A Feature Study for Classification-Based Speech Separation at Low Signal-to-Noise Ratios

Jitong Chen, Yuxuan Wang, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Speech separation can be formulated as a classification problem. In classification-based speech separation, supervised learning is employed to classify time-frequency units as either speech-dominant or noise-dominant. In very low signal-to-noise ratio (SNR) conditions, acoustic features extracted from a mixture are crucial for correct classification. In this study, we systematically evaluate a range of promising features for classification-based separation using six nonstationary noises at the low SNR level of -5 dB, which is chosen with the goal of improving human speech intelligibility in mind. In addition, we propose a new feature called multi-resolution cochleagram (MRCG). The new feature is constructed by combining four cochleagrams at different spectrotemporal resolutions in order to capture both the local and contextual information. Experimental results show that MRCG gives the best classification results among all evaluated features. In addition, our results indicate that auto-regressive moving average (ARMA) filtering, a post-processing technique for improving automatic speech recognition features, also improves many acoustic features for speech separation.**

***Index Terms*—ARMA filtering, classification, multi-resolution cochleagram, speech separation.**

I. Introduction

MONAURAL speech separation aims to separate target speech from background interference given a monaural recording. It has a wide range of applications such as robust speech recognition and hearing aid design. Over the past decades, many approaches have been developed to solve the monaural speech separation problem. For example, speech enhancement approaches [25], such as spectral subtraction and Wiener filtering, make statistical assumptions about the background noise (e.g. stationarity) and do not deal well with nonstationary noises, which are quite common in our daily life.

Computational auditory scene analysis (CASA) represents another approach to speech separation, and it is based on perceptual principles of auditory scene analysis [3]. In CASA, the ideal binary mask (IBM) is often considered as a computational objective [35]. The IBM is a time-frequency (T-F) mask constructed from premixed speech and noise, and it is defined as follows.

$$\mathrm{IBM}(t,f) = \begin{cases} 1, & \text{if } \mathrm{SNR}(t,f) > \mathrm{LC} \\ 0, & \text{otherwise} \end{cases}$$

where t denotes time and f denotes frequency. The IBM assigns the value 1 to a T-F unit if the local SNR within the unit exceeds a local criterion (LC), and 0 otherwise. In subject tests, IBM separation has been shown to dramatically improve speech intelligibility in noise for both normal-hearing and hearing-impaired listeners [4], [24], [37], [1]. The estimation of the IBM amounts to a binary classification problem where supervised learning is employed to predict the label of each T-F unit [8]. Recent studies show that classification-based speech separation produces the first demonstration of speech intelligibility improvements for human listeners in background noise [22], [9].

The two key components of classification-based speech separation are acoustic features extracted from an input mixture and classifiers used for supervised learning. While previous studies have emphasized classifiers, the present study focuses on features. Our goal is to reveal how various features perform in classification-based speech separation. To obtain a fair comparison, we choose and fix a multilayer perceptron (MLP) as the classifier to simplify and speedup training, as we are mainly concerned with the relative performance [18]. In addition, we choose a set of six representative nonstationary noises and fix the evaluation SNR to -5 dB. This very low SNR level is selected with the goal of improving speech intelligibility in mind. It is well known that human listeners, even those with significant hearing loss, perform nearly perfectly unless the SNR is in the negative range [15], [28], [37].

In terms of features chosen for evaluation, since the classification approach is only recently established for speech separation, not many features have been developed for this task. We have therefore performed a systematic literature search for robust features published for automatic speech recognition (ASR) in noise, a task that is expected to be related to speech separation. Feature robustness has been extensively studied in the ASR literature. With low SNR and nonstationary noise in mind, we have selected a subset of promising features in our evaluation, such as relative autocorrelation sequence MFCC (RAS-MFCC), Gabor filterbank (GFB) features and power normalized cepstral coefficients (PNCC). These features, together with those previously investigated for speech separation [38], form the existing feature set. Based on our evaluation, we also propose a new feature called multi-resolution cochleagram (MRCG), specifically

Manuscript received April 20, 2014; revised July 18, 2014; accepted September 05, 2014. Date of publication September 19, 2014; date of current version September 26, 2014. This work was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-12-1-0130, the National Institute on Deafness and Other Communication (NIDCD) under Grant R01 DC012048, and the Ohio Supercomputer Center. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Man-Wai Mak.

J. Chen and Y. Wang are with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: chenjit@cse.ohio-state.edu; wangyuxu@cse.ohio-state.edu).

D. Wang is with the Department of Computer Science and Engineering and the Center for Cognitive Science, The Ohio State University, Columbus, OH 43210 USA (e-mail: dwang@cse.ohio-state.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2359159

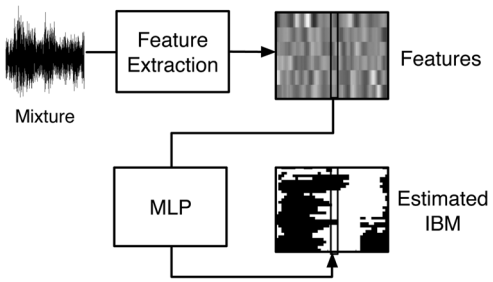

Fig. 1. Diagram of the feature evaluation framework.

designed to achieve the best separation performance. Additionally, we investigate auto-regressive moving average (ARMA) filtering as a post-processing technique to enhance feature robustness for further improving speech separation performance.

We should point out that a recent study has evaluated several features for classification-based speech separation [38]. Our study goes beyond [38] in several aspects. First, our evaluation is conducted on more challenging noisy mixtures using a different classifier (MLP instead of support vector machine). More importantly, features are chosen more systematically in our study, which results in a significantly more expansive list. Finally, while the study in [38] emphasizes feature combination, our study results in a new, effective feature which performs better than the complementary feature set suggested in [38].

This paper is organized as follows. Section II describes feature evaluation framework for classification-based speech separation. The existing features are described in Section III. We introduce the proposed MRCG feature in Section IV. Section V covers feature post-processing and feature combination. We present experimental results in Section VI. Section VII concludes the paper. A preliminary version of this paper is included in [6].

II. Evaluation Framework

In classification-based speech separation, the computational goal typically is to estimate the IBM that is created from premixed signals. The time-frequency representation of a cochleagram is frequently used to construct the IBM. In this study, we use a 32-channel cochleagram with 20 ms frame length and 10 ms frame shift. The local SNR criterion (LC) of the IBM is set to -10 dB to preserve enough speech information (see [9]). Note that, once a binary mask is computed, it can be used to synthesize a time-domain signal by weighting T-F unit signals in an appropriate way (see Chapter 1 of [36] for more details).

Fig. 1 shows the diagram of the evaluation system, which consists of the feature extraction component and the MLP classification component. All mixtures are sampled at 16 kHz. We extract acoustic features from an input signal at the frame level, which are sent to an MLP classifier for IBM estimation. We use a full-band input signal for feature extraction and one MLP for predicting a mask across all channels. In other words, the MLP is trained to predict a T-F mask frame by frame as opposed to sub-band classification in [38].

The features are evaluated based on the mask estimation quality. There are several criteria for measuring the quality of an estimated IBM. One straightforward criterion is to compute classification accuracy, where the percentage of correctly labeled T-F units is calculated for the whole mask. However, this criterion is agnostic to different classification errors. Recent work shows that the HIT–FA criterion well correlates with human intelligibility [22], where HIT refers to the percentage of correctly classified target-dominant T-F units and FA refers to false alarm or the percentage of wrongly classified interference-dominant T-F units. A good IBM estimate should have high HIT and low FA, which leads to high HIT–FA rate. We use both classification accuracy and HIT–FA rate in this study.

III. Existing Features

We evaluate an extensive list of existing acoustic features, consisting of widely used and promising robust speech recognition and separation features. Below we briefly describe a set of 16 such features, and more details can be found in the references.

A. Mel-Frequency Cepstral Coefficient (MFCC)

To compute MFCC, an input signal is divided into 20 ms frames with 10 ms frame shift. We apply a Hamming window to each frame and derive power spectrum using short-time Fourier transform. Then we convert power spectrum into mel scale. Finally, log compression and discrete cosine transform (DCT) are applied to compute 31-dimensional (31-D) MFCC.

B. Perceptual Linear Prediction (PLP)

PLP is designed to minimize the differences between speakers while keeping important formant structure [10]. To compute PLP, The power spectrum of an input signal is converted into bark scale, followed by loudness preemphasis and applying intensity loudness law. Then we derive linear prediction coefficients, which are then converted to cepstral coefficients. By using the 12th order linear prediction model, we end up with 13-D PLP.

C. Relative Spectral Transform PLP (RASTA-PLP)

RASTA-PLP introduces RASTA filtering to PLP [11]. To compute RASTA-PLP, the power spectrum of an input signal is wrapped to the bark scale. The resulting spectrum is log-compressed and filtered with the RASTA filter, which emphasizes the modulation frequencies that are relevant to human speech. The filtered log-spectrum is then expanded by an exponential function. Finally, we perform linear prediction analysis to derive 13-D RASTA-PLP.

D. Gammatone Frequency Cepstral Coefficient (GFCC)

To compute GFCC [33], [42], we pass an input signal through a 64-channel gammatone filterbank to derive sub-band signals. Each sub-band signal is decimated to 100 Hz, amounting to 10 ms frame shift. We then apply cubic root compression to the magnitude of the decimated signals and perform DCT to derive 31-D GFCC.

E. Gammatone Frequency Modulation Coefficient (GFMC)

To compute GFMC [26], we first follow the GFCC procedure to compute 31-D GFCC. Then we calculate the modulation

spectrum of each coefficient. The modulation spectrum corresponds to the Fourier transform of the temporal trajectory of each coefficient. We use 160 ms frame length and 10 ms frame shift to calculate the modulation spectrum. For each modulation spectrum, we calculate the energy for 2–16 Hz modulation frequencies, which are mostly relevant to speech signals [26]. Finally, we concatenate the energy calculated from each coefficient to form 31-D GFMC.

F. Gammatone Feature (GF)

We compute 64-D GF by following the GFCC procedure except that the DCT step is skipped.

G. Zero-Crossings with Peak-Amplitudes (ZCPA)

ZCPA is a speech recognition feature based on zero-crossings [21]. To compute ZCPA, an input signal is decomposed into sub-band signals by a 32-band gammatone filterbank. We divide each sub-band signal into 100 ms frames with 10 ms frame shift. For each frame, we calculate the intervals between every two upward zero-crossings. We classify each interval into 31 frequency bins where the frequency of an interval is the inverse of the interval. Then we identify the peak amplitude within each interval and add a nonlinear-compressed peak amplitude to the corresponding frequency bin. The frequency bins are accumulated across all sub-bands and form a histogram, i.e. 31-D ZCPA.

H. Relative Autocorrelation Sequence MFCC (RAS-MFCC)

RAS-MFCC is designed to suppress background noise by filtering in the autocorrelation domain [40]. To compute RAS-MFCC, we calculate one autocorrelation sequence for each frame of an input signal. A high pass filter is applied to the temporal trajectory of each dimension of autocorrelation sequences to suppress slow-varying components. The filtered autocorrelation sequences are treated as the input to the standard MFCC procedure to derive 31-D RAS-MFCC.

I. Autocorrelation Sequence MFCC (AC-MFCC)

AC-MFCC is also an autocorrelation-domain feature. AC-MFCC is designed to reduce the interference from background noise by discarding low-lag autocorrelation coefficients [32], by assuming that the effect of the noise is usually concentrated in low-lag autocorrelation coefficients. To compute AC-MFCC, an input signal is divided into frames where the autocorrelation of each frame is computed. We discard low-lag, i.e. less than 2 ms, autocorrelation coefficients. Hamming window is applied to high-lag autocorrelation coefficients and the corresponding magnitude spectrum is computed. The remaining steps follow the MFCC procedure to derive 31 cepstral coefficients.

J. Phase Autocorrelation MFCC (PAC-MFCC)

PAC-MFCC is an ASR feature similar to RAS-MFCC. PAC-MFCC computes the angle between a signal and its shifted version [17]. It is assumed that angle sequences are less variant than autocorrelation sequences in the presence of background noise. The standard MFCC procedure is applied to the resulting angle sequences to compute 31-D PAC-MFCC.

K. Power Normalized Cepstral Coefficients (PNCC)

PNCC is a recent ASR feature that utilizes medium-time processing to mitigate noise corruption and employ power-law compression instead of log compression in traditional features [19]. First, the power spectrum of an input signal is integrated using gammatone frequency integration. Then, based on medium-duration temporal analysis, we perform asymmetric filtering and temporal masking to subtract background noise. Finally we apply power-law nonlinearity and DCT to derive 31 coefficients.

L. Gabor Filterbank (GFB) Features

GFB is a recent feature designed for robust ASR by taking into account the spectrotemporal modulation frequencies [31]. To derive GFB, we compute the log mel-spectrum from an input signal. The spectrum is filtered by a Gabor filterbank which consists of 41 carefully designed Gabor filters. Representative channels of each filtered spectrum are selected and concatenated to form 311-D GFB.

M. Amplitude Modulation Spectrogram (AMS)

The AMS feature is a feature used in speech separation [22]. To compute AMS, the full-wave rectified envelope of an input signal is decimated by a factor of 4. As in [22], AMS features are extracted from 32-ms frames (frame shift is still 10 ms). We apply Hamming window and 256-point FFT. Finally, the 15-D feature is derived by integrating the FFT magnitudes using 15 triangular windows uniformly centered from 15.6 to 400 Hz.

N. Pitch-Based Features (PITCH)

Pitch-based features are used in a recent separation study [38]. These are T-F unit level features derived from pitch analysis. We calculate a cochleagram for an input signal and derive six features described in [38] (see also [14]) for each T-F unit. These features capture how likely a T-F unit is dominated by the target speech by utilizing periodicity and instantaneous frequency. In our classification-based speech separation, the ground truth pitch is used during training while the pitch estimated by a recently proposed robust pitch tracker, PEFAC [7], is used during testing.

O. Delta-Spectral Cepstral Coefficient (DSCC)

DSCC is an ASR feature very similar to MFCC except that a delta operation is applied to the spectrum [23]. To compute DSCC, we first follow the standard MFCC procedure to compute the mel-spectrum. Then a delta operation is applied to derive delta spectral features, whose histogram is normalized to give a Gaussian distribution. DCT is applied to compute 31 cepstral coefficients, based on which we further derive 31-D delta cepstral coefficients. Finally, we add traditional MFCC cepstral coefficients to form 93-D DSCC.

P. Suppression of Slowly-varying Components and the Falling Edge of The Power Envelope (SSF)

SSF has been designed to remove slowly-varying components to reduce noise interference and suppress the falling edge of power envelope in order to mitigate reverberation [20]. An input signal is divided into 50 ms medium-duration frames with 10 ms frame shift. The FFT of each frame is integrated across frequencies using gammatone weighting functions. Then we apply SSF processing to the resulting power spectrum. The SSF procedure produces an enhanced version of the original signal. We apply the MFCC procedure to the enhanced version to derive 31-D SSF.

IV. Multi-Resolution Cochleagram Feature

Besides the existing features, we propose a new acoustic feature called the Multi-Resolution Cochleagram (MRCG), which encodes multi-resolution power distributions in the time-frequency representation of a signal. We combine four cochleagrams at different resolutions to construct the MRCG feature. A high resolution cochleagram captures the local information while three low resolution cochleagrams capture spectrotemporal contexts at different scales.

A. Construction of MRCG

The construction of MRCG is based on the cochleagram representation, which is widely used in the CASA literature [36]. To compute the cochleagram, we first pass an input signal to a gammatone filter bank, where the impulse response of a particular gammatone filter is [30],

$$g_{f_c}(t) = t^{N-1}\exp[-2\pi t b(f_c)]cos(2\pi f_c t)u(t), \qquad (1)$$

where f_c denotes the center frequency, N the filter order, and $u(t)$ the step function. The function $b(f_c)$ decides the bandwidth given f_c. To imitate human auditory filters, the center frequencies f_c are uniformly spaced on the equivalent rectangular bandwidth (ERB) scale. The relation between $b(f_c)$ and f_c is shown in Equation (2).

$$b(f_c) = 1.019 * ERB(f_c) = 1.019 * 24.7 * (4.37 * f_c/1000 + 1). \qquad (2)$$

The bandwidth $b(f_c)$ increases as f_c increases, leading to higher resolutions at low frequencies and lower resolutions at high frequencies. After getting response signals from the gammatone filterbank, we divide each response signal into 20 ms frames with a 10 ms frame shift. We derive the cochleagram by computing the power of each frame at each channel [36].

Each T-F unit in the cochleagram contains only local information, which may not be sufficient for estimating the mask. To compensate for this, the MRCG feature provides contextual information by including the power distribution in the neighborhood of each T-F unit. The MRCG feature is similar to the GFB feature in the sense that both are designed to encode the spectrotemporal context systematically (see also [12], [29]).

The steps for computing MRCG are described as follows.

(1) Given an input mixture, compute the first 64-channel cochleagram, CG1. A log operation is applied to each T-F unit.

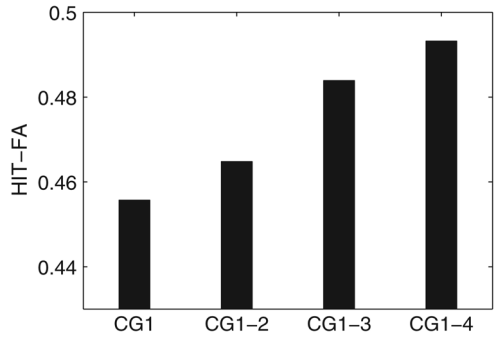

Fig. 2. Effects of adding contextual information for speech separation with −5 dB babble.

(2) Similarly, compute CG2 with the frame length of 200 ms and frame shift of 10 ms.
(3) CG3 is derived by averaging CG1 across a square window of 11 frequency channels and 11 time frames centered at a given T-F unit. If the window goes beyond the given cochleagram, the outside units take the value of zero (i.e. zero padding).
(4) CG4 is computed in a similar way to CG3, except that a 23×23 square window is used.
(5) Concatenate CG1-4 to obtain the MRCG feature, which has 64×4 dimensions for each time frame.

Note that, although the IBM is defined using a 32-channel cochleagram, features can be extracted from a different sized cochleagram (see Section II). We found that 64-channel features extracted in Step 1 perform a little better than 32-channel features. Also, using zero padding in Step 3 for outside T-F units leads to slightly better results than simply averaging the units inside a window.

B. Analysis of MRCG

In the MRCG feature, CG1 contains the local information embedded in a typical cochleagram while CG2-4 provide fine-grain and coarse-grain contexts. The parameters used in the construction of MRCG are decided experimentally as follows. First, the frame length of CG1 is chosen to match the frame length of the IBM. Then we fix CG1 and determine CG2 by expanding to different frame lengths to select the best length. Similarly, we decide the size of the averaging window for CG3, and then for CG4. After obtaining CG1-4, we find that adding more cochleagrams does not provide further performance improvements. Fig. 2 illustrates the effects of adding T-F contexts on the separation results. As shown in Fig. 2, adding CG2-4 consistently improves the results for babble noise at -5 dB SNR. Similar trends are observed for the other noises.

A visualization of the MRCG feature is given in Fig. 3, where the left plots features extracted from a babble mixture at -5 dB SNR and the right from the corresponding clean speech. As shown in Fig. 3, CG1 is the regular cochleagram, CG2 captures temporal context, CG3 encodes relatively small spectrotemporal context and CG4 encodes relatively large spectrotemporal context. The broad rationale behind MRCG is that a T-F unit is more likely to be speech-dominant if it resides in a cluster of many speech-dominant T-F units. In other words, a speech-dominant T-F unit not likely appears alone in a cochleagram.

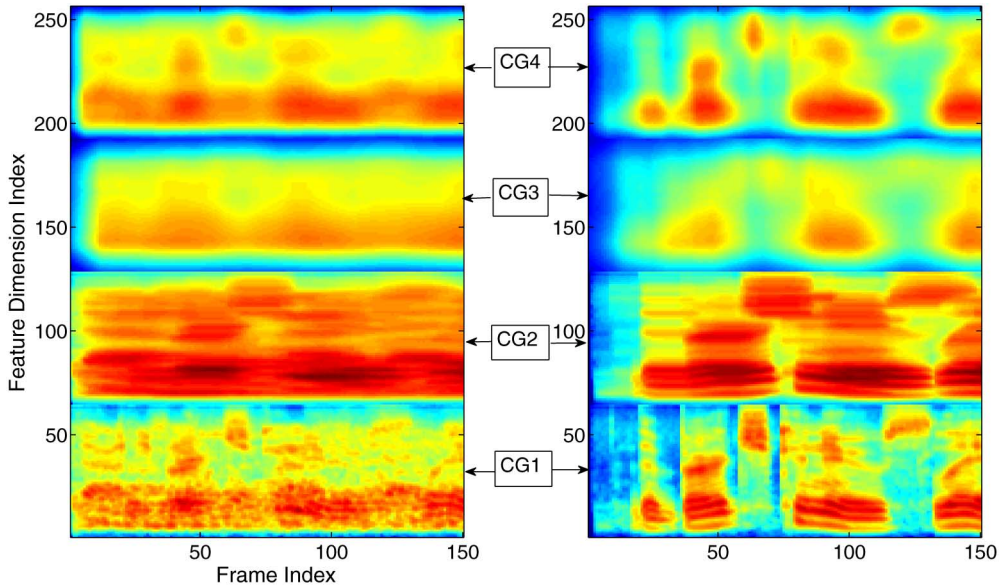

Fig. 3. (Color online) Visualization of the MRCG feature. Left side shows MRCG features extracted from a mixture, while the right side shows MRCG features extracted from premixed clean speech. In CG2-4, feature patterns of the mixture resemble the ones of clean speech to some extent, indicating the MRCG feature could partially retain spectrotemporal patterns of speech in the presence of noise.

V. Feature Post-Processing and Combination

A. Feature Post-Processing

In speech processing, delta (Δ) and double-delta ($\Delta\Delta$) features are widely used to capture temporal dynamics. Adding those features is a popular feature post-processing technique. For example, $\Delta + \Delta\Delta + \text{MFCC}$ yields better speech recognition results than MFCC alone. Recent research shows that Δ and $\Delta\Delta$ features also improve speech separation results [38]. In this study, we thus expand each feature by adding Δ and $\Delta\Delta$ features.

It has been suggested that applying ARMA filtering to mean variance normalized features improves speech recognition results [5]. The ARMA filter is defined below,

$$\begin{aligned}\bar{C}(m) =&[\bar{C}(m-M)+\cdots+\bar{C}(m-1)+C(m)\\ &+\cdots+C(m+M)]/(2M+1)\end{aligned} \tag{3}$$

where $C(m)$ denotes the feature vector at frame m, $\bar{C}(m)$ denotes the filtered feature vector at frame m and M denotes the order of the filter. The idea behind ARMA filtering is to smooth temporal trajectory of each feature dimension so that the interference of background noise is reduced. However, the effect of ARMA filtering in classification-based speech separation is unknown. In this study, we add ARMA filtering as an optional post-processing step and evaluate if it improves speech separation results.

B. Feature Combination

A recent study shows that a proper combination of features can lead to better performance in classification-based speech separation [38]. A straightforward way of finding complementary features is to try all combinations of features. However, the number of combinations is exponential with respect to the number of features. As in [38], we utilize group Lasso (least absolute shrinkage and selection operator) to quickly identify

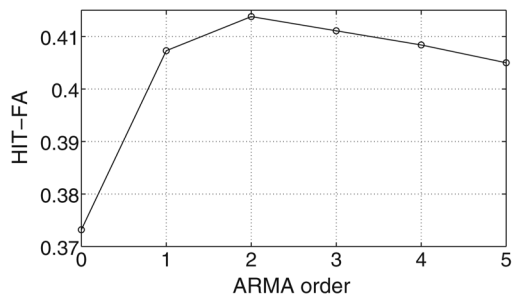

Fig. 4. Effect of the ARMA post-processing order for the PLP feature with babble noise at -5 dB SNR.

complementary features. The idea of group Lasso is to impose ℓ_1/ℓ_2 mixed norm regularization on logistic regression. It is known that ℓ_1/ℓ_2 regularization leads to sparsity between groups (i.e. feature types) [27]. Group Lasso solves the following optimization problem:

$$\begin{aligned}\hat{\beta}_\lambda = \arg\min_{\beta,\alpha} \sum_i \log\left(1+\exp\left(-y_i\left(\beta^T x_i+\alpha\right)\right)\right)\\ +\lambda\sum_{g=1}^{G}\left|\beta_{\mathcal{I}_g}\right|_2\end{aligned} \tag{4}$$

where x_i is an input feature vector, y_i is its label (taking value of 1 or -1), β denotes the response coefficients which we use to identify complementary groups, $\mathcal{I}_g$ denotes the index set of the gth group, $\|\cdot\|_2$ refers to ℓ_2 norm, and λ controls group sparsity. We minimize both the first term, which represents the classification error, and the second term, which imposes ℓ_1/ℓ_2 mixed norm regularization. The input to the logistic regression is the concatenation of all feature types where the training labels are provided by the IBM. The regression is carried out channel by channel. The resulting response coefficients are averaged across channels. The features that have relatively large responses are selected as the complementary features.

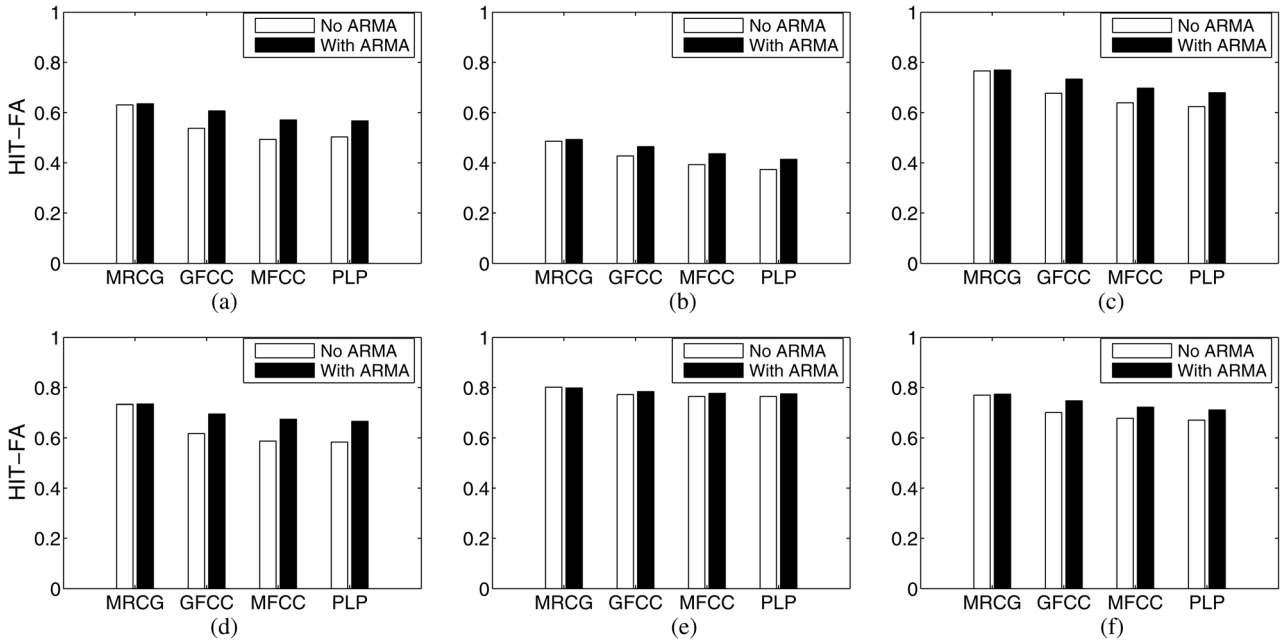

Fig. 5. Effects of ARMA filtering in terms of HIT–FA rate (a) Factory (b) Babble (c) Engine (d) Cockpit (e) Vehicle (f) Tank.

VI. Experimental Results

A. Experimental Setup

In our experiments, we create mixtures using the IEEE corpus recorded by a male speaker [16] and six types of nonstationary noise from the NOISEX corpus [34]. The noise types include factory floor noise (Factory), speech babble (Babble), jet cockpit noise (Cockpit), destroyer engine room noise (Engine), military vehicle noise (Vehicle), and tank noise (Tank). Each mixture is created from one IEEE sentence and one noise type at -5 dB SNR. To create the training set, we use 480 IEEE sentences and the first half of each noise. As for the test set, we use another 50 IEEE sentences and the second half of the noises. Using different parts of a nonstationary noise ensures that the noise segments used in the test set are different from those in the training set. We train and test on the same type of noise. An MLP with one hidden layer is used as the classifier for IBM estimation. The hidden layer includes 300 sigmoidal activation units. We set aside 50 mixtures from the training set as a cross validation set for early stopping.

B. Effect of ARMA Filtering

We first examine the effect of ARMA filtering, a feature post-processing technique, on every feature type. The only tunable parameter in the ARMA filter is the filter order. The experimental results show that 2nd order ($M = 2$) ARMA filtering improves the HIT–FA rate for most feature and noise types. For example, the effect of filter order for the PLP feature with babble noise is shown in Fig. 4, where one can see the HIT–FA rate peaks when $M = 2$, and is significantly better than without using ARMA ($M = 0$). In the following experiments, we set ARMA filter order to 2.

Fig. 5 shows the effects of ARMA filtering on MRCG, GFCC, MFCC and PLP in each noise condition. The MRCG feature does not benefit from ARMA filtering, likely because the averaging windows used in MRCG have already embodied spectrotemporal smoothing. On average we observe 4% improvement in HIT–FA due to ARMA filtering for all noise types.

C. Comparison among Individual Features

Due to its effectiveness, we apply ARMA filtering to all 16 feature types plus MRCG in our comparisons. For the 50 test sentences, the overall classification accuracy and the overall HIT$-$FA rate of each feature are shown in Table I and Table II, respectively, in decreasing order of average performance. In addition, Fig. 6 shows the median and interquartile range for the test sentences for the top four features from Tables I and II. The features can be roughly categorized into the following groups:

(1) Gammatone-domain features: MRCG, GF and GFCC.
(2) Autocorrelation-domain features: RAS-MFCC, PAC-MFCC and AC-MFCC.
(3) Modulation-domain features: GFMC, AMS, GFB, and RASTA-PLP.
(4) Linear prediction features: PLP.
(5) MFCC variants: MFCC and DSCC.
(6) Medium-time processing features: PNCC, SSF.
(7) Zero-crossing feature: ZCPA.
(8) Pitch-based feature: PITCH.

The results indicate that the gammatone-domain features (MRCG, GF, GFCC) perform better than other features. It is interesting to note that, although the modulation-domain feature GMFC is derived from GFCC, it does not perform as well as GFCC. Also interesting is that GFCC is a compact representation of GF, but the latter performs better than GFCC, probably because GF contains more information that can be exploited by the MLP classifier. MFCC, perhaps the most widely used feature, performs reasonably well when it is processed with an ARMA filter. Among the autocorrelation-domain features, RAS-MFCC performs the best and PAC-MFCC the worst. The

TABLE I
CLASSIFICATION ACCURACY (IN%) FOR SIX NOISES WITH ARMA POST-PROCESSING AT −5 dB. BOLDFACE INDICATES BEST RESULT

Feature \ Noise	Factory	Babble	Engine	Cockpit	Vehicle	Tank	Average
MRCG	**88.0**	**79.5**	**92.2**	**92.4**	**89.9**	**90.5**	**88.8**
GF	87.6	77.4	91.9	92.1	89.9	90.2	88.2
GFCC	87.7	78.3	91.3	91.9	89.2	89.7	88.0
DSCC	86.6	77.2	90.5	90.9	88.8	88.8	87.1
MFCC	86.5	77.5	90.2	91.1	88.8	88.6	87.1
PNCC	86.6	77.2	90.1	90.9	88.6	88.3	87.0
PLP	86.9	77.4	89.5	90.9	88.7	88.2	87.0
AC-MFCC	86.7	77.0	89.3	90.5	88.7	88.1	86.7
RAS-MFCC	86.9	76.9	89.4	90.9	87.8	88.1	86.7
GFB	86.3	74.5	89.3	90.9	87.6	87.6	86.0
ZCPA	85.4	75.2	89.6	90.5	87.4	87.7	86.0
SSF	85.7	75.6	89.0	89.5	88.2	87.4	85.9
RASTA-PLP	85.9	75.9	88.2	89.7	87.9	86.8	85.7
GFMC	84.1	74.3	87.5	89.1	83.5	83.7	83.7
PITCH	85.5	69.6	84.8	88.9	79.2	82.3	81.7
AMS	82.5	74.0	84.8	87.8	75.4	79.1	80.6
PAC-MFCC	77.9	69.8	78.1	81.1	70.8	67.9	74.3

TABLE II
HIT–FA (IN%) FOR SIX NOISE TYPES WITH ARMA POST-PROCESSING AT −5 dB, WHERE FA IS SHOWN IN PARENTHESES

Feature \ Noise	Factory	Babble	Engine	Cockpit	Vehicle	Tank	Average
MRCG	**63** (7)	**49** (13)	**77** (4)	**73** (4)	**80** (10)	**77** (6)	**70** (7)
GF	61 (7)	45 (15)	75 (4)	71 (3)	80 (10)	76 (6)	68 (8)
GFCC	61 (6)	46 (14)	73 (4)	70 (3)	78 (11)	74 (6)	67 (7)
DSCC	56 (7)	42 (14)	70 (5)	66 (3)	77 (11)	73 (6)	64 (8)
MFCC	57 (7)	43 (14)	69 (5)	67 (4)	77 (11)	72 (7)	64 (8)
PNCC	56 (6)	44 (14)	69 (5)	66 (4)	77 (11)	71 (7)	64 (8)
PLP	56 (6)	41 (12)	68 (5)	66 (4)	77 (11)	71 (7)	63 (8)
AC-MFCC	56 (6)	42 (14)	67 (5)	65 (4)	77 (11)	71 (7)	63 (8)
RAS-MFCC	57 (6)	41 (14)	68 (5)	66 (4)	76 (11)	71 (7)	63 (8)
GFB	57 (7)	41 (18)	67 (5)	66 (4)	75 (12)	70 (7)	63 (9)
ZCPA	55 (8)	40 (16)	68 (5)	65 (4)	75 (13)	70 (8)	62 (9)
SSF	54 (7)	39 (15)	67 (5)	60 (4)	76 (11)	69 (7)	61 (8)
RASTA-PLP	52 (6)	38 (15)	64 (5)	61 (4)	76 (12)	67 (7)	60 (8)
GFMC	48 (7)	35 (15)	61 (6)	60 (5)	67 (17)	59 (9)	55 (10)
PITCH	46 (3)	29 (22)	50 (5)	50 (2)	59 (16)	53 (7)	48 (9)
AMS	40 (6)	27 (9)	49 (5)	52 (4)	50 (31)	45 (11)	44 (11)
PAC-MFCC	17 (5)	11 (8)	30 (9)	29 (7)	40 (48)	21 (17)	25 (16)

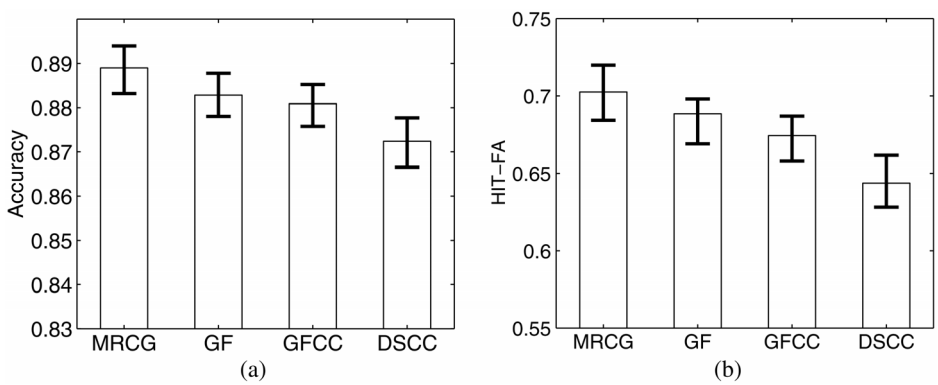

Fig. 6. Median value and interquartile range of 50 test sentences for average performance on six noises. Results are shown for top four features in terms of classification accuracy and HIT–FA rate (a) Accuracy (b) HIT–FA.

performance of the pitch-based feature is poor largely due to the difficulty in pitch estimation at −5 dB.

The proposed MRCG feature performs the best in terms of both classification accuracy and the HIT–FA rate. It is worth mentioning that GFB is also a multi-resolution feature where filters of different sizes are applied to the spectrogram. However, MRCG performs significantly better than GFB.

The differences among various features are more obvious when they are tested on the babble noise or the factory noise, which are more challenging than the other four noises. Observe

TABLE III
HIT–FA (IN%) DURING VOICED INTERVALS

Feature \ Noise	Factory	Babble	Engine	Cockpit	Vehicle	Tank	Average
MRCG	**67**	**46**	**78**	**76**	**73**	**77**	**70**
GF	66	43	76	75	73	76	68
GFCC	66	45	75	73	72	75	68
MFCC	61	41	71	71	71	72	65
RAS-MFCC	61	39	70	70	68	71	63

TABLE IV
HIT–FA (IN%) DURING UNVOICED INTERVALS

Feature \ Noise	Factory	Babble	Engine	Cockpit	Vehicle	Tank	Average
MRCG	**36**	**39**	**63**	**49**	**74**	**62**	**54**
GF	30	33	60	42	74	59	50
GFCC	28	31	56	40	73	55	47
MFCC	26	30	54	38	72	54	46
RAS-MFCC	25	30	50	38	68	51	44

that the relative performance of different features is mostly consistent from one noise to another.

In addition, we examine the performance of features separately during voiced intervals and unvoiced intervals. Unvoiced speech is more susceptible to background noise due to relatively weak energy [13]. Table III and Table IV show the performance of six relatively good features during voiced intervals and unvoiced intervals respectively. Again, the MRCG feature produces the best results during both voiced intervals and unvoiced intervals.

To further validate the relative performance of features, we also evaluate three top features with different classifier—a linear SVM—that performs IBM estimation channel by channel [39]. Note that the input feature vector to each channel SVM is the same across different frequency channels. The average SVM classification accuracy for the six noises is 84.3%, 83.3%, and 82.4%, for MRCG, GF, and GFCC, respectively. The corresponding HIT–FA results are 66%, 63%. and 62%, for MRCG, GF, and GFCC, respectively. These SVM classification results show the same order of feature effectiveness as with MLP classification.

D. Feature Combination Results

We apply group Lasso to select complementary features for each noise type. Each feature type is appended with Δ and $\Delta\Delta$ features, as mentioned in Section V-A. The group Lasso results for the cockpit noise are shown in Fig. 7. The average responses indicate discriminative power of a feature type. A good feature type is expected to show prominent responses. In Fig. 7, MRCG and PITCH have relatively high average responses while others have nearly no response, indicating that MRCG and PITCH are complementary. As for the other noise types, MRCG and PITCH are also identified by group Lasso as complementary features.

Table V and Table VI show the classification accuracy and the HIT–FA rate for the combined feature (MRCG concatenated with PITCH), respectively. When we use ground truth pitch for training and estimated pitch for testing, the combined feature performs worse than MRCG alone. This is mainly because pitch estimation at -5 dB SNR is very challenging and the estimated pitch tends to be very different from the ground truth one. If we use ground truth pitch in both training and testing, the combined feature performs better than MRCG alone, especially for the factory and babble noise. If we use estimated pitch in both training and testing, the combined feature performs almost the same as MRCG alone.

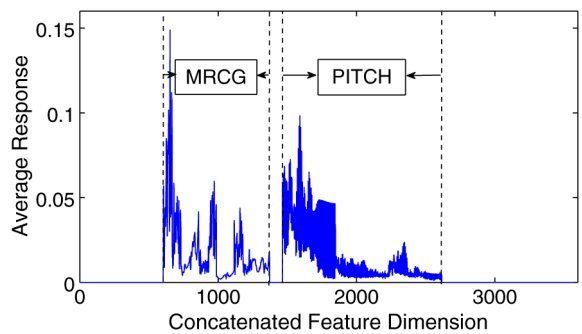

Fig. 7. Average magnitudes of regression coefficients resulted from group Lasso for the cockpit noise.

E. Comparison between MRCG and a Complementary Feature Set

In [38], it is found that AMS, RASTA-PLP, and MFCC form a complementary feature set and their combination outperforms each individual feature alone. Now we compare this complementary feature set and the MRCG feature for the aforementioned six noises at -5 dB SNR. As shown in Fig. 8, MRCG alone outperforms AMS + RASTA-PLP + MFCC. Such improvement mainly comes from the contextual information encoded in MRCG, which is important for separation in very low SNR conditions.

VII. DISCUSSION

In this study, we have evaluated an extensive list of acoustic features specifically for the classification-based speech separation at the very low SNR level of -5 dB—a condition where

TABLE V
CLASSIFICATION ACCURACY (IN%) OF COMBINED FEATURE WITH ARMA POST-PROCESSING AT -5 dB

Feature \ Noise	Factory	Babble	Engine	Cockpit	Vehicle	Tank
MRCG	88.0	79.5	92.2	92.4	89.9	90.5
MRCG + PITCH (Estimated)	87.1	74.6	90.7	91.1	89.1	88.5
MRCG + PITCH (True)	90.8	85.7	92.3	93.2	90.5	90.7

TABLE VI
HIT–FA (IN%) OF COMBINED FEATURE WITH ARMA POST-PROCESSING AT -5 dB

Feature \ Noise	Factory	Babble	Engine	Cockpit	Vehicle	Tank
MRCG	63	49	77	73	80	77
MRCG + PITCH (Estimated)	53	40	71	63	78	71
MRCG + PITCH (True)	70	64	77	76	81	78

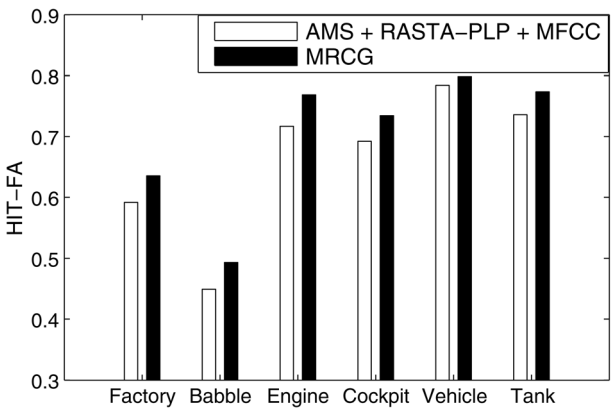

Fig. 8. Comparison of a complementary feature set (AMS + RASTA-PLP + MFCC) and the MRCG feature in terms of HIT–FA.

speech intelligibility is a main concern. In terms of classification accuracy and HIT–FA, we have shown that the gammatone-domain features (including GF, GFCC, MRCG) perform better than other features. The modulation-domain features (including GFMC and AMS) perform worse than most of the features likely because they do not deal with strong nonstationary noises well.

In addition, we have proposed a new feature, MRCG, which captures both local information and spectrotemporal contexts at different scales. The MRCG feature performs the best among the evaluated features. A closer look reveals that MRCG consistently produces the best results during both voiced and unvoiced intervals.

We have explored the effect of ARMA post-processing and found that the second order ARMA filtering improves most of the evaluated features by smoothing the temporal trajectories of feature dimensions. By employing group Lasso, we find that the MRCG feature and the pitch-based features form the best feature combination. Experimental results show that this combination yields the best performance if ground truth pitch is used. However, pitch estimation at -5 dB SNR is very difficult, and hence this insight of feature complementarity is not very useful unless pitch estimation improves substantially in very low SNR conditions. Our systematic study results in a clear recommendation: the simple MRCG feature without ARMA filtering should be considered as a benchmark in future speech separation studies, particularly at low SNR levels where human speech intelligibility is less than perfect.

It is noteworthy that PITCH and AMS features are among the first used in classification-based speech separation [18], [22]; a subsequent study combines these two [8]. Our investigation demonstrates that these are among the worst features for speech separation.

Features are of foundational importance for supervised separation. As embodied by the popularity of MFCC, progress in uncovering new and effective features often lifts performance for a variety of tasks. Another example is GFCC which was first introduced for robust speaker identification [33] but has since been shown to be effective for robust ASR [2] and speech separation in [38] and here. Indeed a recent study found that MRCG outperforms a combination of 11 commonly used features for voice activity detection (VAD) [41]. Given the relationship between speech separation and robust ASR, we conjecture that MRCG is an effective feature for ASR in very noisy conditions. This conjecture obviously remains to be verified in future study.

Finally we emphasize that the focus of this study is on features, not classifiers. The MLP with one hidden layer unlikely represents the state-of-the-art in supervised speech separation, and deep neural networks (DNNs) with multiple hidden layers likely perform better [39]. Producing the best performing speech separation system is not the direct objective of this study, and such a system would require both effective features and effective classifiers. With that said, it is worth noting that the superior VAD performance of MRCG shown in [41] is consistently demonstrated with different DNN classifiers. It will be interesting to study how features like MRCG interact with DNN classifiers for optimal classification.

REFERENCES

[1] M. Ahmadi, V. L. Gross, and D. G. Sinex, "Perceptual learning for speech in noise after application of binary time-frequency masks," *J. Acoust. Soc. Amer.*, vol. 133, pp. 1687–1692, 2013.

[2] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Comput. Speech Lang.*, vol. 27, pp. 621–633, 2013.
[3] A. S. Bregman, "Auditory scene analysis," in *The Perceptual Organization of Sound*. Cambridge, MA, USA: MIT Press, 1994.
[4] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal timefrequency segregation," *J. Acoust. Soc. Amer.*, vol. 120, pp. 4007–4018, 2006.
[5] C. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 257–270, Jan. 2007.
[6] J. Chen, Y. Wang, and D. L. Wang, "A feature study for classification-based speech separation at very low signal-to-noise ratio," in *Proc. ICASSP*, 2014, pp. 7039–7043.
[7] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Proc. Euro. Sig. Process. Conf.*, 2011, pp. 451–455.
[8] K. Han and D. L. Wang, "A classification based approach to speech segregation," *J. Acoust. Soc. Amer.*, vol. 132, pp. 3475–3483, 2012.
[9] E. W. Healy, S. E. Yoho, Y. Wang, and D. L. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Amer.*, vol. 134, pp. 3029–3038, 2013.
[10] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, 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] G. Hu and D. L. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 396–405, Feb. 2007.
[13] G. Hu and D. L. Wang, "Segregation of unvoiced speech from nonspeech interference," *J. Acoust. Soc. Amer.*, vol. 124, pp. 1306–1319, 2008.
[14] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2067–2079, Nov. 2010.
[15] Y. Hu and P. C. Loizou, "A comparative intelligibility study of singlemicrophone noise reduction algorithms," *J. Acoust. Soc. Amer.*, vol. 122, pp. 1777–1786, 2007.
[16] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.
[17] S. Ikbal, H. Misra, and H. Bourlard, "Phase autocorrelation (PAC) derived robust speech features," in *Proc. ICASSP*, 2003, pp. 133–136.
[18] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 625–638, May 2009.
[19] C. Kim and R. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. ICASSP*, 2012, pp. 4101–4104.
[20] C. Kim and R. M. Stern, "Nonlinear enhancement of onset for robust speech recognition," in *Proc. Interspeech*, 2010, pp. 2058–2061.
[21] D.-S. Kim, S.-Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 55–69, Jan. 1999.
[22] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 126, pp. 1486–1494, 2009.
[23] K. Kumar, C. Kim, and R. M. Stern, "Delta-spectral cepstral coefficients for robust speech recognition," in *Proc. ICASSP*, 2011, pp. 4784–4787.
[24] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Amer.*, vol. 123, pp. 1673–1682, 2008.
[25] P. C. Loizou, "Speech enhancement," in *theory and practice*. Boca Raton, FL, USA: CRC, 2007.
[26] H. K. Maganti and M. Matassoni, "An auditory based modulation spectral feature for reverberant speech recognition," in *Proc. Interspeech*, 2010, pp. 570–573.
[27] L. Meier, S. V. D. Geer, and P. Bühlmann, "The group lasso for logistic regression," *J. Roy. Stat. Soc. Ser. B*, vol. 70, pp. 53–71, 2008.
[28] B. C. Moore, *Cochlear hearing loss: Physiological, psychological and technical issues*. Sussex, U.K.: Wiley, 2007.
[29] S. K. Nemala, K. Patil, and M. Elhilali, "A multistream feature framework based on bandpass modulation filtering for robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 416–426, Feb. 2013.
[30] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," Applied Psychology Unit Rep. 2341, 1988.
[31] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 131, pp. 4134–4151, 2012.
[32] B. J. Shannon and K. K. Paliwal, "Feature extraction from higher lag autocorrelation coefficients for robust speech recognition," *Speech Commun.*, vol. 48, pp. 1458–1485, 2006.
[33] Y. Shao and D. L. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Proc. ICASSP*, 2008, pp. 1589–1592.
[34] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, 1993.
[35] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, P. Divenyi, Ed. Boston, MA, USA: Kluwer, 2005, pp. 181–197.
[36] D. L. Wang and G. J. Brown, "Computational auditory scene analysis," in *Principles, algorithms and applications.*. Hoboken, NJ, USA: Wiley-IEEE Press, 2006.
[37] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Amer.*, vol. 125, pp. 2336–2347, 2009.
[38] Y. Wang, K. Han, and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 270–279, Feb. 2013.
[39] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.
[40] K. Yuo and H. Wang, "Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences," *Speech Commun.*, vol. 28, pp. 13–24, 1999.
[41] X.-L. Zhang and D. L. Wang, "Boosted deep neural networks and multiresolution cochleagram features for voice activity detection," in *Proc. Interspeech*, 2014, pp. 1534–1538.
[42] X. Zhao, Y. Shao, and D. L. Wang, "CASA-based robust speaker identification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1608–1616, Jul. 2012.

Jitong Chen received his B.E. degree in information security from Northeastern University, China, in 2011. He is currently pursuing his Ph.D. degree at The Ohio State University. He is interested in machine learning, speech separation and signal processing.

Yuxuan Wang, photograph and biography not provided at the time of publication.

DeLiang Wang, photograph and biography not provided at the time of publication.