backend.

G. Understanding the effect of frequency and amplitude on SID performance

In the CARFAC model, the response shows a frequency dependent amplitude, and its frequency scale is arranged with the Greenwood function (Greenwood, 1961), which are different from the widely used log mel-spectrum and mel-cepstrum. We use the mel-filter bank output (mel-spectrum) to investigate the effect of frequency scale, and the log mel-filter bank output (log mel-spectrum) to understand the effect of amplitude compression. The generated result is shown in Fig. 17.

In this work, we compare the CARFAC with the log mel-spectrum and mel-spectrum accuracy in the SID task. The linear mel-spectrum produces poorest performance under clean condition. The log mel-spectrum produces better performance than the mel-spectrum, but poorer performance under noisy conditions, as shown in Fig. 17. This poor performance due to the low variance of log output for a high change of amplitude in an input. The application of DCT on the log mel-spectrum improves SID performance

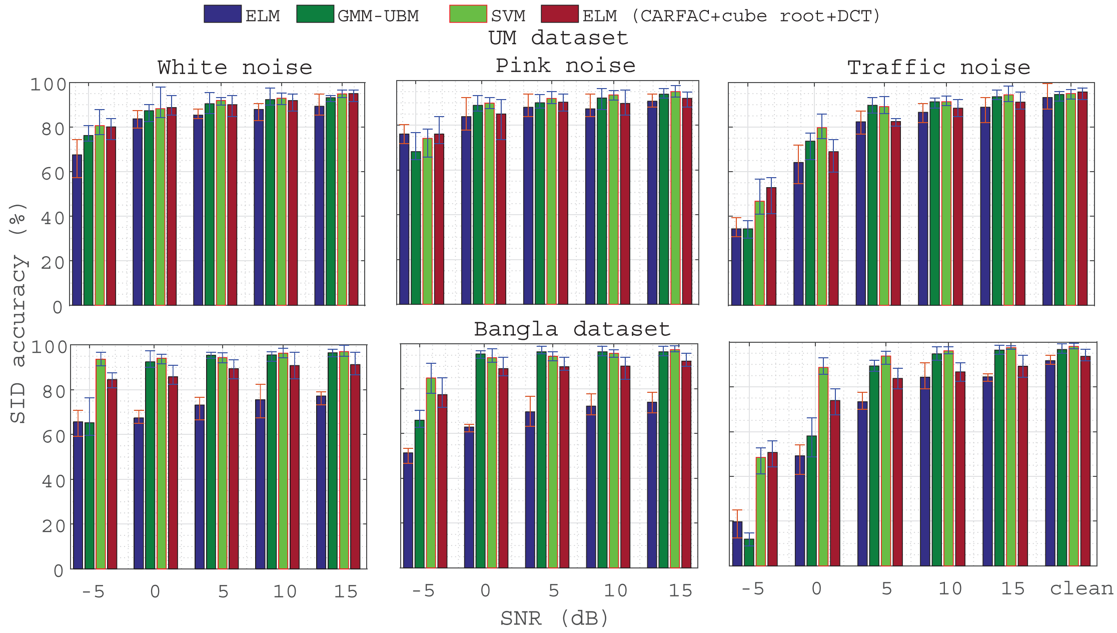

FIG. 16. (Color online) Results for the UM (top row) and Bangla (bottom row) datasets using the CARFAC as the front-end and the ELM (blue bars), the GMM-UBM (dark green bars), and the SVM (light green bars) as a back-end. We also incorporated additional cube root and DCT operations to CARFAC's output BM features (red bars).

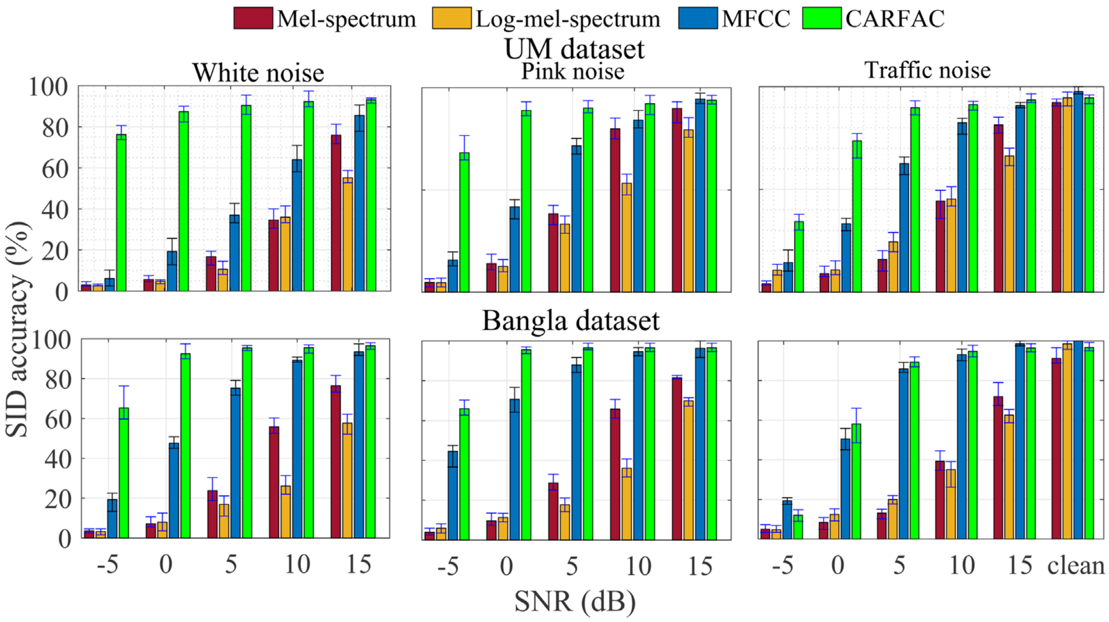

FIG. 17. (Color online) The performance of the MFCC front-end applying only Mel-filter-bank outputs (red bars), log-Mel-filter-bank outputs (yellow bars), and MFCCs (blue bars). The CARFAC results are added to show their noise-robustness compared to Mel-spectrum variants. Results are shown for UM (top row) and Bangla (bottom row) datasets using the GMM-UBM as a classifier.

substantially, as shown in Fig. 17. The mel-spectrum shows more logarithmic in frequency scale and keeps a unit gain in amplitude. The log mel-spectrum shows more logarithmic in both frequency scale and amplitude. As shown in Fig. 17, we found in this experiment, the CARFAC outperforms all variants of mel-spectrum or mel-cepstrum with the inherent frequency mapping and frequency gain.

IV. DISCUSSION

Humans are excellent at identifying speakers, even in noisy environments. We investigated whether cochlea-inspired front-ends could outperform more conventional approaches to SID tasks when inputs are corrupted by noise. We showed that the CARFAC front-end is very effective at filtering out stationary noise, even at low SNR. So, its back-end classifier learns accurate representations of speakers from noisy input, even from small (text-dependent) input datasets. We showed that traditional SID front-ends, i.e., MFCCs and FLDP, comparatively struggle at this task. We showed these results on two datasets, with two back-end classifiers, with various noise types and amplitudes, with different channel numbers, and with different speaking speeds.

We also investigated the impact of cochlear nonlinearities in SID performance, particularly if the corrupting noise was nonstationary. Compression, two-tone suppression, and level-dependent critical bandwidth variation (emulated by AGC) are apparently more important to SID tasks than instantaneous nonlinearities (NLF), particularly for time-varying noise at low SNRs. However, combining a compressive nonlinearity and an instantaneous nonlinearity is more effective than either in isolation. Other nonlinearities such as the cube root exponent can compress the input signal. When we applied the cube root to the linear CAR section, we found that the resultant SID performance rivaled or substantially exceeded that of CARFAC on noisy nonstationary data. Presumably, the cube root can optimize loudness more effectively than the FAC section of CARFAC. Then perhaps we can construct better models of cochlear loudness compression that would further improve the SID performance of the CARFAC model. For example, we could use the cube root to emulate the function of the outer hair cells and an instantaneous nonlinearity. In future work, we can investigate the effect of the CARFAC amplitude and the frequency scale on a SID task. An investigation of SID performance applying the CARFAC using per-channel energy normalization (Lyon, 2011a) and learnable audio front-end (Zeghidour *et al.*, 2021) can be done in the future.

We used simple classifiers to focus our experiments on the relationship between nonlinearities in cochlear front-ends and SID accuracy. We generally obtained higher SID accuracies when we paired an SVM with a nonlinear kernel with our cochlea models. This observation suggests that nonlinearities in the back-end can also enhance SID performances, as certain nonlinearities in our cochlea models do. A channel decorrelation technique such as DCT in the front-end features can further enhance the performance of back-end classifiers, particularly in noisy conditions. A differencing operation between adjacent channels, such as principal component analysis could also be helpful for feature decorrelation purposes and will be investigated in future work. Incorporating noisy data in speaker training or speech enhancement (Taherian *et al.*, 2020) may also help to achieve noise-robust SID performance. The application of a neural network as a back-end (Chen and Salman, 2011) for

CARFAC may further enhance SID performance due to its nonlinear operations, particularly when a large training dataset is available.

The CARFAC front-end, and cochlea-inspired algorithms in general, offer promising approaches to perform text-dependent SID tasks in real-world (i.e., noisy) conditions (Islam *et al.*, 2016). One extension of this study would be to compare performances of similar algorithms on text-independent datasets. Our preliminary investigations found that CARFAC with the cube root and DCT improves performance over the CARFAC-only version for a text-independent SID task using the TIMIT dataset (results not shown). Applying joint speech separation (Mowlaee *et al.*, 2012; Mowlaee *et al.*, 2010) in a text-independent SID task can convert it to a text-dependent SID task by adding a speech separation block (Rix *et al.*, 2001). Then CARFAC might achieve noise-robust performance in this converted task. A low-powered, real-time implementation of the CARFAC model is available (Xu *et al.*, 2018). If we can tweak it to achieve robust performance on text-independent SID tasks, we could implement real-time recognition systems with direct applications to e.g., smartphone access, crime investigation, and telephone banking. Additionally, we could apply the implemented SID system in edge computing systems with restrictions on hardware, power, training data, and bandwidth. We can also apply it in situations where noise is significant and large deep neural networks struggle from a lack of training data. Biology and neuroscience have a history of inspiring machine learning algorithms for a wide variety of applications (Monk *et al.*, 2016, 2018). Cochlear models for SID tasks should be added to that list.

Al-Kaltakchi, M. T., Abdullah, M. A., Woo, W. L., and Dlay, S. S. (**2021**). "Combined i-vector and extreme learning machine approach for robust speaker identification and evaluation with SITW 2016, NIST 2008, TIMIT Databases," Circuits Syst. Sign. Process. **40**, 4903–4921.

Alam, M. S., and Zilany, M. S. (**2019**). "Speaker identification system under noisy conditions," in *Paper Presented at the 2019 5th International Conference on Advances in Electrical Engineering (ICAEE)*.

Allen, J. (**2001**). "Nonlinear cochlear signal processing," in *Physiology of the Ear*, 2nd ed. (Singular Thompson, Norwich), pp. 393–442.

Bruce, I. C., Erfani, Y., and Zilany, M. S. (**2018**). "A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites," Hear. Res. **360**, 40–54.

Chang, C.-C., and Lin, C.-J. (**2011**). "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol. **2**(3), 27.

Chen, K., and Salman, A. (**2011**). "Learning speaker-specific characteristics with a deep neural architecture," IEEE Trans. Neural Netw. **22**(11), 1744–1756.

Cortes, C., and Vapnik, V. (**1995**). "Support-vector networks," Mach. Learn. **20**(3), 273–297.

Cristianini, N., and Shawe-Taylor, J. (**2000**). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods* (Cambridge University Press, Cambridge).

Davies, M., Srinivasa, N., Lin, T.-H., Chinya, G., Cao, Y., Choday, S. H., Dimou, G., Joshi, P., Imam, N., Jain, S., Liao, Y., Lin, C.-K., Lines, A., Liu, R., Mathaikutty, D., McCoy, S., Paul, A., Tse, J., Venkataramanan, G., Weng, Y.-H., Wild, A., Yang, Y., and Wang, H. (**2018**). "Loihi: A neuromorphic manycore processor with on-chip learning," IEEE Micro **38**(1), 82–99.

Davis, S. B., and Mermelstein, P. (**1990**). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *Readings in Speech Recognition* (Elsevier, Amsterdam), pp. 65–74.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (**1977**). "Maximum likelihood from incomplete data via the EM algorithm," J. R. Statistical Soc. Ser. B (Method.) **39**(1), 1–22.

Ellis, D. P. (**2005**). PLP and RASTA and MFCC, and inversion in Matlab.

Franke, J. (**1985**). "A Levinson-Durbin recursion for autoregressive-moving average processes," Biometrika **72**(3), 573–581.

Ganapathy, S., Thomas, S., and Hermansky, H. (**2012**). "Feature extraction using 2-D autoregressive models for speaker recognition," paper presented at *Odyssey 2012-The Speaker and Language Recognition Workshop*.

Ghazanfar, A. A., and Rendall, D. (**2008**). "Evolution of human vocal production," Curr. Biol. **18**(11), R457–R460.

Greenwood, D. D. (**1961**). "Critical bandwidth and the frequency coordinates of the basilar membrane," J. Acoust. Soc. Am. **33**(10), 1344–1356.

Hansen, J. H., and Hasan, T. (**2015**). "Speaker recognition by machines and humans: A tutorial review," IEEE Sign. Process. Mag. **32**(6), 74–99.

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (**2006**). "Extreme learning machine: Theory and applications," Neurocomputing **70**(1-3), 489–501.

Islam, M. (**2016**). "Frequency domain linear prediction-based robust text-dependent speaker identification," paper presented at *2016 International Conference on Innovations in Science, Engineering and Technology (ICISET)*.

Islam, M., Zilany, M., and Wissam, A. (**2015**). "Neural-Response-Based Text-Dependent speaker identification under noisy conditions," paper presented at *International Conference for Innovation in Biomedical Engineering and Life Sciences*.

Islam, M. A. (**2017**). "Modified mel-frequency cepstral coefficients (MMFCC) in robust text-dependent speaker identification," paper presented at *2017 4th International Conference on Advances in Electrical Engineering (ICAEE)*.

Islam, M. A., Jassim, W. A., Cheok, N. S., and Zilany, M. S. A. (**2016**). "A robust speaker identification system using the responses from a model of the auditory periphery," PloS One **11**(7), e0158520.

Islam, M. A., and Sakib, A.-N. (**2019**). "Bangla dataset and MMFCC in text-dependent speaker identification," Eng. Appl. Sci. Res. **46**(1), 56–63.

Islam, M. A., Xu, Y., Monk, T., Afshar, S., and van Schaik, A. (**2022**). "Text dependent SID: Noise-robust text-dependent speaker identification using cochlear models," SID results showing the effect of damping factor, pole-zero distance, and other types of nonstationary noise, https://www.westernsydney.edu.au/icns/reproducible_research/publication_support_materials/text_dependent_sid.

Li, Q., and Huang, Y. (**2011**). "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," IEEE Trans. Audio Speech Lang. Process. **19**(6), 1791–1801.

Lyon, R. F. (**2011a**). "Cascades of two-pole–two-zero asymmetric resonators are good models of peripheral auditory function," J. Acoust. Soc. Am. **130**(6), 3893–3904.

Lyon, R. F. (**2011b**). "Using a cascade of asymmetric resonators with fast-acting compression as a cochlear model for machine-hearing applications," in *Paper Presented at the Autumn Meeting of the Acoustical Society of Japan*.

Lyon, R. F. (**2017**). *Human and Machine Hearing* (Cambridge University Press, Cambridge).

Makhoul, J. (**1975**). "Linear prediction: A tutorial review," Proc. IEEE **63**(4), 561–580.

Modha, D. S. (**2014**). *Introducing a Brain-Inspired Computer: TrueNorth's Neurons to Revolutionize System Architecture* (IBM Research, Cambridge).

Monk, T., Savin, C., and Lücke, J. (**2016**). "Neurons equipped with intrinsic plasticity learn stimulus intensity statistics," paper presented at *Advances in Neural Information Processing Systems*.

Monk, T., Savin, C., and Lücke, J. (**2018**). "Optimal neural inference of stimulus intensities," Sci. Rep. **8**(1), 1–10.

Mowlaee, P., Saeidi, R., Christensen, M. G., Tan, Z.-H., Kinnunen, T., Franti, P., and Jensen, S. H. (**2012**). "A joint approach for single-channel speaker identification and speech separation," IEEE Trans. Audio Speech Lang. Process. **20**(9), 2586–2601.

Mowlaee, P., Saeidi, R., Tan, Z.-H., Christensen, M. G., Fränti, P., and Jensen, S. H. (**2010**). "Joint single-channel speech separation and speaker identification," *Paper Presented at the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Nassif, A. B., Shahin, I., Hamsa, S., Nemmour, N., and Hirose, K. (**2021**). "CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions," Appl. Soft Comput. **103**, 107141.

Nayana, P., Mathew, D., and Thomas, A. (**2017**). "Comparison of text independent speaker identification systems using GMM and i-vector methods," Procedia Comput. Sci. **115**, 47–54.

Poddar, A., Sahidullah, M., and Saha, G. (**2015**). "Performance comparison of speaker recognition systems in presence of duration variability," in *Paper Presented at the 2015 Annual IEEE India Conference (INDICON)*.

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (**2000**). "Speaker verification using adapted Gaussian mixture models," Dig. Sign. Process. **10**(1-3), 19–41.

Rhode, W. S. (**1971**). "Observations of the vibration of the basilar membrane in squirrel monkeys using the Mössbauer technique," J. Acoust. Soc. Am. **49**(4B), 1218–1231.

Rhode, W. S. (**1978**). "Some observations on cochlear mechanics," J. Acoust. Soc. Am. **64**(1), 158–176.

Rindel, J. (**2019**). "Restaurant acoustics–Verbal communication in eating establishments," Acoust. Practice **7**, 1–14.

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (**2001**). "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," paper presented at *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Cat. No. 01CH37221).

Saremi, A., Beutelmann, R., Dietz, M., Ashida, G., Kretzberg, J., and Verhulst, S. (**2016**). "A comparative study of seven human cochlear filter models," J. Acoust. Soc. Am. **140**(3), 1618–1634.

Saremi, A., and Lyon, R. F. (**2018**). "Quadratic distortion in a nonlinear cascade model of the human cochlea," J. Acoust. Soc. Am. **143**(5), EL418–EL424.

Saremi, A., and Stenfelt, S. (**2013**). "Effect of metabolic presbyacusis on cochlear responses: A simulation approach using a physiologically-based model," J. Acoust. Soc. Am. **134**(4), 2833–2851.

Shao, Y., Srinivasan, S., and Wang, D. (**2007**). "Incorporating auditory feature uncertainties in robust speaker identification," paper presented at *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007, ICASSP 2007*.

Shao, Y., and Wang, D. (**2008**). "Robust speaker identification using auditory features and computational auditory scene analysis," paper presented at *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (**2018**). "X-vectors: Robust dnn embeddings for speaker recognition," in *Paper Presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Stemple, J. C., Roy, N., and Klaben, B. K. (**2018**). *Clinical Voice Pathology: Theory and Management* (Plural Publishing, San Diego).

Stevens, S. (**1972**). "Perceived level of noise by Mark VII and decibels (*E*)," J. Acoust. Soc. Am. **51**(2B), 575–601.

Stevens, S. S. (**1957**). "On the psychophysical law," Psychol. Rev. **64**(3), 153–181.

Sztahó, D., Szaszák, G., and Beke, A. (**2019**). "Deep learning methods in speaker recognition: A review," arXiv:1911.06615.

Taherian, H., Wang, Z.-Q., Chang, J., and Wang, D. (**2020**). "Robust speaker recognition based on single-channel and multi-channel speech enhancement," IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 1293–1302.

Togneri, R., and Pullella, D. (**2011**). "An overview of speaker identification: Accuracy and robustness issues," IEEE Circuits Syst. Mag. **11**(2), 23–61.

Turchin, A. (**2019**). "Assessing the future plausibility of catastrophically dangerous AI," Futures **107**, 45–58.

Verhulst, S., Dau, T., and Shera, C. A. (**2012**). "Nonlinear time-domain cochlear model for transient stimulation and human otoacoustic emission," J. Acoust. Soc. Am. **132**(6), 3842–3848.

Von Helmholtz, H. (**1863**). *Die Lehre Von Den Tonempfindungen* (*The Theory of Tonal Sensations*) (Friedrich Vieweg und Sohn, Brunswick, Germany).

Xu, Y., Thakur, C. S., Singh, R. K., Hamilton, T. J., Wang, R. M., and van Schaik, A. (**2018**). "A FPGA implementation of the CAR-FAC cochlear model," Front. Neurosci. **12**, 198.

Zeghidour, N., Teboul, O., Quitry, F. d C., and Tagliasacchi, M. (**2021**). "Leaf: A learnable frontend for audio classification," arXiv:2101.08596.

Zhang, M., Kang, X., Wang, Y., Li, L., Tang, Z., Dai, H., and Wang, D. (**2018**). "Human and machine speaker recognition based on short trivial events," paper presented at *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Zhao, X., Shao, Y., and Wang, D. (**2012**). "CASA-based robust speaker identification," IEEE Trans. Audio Speech Lang. Process. **20**(5), 1608–1616.

Zhao, X., and Wang, D. (**2013**). "Analyzing noise robustness of MFCC and GFCC features in speaker identification," paper presented at *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Zilany, M. S. (**2018**). "A novel neural feature for a text-dependent speaker identification system," Eng. Appl. Sci. Res. **45**(2), 112–119.

Zilany, M. S., and Bruce, I. C. (**2006**). "Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery," J. Acoust. Soc. Am. **120**(3), 1446–1466.

Zilany, M. S., Bruce, I. C., and Carney, L. H. (**2014**). "Updated parameters and expanded simulation options for a model of the auditory periphery," J. Acoust. Soc. Am. **135**(1), 283–286.