SPEECH DEREVERBERATION AND DENOISING USING COMPLEX RATIO MASKS

Donald S. Williamson[1], *and DeLiang Wang*[2,3]

[1]School of Informatics and Computing, Indiana University, USA
[2]Department of Computer Science and Engineering, The Ohio State University, USA
[3]Center for Cognitive and Brain Sciences, The Ohio State University, USA
williads@indiana.edu, dwang@cse.ohio-state.edu

ABSTRACT

Traditional speech separation systems enhance the magnitude response of noisy speech. Recent studies, however, have shown that perceptual speech quality is significantly improved when magnitude and phase are both enhanced. These studies, however, have not determined if phase enhancement is beneficial in environments that contain reverberation as well as noise. In this paper, we present an approach that jointly enhances the magnitude and phase of reverberant and noisy speech. We use a deep neural network to estimate the real and imaginary components of the complex ideal ratio mask (cIRM), which results in clean and anechoic speech when applied to a reverberant-noisy mixture. Our results show that phase is important for dereverberation, and that complex ratio masking outperforms related methods.

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

1. INTRODUCTION

Reverberation adversely affects perceptual speech quality and intelligibility, because sound reflections smear speech structure across time and frequency. This presents challenges for many applications, such as, automatic speech recognition (ASR) [1], speaker identification [2], and hearing aid design. Reverberation is also debilitating for individuals with hearing impairments [3, 4].

Many techniques have been proposed for speech dereverberation. In [5], Weninger *et al.* perform dereverberation with a deep bi-directional Long Short-Term Memory (LSTM) recurrent neural network. They use this network to estimate the log mel-spectral magnitudes of clean speech from the log mel-spectral magnitudes of reverberant speech. Han *et al.* learn a spectral mapping to clean spectral magnitudes, using a deep neural network (DNN) [6]. Other spectral-magnitude based approaches employ inverse filtering [7] or non-negative matrix factorization (NMF) [8].

The above approaches address the magnitude response of reverberant speech. A study by Paliwal *et al.*, however, shows that the phase response is important for improving the perceptual quality of noisy speech [9]. Different phase enhancement approaches are discussed in [10, 11, 12]. Phase enhancement only addresses the phase response, so separate enhancement of the magnitude response is needed. Our recent approach estimates the complex ideal ratio mask (cIRM), which jointly enhances the magnitude and phase of noisy speech [13]. It has been shown to perform very well in various noisy environments, and it substantially outperforms related methods. The benefit of complex ratio masking is that anechoic speech results when the ideal mask is applied. Complex ratio masking, however, has not be investigated in environments that contain reverberation.

Although many approaches have been proposed for dereverberation, their performance is limited since they cannot fully reconstruct anechoic speech. Additionally, reverberation and noise are both present in real-world environments, which compounds an already challenging situation. In fact, it has been shown that the speech intelligibility for normal hearing and hearing impaired listeners is worsened under this condition [14, 15].

In this paper, we propose to use the cIRM for speech dereverberation and denoising. Features are extracted from reverberant and noisy speech, where these features are supplied to a DNN for cIRM estimation. More specifically, the DNN is trained to jointly estimate the real and imaginary components of the cIRM. The definition of the cIRM is modified to deal with reverberant and noisy spectra. The desired output is the anechoic speech spectra.

The rest of this paper is organized as follows. Section 2 discusses the relation to prior work. A detailed description of our approach is given in Section 3. The experiments and results are given in Section 4. Finally, a conclusion is given in Section 5.

2. RELATION TO PRIOR WORK

The work presented here focuses on speech dereverberation and denoising in the complex domain. Previous studies on this topic perform dereverberation and denoising in the spectral-magnitude domain [5, 6]. Although complex domain dereverberation is presented in [16, 17], their approach is unsupervised and does not handle background noise. It also is an utterance based approach that repeatedly processes the entire test signal. Our approach on the other hand, only requires small time segments.

3. ALGORITHM DESCRIPTION

Our algorithm uses a deep neural network to spectrally map features extracted from reverberant and noisy speech to the cIRM. This section begins by describing the cIRM. We then describe the feature extraction process and give details about the DNN.

3.1. Complex Ideal Ratio Mask (cIRM)

The complex ideal ratio mask is generated from reverberant (and noisy) speech and the direct (anechoic) speech signal. It is defined so that the product of the cIRM and reverberant observation results in direct speech. This occurs in the time-frequency (T-F) domain, so T-F representations for the reverberant observation and direct speech are needed. Given the short-time Fourier transform (STFT) of reverberant speech, $Y(t, f)$, and the cIRM, $M(t, f)$, direct speech, $D(t, f)$, is computed as follows:

$$D(t, f) = M(t, f) * Y(t, f) \tag{1}$$

where t and f index time and frequency, respectively. Since the STFT is complex, '*' indicates complex multiplication. From Eq. (1), it is clear that the cIRM is computed by dividing the STFT of direct speech, with the STFT of reverberant speech:

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

The exact calculation at each T-F unit is shown after expanding $Y(t, f)$ and $D(t, f)$ into their complex representations, where subscripts r and i indicate the real and imaginary components, respectively.

The cIRM can also be written in polar form, as shown below:

$$\begin{aligned} M(t, f) &= \frac{|D(t, f)|e^{j\phi_d(t,f)}}{|Y(t, f)|e^{j\phi_y(t,f)}} \\ &= \frac{|D(t, f)|}{|Y(t, f)|}e^{j(\phi_d(t,f)-\phi_y(t,f)} \end{aligned} \tag{3}$$

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, indicating that magnitude and phase are both enhanced when it is applied.

The real and imaginary components of the cIRM, M_r and M_i, may have large values in the range $(-\infty, \infty)$. This may be problematic for supervised learning with deep neural networks. To alleviate this problem, we compress the components of the cIRM using the following hyperbolic tangent.

$$M_x'(t, f) = Q\frac{1 - e^{-C \cdot M_x(t,f)}}{1 + e^{-C \cdot M_x(t,f)}} \tag{4}$$

where $x \in \{r, i\}$, denoting the real or imaginary component. M' is the compressed cIRM. After compression, the complex components are within $[-Q, Q]$. C controls the steepness of the hyperbolic tangent.

3.2. Feature Extraction

A complementary feature set is computed from the reverberant (and noisy) signal [18]. This set includes amplitude modulation spectrogram (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP), 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 [19], and they work well for cIRM estimation in noisy speech [13].

We use temporal dynamics to capture the correlations between adjacent frames of the feature set, $\boldsymbol{F}$. Specifically, we join adjacent frames into a single feature vector. The augmented feature vector, $\tilde{\boldsymbol{F}}$, centered at the t^{th} time frame is as follows:

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

where p denotes the number of adjacent frames to include on each side. The augmented feature set is then normalized to have zero mean and unit variance. After normalization, auto-regressive moving average (ARMA) filtering is performed [20].

3.3. cIRM Estimation

We use a deep neural network to estimate the cIRM. The DNN is trained to spectrally map the reverberant (and noisy) features to the cIRM. Figure 1 shows the network structure of the DNN.

The DNN is trained to map a single frame of the augmented feature vector to a single frame of the cIRM (real and imaginary). This is accomplished with a four layer DNN, where each of the hidden layers has 1024 units. Rectified linear activation functions are used in the hidden layer. Two

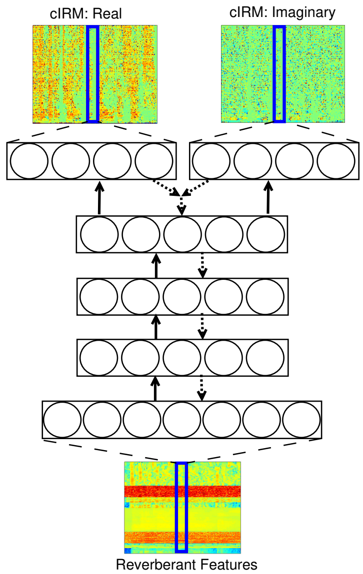

Fig. 1. (Color Online). DNN network structure for cIRM estimation.

separate sub-output layers are used for the real and imaginary components of the cIRM, respectively. Linear activation functions are used in the output layer. The DNN is trained using the standard back propagation algorithm with mean-square error cost function

$$\frac{1}{2N}\sum_t\sum_f[(\hat{M}_r^{'}(t,f)-M_r^{'}(t,f))^2+(\hat{M}_i^{'}(t,f)-M_i^{'}(t,f))^2] \tag{6}$$

where $\hat{M}_r^{'}(t,f)$ and $\hat{M}_i^{'}(t,f)$ are the estimated real and imaginary components that are generated by the DNN. N is the number of time frames for the input.

The DNN estimates compressed values for the real and imaginary components of the cIRM (see Eq. 4). During testing, these values are uncompressed using the following:

$$\hat{M}_x = -\frac{1}{C}\log(\frac{Q-\hat{M}_x^{'}}{Q+\hat{M}_x^{'}}) \tag{7}$$

where $\hat{M}_x$ is an estimate of the uncompressed component (i.e. $\hat{M}_r$ or $\hat{M}_i$). The uncompressed estimates are then used to extract an estimate of the direct or anechoic speech.

4. EXPERIMENTS

Three different experiments are conducted to evaluate the performance of our proposed approach. For each experiment, our proposed approach is compared to ideal ratio mask (IRM) estimation (denoted as RM) [19], phase sensitive mask (PSM) estimation [21], and direct estimation of spectral magnitudes (DSM) [6]. The IRM is a magnitude domain approach, whereas PSM incorporates magnitude and phase information. The PSM is equivalent to the real component of the cIRM. DNNs are trained to estimate these targets, where the basic DNN configuration (features and network structure) match that described in Sections 3.2 and 3.3. Only the DNN for DSM uses different input features (i.e. log spectral-magnitudes), since we found that this works best. We also compare to weighted error prediction (WPE), which also operates in the complex domain [16, 17]. Note that PSM estimation has not been previously evaluated for dereverberation.

The STFTs for each approach are computed by divding a signal into 32ms time frames with 75% overlap between adjacent frames. The fast Fourier transform (FFT) is then computed within each time frame using a 512-point FFT. The sampling rate for all test signals is 16 kHz. For feature augmentation, it is empirically determined that p be set to 2. Similarly, we set Q to 1 and C to 0.5 for cIRM compression.

The perceptual evaluation of speech quality (PESQ) [22] and the frequency-weighted segmental signal-to-noise ratio (SNR_{fw}) [23] are used to evaluate performance.

4.1. Experiment 1: One room and one speaker

We first evaluate dereverberation performance using simulated room impulse responses (RIRs), where the simulated RIRs are generated using the imaging method [24]. RIRs are generated by placing a target speaker and microphone in random positions throughout a simulated room of size 9 x 8 x 7m. The distance between the speaker and microphone is fixed at 1m. Eleven RIRs are generated at T_{60}s (the time taken for a direct sound to attenuate by 60 dB) of 0.3, 0.6, and 0.9 seconds, resulting in 33 RIRs. During training, 30 of the RIRs (10 for each T_{60}) are convolved with 500 utterances from the IEEE corpus [25]. These utterances are spoken by a single male speaker. During testing, 100 different utterances are convolved with the remaining 3 RIRs, resulting in 300 test signals.

The average PESQ and SNR_{fw} results at each T_{60} are shown in Table 1, where the best performing systems are shown in **bold**. In terms of PESQ, at each T_{60} cRM clearly outperforms DSM and RM, whereas its performance is identical to PSM. At T_{60}s of 0.6 and 0.9s, cRM noticeably outperforms WPE. In terms of SNR_{fw}, cRM outperforms all other approaches, except for RM at 0.3 and 0.6.

Table 1. Average PESQ and SNR_{fw} scores for experiment 1.

	PESQ			SNR_{fw}		
	0.3	0.6	0.9	0.3	0.6	0.9
Mixture	3.67	2.82	2.53	15.75	10.63	8.71
DSM	3.57	3.15	2.88	13.23	11.43	10.19
RM	3.85	3.22	2.90	**16.64**	**12.85**	11.02
cRM	**3.91**	**3.33**	**3.00**	16.54	12.70	**11.07**
PSM	3.90	**3.33**	**3.00**	15.80	12.21	10.70
WPE	**3.91**	3.10	2.70	15.74	12.43	10.15

Table 2. Average PESQ and SNR_{fw} scores for experiment 2.

	PESQ			SNR_{fw}		
	0.3	0.6	0.9	0.3	0.6	0.9
Mixture	3.54	2.30	2.60	18.16	8.31	8.94
DSM	3.33	2.39	2.57	11.24	7.70	8.49
RM	3.73	2.62	2.89	17.54	9.90	9.98
cRM	**3.86**	**2.78**	**3.02**	17.63	**10.35**	**10.07**
PSM	**3.86**	2.75	3.01	17.73	10.32	9.99
WPE	3.85	2.47	2.86	**18.35**	9.03	9.66

Table 3. Average PESQ and SNR_{fw} scores for experiment 3.

	PESQ		SNR_{fw}	
	SSN	Factory	SSN	Factory
Mixture	1.93	1.69	3.45	3.16
DSM	2.19	2.08	**7.29**	6.05
RM	2.25	2.26	5.62	**6.38**
cRM	**2.47**	**2.39**	7.15	6.04
PSM	2.38	2.33	6.41	6.21
WPE	1.93	1.71	3.49	3.35

4.2. Experiment 2: Three rooms and many speakers

In this section, dereverberation performance is evaluated using utterances from many speakers and multiple rooms. Specifically, simulated RIRs are generated in three different rooms. The first room has dimensions of 9 x 8 x 7m, the second room has dimensions of 6 x 6 x 10m, and the third room's dimensions are 8 x 10 x 4m. Our DNN in this case is trained from RIRs generated in the first two rooms. Fifteen RIRs are generated in each of these rooms using T_{60}s of 0.3, 0.6, and 0.9 seconds (5 per T_{60}). This results in 30 RIRs that are used for training. These RIRs are convolved with 500 utterances from 50 different speakers (i.e. 10 utterances per speaker) using the TIMIT corpus [26]. Of the 50 speakers, 35 are male and 15 are female. During testing, three unseen RIRs are generated, one from each room (rooms 1, 2, and 3). The RIR from room 1 has a T_{60} of 0.3, the RIR from room 2 has a T_{60} of 0.9, and the RIR from room 3 has a T_{60} of 0.6. These three RIRs are convolved with 100 utterances that are generated from 10 different speakers (10 utterances per speaker), where 7 male and 3 female speakers are used. Thus, this experiment tests on unseen RIRs, rooms, and speakers.

Table 2 shows the results for this experiment. The PESQ performance is very much similar to the PESQ results from Section 4.1, where cRM and PSM perform best. In terms of SNR_{fw}, cRM performs best at T_{60}s of 0.6 and 0.9.

4.3. Experiment 3: One room, one speaker, and noise

This last experiment evaluates dereverberation and denoising performance. A set of simulated RIRs are generated for speech and noise, where the position of the speech and noise are randomly placed on a 1m radius from a microphone in a single room. This room has the same dimensions as in Experiment 1. Eleven pairs of RIRs are generated at T_{60}s of 0.3, 0.6, and 0.9s, resulting in 33 RIR pairs. Of these pairs, 30 (10 per T_{60}) are used during training, while the other 3 (1 per T_{60}) are used during testing. The same 500 training and 100 testing utterances used in Experiment 1 are also used here. For noises, speech-shaped noise (SSN) and factory noise are used. These noises are approximately 4 minutes in length, where random cuts from the first half of the signal are used for training and random cuts from the later half of each signal are used during testing. The utterances and random cuts of noise are each convolved with the corresponding RIR from the pair of 30 for training, and the remaining 3 for testing. The SNR is set to 0 dB, where SNR is the ratio of energy between reverberant speech and reverberant noise.

The dereverberation and denoising results are shown in Table 3, where the average PESQ and SNR_{fw} scores are shown for each noise type. Our proposed approach, cRM, outperforms all other approaches, in terms of PESQ, for both noises.

5. CONCLUSIONS

We have proposed a deep learning approach for speech dereverberation and denoising. This approach enhances the magnitude and phase of reverberant-noisy speech by operating in the complex domain. This enables the complex ratio mask to fully reconstruct anechoic speech. A deep neural network estimates the real and imaginary components of the cIRM. Our results show that cIRM estimation consistently outperforms directly estimating spectral magnitudes (i.e. DSM) and ratio masking in the magnitude domain (i.e. RM). When simultaneously performing dereverberation and denoising, complex ratio masking also outperforms WPE and PSM approaches. It is also worth noting that the cIRM is capable of producing maximum PESQ and SNR_{fw} scores. The challenge of estimating the imaginary component of the cIRM, which is less structured, likely causes PSM estimation to perform similarly, suggesting room for further improvement.

6. REFERENCES

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

[2] 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.

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

[4] 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.

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

[6] 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.

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

[8] 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.

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

[10] P. Mowlaee, R. Saeidi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation.," in *Proc. INTERSPEECH*, 2012, pp. 1–4.

[11] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio Speech Lang Process.*, vol. 22, pp. 1931–1940, 2014.

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

[13] 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.

[14] E. L. J. George, S. T. Goverts, J. M. Festen, and T. Houtgast, "Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners," *J. Speech, Lang., and Hear. Res.*, vol. 53, pp. 1429–1439, 2010.

[15] N. Roman and John 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

[23] 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.

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

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

[26] 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.