COMPLEX RATIO MASKING FOR JOINT ENHANCEMENT OF MAGNITUDE AND PHASE

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

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
{williado,wangyuxu,dwang}@cse.ohio-state.edu

ABSTRACT

The phase response of noisy speech has largely been ignored, but recent research shows the importance of phase for perceptual speech quality. A few phase enhancement approaches have been developed. These systems, however, require a separate algorithm for enhancing the magnitude response. In this paper, we present a novel framework for performing monaural speech separation in the complex domain. We show that much structure is exhibited in the real and imaginary components of the short-time Fourier transform, making the complex domain appropriate for supervised estimation. Consequently, we define the complex ideal ratio mask (cIRM) that jointly enhances the magnitude and phase of noisy speech. We then employ a single deep neural network to estimate both the real and imaginary components of the cIRM. The evaluation results show that complex ratio masking yields high quality speech enhancement, and outperforms related methods that operate in the magnitude domain or separately enhance magnitude and phase.

Index Terms— Deep neural network, speech separation, speech quality, complex ideal ratio mask

1. INTRODUCTION

Speech separation systems that operate on the short-time Fourier transform (STFT) of noisy speech usually enhance only the magnitude spectrum and use noisy phase during signal reconstruction. This is partially attributed to the findings in [1], which shows that enhancing noisy phase does not lead to significant improvements in equivalent signal-to-noise ratio (SNR). Another study by Ephraim and Malah [2] concludes that the complex exponential of noisy phase is the minimum-mean square error (MMSE) estimate of the complex exponential of clean phase. Indicating that the phase does not need to be altered when the MMSE is used to enhance noisy speech.

Contrary to these studies, Paliwal *et al.* [3] show that enhancing the phase spectrum of noisy speech leads to perceptual quality improvements. Paliwal *et al.* combine the noisy magnitude response with clean phase, noisy phase, and enhanced phase where mismatched analysis windows are used to extract the magnitude and phase spectra. Objective metrics and a listening study are used to assess speech quality, where the listening evaluation involves a preference selection between a pair of signals. The results reveal that significant speech quality improvements are attainable when the clean phase spectrum is applied to the noisy magnitude spectrum, while modest improvements are obtained when the noisy phase is used. In addition, high preference scores are achieved when the MMSE estimate of the clean magnitude spectrum is combined with an enhanced phase response.

The importance of phase to speech quality has led some researchers to develop phase estimation approaches for speech separation [4, 5, 6]. In [4], multiple input spectrogram inversion (MISI) is used to iteratively estimate the time-domain source signal in a mixture given the corresponding estimated STFT magnitude responses. Spectrogram inversion estimates signals by iteratively recovering the missing phase information, while constraining the magnitude response. The average total error between the mixture and the sum of the estimated sources updates the source estimates at each iteration. In [5], Mowlaee *et al.* perform MMSE phase estimation where the phases of two sources in a mixture are estimated by minimizing the square error. This minimization results in several phase candidates, but ultimately the pair of phases with the lowest group delay is chosen. Krawczyk and Gerkmann [6] enhance the phase of voiced speech by reconstructing the phase between harmonic components across frequency and time, given an estimate of the fundamental frequency. Unvoiced frames are left unchanged. The approaches in [4, 5, 6] all show objective quality improvements, but they do not address the noisy magnitude response.

Supervised time-frequency (T-F) mask estimation has recently been shown to improve human speech intelligibility in very noisy conditions (i.e. negative SNRs) [7, 8]. Additionally, a deep neural network (DNN) that estimates the ideal ratio mask (IRM) has been shown to improve objective speech quality and intelligibility [9]. T-F masking operates in the magnitude domain and uses the noisy phase during signal resynthesis. The use of noisy phase becomes more problem-

This research was supported in part by an AFOSR grant (FA9550-12-1-0130), an NIDCD grant (R01 DC012048), and the Ohio Supercomputer Center.

atic in negative SNRs than at higher SNR conditions, since noisy phase reflects more the phase of background noise than that of the target speech [10]. Based on phase enhancement research, masking results should further improve if both the magnitude and phase responses are enhanced. However, our recent attempts to estimate clean phase from noisy phase using deep learning were unsuccessful due to the lack of structure within the phase.

Instead of separately enhancing the magnitude and phase responses of noisy speech, we propose to operate in the complex domain by jointly enhancing the real and imaginary components. More specifically, we define the complex ideal ratio mask (cIRM) and use a DNN to estimate its complex parts. By operating in the complex domain, the cIRM is able to simultaneously enhance both the magnitude and phase responses. We will show that the estimated cIRM leads to objective quality improvements over the estimated IRM and systems that separately enhance the magnitude and phase.

This paper is organized as follows. Section 2 describes the structure within the complex domain. The cIRM is derived in Section 3. Section 4 explains how a DNN estimates the cIRM. Experimental results and system comparisons are presented in Section 5. We conclude with a discussion in Section 6.

2. STRUCTURE OF SHORT-TIME FOURIER TRANSFORM

The relationship between the STFT and its magnitude and phase is shown in (1)

$$S_{t,f} = |S_{t,f}|e^{i\theta_{S_{t,f}}} \tag{1}$$

where $|S_{t,f}|$ is the magnitude response and $\theta_{S_{t,f}}$ is the phase response at time t and frequency f. Each T-F unit in the STFT is a complex number with real and imaginary components. The magnitude and phase responses are computed directly from the real and imaginary components, as given below respectively.

$$|S_{t,f}| = \sqrt[2]{\Re(S_{t,f})^2 + \Im(S_{t,f})^2} \tag{2}$$

$$\theta_{S_{t,f}} = \tan^{-1}\frac{\Im(S_{t,f})}{\Re(S_{t,f})} \tag{3}$$

An example of the magnitude (top-left) and phase (top-right) responses for a clean speech signal is shown in Fig. 1. The magnitude response exhibits clear temporal and spectral structure, while the phase response looks rather random. When a learning algorithm is used to map features to a training target, it is important that there is structure in the mapping function. Fig. 1 shows that using DNNs to predict the clean phase response directly is unlikely effective.

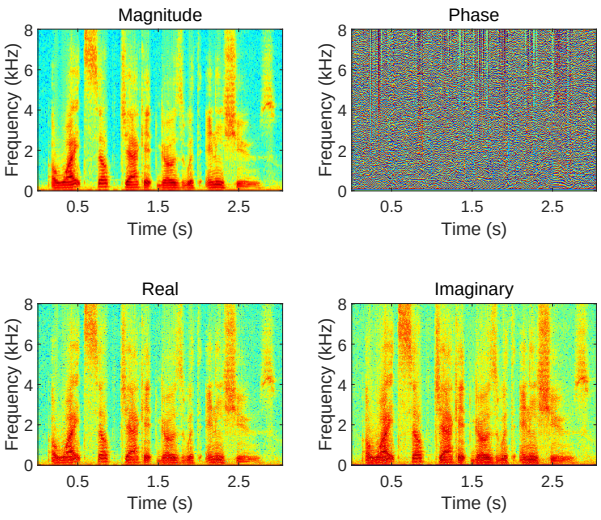

Fig. 1. (Color online) Example magnitude (top-left) and phase (top-right) spectrograms, and real (bottom-left) and imaginary (bottom-right) spectrograms, for a clean speech signal.

An expansion of the complex exponential in (1) leads to the following definitions for the real and imaginary components of the STFT:

$$S_{t,f} = |S_{t,f}|cos(\theta_{S_{t,f}}) + i|S_{t,f}|sin(\theta_{S_{t,f}}) \tag{4}$$

$$\Re(S_{t,f}) = |S_{t,f}|cos(\theta_{S_{t,f}}) \tag{5}$$

$$\Im(S_{t,f}) = |S_{t,f}|sin(\theta_{S_{t,f}}) \tag{6}$$

The lower part of Fig. 1 shows the log compressed, absolute value of the real (bottom-left) and imaginary (bottom-right) spectra of clean speech. Both real and imaginary components show clear structure, similar to magnitude spectrum, and are thus amenable to supervised learning. Based on this structure, a straightforward idea is to use DNNs to predict the complex components of the STFT. However, our recent study shows that directly predicting the magnitude spectrum may not be as good as predicting an ideal T-F mask [9]. Therefore, we propose to predict the real and imaginary components of the complex ideal ratio mask, which is described in the next section.

3. COMPLEX IDEAL RATIO MASK

Our goal is to derive a complex ratio mask that, when applied to the STFT of noisy speech, produces the STFT of clean speech. In other words, given the complex spectrum of noisy speech, $Y_{t,f}$, we get the complex spectrum of clean speech, $S_{t,f}$, as follows:

$$S_{t,f} = M_{t,f} * Y_{t,f} \tag{7}$$

where '$*$' indicates complex multiplication and $M_{t,f}$ is the cIRM. $Y_{t,f}$, $S_{t,f}$ and $M_{t,f}$ are complex numbers, and can be written in rectangular form as:

$$Y = Y_r + iY_i \tag{8}$$
$$M = M_r + iM_i \tag{9}$$
$$S = S_r + iS_i \tag{10}$$

where the subscripts r and i indicate the real and imaginary components, respectively. The subscripts for time and frequency are not shown for convenience, but the definitions are given for each T-F unit. Based on these definitions, Eq. (7) can be extended:

$$\begin{aligned} S_r + iS_i &= (M_r + iM_i) * (Y_r + iY_i) \\ &= (M_r Y_r - M_i Y_i) + i(M_r Y_i + M_i Y_r) \end{aligned} \tag{11}$$

The real and imaginary components of clean speech are then given as

$$S_r = M_r Y_r - M_i Y_i \tag{12}$$
$$S_i = M_r Y_i + M_i Y_r \tag{13}$$

After solving for M_r and M_i using Eqs. (12) and (13), the complex ideal ratio mask M is defined as

$$M = \frac{Y_r S_r + Y_i S_i}{Y_r^2 + Y_i^2} + i\frac{Y_r S_i - Y_i S_r}{Y_r^2 + Y_i^2} \tag{14}$$

The cIRM is closely related to the Wiener filter, which is the complex ratio of the cross-power spectrum of the clean and noisy speech to the power spectrum of the noisy speech [11].

S_r, S_i, Y_r, and $Y_i \in \mathbb{R}$, meaning that M_r and $M_i \in \mathbb{R}$. With this, the complex mask may have large values in the range $(-\infty, \infty)$, which may complicate cIRM estimation. Based on this analysis, we propose to clip M_r and M_i to values in the range $[-L, L]$, where L is a positive integer.

4. DNN-BASED ESTIMATION OF COMPLEX IDEAL RATIO MASK

Fig. 2 depicts the DNN that is used to estimate the cIRM. The DNN has three hidden layers where each has 1024 units [9]. The rectified linear (ReLU) [12] activation function is used for hidden units, while linear units are used for the output layer. The standard backpropagation algorithm using the mean-square error cost function is used to train the DNN. The output layer is separated into two sub-layers, one for the real and imaginary components of the cIRM, respectively. This Y-shaped network structure in the output layer is commonly used to jointly estimate related targets [13].

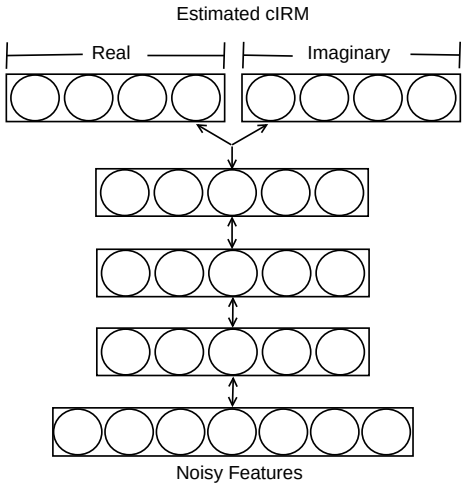

Fig. 2. DNN architecture used to estimate the complex ideal ratio mask.

The following set of complementary features are used as inputs: amplitude modulation spectrogram (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP), mel-frequency cepstral coefficients (MFCC), and cochleagram response, as well as their deltas [9]. A sliding context window is used to splice adjacent frames into a single vector for each time frame [9, 14]. This is employed for the input and output of the DNN.

5. EXPERIMENTAL RESULTS

The system is evaluated on the IEEE database [15], which consists of 720 utterances spoken by a single male speaker. The DNN for estimating the cIRM is trained with 500 utterances and the following noises: speech-shaped noise (SSN), cafeteria (Cafe), speech babble (Babble), and factory floor noise (Factory). The training set for estimating the cIRM is generated by combining ten random cuts from the first half of each noise with each training utterance at SNRs of -3, 0, and 3 dB. The test set is generated by mixing 60 clean utterances with the last half of the above noises at the SNR levels of -3, 0, and 3 dB. Dividing the noises into two halves ensures that the testing noise segments are unseen during training. In addition, a development set determines parameter values for the DNN and STFT. This development set is generated from 50 distinct clean IEEE utterances that are mixed with random cuts from the first half of the above four noises at SNRs of -3, 0, and 3 dB.

The DNN is trained to estimate the cIRM for each training mixture as described in (14). A 40 ms Hann window with 50% overlap between adjacent frames is used to produce the STFTs. The clipping level, L, is set to 10. Other clipping levels were evaluated to estimate the cIRM, but it was deter-

Table 1. Average performance scores for different systems. **Bold** indicates best result.

	PESQ				STOI				SNR_{fw}			
	SSN	Cafe	Babble	Factory	SSN	Cafe	Babble	Factory	SSN	Cafe	Babble	Factory
Noisy Speech	1.97	1.89	1.96	1.83	0.70	0.64	0.66	0.65	2.67	3.77	3.80	2.95
NS-K&G	2.03	1.99	1.99	1.93	0.64	0.61	0.63	0.61	3.51	4.48	4.36	3.78
NS-G&L	1.99	1.91	1.98	1.84	0.69	0.64	0.65	0.65	2.64	3.68	3.76	2.93
RM	2.47	2.34	2.54	2.40	0.83	0.77	**0.85**	0.78	7.53	6.99	8.67	7.07
RM-K&G	2.56	2.41	2.50	2.47	0.81	0.76	0.82	0.77	7.85	7.14	7.98	7.38
RM-G&L	2.47	2.34	2.54	2.40	0.83	0.77	**0.85**	**0.79**	7.54	7.00	8.68	7.08
cRM	**2.71**	**2.50**	**2.69**	**2.57**	**0.84**	**0.78**	0.84	**0.79**	**8.10**	**7.73**	**9.10**	**7.56**
CMF	2.16	2.16	2.10	2.17	0.76	0.70	0.71	0.72	4.75	4.77	5.21	4.40

mined through informal listening that a level of 10 optimizes both perceptual quality and noise reduction when compared against no clipping and a clipping level of 1. A three-frame context window augments each frame of the cIRM for the output layer and a context window covering five frames augments the complementary features.

The estimated cIRM (i.e. cRM) is compared to IRM estimation (i.e. RM) [9] and complex-domain nonnegative matrix factorization (CMF) [16, 17, 18]. In addition, we combine different magnitude spectra with phase spectra to evaluate approaches that separately enhance magnitude and phase. For phase estimation, we use a recent system by Krawczyk and Gerkmann [6] that enhances the phase response of voiced speech and a standard phase enhancing method by Griffin and Lim [19]. Since these approaches only enhance the phase responses, we combine them with the magnitude responses of speech separated by an estimated IRM (denoted as RM-K&G and RM-G&L) and of noisy speech (denoted as NS-K&G and NS-G&L). The perceptual evaluation of speech quality (PESQ) [20], the short-time objective intelligibility (STOI) score [21], and the frequency-weighted segmental SNR (SNR_{fw}) [22] are used to evaluate the quality and intelligibility of the different signals.

Table 1 shows the average performance for each signal for all noise types and at three test SNRs. Boldface indicates the system that performed best within a noise type. Each approach improves PESQ performance when compared to noisy speech. When enhancing only the phase of noisy speech, NS-K&G and NS-G&L slightly improve PESQ, which is consistent with the results from [6]. The estimated IRM and cIRM each produce considerable improvements over the noisy speech, with cRM performing best for each noise.[1] The results show that separately enhancing the magnitude response, with the ratio mask, and the phase response offers little to no improvement over ratio masking alone. Indicating that a joint enhancement of real and imaginary components can be more beneficial than separately enhancing magnitude and phase. CMF performs consistently for each noise, but it offers the smallest PESQ improvement over the noisy speech.

[1]Sound files can be found at http://web.cse.ohio-state.edu/~williado/cIRMdemos.html

When evaluating objective intelligibility with STOI, the systems that only enhance the phase do not improve scores. On the other hand, the methods that enhance magnitude only, or magnitude and phase each show significant improvement. The STOI scores between ratio masking and complex ratio masking are very similar, indicating that phase may not be as important for intelligibility as it is for quality. Overall, the estimated cIRM performs better than the related approaches for SSN, cafe, and factory noise. Similar trends are shown when evaluating with SNR_{fw}, where complex ratio masking performs best for all noises.

6. CONCLUSION

We have presented a framework for jointly enhancing the magnitude and phase of noisy speech by operating in the complex domain. This study shows that there is spectral and temporal structure within the complex components of the STFT. The complex ideal ratio mask is defined and our results show that a DNN can effectively estimate its components. Our experiments reveal that complex ratio masking outperforms ratio masking in the magnitude domain and complex nonnegative matrix factorization. The performance indicates that jointly enhancing the real and imaginary components of the cIRM can be better than independently enhancing the magnitude and phase. Lastly, the results provide further support of the importance of phase to speech quality.

Even though complex ratio masking is shown to outperform ratio masking when evaluated with objective metrics, through informal listening we find that the difference between the two is more evident. The objective metrics may be limited by not using phase information during their calculations.

To our knowledge, this is the first study using deep learning to perform speech separation in the complex domain, so there is likely room for further improvement. A systematic examination of current and new features needs to take place. Likewise, a study of more effective activation functions in the complex domain needs to occur.

7. REFERENCES

[1] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 30, pp. 679–681, 1982.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 32, pp. 1109–1121, 1984.

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

[4] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Signal Process. Lett.*, vol. 17, pp. 421–424, 2010.

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

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

[7] G. Kim, Y. Lu, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Am*, vol. 126, pp. 1486–1494, 2009.

[8] E. W. Healy, S. E. Yoho, Y. Wang, and D. L. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Am*, vol. 134, pp. 3029–3038, 2013.

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

[10] K. Sugiyama and R. Miyahara, "Phase randomization – a new paradigm for single-channel signal enhancement," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, 2013, pp. 7487–7491.

[11] P. C. Loizou, *Speech Enhancement: Theory and Practice*, Boca Raton, FL, USA: CRC, 2007.

[12] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. of Int. Conf. Artif. Intell. Statist.*, 2011, vol. 15, pp. 315–323.

[13] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, pp. 41–75, 1997.

[14] X.-L. Zhang and D. L. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Proc. of INTERSPEECH*, 2014, pp. 1534–1538.

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

[16] R. M. Parry and I. Essa, "Incorporating phase information for source separation via spectrogram factorization," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. 661–664.

[17] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 3437–3440.

[18] B. King and L. Atlas, "Single-channel source separation using complex matrix factorization," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 19, pp. 2591–2597, 2011.

[19] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. ASSP-32, pp. 236–243, 1984.

[20] 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.

[21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time frequency weighted noisy speech," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, pp. 2125–2136, 2011.

[22] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 16, pp. 229–238, 2008.