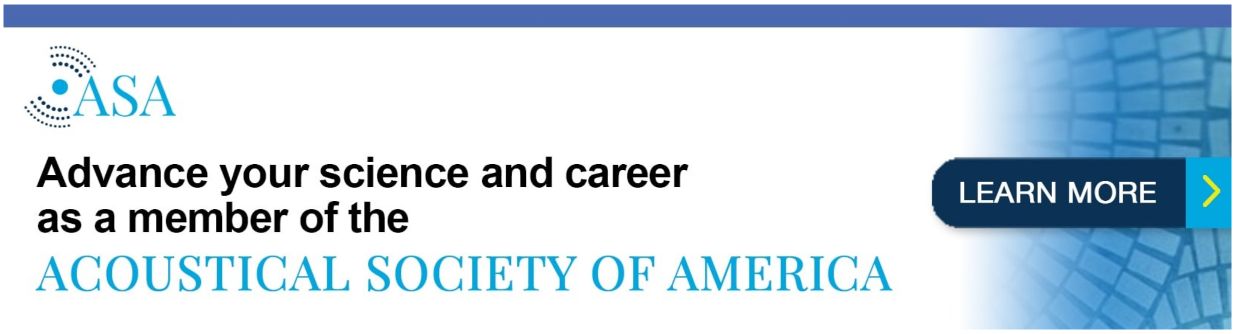
On the importance of power compression and phase estimation in monaural speech dereverberation

Cite as: JASA Express Lett. 1, 014802 (2021); https://doi.org/10.1121/10.0003321
Submitted: 30 September 2020 • Accepted: 21 December 2020 • Published Online: 15 January 2021

Andong Li, Chengshi Zheng, Renhua Peng, et al.



JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

On the importance of power compression and phase estimation in monaural speech dereverberation

Andong Li, Chengshi Zheng, Renhua Peng,[a] and Xiaodong Li
Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
liandong@mail.ioa.ac.cn, cszheng@mail.ioa.ac.cn, pengrenhua@mail.ioa.ac.cn, lxd@mail.ioa.ac.cn

Abstract: Previous studies have shown the importance of introducing power compression on both feature and target when only the magnitude is considered in the dereverberation task. When both real and imaginary components are estimated without power compression, it has been shown that it is important to take magnitude constraint into account. In this paper, both power compression and phase estimation are considered to show their equal importance in the dereverberation task, where we propose to reconstruct the compressed real and imaginary components (cRI) for training. Both objective and subjective results reveal that better dereverberation can be achieved when using cRI.

[Editor: Paul Gendron] https://doi.org/10.1121/10.0003321

Received: 30 September 2020 **Accepted:** 21 December 2020 **Published Online:** 15 January 2021

1. Introduction

In a reverberant enclosure, speech signals often contain lots of reflected components, which can be roughly divided into the early and the late reverberation parts. Previous studies have shown that both speech quality and intelligibility at the receiver may heavily degrade, which can greatly damage the performance of hands-free speech communication.[1] To address the problem, a multitude of approaches have been explored in the past half-century. Conventional dereverberation methods usually include spectral subtraction,[2] inverse filtering,[3] weighted prediction error (WPE),[4] and so on.

Recently, the introduction of deep neural networks (DNNs) has facilitated the rapid development of research toward speech dereverberation.[5–8] When given the reverberant speech feature as input and its corresponding clean counterpart as target, DNN can be utilized as a nonlinear function to learn the spectral mapping from reverberant to anechoic speech. For instance, given a large scale of training pairs with various reverberant time (RT_{60}) and room configurations, the network can gradually grasp the complicated reverberant patterns during the learning process and effectively recover the anechoic component even in very high reverberant environments. This paper focuses on speech dereverberation when only one microphone is available.

When it comes to input features, there are usually two categories for magnitude-based approaches, i.e., the raw magnitude of spectrum (rMS) and the compressed magnitude of spectrum (cMS). For the first type, the magnitude of the reverberant spectrum is directly extracted as the network feature and its corresponding anechoic target is then estimated. For the second one, the range of magnitude value will be restricted with a nonlinear compression function before sent to the network, e.g., logarithm operation[5] or cubic root operation.[8] However, it still remains controversial whether feature compression improves the dereverberation performance or not, especially in subjective speech quality. Meanwhile, no optimal compression function has been investigated yet. More recently, the importance of phase information began to be emphasized in both speech enhancement and dereverberation tasks.[7,9] Reported results showed that when both magnitude and phase were optimized by estimating real and imaginary (RI) components simultaneously, the speech quality could be further improved than magnitude-only approaches.[7]

At this point, one may ask the question: *Can power compression and phase estimation combine together to further improve the performance of dereverberation?* To the best of our knowledge, this article is the first time they have been considered together in dereverberation. The main problem is how to integrate the phase into the compressed magnitude, and thus we can train the model with the compressed magnitude and the original phase. To this end, this article proposes to couple the compressed magnitude and original phase together, where only the magnitude is compressed with the phase unaltered. As a result, the reverberant and anechoic real and imaginary components (cRI) serve as the feature and the target, respectively. Experimental results demonstrate that under the same network configuration, a suitable compression scheme can improve the performance of the rMS-based method in both perceptual evaluation of speech quality (PESQ)[10]

[a] Author to whom correspondence should be addressed.

and frequency-weighted segmental signal-to-noise ratio (SNR_{fw}).[11] Moreover, when phase information is also integrated into optimization, a further performance improvement can be obtained in both objective and subjective tests.

2. Method

2.1 Problem Formulation

In time domain, $s(t)$ and $h(t)$ are assumed to be anechoic speech and room impulse response (RIR), respectively. The reverberant speech can be modeled as

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

where $*$ refers to convolution operation. With short time Fourier transform (STFT), Eq. (1) can be transformed into time-frequency (T-F) domain, given by

$$X(k,l) = S(k,l)H(k,l), \tag{2}$$

where $X(k,l)$, $S(k,l)$, and $H(k,l)$ denote the STFT representations of $x(t)$, $s(t)$, and $h(t)$ in the frequency index of k and time index of l, respectively. For notation simplicity, we omit $(k,\ l)$ when no confusion arises. For a magnitude-based DNN approach, given reverberant speech features, the magnitudes of spectra are estimated by the network, which are subsequently coupled with noisy phase to reconstruct the waveform. For a complex spectral mapping-based (dubbed complex-based for short) DNN approach, with complex reverberant spectra as the network input, the RI components of anechoic speech are estimated, which can implicitly refine the phase information. As a result, the two approaches can be expressed as

$$|\tilde{S}| = \mathcal{F}_M(|X|;\theta_M), \tag{3}$$

and

$$\left(\tilde{S}_r, \tilde{S}_i\right) = \mathcal{F}_C(X_r, X_i;\theta_C), \tag{4}$$

where $\mathcal{F}_M$ and $\mathcal{F}_C$ denote the mapping function for magnitude and for complex-based networks, respectively. θ_M and θ_C denote the set of magnitude and of complex-based network parameters, respectively. Subscripts r and i refer to real and imaginary components, respectively. Tilde notation denotes the network estimation variable.

2.2 Proposed method

Empirically, the energy distribution of the spectrum is usually dramatically unbalanced over frequency. For example, for spectral regions with formants ranging from 0 to 2000 Hz, the spectral values are relatively large, while *vice versa* for silent or unvoiced regions. As such, when the network is trained with the criterion like minimum square error (MSE), if no feature compression is applied toward the spectrum, the regions with larger spectral value tend to be optimized with priority, as the optimization of such regions brings more obvious loss decrease. On the contrary, the regions with smaller value cannot be optimized well, as the contribution toward the final loss calculation becomes trivial, leading to spectral structure blurring in the weak energy regions. Therefore, if a suitable compression function can be applied to decrease the dynamic range and balance the loss gap between different spectral regions, the network is expected to capture more detailed information in the weak regions, which could be helpful for perceptual quality.

Log-power spectral feature is considered a universal approach for noise reduction as it better describes the characteristic of the human ear toward sound intensity level.[12] More recently, Zhao *et al.*[8] adopted a different feature preprocessing method with cubic root compression for dereverberation task, and moderate performance was obtained. Based on that, we propose a generalized compression approach, given as $|X|^\beta$, where $\beta \in (0,1]$ represents a tunable compression parameter. The smaller value that β takes, the stronger compression it will be given.

In polar coordinates, the complex spectrum of the reverberant speech can be written as $X = |X|e^{i\theta_X}$. Since phase spectrum shows little temporal and spectral regularities, it is quite difficult to estimate accurately, we only compress the magnitude while leaving the phase information unaltered here. As a result, the compressed complex spectrum can be given as $X^{\mathscr{C}} = |X|^\beta e^{i\theta_X}$, which can also be rewritten in Cartesian coordinates, given by

$$X^{\mathscr{C}} = X_r^{\mathscr{C}} + iX_i^{\mathscr{C}}, \tag{5}$$

$$X_r^{\mathscr{C}} = |X|^\beta \cos\theta_X, \tag{6}$$

and

$$X_i^{\mathscr{C}} = |X|^\beta \sin\theta_X, \tag{7}$$

where superscript $\mathscr{C}$ refers to compression format. Note that $X_r^{\mathscr{C}}$ and $X_i^{\mathscr{C}}$ become X_r and X_i, respectively, when $\beta = 1$. Because X_r and X_i are known as the RI components, we name $X_r^{\mathscr{C}}$ and $X_i^{\mathscr{C}}$ as cRI due to the fact that β is introduced to compress RI here.

2.3 Network architecture

In this paper, gated complex convolutional recurrent network (GCCRN) is deployed as the mapping network, as it has achieved state-of-the-art (SOTA) performance in the speech enhancement task.[13] The schematic diagram of the network is illustrated in Fig. 1. It is mainly comprised of three components, namely one convolutional encoder, long-short term memory (LSTM) modules, and two decoders for both real and imaginary parts reconstruction. Note that for a magnitude-based network, we take a similar network topology, except only one decoder is provided and the output activation function is set to Rectified Linear Unit (ReLU) to satisfy the output range. Due to space limitation, interested readers may refer to Ref. 13 for network details.

Following Ref. 13, the RI components are estimated by DNN simultaneously. To this end, we use the following loss function to recover the information in both real and imaginary parts, given as

$$\mathcal{L}_{RI} = \|S_r^{\mathscr{C}} - \tilde{S}_r^{\mathscr{C}}\|_2^2 + \|S_i^{\mathscr{C}} - \tilde{S}_i^{\mathscr{C}}\|_2^2, \tag{8}$$

where Wang *et al.*[9] showed that when magnitude loss was also considered for the complex spectral mapping-based method, speech quality could be further improved. As a result, we also add the magnitude-based loss item, that is,

$$\mathcal{L}_{RI+Mag} = \mathcal{L}_{RI} + \left\|\sqrt{|\tilde{S}_r^{\mathscr{C}}|^2 + |\tilde{S}_i^{\mathscr{C}}|^2} - \sqrt{|S_r^{\mathscr{C}}|^2 + |S_i^{\mathscr{C}}|^2}\right\|_2^2. \tag{9}$$

Note that if $\beta = 1$, Eq. (9) will reduce to uncompressed format. For a magnitude-based network, as only the magnitude is estimated by DNN, the loss is defined as

$$\mathcal{L}_{Mag} = \||\tilde{S}^{\mathscr{C}}| - |S^{\mathscr{C}}|\|_2^2. \tag{10}$$

3. The datasets and parament configurations

3.1 Datasets

All the experiments are conducted on the WSJ0 SI-84 dataset (WSJ),[14] which includes 7138 utterances by 83 speakers (42 males and 41 females). To prepare the datasets, we split 5428 and 957 utterances with 77 speakers for training and validation. To test the performance of models, we also split independent 150 utterances with 6 speakers for testing. The testing dataset is comprised of two categories. For the first type, the speaker information is within training set, i.e., seen speaker, and for the second type, the speaker is untrained, i.e., unseen speaker.

We simulate a rectangular room with the size 9 m × 8 m × 5 m, where a microphone is placed in the centre of the room, i.e., (4.5 m, 4 m, 2.5 m). The position of speaker is randomly chosen, where the height is the same as the receiver (microphone) and the distance is fixed to 1.5 m. The image method[15] is deployed to generate different RIRs, which are subsequently convolved with clean speech to generate the reverberant speech. In this study, RT_{60} ranges from 0.3 s to 1.4 s with the interval 0.1 s, in each of which 50 different RIRs are generated. For model testing, the range of RT_{60} is from 0.4 s to 1.0 s with the interval 0.2 s. For each RT_{60}, 1 RIR is generated, which is untrained. In this study, we generate 180 001 800 reverberant-anechoic pairs for training and validation, respectively. For testing, under each RT_{60}, 150 pairs are generated.

3.2 Parameter configurations

All the utterances are sampled at 16 kHz. The window size is set to 20 ms, with 50% overlap between adjacent frames. A 320-point FFT is adopted to generate input features. For a magnitude-based network, rMS and cMS are extracted while for a complex spectrum-based network, RI and cRI are extracted as network inputs. The network is trained with MSE criterion and Adam optimizer.[16] The initialized learning rate is 0.001, and the whole network is trained for 50 epochs. The batch number is set to 8 at an utterance level.

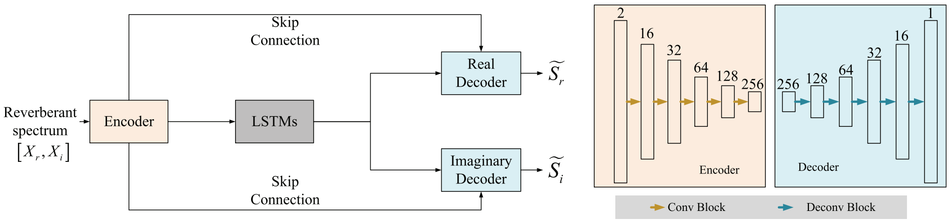


Fig. 1. The architecture illustration of GCCRN. It consists of three principal components, namely convolutional encoder, LSTM modules, and decoders for RI components of the spectrum. The number of channels for each layer is also presented in the right part of the figure.

4. Results and analysis

4.1 Objective comparisons

In this part, both magnitude- and complex-based experiments are conducted. For both type, five conditions are set, namely, no compression ($\beta = 1$), $\beta = 2/3$, $\beta = 1/2$, $\beta = 1/3$, and logarithm (dubbed *log*). Note that for log-based complex compression, when the absolute magnitude value is smaller than 1, the positive and negative signs will reverse, which notably impacts the phase distribution when coupled with phase information. To mitigate the problem, we replace $\log(x)$ with $\log(1+x) \in [0, \infty)$ to keep the consistency of phase sign. PESQ and SNR_{fw} are used as the evaluation metrics for performance comparison.

Table 1 presents the results of different compression parameters in terms of PESQ and SNR_{fw}. From the table, one can get the following conclusions. First, compared with no compression method, when a suitable compression parameter is chosen, better results can be achieved. For example, for seen speaker case, going from $\beta = 1$ to 2/3, average 0.08 and 0.13 PESQ improvements are observed while average 1.23 and 1.19 dB SNR_{fw} improvements are achieved for magnitude and complex based methods, respectively. The reason can be explained as the contributions of weak energy regions, e.g., middle and high-frequency regions, are given relatively more emphasis when suitable compression operation is applied, and, therefore, more detailed structure can be captured. Second, compared with the magnitude-based method, the complex-based method brings pronounced metric improvements. Taking seen speaker case with $\beta = 1$ as an example, when the complex method is adopted, average 0.21 and 0.42 dB PESQ and SNR_{fw} improvements are obtained. This indicates the importance of phase refinement. Third, among different compression methods, when $\beta = 1/2$, consistently better performance is achieved for various cases. Actually, when too much compression is applied, i.e., 1/3, the loss contributions of high energy regions are heavily suppressed. As a result, the information of formants in the low frequency regions may get lost during the optimization process, which is harmful to speech quality. In addition, we notice that no obvious advantage is observed for *log* operation over other compression methods.

Speech spectrograms are presented in Fig. 2. One can observe that cRI with $\beta = 1/2$ better recovers the detailed structure in weak energy regions than the other two methods, as shown in the red box area, which reflects the importance of power compression with a suitable value of β.

4.2 Subjective result comparisons

To further illustrate the advantage of the proposed cRI-based approach, we designed a subjective AB listening test, following the procedure of Ref. 17. Thirteen volunteers with normal hearing (NH) participated in the test (10 males and 3

Table 1. Objective result comparisons among different compression schemes in terms of seen and unseen speakers. PESQ and SNR_{fw} are adopted as two evaluation metrics. Abbreviations "Mag." and "Com." denote magnitude and complex spectrum-based methods, respectively.

		Metrics		PESQ					SNR_{fw} (in dB)				
		RT_{60}(s)		0.4	0.6	0.8	1.0	Avg.	0.4	0.6	0.8	1.0	Avg.
		Rev		2.58	2.23	2.04	1.94	2.20	11.61	9.00	7.90	7.02	8.88
Seen	Com.	RI	$\beta = 1$	3.24	2.94	2.76	2.62	2.89	12.90	11.21	10.55	9.91	11.14
		cRI	$\beta = 2/3$	3.38	3.07	2.89	2.73	3.02	14.29	12.44	11.66	10.92	12.33
			$\beta = 1/2$	**3.43**	**3.12**	**2.94**	**2.76**	**3.06**	**14.40**	**12.59**	**11.80**	**11.09**	**12.47**
			$\beta = 1/3$	3.30	2.97	2.77	2.61	2.91	13.62	11.84	11.13	10.49	11.77
			log	3.32	2.99	2.81	2.65	2.94	13.74	12.13	11.38	10.67	11.98
	Mag.	rMS	$\beta = 1$	3.04	2.71	2.54	2.43	2.68	12.66	10.86	10.05	9.31	10.72
		cMS	$\beta = 2/3$	3.12	2.79	2.62	**2.50**	2.76	13.92	12.05	11.28	10.54	11.95
			$\beta = 1/2$	**3.15**	**2.81**	**2.63**	**2.50**	**2.77**	**14.08**	**12.20**	**11.40**	**10.65**	**12.08**
			$\beta = 1/3$	3.13	2.79	2.62	2.48	2.76	14.01	12.09	11.27	10.52	11.97
			log	3.02	2.68	2.50	2.36	2.64	13.55	11.76	10.98	10.19	11.62
Unseen		Rev		2.57	2.22	2.03	1.95	2.19	11.47	8.79	7.68	6.84	8.69
	Com.	RI	$\beta = 1$	3.17	2.88	2.70	2.56	2.83	12.61	10.94	10.22	9.58	10.84
		cRI	$\beta = 2/3$	3.30	2.98	2.81	2.64	2.93	14.01	12.10	11.30	10.61	12.00
			$\beta = 1/2$	**3.35**	**3.03**	**2.84**	**2.67**	**2.97**	**14.19**	**12.29**	**11.46**	**10.79**	**12.18**
			$\beta = 1/3$	3.18	2.85	2.66	2.49	2.80	13.30	11.49	10.75	10.14	11.42
			log	3.22	2.91	2.71	2.57	2.85	13.56	11.87	11.04	10.39	11.72
	Mag.	rMS	$\beta = 1$	2.99	2.65	2.47	2.35	2.62	12.43	10.64	9.76	9.03	10.47
		cMS	$\beta = 2/3$	3.06	2.73	2.55	**2.43**	2.69	13.69	11.76	10.92	10.18	11.64
			$\beta = 1/2$	**3.08**	**2.75**	**2.57**	**2.43**	**2.71**	**13.83**	**11.90**	**11.03**	**10.30**	**11.76**
			$\beta = 1/3$	3.07	2.73	2.55	2.41	2.69	13.68	11.75	10.85	10.14	11.61
			log	2.95	2.61	2.41	2.28	2.56	13.25	11.42	10.55	9.81	11.26

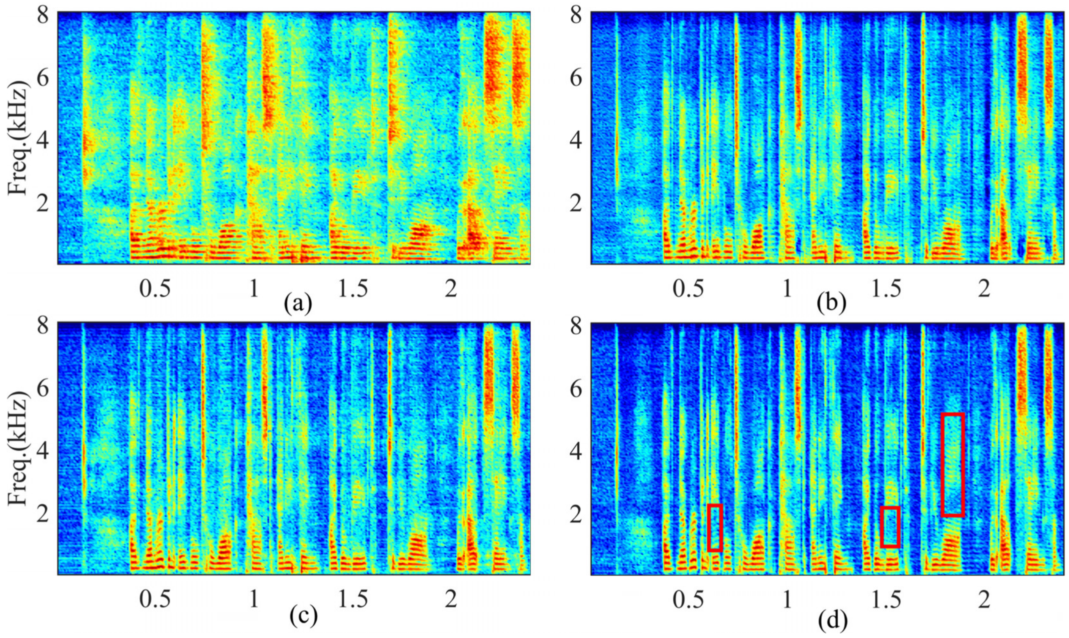


Fig. 2. Spectrum visualization of a processed utterance. (a) Unprocessed reverberant speech, $RT_{60} = 0.7$s. (b) Processed speech with rMS. (c) Processed speech with RI. (d) Processed speech with cRI using $\beta = 1/2$.

females, mean = 25 yr). Four groups of experiments were conducted in all. In the first group, set A and B denoted cMS ($\beta = 1/2$) and rMS, respectively, and for the second group, set A and B were RI and cRI ($\beta = 1/2$), respectively. In the third group, set A denoted the approach with rMS, and set B denoted the approach with RI. In the last group, set A and B denoted cRI ($\beta = 1/2$) and rMS, respectively. Each group consisted of 15 utterance pairs, each pair of which included an unprocessed reverberant speech and two utterances processed by two algorithms. All the unprocessed utterances were randomly sampled from the test dataset with RT_{60} ranges from 0.6 to 1.0 s. The utterance sequence in each group was randomly mixed. For participants, they were placed in a quiet office to finish the listening test. All the utterances were played by a PC and presented over Sennheiser HD 380 Pro (Wedemark, Germany). For listeners, the unprocessed reverberant utterance was first played, followed by two utterances processed by two algorithms. Since many dereverberation algorithms usually led to speech blurring under a high reverberation environment, in this test, the criterion was to choose the processed utterance id with better clearness and naturalness. Choosing "equal" was also a provided option if the listeners could not distinguish the difference between the two utterances.

The AB test results for above four groups are shown in Fig. 3. From the figure, one can reach the following conclusions. First, power compression does improve the subjective quality. For example, from Figs. 3(a) and 3(b), one can find obvious subjective preference toward the sets with power compression in both MS and RI based conditions. Second, phase optimization is quite significant for perceptual quality improvement. This conclusion can be drawn from Figs. 3(a) and 3(b), where the average preference toward the compression-based method increases from 67% in (a) to 87% in (b). Third, when power compression and phase estimation are combined together into optimization, further subjective quality

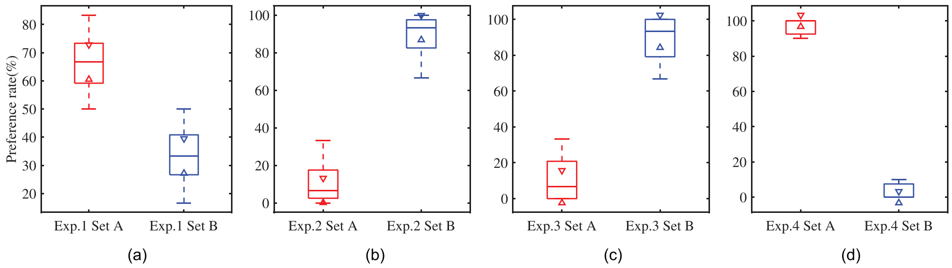


Fig. 3. Results of AB preference test with different configurations. (a) Subjective comparison between cMS (red) and rMS (blue). (b) Subjective comparison between RI (red) and cRI (blue). (c) Subjective comparison between rMS (red) and RI (blue). (d) Subjective comparison between cRI (red) and rMS (blue).

improvement can be obtained. For example, going from Figs. 3(c) and 3(d), the average preference for the compressed-based method increases from 89% to 95%.

5. Conclusion

In this article, we study the importance of phase information and power compression on monaural dereverberation. We also propose a compressed cRI reconstruction method for dereverberation. Compared with previous spectral mapping methods, the proposed method preserves the phase information while effectively reducing the dynamic range of spectral magnitude. In consequence, the loss gap of different spectral regions can be better compensated, which facilitates the recovery of detailed spectral information. Objective and subjective experiments are conducted to validate the effectiveness of proposed approach and the results reveal that when phase information is integrated into optimization, better speech quality is obtained. Moreover, suitable power compression is helpful to reconstruct the detailed information in the weak spectral regions, which is blurred by heavy reverberation. Future work can concentrate on extending cRI to noise suppression and dereverberation in noisy and reverberant environments for practical applications.

Acknowledgments

This work was supported by the National Key Research and Development Program of China. This work was also supported by IACAS Young Elite Researcher Project under Grant No. QNYC201813.

References and links

[1] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation* (Springer Science & Business Medical, 2010).
[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Proc. **27**(2), 113–120 (1979).
[3] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," J. Acoust. Soc. Am. **66**(1), 165–169 (1979).
[4] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," IEEE Trans. Audio Speech Lang. Proc. **18**(7), 1717–1731 (2010).
[5] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," IEEE/ACM Trans. Audio Speech Lang. Proc. **23**(6), 982–992 (2015).
[6] E. W. Healy, M. Delfarah, E. M. Johnson, and D. Wang, "A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation," J. Acoust. Soc. Am. **145**(3), 1378–1388 (2019).
[7] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," IEEE/ACM Trans. Audio Speech Lang. Proc. **25**(7), 1492–1501 (2017).
[8] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," IEEE/ACM Trans. Audio Speech Lang. Proc. **28**, 1598–1607 (2020).
[9] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," IEEE/ACM Trans. Audio Speech Lang. Proc. **28**, 1778–1787 (2020).
[10] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ): A new method for speech quality assessment of telephone networks and codecs," in *Proceedings of ICASSP*, Salt Lake City, Utah (May 7–11, 2001), pp. 749–752.
[11] J. Tribolet, P. Noll, B. McDermott, and R. Crochiere, "A study of complexity and quality of speech waveform coders," in *Proceedings of ICASSP*, Tulsa, Oklahoma (April 10–12, 1978), pp. 586–590.
[12] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," IEEE Signal Process. Lett. **21**(1), 65–68 (2014).
[13] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," IEEE/ACM Trans. Audio Speech Lang. Proc. **28**, 380–390 (2020).
[14] D. B. Paul and J. M. Baker, "The design for the *Wall Street Journal*-based CSR corpus," in *Workshop on Speech and Natural Language*, New York (February 23–26, 1992), pp. 357–362.
[15] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. **65**(4), 943–950 (1979).
[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980.
[17] S.-W. Fu, C.-F. Liao, Y. Tsao, and, and S. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proceedings of ICML*, Long Beach, California (June 9–15, 2019), pp. 2031–2041.