Features for Masking-Based Monaural Speech Separation in Reverberant Conditions

Masood Delfarah, *Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Monaural speech separation is a fundamental problem in speech and signal processing. This problem can be approached from a supervised learning perspective by predicting an ideal time–frequency mask from features of noisy speech. In reverberant conditions at low signal-to-noise ratios (SNRs), accurate mask prediction is challenging and can benefit from effective features. In this paper, we investigate an extensive set of acoustic–phonetic features extracted in adverse conditions. Deep neural networks are used as the learning machine, and separation performance is evaluated using standard objective speech intelligibility metrics. Separation performance is systematically evaluated in both nonspeech and speech interference, in a variety of SNRs, reverberation times, and direct-to-reverberant energy ratios. Considerable performance improvement is observed by using contextual information, likely due to temporal effects of room reverberation. In addition, we construct feature combination sets using a sequential floating forward selection algorithm, and combined features outperform individual ones. We also find that optimal feature sets in anechoic conditions are different from those in reverberant conditions.**

***Index Terms*—Deep neural networks, feature combination, monaural speech separation, room reverberation, speech intelligibility.**

I. Introduction

MONAURAL speech separation refers to separating a target speaker from background interference from single-microphone recordings. In this paper, we perform a systematic feature study for monaural speech separation in noisy and reverberant conditions, with the goal of improving speech intelligibility in human listeners. In real world environments, the received speech signal is usually distorted by both background noise and room reverberation. The reflections from the surfaces in a room smear the structure of sound and weaken the segregation cues. Perceptual studies report a significant loss of speech intelligibility for human listeners, especially those with hearing impairment, when exposed to both background noise and room reverberation [6], [10]. On the other hand, speech separation has numerous applications, including hearing-aid design and mobile communication.

Speech separation has been studied for decades, and several approaches have been proposed. Microphone-array methods perform spatial filtering [2]. This approach is effective only when the sound sources are well-separated in space. Speech enhancement methods [23] such as spectral subtraction are applicable to monaural recordings. To be effective, these methods need to make restrictive assumptions, such as noise stationarity. Computational auditory scene analysis (CASA) [36] utilizes perceptual principles to perform sound separation. However, the detection of grouping cues (such as harmonicity and onset) from the noisy input is difficult, and limits CASA performance.

In recent years, supervised speech separation, particularly with the use of deep neural networks (DNNs), has elevated separation performance to a new level. The first study to introduce DNN for speech separation was conducted by Wang and Wang [39]. They used DNNs to perform feature learning and predict the ideal binary mask (IBM) in the time-frequency (T-F) domain. This DNN-based binary classifier produced the first substantial speech intelligibility improvements for hearing-impaired, as well as normal-hearing listeners in anechoic conditions [9]. Since then, many studies make use of DNN for speech separation. An examination of various training targets demonstrates the advantage of predicting the ideal ratio mask (IRM) over IBM prediction [38]. Xu *et al.* [41], [42] train a DNN to map from the spectral magnitudes of noisy speech to those of clean speech. At about the same time, Han *et al.* [8] train a DNN to learn a spectral mapping function from spectral features of reverberant and noisy speech to clean speech to enhance noisy speech. Huang *et al.* [13] use both DNNs and recurrent neural networks (RNNs) to separate cochannel (i.e. two-talker) speech. They predict the IRM for both target and interfering speech. In a very recent study, Zhang and Wang [45] use a stack of DNNs to predict an ideal mask for target speech in two-talker mixtures, and demonstrate that masking-based (i.e. predicting an ideal mask) separation tends to be more preferable than mapping-based separation (see also [38]).

A next step towards solving the speech separation or cocktail party problem would be to separate the target speech in reverberant conditions. A recent study that applies DNN-based spectral mapping does not lead to a consistent speech intelligibility improvement [47]. In addition, to our knowledge, cochannel speech separation in reverberant conditions has not been explored in the supervised learning framework.

Manuscript received October 18, 2016; revised February 12, 2017 and March 20, 2017; accepted March 20, 2017. Date of publication March 27, 2017; date of current version April 7, 2017. This work was supported in part by the Air Force Office of Scientific Research under Grant FA9550-12-1-0130 and the Ohio Supercomputer Center. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Richard Christian Hendriks. *(Corresponding author: Masood Delfarah.)*

The authors are with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: delfarah.1@osu.edu; 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.2017.2687829

Broadly speaking, supervised speech separation consists of two main components: features and learning machines. While DNNs are powerful learning machines, acoustic features have to be informative and sufficiently discriminative. In this paper, we systematically examine an extensive set of monaural features for supervised speech separation in reverberant mixtures. We investigate features for reverberant speech separation in both speech and nonspeech interference, and in both seen and unseen noise conditions. As individual features reveal certain characteristics of the speech signal, it is important to leverage a set of features. This paper further addresses feature selection problem.

To our knowledge, no feature study has been conducted previously for reverberant speech separation. Two earlier feature studies [3], [37] were done in anechoic conditions and the extent to which their conclusions apply to reverberant conditions is unclear. Moreover, these studies only consider nonspeech noise, and interfering speech is not evaluated. As we will see in this paper, features can behave differently in different room and interference environments.

As noted earlier, feature extraction and learning machine are two integral components of a supervised learning system. Therefore, the choice of the learning machine will affect separation results. To isolate the effects of features, we use a fixed DNN as the supervised learning machine, which is the most commonly used learning algorithm in supervised speech separation.

A preliminary version of this paper is included in [4]. The current work goes substantially beyond the preliminary work. The present study includes cochannel conditions. In addition, new features and many more conditions are evaluated in this paper.

The rest of the paper is organized as follows. In Section II we describe the feature evaluation framework. Section III describes the features to be investigated. The experimental setup and effects of contextual information are explained in Section IV. Performance of each individual feature is evaluated in Section V, and feature combination is studied in Section VI. We offer concluding remarks in Section VII.

II. Evaluation Framework

As noted earlier, in masking-based speech separation, IRM prediction is more preferable than IBM prediction. The IRM is defined on the basis of premixed spectrograms. Given a speech mixture $y(t)$:

$$y(t) = s(t) + n(t) \tag{1}$$

where $s(t)$ and $n(t)$ are premixed reverberant target and interfering signals sampled at 16 kHz, we divide $s(t)$ and $n(t)$ into 20 ms time frames with 10 ms overlap, then apply a Hamming window. Short-time Fourier transformation (STFT) is applied, resulting in 161 frequency bins (or channels). At time frame m and frequency channel c, $S(m,c)$ and $N(m,c)$ represent the magnitude spectrograms of $s(t)$ and $n(t)$, respectively. The IRM is given as follows [38]:

$$IRM(m,c) = \sqrt{\frac{S^2(m,c)}{S^2(m,c) + N^2(m,c)}} \tag{2}$$

In this study, we aim to predict the IRM with the spectrogram of the reverberant target speech treated as the signal (see also [17]). Even though dereverberation is not considered, the IRM defined in Eq. (2) is expected to produce highly intelligible speech as human speech intelligibility does not drop significantly in room reverberation without background noise [25].

For a fair comparison among various features, a fixed DNN is used for IRM estimation in our experiments. The DNN has 2 hidden layers, and there are 512 units in each hidden layer. The output layer has 161 units, corresponding to a frame of the IRM. Rectified linear unit (ReLU) [26] and sigmoid function are used as the activation functions for the hidden and output units, respectively. Dropout rate of 0.2 is used for hidden units for regularization purposes [32]. We choose this relatively straightforward DNN architecture as our focus is on features, not learning machines.

DNN training minimizes the following mean square error loss function:

$$\mathcal{L}(\mathbf{IRM}(m,:), \mathbf{F}(m); \mathbf{\Theta}) = \frac{1}{C}\sum_{c=1}^{C}(IRM(m,c) - g_c(\mathbf{F}(m)))^2 \tag{3}$$

where $\mathbf{F}(.)$ denotes the feature vector, $\mathbf{\Theta}$ corresponds to the model parameters, $C = 161$ is the number of frequency channels, and $g_c(.)$ is the value of the $c\mathrm{th}$ output unit.

We use adaptive stochastic gradient descent [21] in the backpropagation algorithm for DNN training. Mini-batches of size 1000 are used in network training. Learning rate is initialized to 0.001 and decayed by a factor 0.9 every epoch. The algorithm is run for 50 epochs. We set aside 5% of the training data for cross validation. The set of the 50 DNN parameters with the least MSE on the validation set is chosen as the optimal parameters during the test phase.

Feature normalization is shown to facilitate the backpropagation algorithm [22]. We calculate the normalized feature $\mathbf{F}(.)$:

$$F(m,d) = \frac{\widetilde{F}(m,d) - \mu_d}{\sigma_d} \tag{4}$$

where d indexes a feature element, $\widetilde{\boldsymbol{F}}(.)$ is an extracted training or test feature, and μ_d and σ_d represent the mean and standard deviation of the $d\mathrm{th}$ feature element of the entire training set.

Fig. 1 shows an overview of our evaluation framework. Acoustic features are first extracted from the noisy and reverberant mixture, frame by frame. The features are normalized and passed through the trained DNN. The output of the DNN is the estimated IRM. The mixture spectrogram is pointwise multiplied by the DNN output, to get the estimated target spectrogram. Noisy-reverberant signal phase and the estimated magnitude are used together to resynthesize the estimated clean target signal.

III. Features

The features chosen to be studied in this paper have been successfully applied in different areas of speech processing. In this section, we briefly describe each of the features.

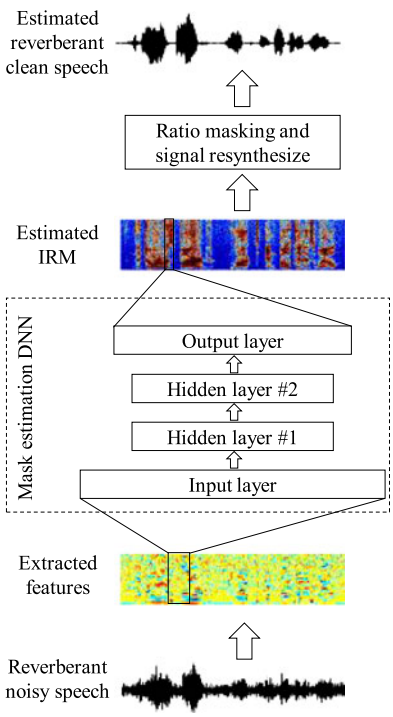

Fig. 1. Diagram of the proposed evaluation framework.

A. Waveform Signal (WAV)

Waveform signal can be directly used without any feature extraction, as done by Sainath *et al.* [28] in automatic speech recognition (ASR). To examine this feature in our framework, we simply use 320 signal samples with 160 sample shift, corresponding to 20 ms time frames with a 10-ms frame shift.

B. Gammatone Frequency Feature (GF)

The mixture signal is passed through a 64-channel gammatone filterbank [36]. To form the gammatone frequency feature, a cubic root operation is applied to the overall energy in each T-F unit, corresponding to 20 ms of each subband signal with a 10-ms frame shift.

C. Gammatone Frequency Cepstral Coefficients (GFCC)

GFCC is a feature designed for automatic speaker identification (SID). We use the first 31 coefficients produced from the discrete cosine transform (DCT) applied to the GF feature to derive the GFCC feature [31].

D. Multiresolution Cochleagram (MRCG)

Chen *et al.* [3] designed the MRCG feature for speech separation in anechoic conditions. This feature benefits from both local and contextual information. Cochleagrams with frame lengths of 20 and 200 ms are computed from the responses of a 64-channel gammatone filterbank. The frame shift in both cochleagrams is 10 ms. Then, a log operation is applied to the two cochleagrams to form CG1 and CG2, with CG1 corresponding to the 20 ms frame length. CG3 is calculated by averaging CG1 across a square window centered at a given T-F unit and window size of 11×11. In a similar way, CG4 is computed from CG2 with the window size of 23×23. Finally, CG1-4 are concatenated to form the MRCG feature.

E. Gammatone Frequency Modulation Coefficients (GFMC)

To mitigate the sensitivity of an ASR system to background noise and reverberation, the GFMC relies on long-term modulation spectrum [24]. The mixture signal first undergoes preemphasis, and then the GFCC is computed. Since human auditory system is most sensitive to the modulation frequencies around 4 Hz, the modulation frequency components concentrated in the range of 2–16 Hz at each dimension of the GFCC is calculated to yield the GFMC feature.

F. Pitch-Based Feature (PITCH)

Pitch is an important cue for auditory scene analysis and has been incorporated in many speech separation studies (see e.g. [46]). To calculate the pitch-based feature, we pass the noisy signal through a 64-channel gammatone filterbank. Then the PEFAC pitch tracking algorithm [7] is applied to each subband signal. With detected pitch, we then extract 6 dimensional features as described in [37]. Finally, we concatenate the features from all 64 channels.

G. Log-Magnitude Spectral Feature (LOG-MAG)

The LOG-MAG feature is computed from the spectrogram of noisy speech. Specifically, a log operation is applied to the magnitude responses of the STFT.

H. Perceptual Linear Prediction Feature (PLP)

The goal of PLP is to suppress the speaker-dependent details in the spectrum [11]. To compute PLP, the power spectrum is converted to the bark scale, and then filtered by the critical-band masking curve, and downsampled. The downsampled spectrum is preemphasized according to the equal-loudness curve, and compressed by an intensity-loudness power law (a cubic root operation). Inverse discrete Fourier transform (IDFT) is applied to the spectrum. The resulting cepstrum is used to solve the autoregressive coefficients of a twelfth order all-pole model, which are then transformed to the cepstral coefficients of the model.

I. Relative Spectral Transform PLP Feature (RASTA-PLP)

Hermansky *et al.* [12] add RASTA filtering to PLP to suppress slowly changing or steady-state factors in noisy speech. RASTA is a filter designed to attenuate background noise by suppressing the high frequency components in the spectrum and mitigate reverberation by suppressing the low frequency components of the spectrum. To compute the RASTA-PLP feature, critical-band power spectrum is computed as in the PLP processing. Then, the magnitude spectrum is log-compressed. RASTA filtering is performed on the log-spectrum, and then the filtered log-spectrum is decompressed. Finally, the cepstral coefficients are calculated from the linear prediction analysis, as in the PLP feature.

J. Amplitude Modulation Spectrogram (AMS)

To compute this modulation feature, we decimate the full-wave rectified envelope of the noisy signal by a factor of 4. Then, the signal is segmented into 32 ms frames, with 10 ms frame shift. The signal in each frame is windowed by a Hann function, and a 256-point FFT is applied. The modulation responses are multiplied by 15 triangular-shaped windows that are uniformly centered in the range of 15.6 and 400 Hz. The resulting 15 responses form the AMS feature [20].

K. Gabor Filterbank Feature (GFB)

Each of the subband signals in the log-mel-spectrogram of the mixture signal is processed with a bank of 41 spectro-temporal Gabor filters [29]. Then a subset of the channels is systematically selected to reduce the correlations among the feature components.

L. Mel-Frequency Cepstral Coefficients (MFCC)

MFCC is a widely used feature in speech processing. To compute MFCC, the spectrogram of the input signal is calculated. Then, the power spectrum is converted to the mel scale and log-compressed. Finally, we apply DCT and the first 31 cepstral coefficients represent the MFCC feature.

M. Log-Mel Filterbank Feature (LOG-MEL)

The LOG-MEL feature has been widely used in ASR as well as in speech separation [13]. The spectrogram of the mixture signal is processed by a 40-channel mel filterbank. A log operation results in the LOG-MEL feature.

N. Relative Autocorrelation Sequence MFCC (RAS-MFCC)

To provide a noise-robust feature, RAS-MFCC computes autocorrelation sequences in each time frame [44]. Then, a high-pass filter is applied. The filtered sequences are fed to the process for MFCC extraction, yielding the RAS-MFCC feature.

O. Phase Autocorrelation MFCC (PAC-MFCC)

PAC-MFCC is based on the phase trajectory of the signal over time. To compute PAC-MFCC [16], we apply the MFCC procedure to the phase angle between the noisy signal and its shifted versions.

P. Autocorrelation MFCC (AC-MFCC)

The main difference between AC-MFCC and RAS-MFCC is that AC-MFCC applies MFCC processing only to high-lag autocorrelation sequences. This feature discards the lower-lag autocorrelation coefficients of speech signal because they are often corrupted in the presence of background noise. Specifically, autocorrelation sequences are computed in each time frame, and coefficients corresponding to the lags less than 2 ms are discarded. The high-lag autocorrelation sequences are Hamming windowed, and finally, undergo the MFCC procedure [30].

TABLE I
LIST OF FEATURES EVALUATED

Feature	Dimension	Frame size (ms)	Extraction time (ms/frame)
AC-MFCC	31	20	2.625
AMS	15	32	0.160
GFB	311	25	1.592
GF	64	20	12.768
GFCC	31	20	13.192
GFMC	31	20	15.234
LOG-MAG	161	20	0.048
LOG-MEL	40	20	0.027
MFCC	31	20	0.030
MRCG	256	420	13.475
PAC-MFCC	31	20	0.086
PITCH	384	10	76.337
PLP	13	20	0.282
PNCC	13	25.6	11.993
RAS-MFCC	31	20	2.332
RASTA-PLP	13	20	0.324
SSF-I	31	50	1.487
SSF-II	31	50	1.480
WAV	320	20	0.000

Q. Power-Normalized Cepstral Coefficients (PNCC)

Kim and Stern [19] proposed PNCC, a modification to MFCC in order to achieve more reverberation and noise robustness. To compute the PNCC feature, magnitude spectrum is integrated by a 40-channel gammatone filterbank. Then asymmetric noise suppression procedure detects a lower envelope of the filtered spectrum, as the noise floor. This lower envelope is then utilized to perform temporal masking on the noisy spectrum. The masked spectrum is compressed by fifteenth-root operation and finally DCT yields in the PNCC feature.

R. Suppression of Slowly Varying Components and the Falling Edge of the Power Envelope (SSF-I and SSF-II)

SSF is another variation of MFCC, which is designed to suppress noise and reduce the mismatch between the anechoic and reverberant signal. To calculate SSF, first, the signal is preemphasized [18]. Then, the magnitude spectrum with the frame length of 50 ms and frame shift of 10 ms is calculated. As in PNCC, gammatone frequency integration is performed on the spectrum. The spectrum is reshaped with the coefficients calculated by the SSF processing. Kim and Stern [18] introduce one type of SSF that has noise-robustness (SSF-I), and an alternative type that performs better in reverberant conditions (SSF-II).

As noted, we use publicly available programs to extract GFB[1], PNCC, SSF-I, SSF-II[2], GF, GFCC, and MRCG[3]. Feature extraction for PLP, RASTA-PLP, and MFCC is performed using the RASTAMAT toolbox[4]. Table I summarizes the features described above, including computational costs and feature dimensions per time frame, and extraction times are averaged

[1]https://github.com/m-r-s/reference-feature-extraction
[2]http://www.cs.cmu.edu/~chanwook/MyAlgorithms
[3]http://web.cse.ohio-state.edu/pnl/software.html
[4]http://labrosa.ee.columbia.edu/matlab/rastamat

from 1000 frames and obtained on a Dell OptiPlex 780 PC with a quad-core processor at 2.66 GHz and 8 GB RAM.

IV. Experimental Setup and Contextual Windows

As the target utterances, we use IEEE sentences [15] uttered by a male speaker. Out of a total of 720 sentences, 400 sentences are used in generating the training and development set, and the rest are reserved for testing. For interference, we use Tank, Cockpit, and Factory noises from NOISEX [35], DWASHING, DLIVING, PSTATION, and TCAR noises from the DEMAND corpus [34], as well as the IEEE sentences uttered by a female speaker and a different male speaker. To mimic speech babble in real environments, we generate a 16-talker babble noise with a symmetric placement of 8 female and 8 male speakers around a virtual microphone, where each speaker is fixed at a 2 m distance to the microphone. The 16 speakers for the babble are randomly picked from the TIMIT corpus [5]. Training and test interference signals utilized in the matched noise case are Cockpit, DLIVING, DWASHING, Tank, and babble; the first half of each noise is used in training and the second half in testing. For the test mixtures in the unmatched case we use only the first noises, while Factory, PSTATION, and TCAR are used in the training mixtures. Finally, in the cochannel case the interference is a male or a female speaker.

The image method [1] is a commonly used technique to simulate reverberant room conditions. We use a room impulse response (RIR) generator[5] to produce reverberant signals at four different reverberation times (T_{60}) of 0.0, 0.3, 0.6, and 0.9 s, where $T_{60} = 0.0$ s corresponds to the anechoic case. The room dimension is chosen to be 10 m $\times$ 9 m $\times$ 8 m, and the microphone is fixed at (3, 4, 1.5) m. The target speaker is placed at a random position on the spheres with the radius of 1, 2, 4, or 8 m centered at the microphone. Note that the speaker, interferer, and microphone may have different elevations in the room.

Among the nonspeech noises in our study, Factory, PSTATION, and DLIVING noises are recorded in reverberant space. Accordingly, we directly mix them with the reverberant target speech signals. For the other noises as well as the competing speaker, we randomly place them in the same simulated reverberant room as the target speaker.

To generate training data, a target utterance is mixed with one interfering signal, at each of -9, -6, -3, 0, and 3 dB SNRs. In our experiments, SNR calculation is based on the energy ratio of target and interference signals without silence removal, since the signals are dense enough. Overall, we generate reverberant training mixtures in 10 (interference) $\times$ 5 (SNR) $\times$ 3 (T_{60}) $\times$ 4 (microphone-target distance) $= 600$ different conditions, and anechoic mixtures in 10 (interference) $\times$ 5 (SNR) $= 50$ conditions. For each of these conditions 250 mixturs are generated.

Experiments for the matched noise, unmatched noise, and cochannel conditions are conducted separately. The total number of training mixtures in the matched and unmatched noise conditions is 81250, and for the cochannel condition is 32500.

[5]https://github.com/ehabets/RIR-Generator

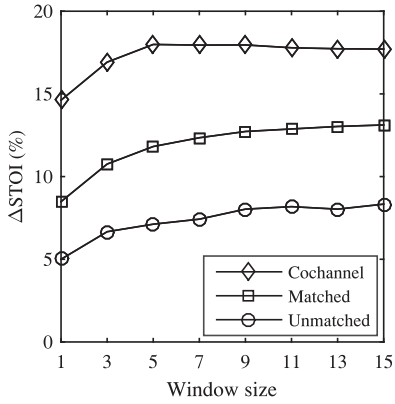

Fig. 2. Effects of contextual window on separation performance using the GF feature.

The average duration of training mixtures is approximately 2.66 seconds (i.e. 266 time frames).

Testing is done at a single SNR of -6 dB. This low SNR level is chosen so that speech intelligibility is a major concern, even for normal-hearing listeners. Thirty test mixtures are generated for each pair of T_{60} and microphone-to-speaker distance. It is worth stressing that the noise segments and speech utterances used in testing are different from those used in training.

We also evaluate the features in recorded RIRs to complement evaluation using simulated RIRs. For this purpose, the RIRs from [14] are used. This corpus consists of recordings in four different real rooms A, B, C, and D with $T_{60} = 0.32, 0.47, 0.68,$ and 0.89 s, respectively.

For speech separation evaluation, we use short-time objective intelligibility (STOI) [33] as the performance metric, which is a standard intelligibility predictor that is highly correlated with human speech intelligibility assessments. A STOI score is between -1 and 1 (typically positive) with a larger score indicating higher intelligibility. The evaluation criterion in this study is the STOI change in percent, defined as:

$$\Delta\text{STOI}\ (\%) = 100 \times (\text{STOI}_{\text{separated}} - \text{STOI}_{\text{mixture}}) \quad (5)$$

As mentioned in Section II, the reference signal for STOI measurement is the reverberant and noiseless target speech.

In a reverberant room, the sound is reflected by the surfaces and these reflections arrive at the microphone with time delays compared to the direct signal. Accordingly, a single time frame may not adequately provide the information for separation. We thus use a window of frames to incorporate temporal aspects of reverberation. Given a feature vector at time frame m, $\mathbf{F}(m)$, we extend the vector to adjacent frames as follows:

$$\mathbf{F}_a(m) = [\mathbf{F}(m-a), \cdots, \mathbf{F}(m), \cdots, \mathbf{F}(m+a)] \quad (6)$$

An interesting question is how incorporating a contextual window affects separation performance in reverberant conditions. We fix the train data size and the DNN structure as described earlier. Then, for different values of a, we calculate the average score over all of the test conditions. Fig. 2 shows the system performance with different window sizes using the GF feature, which is simple and yet very effective for speech separation [3]. In all of the test conditions, the system benefits from

TABLE II
Δ STOI SCORES FOR INDIVIDUAL FEATURES AVERAGED ON ALL OF THE TEST NOISES

Feature	Matched noise			Unmatched noise			Cochannel			Average
	Anechoic	Sim. RIRs	Rec. RIRs	Anechoic	Sim. RIRs	Rec. RIRs	Anechoic	Sim. RIRs	Rec. RIRs	
MRCG	**7.12**	**14.25**	**12.15**	**7.00**	7.28	8.99	21.25 (13.00)	22.93 (13.19)	21.29 (12.81)	**12.92**
GF	6.19	13.10	11.37	6.71	7.87	8.24	22.56 (11.86)	**23.95** (12.31)	22.35 (12.87)	12.71
GFCC	5.33	12.56	10.99	6.32	6.92	7.01	**23.53** (**14.34**)	**23.95** (**14.01**)	**22.76** (**13.90**)	12.50
LOG-MEL	5.14	12.07	10.28	6.00	6.98	7.52	21.18 (13.88)	22.75 (13.54)	21.71 (13.18)	12.08
LOG-MAG	4.86	12.13	9.69	5.75	6.64	7.19	20.82 (13.84)	22.57 (13.40)	21.82 (13.55)	11.91
GFB	4.99	12.47	11.51	6.22	7.01	7.86	19.61 (13.34)	20.86 (11.97)	19.97 (11.60)	11.75
PNCC	1.74	8.88	10.76	2.18	**8.68**	**10.52**	19.97 (10.73)	19.47 (10.03)	19.35 (9.56)	10.78
MFCC	4.49	11.03	9.69	5.36	5.96	6.26	19.82 (11.98)	20.32 (11.47)	19.66 (11.54)	10.72
RAS-MFCC	2.61	10.47	9.56	3.08	6.74	7.37	18.12 (11.38)	19.07 (11.19)	17.87 (10.30)	10.44
AC-MFCC	2.89	9.63	8.89	3.31	5.61	5.91	18.66 (12.50)	18.64 (11.59)	17.73 (11.27)	9.87
PLP	3.71	10.36	9.10	4.39	5.03	5.81	16.84 (11.29)	16.73 (10.92)	15.46 (9.50)	9.46
SSF-II	3.41	8.57	8.68	4.18	5.45	6.00	16.76 (10.07)	17.72 (9.18)	18.07 (8.93)	9.09
SSF-I	3.31	8.35	8.53	4.09	5.17	5.77	16.25 (10.44)	17.70 (9.40)	18.04 (9.35)	8.97
RASTA-PLP	1.79	7.27	8.56	1.97	6.62	7.92	11.03 (6.76)	10.96 (6.06)	10.27 (6.28)	7.46
PITCH	2.35	4.62	4.79	3.36	3.36	4.61	19.71 (9.37)	17.82 (8.45)	16.87 (6.72)	7.03
GFMC	−0.68	7.05	5.00	−0.54	4.44	4.16	5.04 (−0.07)	6.01 (0.33)	4.97 (0.28)	4.40
WAV	0.94	2.32	2.68	0.02	0.99	1.63	11.62 (4.81)	11.92 (6.25)	10.54 (1.05)	3.89
AMS	0.31	0.30	−1.38	0.19	−2.99	−3.40	11.73 (5.96)	10.97 (6.76)	10.20 (4.90)	1.71
PAC-MFCC	0.00	−0.33	−0.82	0.18	−0.92	−0.67	0.95 (0.15)	1.25 (0.26)	1.17 (0.09)	−0.17

"Sim." and "Rec." indicate simulated and recorded, respectively. Best scores are highlighted in boldface in each condition. In cochannel cases, scores for female and male interference are shown separately with the latter in parentheses.

the features in neighboring frames, but the improvement tapers off as more frames are included. The amount of contextual information to use is a trade off between the computational cost and performance gain. From Fig. 2, we set $a = 4$ in Eq. (6).

V. SINGLE FEATURE EVALUATION

In this section, we examine separation performance of individual features. Table II shows Δ STOI scores for different interference and reverberant conditions. The MRCG feature achieves the highest average score, consistent with the previous study in anechoic conditions [3]. This is not surprising given the strong performance of GF features, and the fact that MRCG builds on GF with additional contextual information. MRCG has the best performance in all nonspeech noise cases, except in reverberant and unmatched noise case where, interestingly, PNCC outperforms other features. This is probably due to the temporal masking module in PNCC that can handle reverberation to some extent. Clearly, PNCC does not perform well in anechoic noisy conditions. We also observe that GFCC is the best feature for cochannel speech separation. Note that, in the matched noise case, the STOI results in the anechoic condition are generally worse than the corresponding results in the reverberant conditions. This is mainly due to the inclusion of babble in the matched noise case (absent in the unmatched noise case). The results for anechoic babble are much worse than reverberant babble, likely because more reverberant mixtures are included in training and reverberation tends to make babble more stationary (hence easier to separate).

The information given in Tables I and II does not show any strong relationship between feature performance and its dimensionality or frame size. However, we observe that, in general, gammatone-domain features outperform others. In addition, the LOG-MEL feature is the best among spectrum-based features.

We do not find significant differences between the two types of SSF features. The WAV feature does not perform well, consistent with a previous study in speech separation [40]. On the other hand, we should note that the DNN used in this study may not couple well with the WAV feature, and convolutional and recurrent networks used in [28] may be better for waveform signals. PITCH is a more successful feature for cochannel separation than for speech-noise separation. Our evaluations show that GFMC, AMS, and PAC-MFCC are the worst features for speech separation in reverberant conditions.

To see the effects of the source-receiver distance, Table III shows Δ STOI performance by varying the distance of the target speaker to the microphone. The scores are averaged across all simulated reverberant rooms. As expected, the best performance is achieved at a 1 m distance (closest to the microphone). The results at the distances of 4 m and 8 m are not much different for nonspeech noises, with those at 8 m slightly higher. That the worst performance does not occur at the farthest distance suggests that, beyond a certain distance, the STOI gain due to separation of reverberant speech does not degrade.

VI. FEATURE COMBINATION

Each of the features studied in the previous section extracts certain characteristics of speech. These features may complement each other, and when used in a combination can boost the system performance. How to identify a feature subset with complementary characteristics? In [3], [37], group Lasso is used to find complementarity between features. In the following, we first study feature combinations based on the group Lasso method, and then present a sequential floating forward selection (SFFS) algorithm. We will show that the later method is more effective.

TABLE III
Δ STOI SCORES FOR INDIVIDUAL FEATURES AT DIFFERENT DISTANCES TO THE MICROPHONE

Feature	Matched noise				Unmatched noise				Cochannel			
	$d = 1$ m	$d = 2$ m	$d = 4$ m	$d = 8$ m	$d = 1$ m	$d = 2$ m	$d = 4$ m	$d = 8$ m	$d = 1$ m	$d = 2$ m	$d = 4$ m	$d = 8$ m
MRCG	**14.93**	**14.59**	**13.22**	**14.25**	7.82	7.71	6.76	6.82	23.62 (12.97)	23.11 (12.89)	23.32 (13.09)	21.67 (13.83)
GF	13.89	13.74	11.96	12.79	9.06	8.27	7.31	6.86	24.70 (12.40)	**24.43** (12.44)	**24.13** (11.89)	22.52 (12.53)
GFCC	13.50	12.97	11.51	12.27	8.08	7.36	6.59	5.63	**24.78** (**13.74**)	24.08 (**14.25**)	24.08 (**13.76**)	**22.85** (**14.28**)
LOG-MEL	12.82	12.59	10.90	12.00	8.03	7.25	6.43	6.23	23.46 (13.22)	23.14 (13.69)	23.02 (13.16)	21.39 (14.08)
LOG-MAG	12.79	12.53	11.04	12.14	7.76	6.71	5.94	6.15	23.47 (13.06)	22.85 (13.51)	22.86 (13.21)	21.08 (13.82)
GFB	13.58	12.74	11.43	12.13	8.18	7.30	6.26	6.30	21.45 (11.96)	21.12 (11.80)	21.14 (12.16)	19.73 (11.94)
PNCC	10.04	8.67	8.15	8.71	**9.67**	**8.96**	**7.95**	**8.21**	20.09 (10.08)	19.54 (9.94)	19.86 (9.79)	18.40 (10.29)
MFCC	11.83	11.42	9.93	10.96	7.11	6.16	5.53	5.04	20.97 (11.33)	20.51 (11.51)	20.39 (11.14)	19.41 (11.88)
RAS-MFCC	11.15	10.49	9.52	10.66	7.72	6.60	6.35	6.27	19.54 (10.81)	19.58 (11.21)	19.06 (10.65)	18.08 (12.08)
AC-MFCC	10.62	9.91	8.69	9.36	6.94	5.50	5.17	4.84	18.91 (11.62)	19.25 (11.94)	18.86 (11.06)	17.55 (11.74)
PLP	11.17	10.76	9.06	10.52	5.93	5.52	4.73	3.91	17.14 (10.85)	17.08 (10.94)	17.19 (10.72)	15.52 (11.18)
SSF-II	9.94	8.72	7.19	8.42	6.80	5.38	4.63	5.01	18.46 (9.11)	17.85 (8.97)	17.90 (8.84)	16.67 (9.78)
SSF-I	9.75	8.51	6.96	8.15	6.53	5.11	4.35	4.72	18.45 (9.49)	17.86 (9.11)	17.86 (9.11)	16.63 (9.90)
RASTA-PLP	8.27	7.17	6.13	7.52	7.48	6.64	5.81	6.58	11.18 (5.98)	10.96 (5.96)	11.46 (5.57)	10.25 (6.73)
PITCH	6.51	4.81	3.63	3.56	5.14	3.92	2.44	1.96	18.60 (8.21)	18.09 (8.98)	17.93 (8.03)	16.68 (8.59)
GFMC	7.54	7.68	6.26	6.73	5.37	4.53	4.01	3.86	6.01 (−0.01)	6.71 (−0.02)	6.41 (0.96)	4.92 (0.39)
WAV	3.25	2.06	1.98	2.01	1.80	0.74	0.68	0.77	12.68 (6.39)	12.48 (6.37)	12.61 (5.91)	9.93 (6.33)
AMS	1.20	0.60	−1.56	−1.47	−1.64	−2.32	−3.63	−3.64	11.71 (6.57)	11.15 (6.86)	11.35 (6.80)	9.68 (6.82)
PAC-MFCC	−0.12	−0.12	−0.21	−0.90	−0.80	−0.96	−0.86	−1.09	1.31 (0.24)	1.30 (0.31)	1.22 (0.33)	1.18 (0.17)

Average scores for all reverberation times are presented. In cochannel cases, scores for female and male interference are shown separately with the latter in parentheses.

A. Group Lasso

Group Lasso [43] adds a mixed norm regularization to multiple linear regression, which is shown to promote sparsity in the coefficients corresponding to predefined groups of variables. In this paper, we apply group Lasso in each frequency channel individually and then integrate across all channels:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \|\boldsymbol{\alpha} \cdot \mathbf{x} - y\|_2^2 + \lambda \sum_{k=1}^{K} \|\boldsymbol{\alpha}_{\mathbf{F}_k}\|_2 \tag{7}$$

$$\boldsymbol{\alpha} = [\boldsymbol{\alpha}_{\mathbf{F}_1}, \boldsymbol{\alpha}_{\mathbf{F}_2}, \cdots, \boldsymbol{\alpha}_{\mathbf{F}_K}] \tag{8}$$

$$\mathbf{x} = [\mathbf{x}_{\mathbf{F}_1}, \mathbf{x}_{\mathbf{F}_2}, \cdots, \mathbf{x}_{\mathbf{F}_K}] \tag{9}$$

where $\|\cdot\|_2$ denotes the Euclidean norm, y is the desired response, $K = 19$ is the size of the feature set, and $\boldsymbol{\alpha}_{\mathbf{F}_k}$ indicates the coefficients for kth feature vector, $\mathbf{x}_{\mathbf{F}_k}$.

In Eq. (7), λ is a parameter to control sparsity in groups of coefficients. In practice, often $\lambda_{\max}$ is calculated resulting in the solution $\boldsymbol{\alpha} = \mathbf{0}$, and then a ratio value $\beta \in [0, 1)$ in $\lambda = \beta \times \lambda_{\max}$ is chosen so that the linear regression error is minimized on a development set. To apply group Lasso, in both nonspeech and cochannel separation cases, we downsample the training data by 50.

Fig. 3(a) shows the magnitudes of average group Lasso coefficients for nonspeech noise, where $\beta = 0.4$ is used. As seen, PITCH and LOG-MAG are the only features with significant responses. Fig. 3(b) represents the coefficients when we repeat the method for cochannel speech separation, using the parameter $\beta = 0.6$. In this case, GFB and PITCH are found to be complementary, where all other features have zero or negligible responses.

Accordingly, in the remainder of the paper, we use PITCH+LOG-MAG as the complementary feature set for nonspeech noise, and PITCH+GFB for interfering speech from

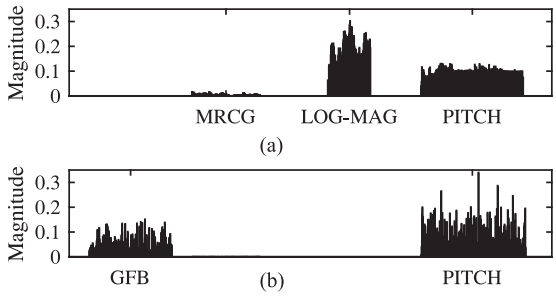

Fig. 3. Average magnitude responses of group Lasso for (a) nonspeech noise and (b) cochannel separation.

group Lasso. Note that this method is agnostic to matched and unmatched noises.

B. SFFS

This method [27] starts with an empty set and systematically adds and drops features until a desired number of features is selected. The algorithm description is given in Procedure 1. Since the number of features in the final set is not known in advance in our case, we modify the algorithm so that it stops when no improvement is achieved by adding the next feature.

We apply the SFFS algorithm to the three cases of speech separation, on the same training set used for group Lasso. Function $J(.)$ in Procedure 1 is the average STOI performance on the entire test set, where $J(\varnothing) = -\infty$.

Fig. 4 shows the state of the selected features set in each step of the SFFS algorithm. The matched noise and unmatched noise separation cases follow different paths and both end up in the same set consisting of GF+PNCC+LOG-MEL. The algorithm results in GFCC+PNCC+LOG-MEL for cochannel separation.

Procedure 1: SFFS Algorithm.

Input: $Y = \{F_k | k = 1, 2, \ldots, K\}$ # Set of all features
Output: $X_j = \{x_j | j = 1, 2, \ldots, k\}$,
$k = 1, 2, \ldots, K$ # Set of selected features

1: $j \leftarrow 0$, $X_j \leftarrow \varnothing$
2: $x^+ \leftarrow \underset{x \in Y - X_j}{\operatorname{argmax}} J(X_j \cup x)$
3: **if** $J(X_j \cup x^+) > J(X_j)$ **then**
4: $X_{j+1} \leftarrow X_j \cup x^+$
5: $j \leftarrow j + 1$
6: **else**
7: **close;**
8: **end if**
9: $x^- \leftarrow \underset{x \in X_j}{\operatorname{argmax}} J(X_j - x)$
10: **if** $J(X_j - x^-) > J(X_j)$ **then**
11: $X_{j-1} \leftarrow X_j - x^-$
12: $j \leftarrow j - 1$
13: **goto** 9.
14: **else**
15: **goto** 2.
16: **end if**

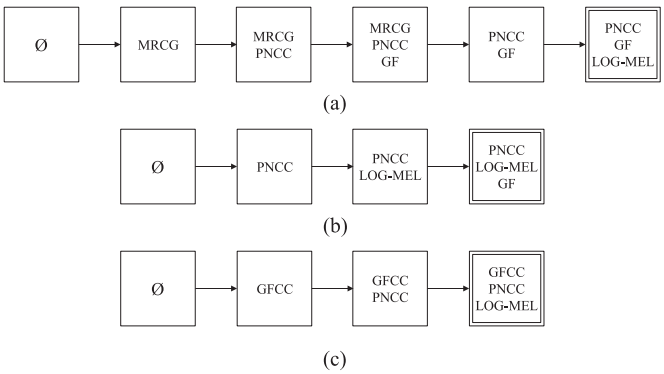

Fig. 4. Steps taken in the SFFS algorithm for (a) matched noise, (b) unmatched noise, and (c) cochannel separation. ∅ indicates an empty set.

TABLE IV
Δ STOI SCORES FOR FEATURE COMBINATIONS IN MATCHED NOISE CONDITIONS

	Anechoic	Sim. RIR	Rec. RIR	Average
Proposed (SFFS)	**7.16**	**14.96**	**15.14**	**14.54**
Group Lasso	5.98	12.88	11.20	12.06
Chen *et al.*	6.41	13.79	12.10	12.94

C. Feature Combination Results

Using group Lasso, Wang *et al.* [37] found complementarity in AMS+RASTA-PLP+MFCC for speech-noise separation in anechoic conditions, and later, Chen *et al.* [3] concluded that PITCH+MRCG forms a complementary feature set.

To see if Chen *et al.*'s combination is still effective in reverberant conditions, we compare this set with the sets derived in the previous two subsections. In Table IV, Δ STOI scores are provided for the three feature combinations in matched noise conditions. The feature set from the SFFS algorithm outperforms the other two in all conditions. Chen *et al.*'s combination performs slightly better than the group Lasso in our study, possibly because their feature set contains MRCG, which is a high quality feature in speech separation. Note that the list of features examined in this study and that in Chen *et al.* are not identical; the training targets are also different.

TABLE V
Δ STOI SCORES FOR FEATURE COMBINATIONS IN UNMATCHED NOISE CONDITIONS

	Anechoic	Sim. RIR	Rec. RIR	Average
Proposed	**7.84**	**10.12**	**11.70**	**10.36**
Group Lasso	6.92	7.97	8.68	8.08
Chen *et al.*	7.68	7.94	9.59	8.32

TABLE VI
Δ STOI SCORES FOR FEATURE COMBINATIONS IN COCHANNEL CONDITIONS

	Anechoic	Sim. RIR	Rec. RIR	Average
Proposed	**24.35** (**15.65**)	**25.01** (14.44)	**24.74** (**14.40**)	**19.71**
Group Lasso	21.98 (14.17)	22.38 (13.46)	21.41 (12.35)	17.69
Chen *et al.*	22.31(15.04)	22.91 (**14.84**)	21.40 (13.43)	18.52

Scores for female and male interference are shown separately with the latter in parentheses.

TABLE VII
FEATURE VECTOR SIZES FOR DIFFERENT FEATURE COMBINATIONS

	Matched	Unmatched	Cochannel
Proposed	1215	1215	918
Group Lasso	4905	4905	6255
Chen *et al.*	5760	5760	5760

STOI scores in unmatched noise conditions are given in Table V. Again, in all conditions, SFFS achieves the best scores. These results are important as they indicate the generalization power of the proposed set to unseen and realistic conditions. Finally, Table VI shows the STOI results in cochannel separation. The proposed set is the best in all of the conditions except female interference case in simulated reverberant conditions.

Shorter feature vectors are desirable since they imply smaller computational costs. Table VII lists feature dimensionality for different sets where 4 succeeding and 4 preceding frames are included for providing contextual information. The proposed set from SFFS has significantly smaller dimensionality compared to Chen *et al.* and group Lasso.

Fig. 5 compares Δ STOI results from Chen *et al.* and our proposed feature combination in four real rooms with different reverberation times. As shown, our method outperforms their results in all recorded reverberant conditions. Fig. 6 shows the same comparison with regard to different distances to the microphone. Again, our feature combination method produces consistently better results. We have also increased the size of the two hidden layers in the fixed DNN from 512 units to 2048 units.

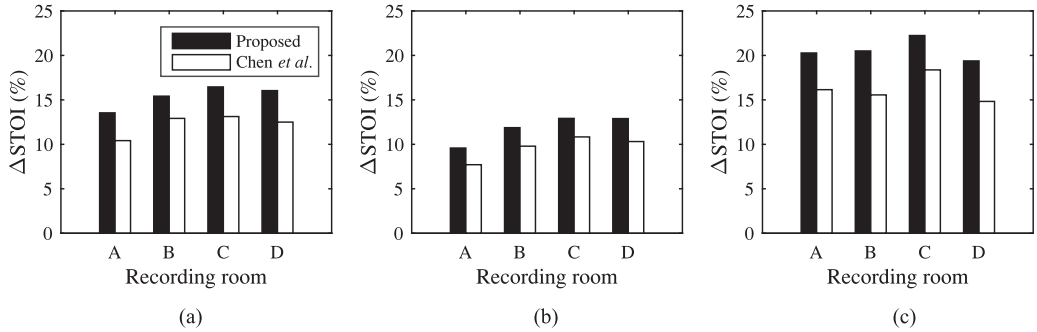

Fig. 5. Δ STOI scores for feature combinations in real rooms in (a) matched noise, (b) unmatched noise, and (c) cochannel separation.

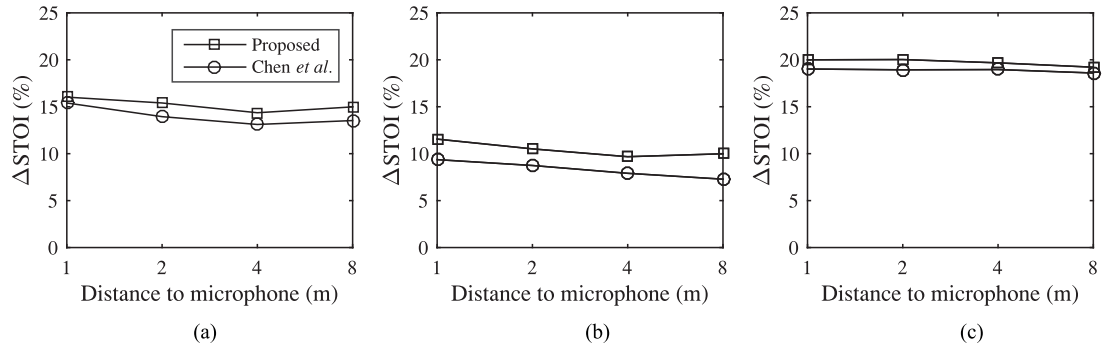

Fig. 6. Δ STOI scores for feature combinations with regard to distance to microphone in (a) matched noise, (b) unmatched noise, and (c) cochannel separation.

While the larger DNN results in slightly higher STOI numbers (less than 1 percentage point) for all three feature combinations, their relative advantages do not change.

VII. Concluding Remarks

In this paper, we have studied a broad range of features for masking-based speech separation in different reverberant conditions in the DNN framework. Both nonspeech and cochannel interference are investigated. We find that contextual information substantially boosts separation performance. Gammatone-domain features perform better than other features (see also [3]). However, not a single feature performs best in every condition. MRCG has the best overall performance in matched noise, PNCC in unmatched noise, and GFCC feature in cochannel condition. Even though the aim of our study is acoustic features, we have provided strong baseline results for future speech separation research in unmatched noises and reverberant conditions.

We have demonstrated that complementary feature sets for speech separation in reverberant conditions are different from those in anechoic conditions. We find that the SFFS algorithm produces better feature sets than the group Lasso method. The best feature combination for both matched and unmatched noise is PNCC+GF+LOG-MEL, while the best combination for cochannel separation is PNCC+GFCC+LOG-MEL. We utilize SFFS to select features in a step-by-step fashion. On the other hand, group Lasso is a multiple linear regression method. A linear model may not be strong enough to handle nonlinear masking-based speech separation. As seen in Section VI-A, group Lasso seems to favor features with large dimensionalities, and it does not differentiate matched and unmatched noise conditions.

Acknowledgment

The authors thank the anonymous reviewers for their helpful comments and Jitong Chen for his help with feature extraction.

References

[1] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.
[2] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer, 2001.
[3] J. Chen, Y. Wang, and D. L. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1993–2002, Dec. 2014.
[4] M. Delfarah and D. L. Wang, "A feature study for masking-based reverberant speech separation," in *Proc. Interspeech*, 2016, pp. 555–559.
[5] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, *NASA STI/Recon Tech. Rep. 4930*, 1993.
[6] E. L. George, S. T. Goverts, J. M. Festen, and T. Houtgast, "Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners," *J. Speech Lang. Hearing Res.*, vol. 53, pp. 1429–1439, 2010.
[7] S. Gonzalez and M. Brookes, "PEFAC-A pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 518–530, Feb. 2014.
[8] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, Jun. 2015.
[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] K. S. Helfer and L. A. Wilber, "Hearing loss, aging, and speech perception in reverberation and noise," *J. Speech Lang. Hearing Res.*, vol. 33, pp. 149–155, 1990.
[11] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.
[12] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech, Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.
[13] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.
[14] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1867–1871, Sep. 2010.
[15] *IEEE Recommended Practice for Speech Quality Measurements,* IEEE No 297–1969, 1969.
[16] S. Ikbal, H. Misra, and H. Bourlard, "Phase autocorrelation (PAC) derived robust speech features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2003, pp. II-133–II-136.
[17] Z. Jin and D. L. Wang, "Reverberant speech segregation based on multipitch tracking and classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2328–2337, Nov. 2011.
[18] C. Kim and R. M. Stern, "Nonlinear enhancement of onset for robust speech recognition," in *Proc. Interspeech*, 2010, pp. 2058–2061.
[19] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 7, pp. 1315–1329, Jul. 2016.
[20] 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.
[21] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Mach. Learn.*, 2014, arXiv:1412.6980.
[22] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Muller, "Efficient backprop," in *Neural Networks: Tricks of the Trade*. Berlin, Germany: Germany: 2012, pp. 9–48.
[23] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2013.
[24] H. K. Maganti and M. Matassoni, "An auditory based modulation spectral feature for reverberant speech recognition," in *Proc. Interspeech*, 2010, pp. 570–573.
[25] A. K. Nábělek and P. K. Robinson, "Monaural and binaural speech perception in reverberation for listeners of various ages," *J. Acoust. Soc. Amer.*, vol. 71, pp. 1242–1248, 1982.
[26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.
[27] P. Pudil, F. Ferri, J. Novovicova, and J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions," in *Proc. 12th IAPR Int. Conf. Pattern Recognit.*, 1994, vol. 2, pp. 279–283.
[28] T. N. Sainath, R. J. Weiss, A. W. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs." in *Proc. Interspeech*, 2015, pp. 1–5.
[29] 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.
[30] 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.
[31] Y. Shao, S. Srinivasan, and D. L. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. IV-277–IV-280.
[32] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.
[33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[34] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multichannel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoust. Soc. Amer.*, vol. 133, pp. 3591–3591, 2013.
[35] 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.
[36] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley-IEEE Press, 2006.
[37] 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.
[38] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.
[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] Y. Wang and D. L. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4390–4394.
[41] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.
[42] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.
[43] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. Ser. B Statist. Methodol.*, vol. 68, pp. 49–67, 2006.
[44] K.-H. Yuo and H.-C. 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.
[45] X.-L. Zhang and D. L. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 5, pp. 967–977, May 2016.
[46] X. Zhang, H. Zhang, S. Nie, G. Gao, and W. Liu, "A pairwise algorithm using the deep stacking network for speech separation and pitch estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 6, pp. 1066–1078, Jun. 2016.
[47] Y. Zhao, D. L. Wang, I. Merks, and T. Zhang, "DNN-based enhancement of noisy and reverberant speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 6525–6529.

Masood Delfarah (S'16) received the B.Sc. degree in computer engineering from the University of Tehran, Tehran, Iran, in 2013. He is currently working toward the Ph.D. degree in computer engineering at The Ohio State University, Columbus, OH, USA. His research interests include speech separation, dereverberation, machine learning, and deep learning.

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