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 (DNNs) 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—Monaural speech separation, speech intelligibility, feature combination, room reverberation, deep neural networks.

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 [10], [6]. 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.

 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 fea-
We offer concluding remarks in Section VII. Section IV describes the features to be investigated. The experimental II we describe the feature evaluation framework. Section III new features and many more conditions are evaluated in this current work goes substantially beyond the preliminary work.

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 \( y(t) = s(t) + n(t) \) (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:

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

where \( F(\cdot) \) denotes the feature vector, \( \Theta \) corresponds to the model parameters, \( C = 161 \) is the number of frequency channels, and \( g_c(\cdot) \) is the value of the \( c^{th} \) 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 \( F(\cdot) \):

\[
F(m, d) = \frac{\tilde{F}(m,d) - \mu_d}{\sigma_d} \tag{4}
\]

where \( d \) indexes a feature element, \( \tilde{F}(\cdot) \) is an extracted training or test feature, and \( \mu_d \) and \( \sigma_d \) represent the mean and standard deviation of the \( d^{th} \) 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 point-wise 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.

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. Multi-resolution 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 $11 \times 11$. In a similar way, CG4 is computed from CG2 with the window size of $23 \times 23$. 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 PEFACT 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 downsamples. 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 [50].

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).

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

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 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 [2] 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 × 9 m × 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, PSTAT-ION, 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 room.

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) × 5 (SNR) × 3 ($T_{60}$) × 4 (microphone-target distance) = 600 different conditions, and anechoic mixtures in 10 (interference) × 5 (SNR) = 50 conditions. For each of these conditions 250 mixtures 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. 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 $T_{60} = 0.32, 0.47, 0.68, \text{ and } 0.89$ 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:

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

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 $m$, $F(m)$, we extend the vector to adjacent frames as follows:

$$F_a(m) = [F(m - a), \cdots, F(m), \cdots, F(m + a)]$$

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 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 III shows $\Delta\text{STOI}$ 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 Table I and Table 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 ∆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 [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.

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:

\[
\min_{\alpha} \frac{1}{2} \|\alpha \cdot x - y\|_2^2 + \lambda \sum_{k=1}^K \|\alpha_{F_k}\|_2
\]

where \(\|\cdot\|_2\) denotes the Euclidean norm, \(y\) is the desired response, \(K = 19\) is the size of the feature set, and \(\alpha_{F_k}\) indicates the coefficients for \(k^{th}\) feature vector, \(x_{F_k}\).

In Eq. (7) \(\lambda\) is a parameter to control sparsity in groups of coefficients. In practice, often \(\lambda_{\text{max}}\) is calculated resulting in the solution \(\alpha = 0\), and then a ratio value \(\beta \in [0, 1]\) in \(\lambda = \beta \times \lambda_{\text{max}}\) 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, GF and PITCH are found to be complementary, while 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+GF for interfering speech from 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 I. 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 \(J(\cdot)\) in Procedure I is the average STOI performance on the entire test set, where \(J(\emptyset) = -\infty\).
TABLE II

ΔSTOI scores for individual features averaged on all of the test noises. “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.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Anechoic</th>
<th>Unmatched noise</th>
<th>Cochannel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rec. RIRs</td>
<td></td>
<td>Rec. RIRs</td>
</tr>
<tr>
<td>MRCG</td>
<td></td>
<td>14.25</td>
<td>12.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE III

ΔSTOI scores for individual features at different distances to the microphone. 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.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Matched noise</th>
<th>Unmatched noise</th>
<th>Cochannel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>d = 1 m</td>
<td>d = 2 m</td>
<td>d = 4 m</td>
</tr>
<tr>
<td>MRCG</td>
<td>14.93</td>
<td>13.22</td>
<td>14.25</td>
</tr>
<tr>
<td></td>
<td>13.89</td>
<td>13.34</td>
<td>13.79</td>
</tr>
<tr>
<td></td>
<td>13.00</td>
<td>12.97</td>
<td>12.97</td>
</tr>
<tr>
<td></td>
<td>12.90</td>
<td>12.59</td>
<td>12.99</td>
</tr>
<tr>
<td></td>
<td>12.79</td>
<td>12.53</td>
<td>12.14</td>
</tr>
<tr>
<td></td>
<td>13.58</td>
<td>12.74</td>
<td>11.43</td>
</tr>
<tr>
<td></td>
<td>10.04</td>
<td>8.67</td>
<td>8.15</td>
</tr>
<tr>
<td></td>
<td>11.83</td>
<td>11.42</td>
<td>9.93</td>
</tr>
<tr>
<td></td>
<td>11.15</td>
<td>10.49</td>
<td>9.52</td>
</tr>
<tr>
<td></td>
<td>10.62</td>
<td>9.91</td>
<td>8.69</td>
</tr>
<tr>
<td></td>
<td>11.17</td>
<td>10.76</td>
<td>9.06</td>
</tr>
<tr>
<td></td>
<td>9.94</td>
<td>8.72</td>
<td>7.19</td>
</tr>
<tr>
<td></td>
<td>9.75</td>
<td>8.51</td>
<td>6.96</td>
</tr>
<tr>
<td></td>
<td>8.27</td>
<td>7.17</td>
<td>6.13</td>
</tr>
<tr>
<td></td>
<td>6.51</td>
<td>4.81</td>
<td>3.63</td>
</tr>
<tr>
<td></td>
<td>7.54</td>
<td>7.68</td>
<td>6.26</td>
</tr>
<tr>
<td></td>
<td>3.25</td>
<td>2.06</td>
<td>1.98</td>
</tr>
<tr>
<td></td>
<td>1.20</td>
<td>0.60</td>
<td>-1.56</td>
</tr>
</tbody>
</table>

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.

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. [35] 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 ΔSTOI 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

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

Fig. 4 shows the state of the selected features set in each step of the SFFS algorithm. The matched noise and unmatched

![Magnitude](a)

![Magnitude](b)
Fig. 4: Steps taken in the SFFS algorithm for (a) matched noise, (b) unmatched noise, and (c) cochannel separation. $\varnothing$ indicates an empty set.

**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. $x^+ \leftarrow \text{argmax} \ J(X_j \cup x)$
   $x \in \mathbb{Y} - X_j$
3. if $J(X_j \cup x^+) > J(X_j)$ then
   4. $X_{j+1} \leftarrow X_j \cup x^+$
   5. $j \leftarrow j + 1$
7. **else**
8. **end if**
9. $x^- \leftarrow \text{argmax} \ J(X_j - x)$
   $x \in X_j$
10. if $J(X_j - x^-) > J(X_j)$ then
11. $X_{j-1} \leftarrow X_j - x^-$
12. $j \leftarrow j - 1$
14. **else**
15. **goto** 2.
16. **end if**

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.

STOI scores in unmatched noise conditions are given in Table IV. 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 V 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. V compares $\Delta$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. 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

We thank the anonymous reviewers for their helpful comments and Jitong Chen for his help with feature extraction. This research was supported in part by an AFOSR grant (FA9550-12-1-0130) and the Ohio Supercomputer Center.

REFERENCES


![Fig. 5: ΔSTOI scores for feature combinations in real rooms in (a) matched noise, (b) unmatched noise, and (c) cochannel separation.](image)

![Fig. 6: ΔSTOI scores for feature combinations with regard to distance to microphone, in (a) matched noise, (b) unmatched noise, and (c) cochannel separation.](image)


Massoud Delfarbah (’16) received the B.Sc. degree in computer engineering from the University of Tehran, Iran, in 2013. He is currently pursuing the Ph.D. degree in computer engineering at The Ohio State University. His research interests include speech separation, dereverberation, machine learning, and deep learning.

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