This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2696307, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Time-Frequency Masking in the Complex Domain for Speech Dereverberation and Denoising

Donald S. Williamson, *Member, IEEE,* and DeLiang Wang, *Fellow, IEEE*

***Abstract*—In real-world situations, speech is masked by both background noise and reverberation, which negatively affect perceptual quality and intelligibility. In this paper, we address monaural speech separation in reverberant and noisy environments. We perform dereverberation and denoising using supervised learning with a deep neural network. Specifically, we enhance the magnitude and phase by performing separation with an estimate of the complex ideal ratio mask. We define the complex ideal ratio mask so that direct speech results after the mask is applied to reverberant and noisy speech. Our approach is evaluated using simulated and real room impulse responses, and with background noises. The proposed approach improves objective speech quality and intelligibility significantly. Evaluations and comparisons show that it outperforms related methods in many reverberant and noisy environments.**

***Index Terms*—Deep neural networks, speech separation, speech quality, complex ideal ratio mask, dereverberation.**

I. INTRODUCTION

Room acoustics affect the speech signal transmitted inside a room. When someone is having a conversation, they hear not only the sound that directly reaches their ears, but also reflections off the walls, ceiling and furniture. These reflections, termed reverberation, are altered versions of the original speech. In fact, reverberant speech consists of three components: the direct sound, early and late reflections. The direct sound is the anechoic part corresponding to the first wavefront, early reflections typically arrive up to 50ms after the direct sound, and late reflections come anytime thereafter.

Reverberation is problematic because the reflections cause smearing across time and frequency, which interferes with the direct sound. This is particularly challenging for hearing-impaired listeners, since the smearing affects their ability to recognize speech [2], [28]. Additionally, the performance of speech processing applications is degraded in reverberant environments, where reverberation causes automatic speech recognition (ASR) [22] and speaker identification systems [45] to become less accurate. The problem is worsened when background noise is present. Roman and Woodruff [34] show that reverberation combined with additive noise can be detrimental to the speech intelligibility of normal hearing listeners. A solution for removing reverberation and noise would be beneficial for a variety of speech processing tasks.

Many approaches have been developed to remove reverberation. Delcroix *et al.* use a weighted prediction error (WPE) algorithm and beamforming to remove room reverberation [6]. Reverberant speech corresponds to convolving a room impulse response (RIR) with anechoic speech (i.e. direct sound). WPE is an unsupervised approach that operates in the complex time-frequency (T-F) domain and uses linear prediction to shorten the RIR, which in effect removes late reverberation [44]. Although WPE helps with dereverberation, it does not address noise that is typically present in real situations. Inverse filtering is another technique for dereverberation. Inverse filters attempt to undo the effects of the RIR, since the convolution of the inverse filter with the reverberant signal results in anechoic speech. Inverse filters, however, cannot be fully realized, since the RIR is unstable due to its nonminimum phase nature [29]. Miyoshi and Kaneda [26] address the invertibility of the inverse filter by utilizing multiple finite impulse response (FIR) filters. In [21] and [35], the T-F magnitude response of the RIR is estimated. Another approach uses the RIR magnitude response and nonnegative matrix factorization (NMF) to remove reverberation [27]. A two-stage algorithm for enhancing reverberant speech is described by Wu and Wang [43], where the first stage estimates an inverse filter and the second stage uses spectral subtraction to minimize long-term reverberation. A monaural pitch-based method that estimates an inverse filter [33] has also been investigated. It should also be noted that inverse filtering is fundamentally sensitive to RIRs, which strongly limits the robustness of this approach [32], [20].

More recent studies perform dereverberation in a supervised manner. In [20], Jin and Wang use a multi-layer perceptron (MLP) to learn a mapping from pitch-based features to grouping cues that encode the *posterior* probability of a T-F unit being speech dominant given the reverberant observation. The mapping results in a binary mask that is used to retain the speech dominant units. Evaluations show that this system generalizes well in various reverberant environments. Jiang *et al.* [19] use deep neural networks (DNNs) to estimate the ideal binary mask (IBM), where binaural and monaural features are used to train a DNN. Weninger *et al.* [40] use deep bidirectional Long Short-Term Memory (LSTM) recurrent neural networks (RNNs) to dereverberate features that are inputed to an ASR system. Very recently, Han *et al.* [13] learn a spectral mapping from the log-magnitude spectra of noisy and reverberant speech to the log-magnitude spectra of clean speech using a DNN. Although each of these approaches produces improvements in various conditions, their

Manuscript received May 13, 2016.

D. S. Williamson was with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: williado@cse.ohio-state.edu). He is now with the School of Informatics and Computing, Indiana University, Bloomington, IN 47405 USA (e-mail: williads@indiana.edu).

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

2329-9290 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2696307, IEEE/ACM Transactions on Audio, Speech, and Language Processing

performance is limited since they only enhance the magnitude response, and use reverberant and noisy phase during signal reconstruction. As a result, the quality of separated speech is not good when interference is strong, and there is a strong need to produce speech estimates with high quality in reverberant and noisy environments.

When dealing with background noise, we recently found that performing T-F masking in the complex domain is very beneficial [42]. This approach jointly enhances the magnitude and phase response of noisy speech by estimating the complex ideal ratio mask (cIRM) in the real and imaginary domains. The performance of complex domain processing is not bounded since a full (magnitude and phase) reconstruction of speech is possible in the ideal case. Results show that the estimated cIRM substantially outperforms directly estimating speech in the time domain, traditional ideal ratio mask (IRM) estimation in the magnitude domain, and other related methods. Furthermore, cIRM estimation is shown to outperform methods that separately enhance the magnitude and phase of noisy speech. More details about different phase enhancement techniques can be found in [11].

Complex ratio masking, however, has not been investigated in adverse conditions with both room reverberation and background noise. In this paper, we propose to use DNNs to learn a mapping from reverberant (and noisy) speech to the cIRM. We extend the definition of the cIRM to deal with reverberant (and noisy) spectra, where the desired output is the spectra of the direct sound source. Unlike previous approaches, applying the cIRM enables the complete reconstruction of the clean and anechoic speech, since it jointly enhances the magnitude and phase. To our knowledge, this is the first supervised separation study that addresses dereverberation and denoising in the complex domain. A preliminary version of this work is published in [41].

This paper is organized as follows. Section II provides notations and definitions. A description of our algorithm is given in Section III. The evaluation criteria and experimental results are given in Section IV. A discussion of related issues and a conclusion are given in Section V.

II. NOTATION AND DEFINITIONS

As mentioned earlier, reverberation can be modeled as the convolution of speech with an RIR:

$$y(t) = h(t) * s(t) \tag{1}$$

where '$*$' indicates convolution, and t indexes a time sample. $y(t)$ denotes reverberant speech, and $s(t)$ clean anechoic speech. $h(t)$ denotes the RIR, which models every aspect of sound propagation from the source to the receiver. In this case, it models the direct sound (delayed and attenuated speech) that reaches the ears, as well as the early and late reflections. These terms can be modeled with $h(t)$, by dividing it into three components (one for each signal) and using the distributive property of convolution [30]. In other words, the RIR can be represented as the sum of impulse responses for the direct sound, early and late reflections:

$$h(t) = h_d(t) + h_e(t) + h_l(t) \tag{2}$$

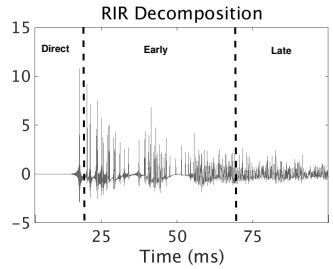


Fig. 1. A depiction of the decomposition of a room impulse response into its three components: direct, early, and late.

where $h_d(t)$, $h_e(t)$, and $h_l(t)$ are the impulse responses for the direct sound, early and late reflections, respectively. An example of this decomposition is given in Figure 1. The direct sound impulse response, $h_d(t)$, ranges from the start of $h(t)$ and ends approximately 1ms after the first impulse. The early reflection impulse response, $h_e(t)$, extends 50ms after the end of the direct sound impulse response [3], and the late reflection impulse response extends from the end of $h_e(t)$ to the end of $h(t)$. Note that the length of each of the impulse responses is the same as $h(t)$, but the component impulse responses are zero outside of the regions defined above. The distributive property of convolution says that the three components of reverberant speech can be computed by convolving their corresponding impulse response with speech.

$$\begin{aligned} y(t) &= h_d(t) * s(t) + h_e(t) * s(t) + h_l(t) * s(t) \\ &= d(t) + y_e(t) + y_l(t) \end{aligned} \tag{3}$$

with $d(t)$ corresponding to the direct sound, and $y_e(t)$ and $y_l(t)$ corresponding to the early and late reflections.

When reverberation and noise are present, reverberant and noisy speech, $y_{rn}(t)$, is defined as

$$y_{rn}(t) = h_s(t) * s(t) + \beta h_n(t) * n(t) \tag{4}$$

where $n(t)$ corresponds to the noise at time t. The RIR for reverberant speech and noise are represented with $h_s(t)$ and $h_n(t)$, respectively. The parameter β controls the signal-to-noise ratio (SNR) between the reverberant noise and speech.

Our goal in this study is to estimate the short-time Fourier transform (STFT) of the direct sound D, since it is clean and anechoic. It is a delayed and attenuated version of the true speech, but it is time aligned with the reverberant speech. This time alignment assists in learning a mapping from noisy speech features to a training target. An exact description of this mapping is presented in the next section.

III. ALGORITHM DESCRIPTION

We propose to use a DNN to learn a spectral mapping from reverberant (and noisy) speech to the cIRM. We begin this section by describing the spectral features. We then define the cIRM in this domain space, and conclude by providing details about the DNN.

A. Features

A complementary set of features is computed from the given signal [37]. These features include amplitude modu-

2329-9290 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2696307, IEEE/ACM Transactions on Audio, Speech, and Language Processing

lation spectrogram (AMS) [23], relative spectral transform and perceptual linear prediction (RASTA-PLP) [14], [15], mel-frequency cepstral coefficients (MFCC), as well as their deltas. Gammatone filterbank energies and their deltas are also appended to the feature vector. The features are computed for each time frame of the signal. A variant of this feature set has been shown to be effective for speech separation [38], and they have recently been shown to work well for cIRM estimation [42].

Since speech is correlated from frame to frame, we incorporate temporal dynamics by joining adjacent frames into a single feature vector. The feature vector centered at the kth time frame is as follows:

$$\tilde{\boldsymbol{F}}(k) = [\boldsymbol{F}(k-p), \ldots, \boldsymbol{F}(k), \ldots, \boldsymbol{F}(k+p)]^T \tag{5}$$

where p denotes the number of adjacent frames to include on each side.

B. Complex ideal ratio mask (cIRM)

The complex ideal ratio mask is a T-F mask constructed from the reverberant (and noisy) signal and the targeted speech. The cIRM is defined so that when it is applied to the reverberant observation, the targeted signal results [42]. In other words, $D(k,f) = M(k,f) \times Y(k,f)$, where $D(k,f)$, $M(k,f)$, and $Y(k,f)$ are the STFTs of the targeted speech, the cIRM, and the reverberant speech at time frame k and frequency channel f. In this case, the targeted speech is the spectra of the direct sound source, D. These STFTs are complex, so they have real and imaginary components. The traditional IRM can be defined as the ratio between the spectral magnitudes of the direct and reverberant speech (i.e. $M_{IRM} = |D|/|Y|$). On the other hand, the cIRM is defined as follows:

$$\begin{aligned} M(k,f) &= \frac{D(k,f)}{Y(k,f)} \\ &= \frac{Y_r(k,f)D_r(k,f) + Y_i(k,f)D_i(k,f)}{Y_r(k,f)^2 + Y_i(k,f)^2} + \\ &\quad j\frac{Y_r(k,f)D_i(k,f) - Y_i(k,f)D_r(k,f)}{Y_r(k,f)^2 + Y_i(k,f)^2} \end{aligned} \tag{6}$$

where subscripts r and i indicate the real or imaginary components, respectively. In essence, the cIRM can be thought of as an inverse filter, since it reverses the effects of reverberation. A depiction of the real and imaginary components of the direct speech, reverberant speech, and cIRM is shown in Figure 2.

Eq. (6) shows that M has real and imaginary components, but it can also be defined in polar coordinates.

$$M(k,f) = \frac{|D(k,f)|}{|Y(k,f)|} e^{j(\phi_d(k,f) - \phi_y(k,f))} \tag{7}$$

where ϕ_d and ϕ_y are the phases of the direct speech and reverberant observation, respectively. This equation shows that the cIRM is based on the magnitude and phase of the targeted and reverberant signals. This is important since it means that when it is applied to the reverberant speech, both the magnitude and phase are enhanced, which is crucial for speech quality [31]. Recently, a phase-sensitive mask (PSM) has been defined [8], which amounts to the real portion of the cIRM (i.e. $M_{PSM} = (|D|/|Y|)cos(\phi_d - \phi_y)$). Unlike the cIRM, the PSM does not completely enhance reverberant speech, since it cannot completely restore the phase.

The real and imaginary components of the target and reverberant speech have large values in the range $(-\infty, \infty)$. Since a smaller range is more favorable for supervised learning with DNNs, we compress the components of the cIRM using the following hyperbolic tangent.

$$M_x^{'} = Q\frac{1 - e^{-C \cdot M_x}}{1 + e^{-C \cdot M_x}} \tag{8}$$

where $x \in \{r, i\}$, denoting the real or imaginary components of the compressed cIRM, $M^{'}$. The mask values are compressed to be within [-Q,Q], and C is a steepness constraint.

C. cIRM estimation

We train a deep neural network to learn the spectral mapping from reverberant, or reverberant and noisy signals to the cIRM. A depiction of the DNN is shown in Figure 3.

The DNN is given the complementary set of features that are defined in Section III-A. Before adding temporal correlations, the feature vector has dimensionality of R units. After augmenting the feature vector to include temporal correlations, the feature vector has dimensionality of $R(2p+1)$. The input is normalized to have zero mean and unit variance. After normalization, auto-regressive moving average (ARMA) filtering is performed on the input features [5]. The output layer of the DNN is divided into two sublayers. The sublayers are for the real and imaginary components of the cIRM. Since the real and imaginary components of the cIRM are related, it is important that the network structure jointly estimate them [4]. Linear activation functions are used in the output layer, whereas rectified linear functions are used in the hidden layer.

Back propagation based on the mean-square error is used to train the DNN. Eq. (9) is the cost function for each training utterance:

$$\frac{1}{2N}\sum_k \sum_f [(\hat{M}_r^{'}(k,f) - M_r^{'}(k,f))^2 + (\hat{M}_i^{'}(k,f) - M_i^{'}(k,f))^2] \tag{9}$$

where $\hat{M}_r^{'}(k,f)$ and $\hat{M}_i^{'}(k,f)$ are the estimated real and imaginary components that are generated by the DNN. N is the number of time frames for the input. Adaptive gradient descent [7] with a momentum term is used.

The output of the DNN is an estimate of the compressed mask values of the cIRM. During testing, we uncompress these values using the following:

$$\hat{M}_x = -\frac{1}{C}\log(\frac{Q - \hat{M}_x^{'}}{Q + \hat{M}_x^{'}}) \tag{10}$$

The uncompressed estimates for the real and imaginary components are then used to extract an estimate of the direct speech (i.e. $\hat{D} = \hat{M}Y$, where $\hat{M} = \hat{M}_r + j\hat{M}_i$).

2329-9290 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2696307, IEEE/ACM Transactions on Audio, Speech, and Language Processing

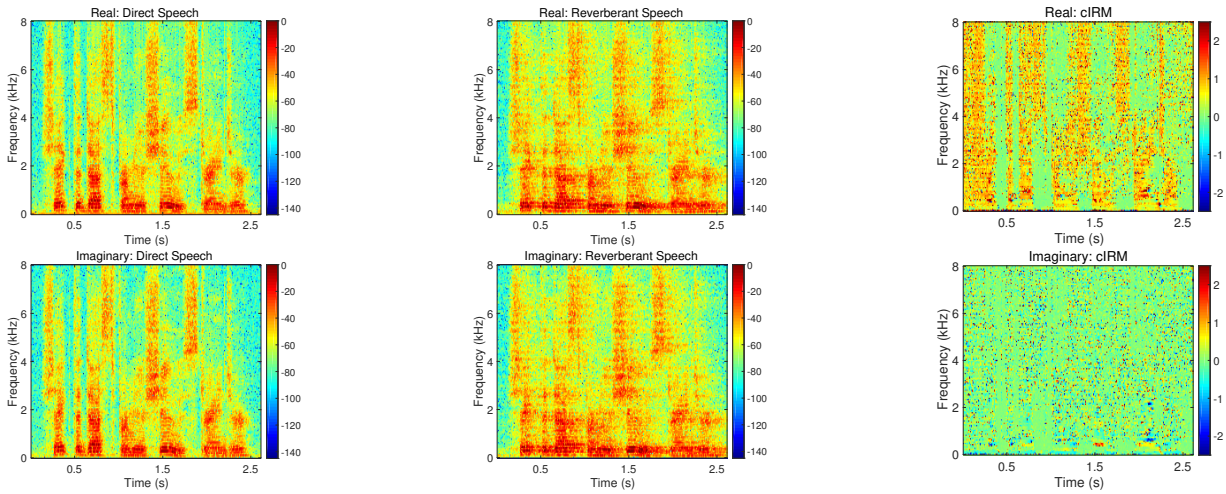


Fig. 2. (Color Online). Spectrogram plots of the real (top) and imaginary (bottom) STFT components of direct speech, reverberant speech, and the complex ideal ratio mask. The reverberant speech is generated using a T_{60} of 0.9s and a 1m distance exists between the speaker and microphone.

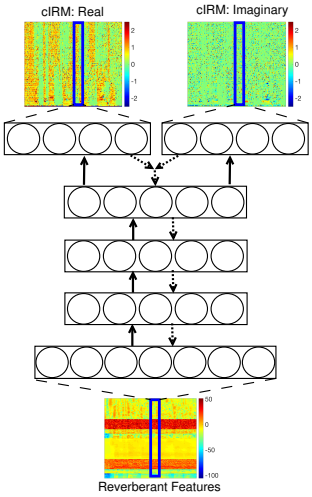


Fig. 3. (Color Online). Network structure of the DNN that estimates the complex ideal ratio mask.

IV. EVALUATIONS AND RESULTS

A. Comparisons and metrics

We compare cIRM estimation with two dereverberation algorithms. Yoshioka and Nakatani [44] use weighted error prediction to develop a filter that removes late reverberation. This approach is used by Delcroix *et al.* [6]. The filter shortens the RIR by leveraging the temporal correlations of speech. The filter is defined in the complex domain, but it is estimated in an unsupervised manner. This approach is denoted as WPE. We also compare to a recent approach by Han *et al.* [13], which uses a deep neural network to spectrally map the log-magnitude response of reverberant speech to the log-magnitude response of clean speech. This approach is denoted as DSM. For this study, the DNN uses the log-magnitude response of reverberant speech as input, and estimates the log-magnitude response of the direct speech signal.

In addition to the above comparisons, we compare cIRM estimation to other supervised T-F masking based approaches. The approaches described below have previously been evaluated for denoising only and not dereverberation. This study shows their performance in reverberant and noisy environments. We compare our approach to IRM estimation [38] to determine the significance of complex masking. The IRM gives the proportion of speech energy in each T-F unit, where speech energy is based solely on the magnitude responses of the direct sound and the reverberant (and noisy) observation. Unlike the cIRM, the IRM does not address phase and it uses the phase from the unprocessed signal for reconstruction. We also compare our approach to phase-sensitive mask (PSM) estimation [8] and time-domain reconstruction (TDR) [39]. PSM corresponds to the real component of the cIRM. TDR uses a DNN to map features to a time-domain signal using a ratio masking subnet and noisy phase. We modify TDR to use the enhanced phase from cIRM estimation when mapping to the time-domain signal, since we find that this gives a slight improvement boost. DNNs are separately trained to estimate each of these targets using the same network structure and cost function as described in previous sections. In each case (cIRM, IRM, PSM and TDR), the input to the DNN is the complementary feature set defined in Section III-A. Note that we evaluated DSM with the complementary feature set as input, but this did not perform as well as the log-magnitude response of reverberant speech.

STFTs are computed by first dividing a signal into 32ms time frames with an 8ms frame shift (i.e. 75% overlap). The fast Fourier transform (FFT) is then computed within each time frame using a 512-point FFT. A 16 kHz sampling rate is used for each signal, so each time frame of the STFT consists of 257 elements.

The DNN is given the complementary feature set, which contains 246 units (i.e. $R = 246$). After including temporal correlations, the feature vector has the dimensionality of $246 \times (2p + 1) = 246 \times 5 = 1230$ (p is set to 2 based on our prior study [42]). Therefore, the input layer of the DNN has 1230 units. Mean and variance normalization is performed once for the entire feature set during training,

2329-9290 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2696307, IEEE/ACM Transactions on Audio, Speech, and Language Processing

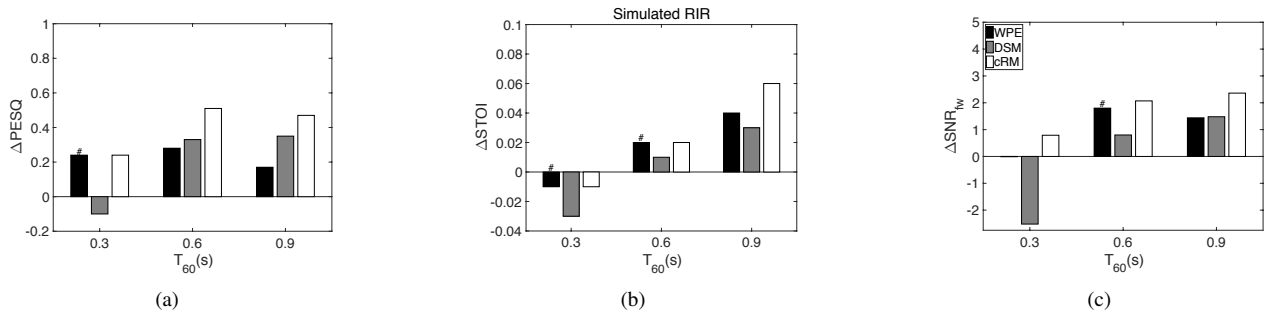


Fig. 4. (a) ΔPESQ, (b) ΔSTOI, (c) ΔSNR_{fw} results using simulated RIRs. The improvement relative to the unprocessed reverberant speech is shown. '#' indicates that the differences from cRM results are not statistically significant according to a one-way ANOVA test with 5% confidence interval.

and once per utterance during testing. Each output sublayer consists of 257 units, where linear activation functions are used. Each hidden layer has 1024 units and three hidden layers are used. The momentum rate of the DNN is set to 0.5 for the first 5 epochs, and 0.9 thereafter. A total of 80 epochs are used. The weights of the DNN are randomly initialized. When compressing the cIRM for training, we set Q to 1 and C to 0.5. Other values were evaluated, but this combination performed best empirically.

We evaluate our approach using objective metrics that give scores for speech quality and intelligibility. The perceptual evaluation of speech quality (PESQ) gives a speech quality score by comparing an enhanced signal to the direct speech signal [18]. PESQ gives scores in the range of $[-0.5, 4.5]$, where higher scores indicate higher quality. In terms of intelligibility, we use short-time objective intelligibility (STOI) [36]. STOI computes the correlation between the temporal envelopes of reference and processed speech signals over short-time segments. It returns a score between 0 and 1, where higher scores indicate better intelligibility. It is important to know that both PESQ and STOI have been shown to be highly correlated with speech quality and intelligibility of human listeners, respectively. In addition, we evaluate the frequency-weighted segmental signal-to-noise ratio (SNR_{fw}) [25], which computes and then averages the weighted signal-to-noise ratio in each critical band. The direct speech is used as the reference for each metric. The improvement score for each metric, relative to the unprocessed reverberant (and noisy) speech, is used to evaluate each approach.

We start by evaluating cIRM estimation, DSM and WPE in reverberant environments and environments that contain reverberation and noise. Afterwards, we compare cIRM estimation with other supervised T-F masking approaches.

B. Reverberation: simulated RIRs

We first evaluate the dereverberation approaches using simulated RIRs. Simulated RIRs are generated using the imaging method [1], which is implemented in [12]. The RIR is generated by placing the target speaker and microphone in random positions in a simulated room of size 9m x 8m x 7m, where the distance between the speaker and microphone is fixed at 1m. The elevations of the speaker and microphone are identical. With this configuration, sets of 11 room impulse responses are generated using T_{60} times of 0.3, 0.6, and 0.9s, respectively. At each T_{60}, 10 of the RIRs are used for training, while the other 1 is used for testing. So in total, 30 RIRs are used for training and 3 are used for testing. The average direct-to-reverberant ratio (DRR) at each T_{60} for the training RIRs is 8.6, 3.2, and 1.1 dB, while the DRR for the testing RIRs is 7.8, 2.7 and 0.8 dB, respectively.

We use the IEEE corpus [17] to train and test our system. This corpus contains 720 utterances spoken by a single male speaker. Our DNN is trained by convolving 500 of these utterances with the 30 training RIRs, resulting in a set of 15000 reverberant signals. For testing, 100 utterances that are not used during training are convolved with the 3 testing RIRs, resulting in 300 test signals. A development set of 100 different utterances is also convolved with the 30 training RIRs for parameter tuning and early stopping.

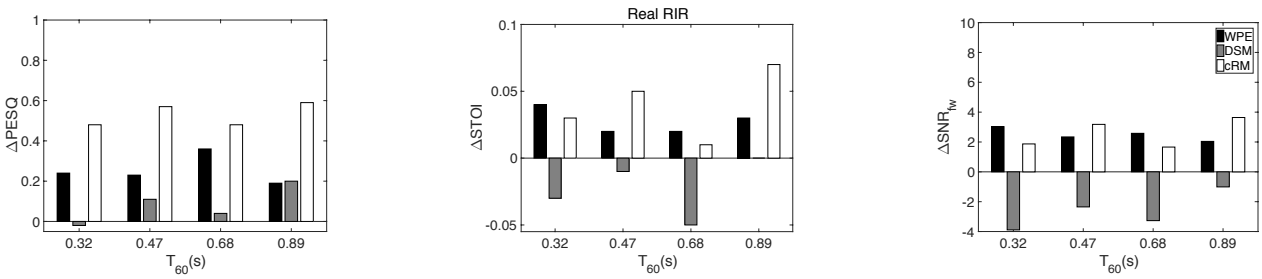
The results using these utterances and simulated RIRs are shown in Figure 4. For PESQ, shown in Figure 4(a), cIRM estimation (denoted as cRM) significantly improves performance relative to the unprocessed reverberant speech. The average improvement is 0.41 points. On the other hand, the improvement over unprocessed reverberant speech is not as high using WPE and DSM algorithms. Note that PESQ improvement for DSM in reverberation is lower in [13], since they predict the clean speech signal as opposed to the direct speech. STOI evaluation results are shown in Figure 4(b). The STOI performance for each approach increases as T_{60} increases, but each approach lowers STOI at 0.3s. STOI performances for WPE and cIRM estimation are approximately equal at 0.3 and 0.6s, but cIRM estimation performs best at 0.9s, which is the most challenging case. SNR_{fw} results in Figure 4(c) show that cIRM estimation increases SNR the most, with an average improvement of 1.74 dB. In fact, cIRM estimation is the only approach to increase SNR at a 0.3s T_{60}.

TABLE I
CHARACTERISTICS OF THE ROOMS USED TO CAPTURE THE REAL RIRS.

Room	Dimensions	T_{60} [s]	DRR [dB]
A	6.64m x 5.72m x 2.31m	0.32	6.09
B	4.65m x 4.65m x 2.68m	0.47	5.31
C	18.8m x 23.5m x 4.6m	0.68	8.82
D	8.72m x 8.02m x 4.25m	0.89	6.12

2329-9290 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2696307, IEEE/ACM Transactions on Audio, Speech, and Language Processing



(a) (b) (c)

Fig. 5. (a) ΔPESQ, (b) ΔSTOI, (c) ΔSNR_{fw} results using real RIRs. The improvement relative to the unprocessed reverberant speech is shown. All differences from cRM results are statistically significant according to a one-way ANOVA test with 5% confidence interval.

C. Reverberation: real RIRs

Although simulated RIRs are important for evaluation purposes, it is necessary to assess performance in real room environments. To that end, we also evaluate our system using real RIRs from the Surrey binaural RIR (BRIR) database [16]. These RIRs are captured in real rooms from sine sweeps played through a loudspeaker, where the responses are deconvolved to produce the impulse response. The loudspeaker is placed along a radius of 1.5m away from the Head and Torso Simulator (HATS). The position of the loudspeaker is varied in 5° increments along the radius, where the center of the loudspeaker is placed at the same elevation as the ears of the HATS. For this study, we are focused on the monaural case, so the RIR of one of the ears is used. Specifically, when the loudspeaker is closer to the right ear, the right RIR is used and vice versa for the left ear. When the loudspeaker is at equal distance to the right and left ears, the left ear response is used. The RIRs are captured in four different room types. The dimensions of each room, the resulting T_{60} and DRR are shown in Table I.

Seven RIRs for each room (i.e. 28 total RIRs) are used to train a DNN. These real RIRs are convolved with the same 500 IEEE training utterances that are used in Section IV-B, resulting in 14000 total training utterances. The same 100 IEEE testing utterances from Section IV-B are convolved with 8 unseen real RIRs (2 per room) to produce a testing set of 800 reverberant signals.

The average results for these real RIRs are shown in Figure 5. Figure 5(a) shows the improvement in terms of PESQ. For each T_{60}, cIRM estimation produces the greatest improvement, and it substantially outperforms WPE and DSM. The STOI results are shown in Figure 5(b). cIRM estimation produces the largest increase in STOI at T_{60}s of 0.47 and 0.89s, while WPE performs best at 0.32 and 0.68s. DSM lowers the objective intelligibility of the reverberant speech for three of the T_{60}s. Lastly, Figure 5(c) shows the improvement in SNR_{fw}. cIRM estimation produces the highest SNR gain at 0.47 and 0.89s, whereas WPE performs best at 0.32 and 0.68s. Overall, the improvement of cRM over reverberant (and noisy) speech is slightly higher for real RIRs than for simulated RIRs.

D. Reverberation and noise

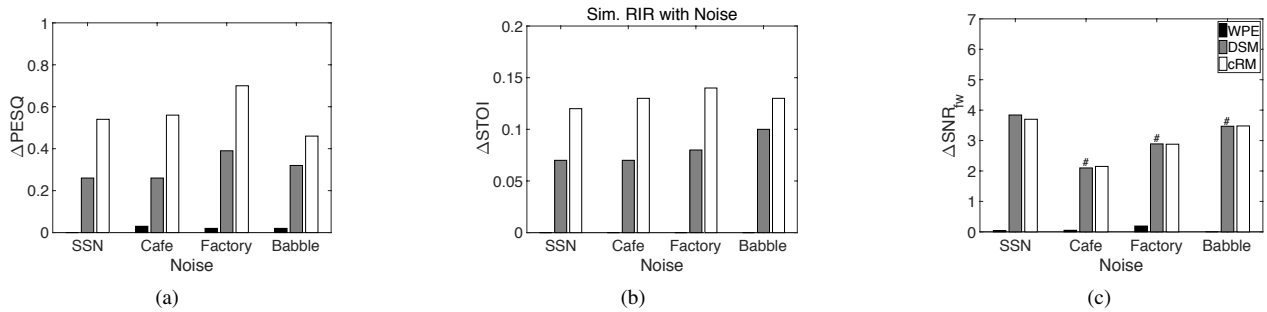
In real environments, reverberation and noise are both present. We test each system's ability to simultaneously perform dereverberation and denoising. For this scenario, the input to each system is reverberant and noisy speech features. The output target for the supervised systems are based on the direct speech.

To evaluate performance in this environment, we generate a set of RIRs for the speech (i.e. $h_s(t)$) and the noise (i.e. $h_n(t)$). The position of the speech and noise are randomly placed on a 1m radius from the microphone, where the elevations of the three components are equal. Eleven pairs of RIRs are generated for $h_s(t)$ and $h_n(t)$ at T_{60}s of 0.3, 0.6, and 0.9s, resulting in a total of 33 RIR pairs. Of these 33 RIR pairs, 30 (10 per T_{60}) are used in the training set, while the remaining 3 (1 per T_{60}) are used in the testing set. Four noise types are used: speech-shaped noise (SSN), cafe noise, factory noise, and babble noise. These noises are approximately 4 minutes in length. For training, random cuts from the first 2 minutes of the signals are used. The SNR in each case is set to 0 dB, where SNR is the ratio of energy between the reverberant speech and the reverberant noise. The training signals are mixed by combining 500 utterances with the 30 training RIRs and 4 noises (500 x 30 x 4 = 60000 training signals). Testing signals are generated by combining 100 utterances with the 3 testing RIRs and 4 noises (100 x 3 x 4 = 1200 testing signals). The testing noise signals are generated from random cuts of the last 2 minutes of the mentioned noises.

Figure 6 displays the performance by noise type, averaged over all T60s, for each system in noisy and reverberant conditions. In terms of PESQ, Figure 6(a), directly mapping to log-magnitude spectra using DSM improves PESQ by 0.26 points on average over the unprocessed noisy reverberant speech. Under these conditions cIRM estimation produces the largest gain of 0.54 points over the unprocessed speech on average. Note that WPE barely improves performance over the unprocessed speech, but this is partially expected since WPE is designed to deal with reverberation and not noise. Figure 6(b) shows the STOI improvement. cIRM estimation produces an improvement score of 0.13 on average which is clearly higher than the other approaches. For SNR_{fw}, Figure 6(c), DSM and cIRM estimation produce very similar improvements for each noise type.

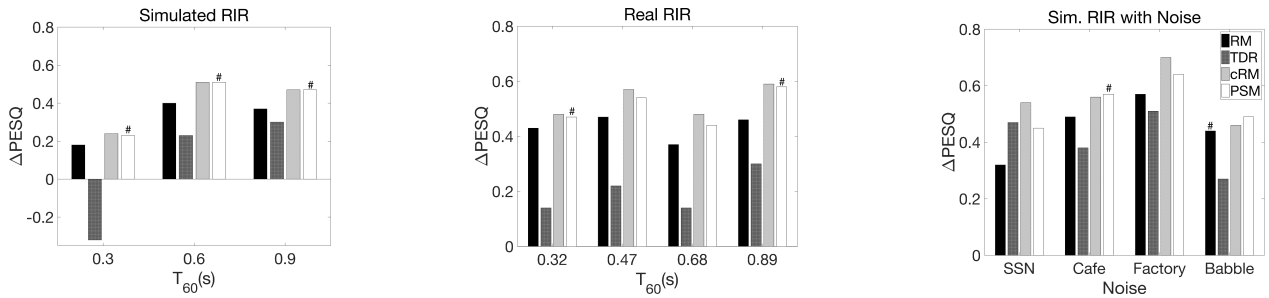
2329-9290 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2696307, IEEE/ACM Transactions on Audio, Speech, and Language Processing



(a) (b) (c)

Fig. 6. (a) ΔPESQ, (b) ΔSTOI, (c) ΔSNR$_{fw}$ results using simulated RIRs and background noise. The improvement relative to the unprocessed noisy-reverberant speech is shown. '#' indicates that the differences from cRM results are not statistically significant according to a one-way ANOVA test with 5% confidence interval.



(a) (b) (c)

Fig. 7. Supervised masking-based approaches are compared. The PESQ improvement is shown for (a) simulated RIRs, (b) real RIRs, and (c) simulated RIRs plus noise. '#' indicates that the differences from cRM results are not statistically significant according to a one-way ANOVA test with 5% confidence interval.

E. Supervised T-F mask comparisons

The PESQ results when cIRM estimation is compared to other supervised T-F masking approaches are shown in Figure 7. Figure 7(a) shows the PESQ improvement over the unprocessed reverberant speech when simulated RIRs are used. In this case, each approach, except TDR at 0.3s, improves objective quality. TDR produces the smallest gain on average followed by IRM estimation (RM). The benefit of enhancing the magnitude and phase spectra is shown in the results for cIRM estimation. The complex ratio mask improves performance over ratio masking at each T_{60}. PSM estimation performs similarly to cIRM estimation in all cases, except at 0.3s where cIRM estimation performs slightly better.

The PESQ improvement results using real RIRs are shown in Figure 7(b). All approaches improve objective speech quality over the unprocessed reverberant speech in this case. As with the simulated RIRs, TDR and IRM estimation offer the lowest gains. In each case, ratio masking in the complex domain (i.e. cRM) outperforms ratio masking in the magnitude domain (i.e. RM). Estimating the cIRM performs best for each T_{60} as well. A similar trend is exhibited when noise and simulated RIRs are used to generate noisy-reverberant speech, where cIRM estimation offers the largest improvement for SSN and Factory noise. In terms of STOI, ratio masking (i.e. RM) produces the highest improvements ($\sim$0.04) when simulated and real RIRs are used, while cRM and PSM closely follow (difference of $\sim$0.1). When simulated RIRs are used with noise, complex ratio masking produces the largest gain (0.13 compared to 0.12 for PSM). The cRM improvements over TDR in STOI are statistically significant when simulated and real RIRs are used. In simulated room responses with noise, cRM STOI improvements over TDR and RM are statistically significant. The SNR$_{fw}$ improvement results follow a similar trend.

F. Reverberation: unseen speakers and simulated rooms

To further test generalizability, we test cRM's ability to perform dereverberation in unseen rooms using utterances from unseen speakers. To accomplish this, we employ the training and testing setup as shown in Table II. The boldface rows indicate training rooms (4 in total), and the remaining rows represent unseen data (room or T_{60}). Six RIRs (2 per T_{60}) are generated for each of the training rooms (24 in total), where the corresponding average DRRs are shown in the third column of Table II. This allows for testing within and beyond the critical distance. The distance between the speaker and microphone is fixed at 1m, but the positions are randomly placed in the rooms. Each of the training RIRs is convolved with 500 utterances from the TIMIT speech corpus [9], using 10 utterances from each of 50 different speakers. For testing, six new RIRs are generated for each of the rooms (2 per T_{60}). The testing RIRs are convolved with 100 different utterances from 10 different speakers (10 utterances per speaker) from the TIMIT speech corpus.

We train and test the DNN for cIRM estimation in the environments described above, and we compare it to WPE

2329-9290 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2696307, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE II
CHARACTERISTICS OF THE SIMULATED ROOMS USED FOR TRAINING AND TESTING (DIM. STANDS FOR DIMENSION).

Dim. (meter)	T_{60} (second)	Avg. DRR (dB)
5 x 4 x 3	**[0.2, 0.3, 0.4]** [0.3, 0.5, 0.7]	**[2.62, -0.66, -2.69]** [-0.64, -3.90, -5.82]
6 x 4 x 3	[0.3, 0.8, 1.1]	[2.25, -0.85, -6.21]
7 x 5 x 4	**[0.3, 0.5 0.8]** [0.5, 0.8, 1.0]	**[3.11, -0.30, -2.90]** [-0.49, -3.15, -4.32]
7 x 6 x 4	[0.5, 0.9, 1.2]	[7.14, -0.15, -3.17]
8 x 7 x 5	**[0.4, 0.75, 1.1]** [0.75, 1.1, 1.3]	**[4.30, 0.43, -1.62]** [0.38, -1.65, -2.49]
8 x 7 x 6	[0.6, 1.2, 1.4]	[4.89, 2.28, -1.42]
9 x 8 x 7	**[0.6, 1.0, 1.2]** [1.0, 1.2, 1.4]	**[3.18, 0.75, -0.07]** [0.42, -0.39, -1.06]
10 x 9 x 7	[0.8, 1.2, 1.5]	[7.00, 3.88, 1.59]

TABLE III
AVERAGE PESQ RESULTS.

	T_{60}/DRR (1)		T_{60}/DRR (2)		T_{60}/DRR (3)	
	Seen	Unseen	Seen	Unseen	Seen	Unseen
Mixture	2.44	2.59	2.18	2.17	2.13	2.06
WPE	2.64	2.87	2.31	2.30	2.25	2.16
cRM	**2.84**	**2.99**	**2.56**	**2.52**	**2.49**	**2.37**

since it is an unsupervised approach. The average PESQ results for the unprocessed mixtures, WPE, and cRM are shown in Table III. The average results are shown by the type of room (seen during training, or unseen), and the T_{60} and average DRR combination. The '(i)' in Table III refers to the average results over the ith T_{60} value and the ith Avg. DRR value across all rooms from Table II. Notice that the proposed cRM clearly outperforms WPE and the unprocessed mixtures in all cases, indicating its ability to generalize to unseen rooms and speakers. The differences are statistically significant in each case.

G. Reverberation: real RIRs and multiple speakers

In addition to the above tests, we further evaluate our approach using real RIRs and multiple speakers. Each method is trained using the 500 utterances (10 from each of 50 speakers) and tested using 100 utterances (10 utterances from each of 10 different speakers) as mentioned previously. The DNN is trained using 15000 reverberant mixtures generated by convolving the training utterances with 30 different real RIRs. The training RIRs are captured from three of the four rooms in Table I using the Surrey BRIR database [16]. RIRs from the fourth room are held out for testing, and we rotate the room that is unseen during training. Ten different RIRs are used for training in each room. The testing utterances are convolved with 8 RIRs, 2 from each of the four rooms listed in Table I. Therefore, this test set evaluates each approach using unseen speakers, unseen RIRs from seen rooms, and unseen RIRs from an unseen room. We compare with several approaches, which are trained as previously described.

Table IV shows the average PESQ scores for each approach across each room and training set, where one room (indicated by **BOLD**) is held out during training. The results reveal that cRM and PSM perform similarly and the best overall for seen and unseen rooms. A one-way ANOVA test (5% confidence interval) shows that cRM improvements over the other comparison methods are statistically significant.

V. DISCUSSION AND CONCLUSION

Our approach significantly improves dereverberation and denoising performance over unprocessed signals. It also outperforms most methods in terms of objective speech quality and intelligibility metrics. Our informal listening to enhanced signals indicates that perceptual quality is consistent with objective results. Most importantly, the results reveal that magnitude and phase are both important for quality, so they both should be enhanced. The joint enhancement of magnitude and phase is the main reason cIRM estimation outperforms IRM estimation and DSM. Incorporating magnitude and phase information is the main reason why PSM estimation performs well.

A. Ideal performance of T-F masking approaches

An important comparison is between the ideal performance of the T-F masking approaches. The average ideal PESQ results for IRM, cIRM, and PSM are 3.53, 4.5, and 3.61, respectively. Notice that only the cIRM is capable of producing the maximum attainable PESQ score, due to its enhancement of magnitude and phase. PSM is not capable of fully recovering the speech. PSM estimation is close to cIRM estimation likely due to the challenge of estimating the imaginary portion of the cIRM, which is less structured than the real component. This indicates that refinements for estimating the imaginary component should be developed.

B. Complex-domain DNN

Section III-C describes how a standard DNN with real components (weights, biases, activation function) is used to jointly estimate the complex components of the cIRM. Since the real and imaginary components of the cIRM are related, it is important to determine if a DNN can further capitalize on this relationship. One way to take advantage of this relationship is to utilize a complex-domain DNN, where the inputs, weights, biases, activation functions, and outputs are all complex. For this purpose, we have defined a complex-domain DNN and used it to either estimate the cIRM or the STFT of direct speech. The structure of the complex-domain DNN matches that of the standard DNN (see Figure 3), except a single layer in the output layer is used. The complex weights are randomly initialized. A complex hyperbolic tangent function is defined and used as the activation function in each layer, where the real and imaginary components of this activation function are defined similarly to Eq. (8). Complex domain backpropagation is used [10], [24]. This complex domain DNN is evaluated using the same experimental setup as defined in Sections IV-B to IV-D. The input to the complex DNN is the STFT of reverberant (and noisy) speech.

The experimental results for estimating the STFT of direct speech and the cIRM are shown in Table V. It is clear from

2329-9290 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2696307, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE IV
AVERAGE PESQ SCORES FOR EACH APPROACH WHEN THE SYSTEMS ARE TRAINED AND TESTED USING MULTIPLE SPEAKERS. **BOLD** SCORE IDENTIFIES THE BEST PERFORMING SYSTEM, WHILE **BOLD** ROOM INDICATES THE ONE NOT SEEN DURING TRAINING.

	Set 1				**Set 2**			
	Room A	Room B	Room C	**Room D**	**Room A**	Room B	Room C	Room D
Mixture	2.93	2.64	2.92	2.39	2.93	2.64	2.92	2.39
RM	3.27	2.98	3.21	2.64	3.27	3.04	3.26	2.74
cRM	3.41	3.12	**3.38**	2.79	3.38	**3.20**	**3.43**	2.93
PSM	**3.42**	**3.13**	3.37	**2.80**	**3.39**	**3.20**	**3.43**	**2.94**
WPE	3.28	2.95	3.36	2.61	3.28	2.95	3.36	2.61
	Set 3				**Set 4**			
	Room A	**Room B**	Room C	Room D	Room A	Room B	**Room C**	Room D
Mixture	2.93	2.64	2.92	2.39	2.93	2.64	2.92	2.39
RM	3.28	2.99	3.25	2.71	3.31	3.02	3.23	2.74
cRM	**3.42**	3.14	**3.40**	2.88	3.42	3.15	3.32	2.88
PSM	**3.42**	**3.15**	3.39	**2.90**	**3.43**	**3.16**	**3.33**	**2.90**
WPE	3.28	2.95	3.36	2.61	3.28	2.95	3.36	2.61

TABLE V
AVERAGE PESQ RESULTS WHEN A COMPLEX-DOMAIN DNN IS USED TO ESTIMATE THE STFT OF DIRECT SPEECH AND THE CIRM.

	Sim. RIR	Real RIR	Sim. RIR + Noise
cRM - stand. DNN	3.42	3.35	2.39
STFT - complex DNN	1.86	1.76	1.71
cRM - complex DNN	2.90	2.80	2.06

TABLE VI
AVERAGE PESQ IMPROVEMENT SCORES WHEN THE TARGET IS DIRECT SIGNAL WITH EARLY REFLECTIONS. IMPROVEMENT IS RELATIVE TO THE UNPROCESSED MIXTURE. **BOLD** IDENTIFIES THE SYSTEM THAT PERFORMED BEST.

	Sim. RIR	Real RIR
WPE	0.23	0.26
RM	0.19	0.23
cRM	**0.31**	**0.33**
PSM	0.30	**0.33**

the results that cIRM estimation using a standard DNN is superior. The result for STFT estimation in reverberant and noisy conditions is similar to the result for STFT in noisy environments (see [42]), where both approaches are not effective. Estimating the cIRM with the complex DNN performs better than STFT estimation, but not as well as when the cIRM is estimated using the DNN defined in Section III-C. This may be attributed to a few reasons, but one may be the activation function that is used. The hyperbolic tangent function suffers from vanishing gradient that affects performance, and this may be the case here. We also ran experiments using the rectified linear activation function defined for the real and imaginary components, but this does not perform any better.

Although WPE is complex, its worth noting that cIRM estimation outperforms it largely due to the benefit of supervised learning. It must be pointed out that WPE only deals with late reverberation, so the comparison might not be truly fair. To address this, we also define the cIRM (and other approaches) with the direct sound plus early reverberation as the target, so it only removes late reverberation. Table VI shows PESQ improvement scores for each approach, respectively. The PESQ results show that cIRM estimation still outperforms WPE when simulated and real RIRs are used, but these results are not as good as when early and late reverberation are removed (see Section IV). Its also worth pointing out that WPE is an utterance based approach, meaning that it processes the entire utterance multiple times before the final dereverberant signal is produced. This differs from the DNN based approaches, which only use a small sliding window to generate speech estimates for a single time frame.

We investigated other approaches for estimating the cIRM. We separately train DNNs to estimate the real and imaginary components, and we jointly estimate the absolute value (i.e. instantaneous amplitude without sign) and sign (positive or negative) of the cIRM. Additionally, we experimented with computing the imaginary component from an estimated IRM and the real component. These cases, however, did not perform as well as the proposed approach. We also conducted experiments using the following features that contain phase information: magnitude and phase, real and imaginary components, or the complementary feature set extracted from the real and imaginary components of reverberant speech. These features, however, did not perform as well as the complementary set. The cIRM amounts to scaling the IRM by factors between -1 and 1 based on the cosine and sine of the phase difference (see Eq (7)). We think that the nonlinear nature of the DNN and the usage of backpropagation enable the DNN to jointly estimate the scaled versions of the IRM without including phase in the input feature set.

In conclusion, we have proposed a supervised learning approach to separate speech in reverberant and noisy environments. We show how the cIRM can be used, where it enhances the magnitude and phase response of an observation. By addressing the magnitude and phase, the cIRM is capable of producing clean and anechoic speech estimates. We train a deep neural network to estimate the cIRM from noisy and reverberant speech, and its performance is consistent using simulated and real room impulse responses and when reverberant noise is present.

ACKNOWLEDGMENT

We would like to thank Takuya Yoshioka and Tomohiro Nakatani for answering questions about their WPE approach,

2329-9290 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2696307, IEEE/ACM Transactions on Audio, Speech, and Language Processing

and Kun Han for providing the DSM implementation. We also thank the anonymous reviewers for their helpful comments and suggestions. This research was supported in part by an AFOSR grant (FA9550-12-1-0130), an NIH grant (R01 DC012048), and the Ohio Supercomputer Center.

REFERENCES

[1] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, 1979.

[2] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *J. Acoust. Soc. Am*, vol. 21, pp. 577–580, 1949.

[3] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Am*, vol. 113, pp. 3233–3244, 2003.

[4] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, pp. 41–75, 1997.

[5] C. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, pp. 257–270, 2007.

[6] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, and A. Nakamura, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," in *Proc. REVERB Challenge*, 2014.

[7] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, 2010.

[8] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, 2015, pp. 708–712.

[9] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus," *Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html*, 1993.

[10] G. M. Georgiou and C. Koutsougeras, "Complex domain backpropagation," *IEEE Trans. Circuits and Systems II: Analog and Digital Signal Processing*, vol. 39, pp. 330–334, 1992.

[11] T. Gerkmann, M. Krawczyk-Becker, and J. L. Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, pp. 55–66, 2015.

[12] E. Habets, "Room impulse response generator (http://home.tiscali.nl/ehabets/rir generator.html)," 2010.

[13] 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 and Lang. Process.*, vol. 23, pp. 982–992, 2015.

[14] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, pp. 1738–1752, 1990.

[15] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 2, pp. 578–589, 1994.

[16] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Trans. Audio Speech and Lang. Process.*, vol. 18, pp. 1867–1871, 2010.

[17] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.

[18] ITU-R, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," p. 862, 2001.

[19] Y. Jiang, D. L. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Trans. Audio Speech and Lang. Process.*, vol. 22, pp. 2112–2121, 2014.

[20] Z. Jin and D. L. Wang, "Supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio Speech and Lang. Process.*, vol. 17, pp. 625–638, 2009.

[21] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. ICASSP*, 2009, pp. 45–48.

[22] B. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," in *Proc. ICASSP*, 1997, pp. 1259–1262.

[23] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *J. Acoust. Soc. Am.*, vol. 95, pp. 1593–1602, 1994.

[24] H. Leung and S. Haykin, "The complex backpropagation algorithm," *IEEE Trans. on Sig. Process.*, vol. 39, pp. 2101–2104, 1991.

[25] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am*, vol. 125, pp. 3387–3405, 2009.

[26] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acous. Speech Sig. Proc.*, vol. 36, pp. 145–152, 1988.

[27] N. Mohammadiha and S. Doclo, "Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling," *IEEE/ACM Trans. Audio Speech and Lang. Process.*, vol. 24, pp. 276–289, 2016.

[28] A. K. Nabelek and J. M. Picket, "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *J. Speech Hear. Res.*, vol. 17, pp. 724–739, 1974.

[29] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am*, vol. 66, pp. 165–169, 1979.

[30] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1999.

[31] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, pp. 465–494, 2010.

[32] B. D. Radlovic, R. C. Williamson, and R. A. Kennedy, "Equalization in an acoustic reverberant environment: robustness results," *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 311–319, 2000.

[33] N. Roman and D. L. Wang, "Pitch-based monaural segregation of reverberant speech," *J. Acoust. Soc. Am*, vol. 120, pp. 458–469, 2006.

[34] N. Roman and J. Woodruff, "Speech intelligibility in reverberation with ideal binary masking: effects of early reflections and signal-to-noise ratio threshold," *J. Acoust. Soc. Am.*, vol. 133, pp. 1707–1717, 2013.

[35] R. Singh, B. Raj, and P. Smaragdis, "Latent-variable decomposition based dereverberation of monaural and multi-channel signals," in *Proc. ICASSP*, 2010, pp. 1914–1917.

[36] 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 and Lang. Process.*, vol. 19, pp. 2125–2136, 2011.

[37] Y. Wang, K. Han, and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. Audio Speech and Lang. Process.*, vol. 21, pp. 270–279, 2013.

[38] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio Speech and Lang. Process.*, vol. 22, pp. 1849–1858, 2014.

[39] Y. Wang and D. L. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. ICASSP*, 2015, pp. 4390–4394.

[40] F. Weninger, S. Watanabe, Y. Tachioka, and B. Schuller, "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition," in *Proc. ICASSP*, 2014, pp. 4623–4627.

[41] D. S. Williamson and D. L. Wang, "Speech dereverberation and denoising using complex ratio masks," in *Proc. ICASSP*, 2017, pp. 5590–5594.

[42] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio Speech and Lang. Process.*, vol. 24, pp. 483–492, 2016.

[43] M. Wu and D. L. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio Speech and Lang. Process.*, vol. 14, pp. 774–784, 2006.

[44] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio Speech and Lang. Process.*, vol. 20, pp. 2707–2720, 2012.

[45] X. Zhao, Y. Wang, and D. L. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE Trans. Audio Speech Lang. Process.*, vol. 22, pp. 836–845, 2014.

Donald S. Williamson received the B.E.E degree in electrical engineering from the University of Delaware, Newark, in 2005 and the M.S. degree in electrical engineering from Drexel University, Philadelphia, PA, in 2007. He received the Ph.D. degree in computer science and engineering at The Ohio State University, Columbus in 2016. He is currently an Assistant Professor in the Department of Computer Science at Indiana University. His research interests include speech separation, robust automatic speech recognition, and music processing.

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

2329-9290 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.