INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

An Iterative Phase Recovery Framework with Phase Mask for Spectral Mapping with An Application to Speech Enhancement

Kehuang Li[1], *Bo Wu*[2]*, *Chin-Hui Lee*[1]

[1]Georgia Institute of Technology
[2]National Laboratory of Radar Signal Processing, Xidian University
kehle@gatech.edu, rambowu11@gmail.com, chl@ece.gatech.edu

Abstract

We propose an iterative phase recovery framework to improve spectral mapping with an application to improving the performance of state-of-the-art speech enhancement systems using magnitude-based spectral mapping with deep neural networks (DNNs). We further propose to use an estimated time-frequency mask to reduce sign uncertainty in the overlap-add waveform reconstruction algorithm. In a series of enhancement experiments using a DNN baseline system, by directly replacing the original phase of noisy speech with the estimated phase obtained with a classical phase recovery algorithm, the proposed iterative technique reduces the log-spectral distortion (LSD) by 0.41 dB from the DNN baseline, and increases the perceptual evaluation speech quality (PESQ) by 0.05 over the DNN baseline, averaging over a wide range of signal and noise conditions. The proposed phase mask mechanism further increases the segmental signal-to-noise ratio (SegSNR) by 0.44 dB at an expense of a slight degradation in LSD and PESQ comparing with the algorithm without using any phase mask.

Index Terms: speech enhancement, spectral mapping, phase recovery, deep neural network, time-frequency mask

1. Introduction

In today's mobile speech communication era, speech enhancement to improve the hearing quality and intelligibility [1] is emerging again to attract a lot of research attentions. It is also a preprocessing vehicle to improve the robustness and accuracy of automatic speech recognition (ASR) [2, 3, 4]. Recent studies showed that deep neural networks (DNNs) have an excellent nonlinear regression capability [5] in dealing with classical signal processing problems, such as speech enhancement [6], source separation [7], bandwidth expansion [8], and speech dereverberation [9, 10]. Spectral mapping solutions are mostly adopted there to map noisy log-power spectra (LPS) to clean LPS features. Nonetheless, only the phase information in noisy speech is utilized in waveform reconstruction (e. g., [5]).

In the early 80's there were some studies looking into methods to reconstruct discrete time signal by using spectral magnitudes [11, 12, 13]. It is known that in a minimum-phase system, spectral magnitude can be related to phase with the Hilbert transform [14, 15]. However, there are no such systems in practice. These early studies mostly focused on some theoretical properties that cannot be used in practice due to some strong restrictions. On the other hand, [16] proved that with a sufficient window overlap, signals can always be reconstructed from their spectral magnitudes.

Some work tried to take advantage of speech harmonics. By using different window functions and frame shift sizes, it is possible to enhance the harmonic structure in instantaneous frequency, group delay, and baseband phase difference (BPD) [17, 18, 19]. However, these methods rely on the detection of voiced segments and fundamental frequencies of the voiced parts in speech, and cannot help on unvoiced segments, such as fricatives, which is critical for speech intelligibility.

There are other perspectives. For example, in [20], spectral magnitudes are enhanced with given phases. It showed there is some information in phase that can help with spectral magnitude estimation. Another thought is to work on complex spectrum or to learn complex masks [21]. That type of methods meet the restriction of model training, since most powerful models were designed to work on real numbers.

Among those studies we were attracted to one branch of methods. In [11], an iteratively reconstructing algorithm for interferometer images with only spectral magnitude was introduced. In [22], it was proved that the difference between the reconstructed signals in successive iterations will always converge, and a speed-up version upon Griffin and Lim's algorithm was given in [23]. A detailed discussion on Griffin and Lim's algorithm was highlighted in [24]. This family of techniques imposes no restriction on the spectral magnitudes. It actually recovers phase to compensate for some performance loss in the waveform reconstruction.

We therefore propose an iterative phase recovery framework in the spirit of a classical algorithm [22] for DNN based enhancement where the spectral magnitudes are well predicted. We further propose a phase mask to improve the segmental signal-to-noise ratio (SegSNR) [25] performance and reduce the stagnation ([24, 26]) problem due to sign uncertainty in phase estimation over neighbouring speech frames.

In a series of DNN based speech enhancement experiments, the proposed iterative phase recovery technique indeed improves the system performance of the baseline DNN system over a wide range of signal and noise conditions. The proposed mask-based mechanism further increases SegSNR at an expense of a slight degradation in log-spectral distortion (LSD) [1] and perceptual evaluation speech quality (PESQ) [27].

2. Spectral Mapping System

The framework we use in this study is similar to [5, 8]. Given log-power spectra Z^{LPS} (as in [28]) of distorted speech z, DNNs were trained to map Z^{LPS} to the LPS of parallel clean speech x, X^{LPS}. Denote the output of DNN as Y, it is to minimize the square error between the prediction and ground-truth,

*This work was done during Bo Wu's visiting at Georgia Institute of Technology in 2014-2016

http://dx.doi.org/10.21437/Interspeech.2016-494

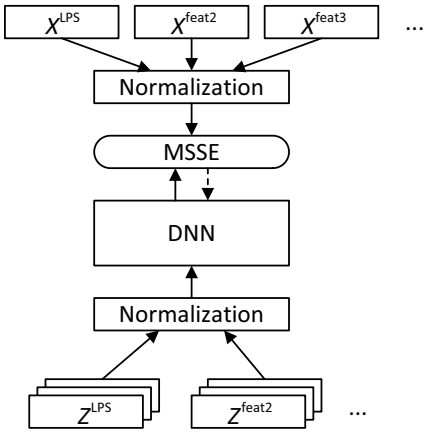

Figure 1: *A DNN based speech enhancement system.*

known as minimum sum of square error (MSSE) [29] criterion,

$$\min \frac{1}{2} \left\| Y - (X - \mu) \Sigma^{-1} \right\|_2^2, \tag{1}$$

where μ and Σ are mean and variance used to normalize the feature vectors. Recent study shows that multi-objective learning can improve the system performance [30], and thus more specifically, we would have the objective function as,

$$\begin{aligned} \min \quad & \frac{\alpha}{2} \left\| Y^{\mathrm{LPS}} - \left(X^{\mathrm{LPS}} - \mu_{\mathrm{LPS}} \right) \Sigma_{\mathrm{LPS}}^{-1} \right\|_2^2 + \\ & \frac{\beta}{2} \left\| Y^{\mathrm{feat2}} - \left(X^{\mathrm{feat2}} - \mu_{\mathrm{feat2}} \right) \Sigma_{\mathrm{feat2}}^{-1} \right\|_2^2 + \\ & \frac{\gamma}{2} \left\| Y^{\mathrm{feat3}} - \left(X^{\mathrm{feat3}} - \mu_{\mathrm{feat3}} \right) \Sigma_{\mathrm{feat3}}^{-1} \right\|_2^2 + \ldots, \end{aligned} \tag{2}$$

where α, β and γ are ratios among different features, and feat2 and feat3 are other features than LPS. In this work, feat2 is Mel-frequency cepstral coefficients (MFCCs) and feat3 is ideal ratio masks (IRMs) [31] if without other note. When the estimated LPS is gathered from DNN,

$$\hat{X}^{\mathrm{LPS}} = Y^{\mathrm{LPS}} \Sigma_{\mathrm{LPS}} + \mu_{\mathrm{LPS}}, \tag{3}$$

an estimation of spectral phase, $\hat{X}^P$, is required to reconstruct the waveform with inverse discrete Fourier transform (IDFT) and overlap-add [8, 22]. In most cases, phase of the distorted speech, Z^P, is used as such estimation [5].

3. Phase Recovery

3.1. Effect of Phase in Waveform Reconstruction

Given $\hat{X}^M = \exp\left(0.5 \hat{X}^{\mathrm{LPS}}\right)$, an estimated spectral magnitude, different spectral phase $\hat{X}^P$ implies various reconstructed waveforms, $\tilde{x}$. Figure 2 shows an example. It can be found that, compared with Figure 2c, Figure 2d has more precise structure in the harmonics as highlighted in the ellipse area. Specifically, Figure 2d even recovers more harmonic structure in the upper part of the ellipse area when compared with Figure 2b. A reason why phase makes such difference is that the spectral features are extracted from overlap windowed frames, which will lead to an inconsistency between reconstructed frames. And different phase will have different effect on the reconstructed waveform when such an inconsistency happens.

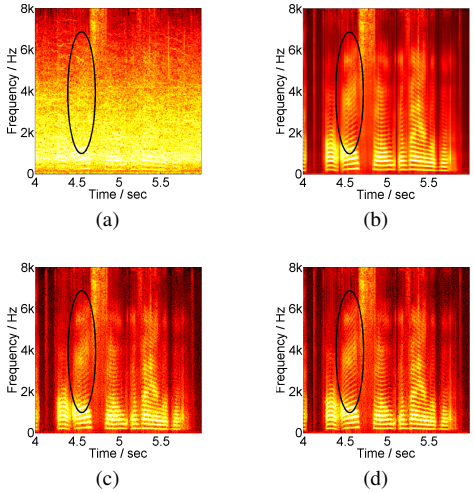

Figure 2: *Spectrograms of an example utterance showing the effects of phase: (a) noisy speech. (b) DNN estimated. (c) reconstructed with DNN estimated LPS and noisy phase. (d) reconstructed with DNN estimated LPS and oracle phase.*

3.2. Iterative Phase Recovery

DNN based enhancement system generates outstanding spectral magnitude, yet loses some performance in reconstruction as illustrated in Section 3.1. To overcome the loss, iterative reconstructing method given in [22] has great performance that it will be shown in experiment session Griffin and Lim's method can take back the loss on the measure of log-spectral-distortion (LSD). As indicated in Figure 3, the spectral phase of the reconstructed waveform will be used in the next iteration together with the predicted spectral magnitude. Iteratively, we make the phase to fit with the magnitude.

3.3. Phase Mask

In [30], it was shown that an estimated ideal binary mask (IBM) [32] can further improve the performance of DNN based speech enhancement. It motivated us to use the mask not only on the magnitude but also on the phase. A linear combination is not suitable due to the phase's cyclic nature. We propose to use a binary mask that the phase of some highly confident frequency bins are masked in the iterative recovery procedure. As shown

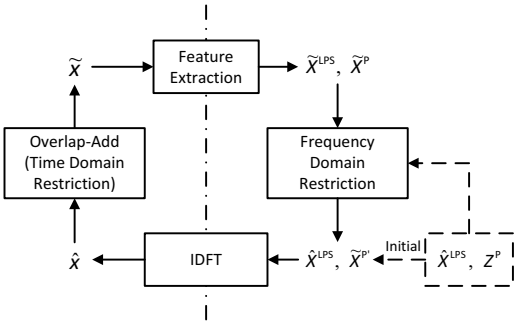

Figure 3: *An iterative waveform reconstruction system.*

in Figure 3, the phase mask is to modify $\tilde{X}^P$ as,

$$\tilde{X}^{P'}_{\ell,k} = \begin{cases} \tilde{X}^P_{\ell,k} & ,\mathrm{IRM}_{\ell,k} \leq \rho; \\ Z^P_{\ell,k} & ,\mathrm{IRM}_{\ell,k} > \rho. \end{cases} \tag{4}$$

Thus those masked frequency bins will keep their original phase and will affect their neighbour areas in spectrograms as highlighted in a spectral domain insight to be discussed next.

3.4. A Discussion on Phase Recovery in Spectral Domain

As shown in Figure 3, overlap-add plays an important role in waveform reconstruction. Actually, the window function h and the frame shift D define how consecutive frames contribute to the reconstructed waveform,

$$\tilde{x}(n) = \frac{\sum_{\ell=\lceil \frac{n-N+1}{D} \rceil}^{\lfloor \frac{n}{D} \rfloor} \hat{x}_\ell(n - D\ell)h(n - D\ell)}{\sum_{\ell=\lceil \frac{n-N+1}{D} \rceil}^{\lfloor \frac{n}{D} \rfloor} h(n - D\ell)^2}, \tag{5}$$

where $\lceil \cdot \rceil$ is the ceiling function and $\lfloor \cdot \rfloor$ is the flooring function, n is the discrete time index starting from 0, and $\hat{x}_\ell$, ℓ starts from 0, is the IDFT of ℓ-th frame's spectral magnitude and phase.

If D is set to half of the window length N, the spectrum of the ℓ-th frame after reconstruction is only affected by its left and right, $(\ell-1)$-th and $(\ell+1)$-th, frames. And thus we can have a simplified version of Eq. (5) in the spectral domain,

$$\begin{aligned} \tilde{X}_\ell \quad = \quad & CH_1C^{-1}\hat{X}_\ell + C_{\text{left}}H_2\left(C^{-1}\right)_{\text{lower}}\hat{X}_{\ell-1} \\ & + C_{\text{right}}H_2\left(C^{-1}\right)_{\text{upper}}\hat{X}_{\ell+1}, \end{aligned} \tag{6}$$

where C is the coefficient matrix of DFT, that is $C(p,q) = \exp(-j\frac{2\pi}{N}pq)$, and subscript 'left' means left half of the matrix, 'upper' means upper half of the matrix, etc. And H_1 is a diagonal matrix that $H_1(p,p) = h(p)^2/\left(h(p)^2 + h(p+\frac{N}{2})^2\right)$ for $p = 0, ..., \frac{N}{2}-1$ and $H_1(p,p) = h(p)^2/\left(h(p)^2 + h(p-\frac{N}{2})^2\right)$ for $p = \frac{N}{2}, ..., N$, H_2 is a diagonal matrix with $H_2(p,p) = h(p)h(p+\frac{N}{2})/\left(h(p)^2 + h(p+\frac{N}{2})^2\right)$ for $p = 0, ..., \frac{N}{2}-1$. Due to the conjugate symmetric property, Eq. (6) can be further written as,

$$\tilde{X}_\ell = A\hat{X}_\ell + B\hat{X}_{\ell-1} + B^*\hat{X}_{\ell+1}, \tag{7}$$

where B^* is the conjugate transpose of B. There might be more terms when using higher percentage frame overlap, but they will have similar forms and effects as B.

Figure 4 shows an example of matrices A and B, where the Hamming window with a frame length of 512 samples and a frame shift of 256 samples were used. Matrix A demonstrates how neighbouring frequency bins affect the central frequency bin. Such an effect comes from the window function, and due to same window function is used in feature extraction and reconstruction we believe it has no other effect than make the spectrogram more smooth along frequency axis. In [17], leaked energy in neighbouring frequency bins was used to help on phase enrichment, yet it doesn't work on our case as will be shown in experiment section. On the other hand, as shown in Figure 4, B is smooth in the central part with Δk between -5 to 5, where k is the discrete frequency index. Note that a baseband phase shift $\frac{2\pi}{N}Dk$ [17], which equals to πk when $D = N/2$, happens to have no effect on the central row of B where $k = 256$. Since B will look into the neighbour area in the consecutive frames

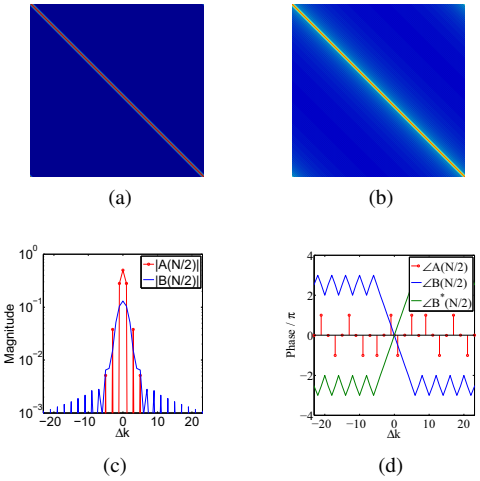

Figure 4: *Representing overlap-add as matrix. (a) log-magnitude of matrix A. (b) log-magnitude of matrix B. (c) magnitude of the central row of A and B. (d) phase of the central row of A and B.*

of the current frequency bin, it will thus recover the phase of the current frequency bin or at least make the phase more consistent between frames in the harmonic areas regardless of the fundamental frequency migration. When the phase of some frequency bins is locked, it not only prevents being affected by neighbouring high energy frequency bins, but also on the ability to recover neighbouring phases. Besides, masking partial phases won't break the convergence of the iterative method following the proof in [22, 23].

4. Experiments and Discussion

4.1. Speech Enhancement Experimental Setup

We experimented on the TIMIT corpus [33] with microphone speech sampled at 16 kHz in 16 bits resolution. It has 4620 training and 1680 test utterances. The window size of STFT [34] was 512 samples with a shift length of 256 samples, and the Hamming window was used in feature extraction. In speech enhancement experiments, we added 100 noise types [35] with 6 SNRs (-5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB) to all training utterances and randomly selected 150,000 out of 277,200 utterances for training (about 116 hours), and 1500 utterances were randomly selected from all noise added test utterances at the same 6 SNR levels to form the test set. During training 1500 randomly selected clean utterances were added to the training set, and 15150 utterances were set aside from the training set for validation. It was guaranteed that all noise types have the same number of utterances in the same set.

DNNs in experiments all had 3 hidden layers and 2500 sigmoid hidden nodes per layer. The base learning rate [36] of MSSE training was set to 10^{-5}, and the "newbob" method [37] was applied to halve the learning rate when the decrease of the mean squared error is less than 0.5%, and stops when it's less than 0.5‰. Mini-batch training [38] with a batch size of 32 utterances and a momentum rate of 0.9 was adopted. The input features are 257-dim LPS and 93-dim MFCCs (30 coefficients from 40 filter bins together with C0, appending first and second order dynamic coefficients [39]), both have 3 previous and 3 fol-

	LSD (dB)						SegSNR (dB)				PESQ				
SNR	N	M	NP	KG	GL	RP	NP	KG	GL	RP	N	NP	KG	GL	RP
-5	19.38	6.56	7.16	7.29	**6.61**	6.62	6.25	5.40	5.98	**6.30**	1.55	2.93	2.91	**2.97**	2.96
0	16.77	6.10	6.67	6.80	**6.15**	6.17	8.04	6.96	7.77	**8.05**	1.77	3.17	3.14	**3.22**	3.21
5	13.21	5.39	5.86	6.07	**5.41**	5.44	10.24	8.69	9.94	**10.29**	2.16	3.43	3.36	**3.49**	3.47
10	10.94	4.90	5.30	5.59	**4.91**	4.95	12.84	10.55	12.39	**12.86**	2.45	3.60	3.51	**3.66**	3.64
15	9.04	4.50	4.84	5.20	**4.51**	4.55	15.41	12.22	14.90	**15.40**	2.81	3.76	3.62	**3.81**	3.79
20	6.98	3.94	4.18	4.64	**3.94**	3.98	18.15	13.95	17.44	**18.11**	3.12	3.90	3.71	**3.93**	3.91
Avg.	12.65	5.22	5.65	5.92	**5.24**	5.27	11.88	9.67	11.46	**11.90**	2.32	3.47	3.38	**3.52**	3.50

Table 1: *Objective Measure on Reconstructed Signals. 'N': noisy speech, 'M': DNN predicted magnitude, 'NP': reconstructed with noisy phase, 'KG': reconstructed with enhanced phase [17], 'GL': Griffin and Lim's method [22], and 'RP': proposed phase recovery.*

lowing context frames, together with an extra 257-dim LPS and 93-dim MFCCs of the estimated noise background appended to input features as presented in [30]. The output features are 257-dim LPS and 93-dim MFCCs of clean speech and 257-dim IRM. All input and output features, except IRM, were normalized to zero mean and unit variance in training. IRMs were not normalized since they're already in the range of [0, 1],

$$\mathrm{IRM}_{\ell,k} = \frac{X^M_{\ell,k}}{\sqrt{\left(X^M_{\ell,k}\right)^2 + \left(\Phi^M_{\ell,k}\right)^2}}, \tag{8}$$

where Φ^M is the spectral magnitude of $\phi = z - x$. The weights between output features in the objective function in Eq. (2) were $\alpha = 0.327$, $\beta = 0.542$, and $\gamma = 0.131$.

4.2. Results and Discussions

We did iterative evaluation of speech enhancement, where the proposed phase recovery method with phase masks was used. For the predicted IRM > 0.75 ($\rho = 0.75$), the corresponding phase would be kept as they are in Z^P. And the LPS used in reconstruction was the combination of the DNN prediction and the LPS of noisy speech,

$$\hat{X}^{\mathrm{LPS,mask}}_{\ell,k} = (1 - \mathrm{IRM}^2_{\ell,k})\hat{X}^{\mathrm{LPS}}_{\ell,k} + \mathrm{IRM}^2_{\ell,k} Z^{\mathrm{LPS}}. \tag{9}$$

In our experiments, different noise levels have the same trend. It was found that the LSDs were all getting better iteratively, and that they dropped down very fast in the first five iterations and converged in about 20 iterations. Here we got a performance very close to that of Griffin and Lim's algorithm, and the performance gap was not growing with more iterations. However, in case of segmental SNR as shown in Figure 5 averaged over 6 SNR levels, Griffin and Lim's method got worse rapidly, while the proposed technique got a slight improvement

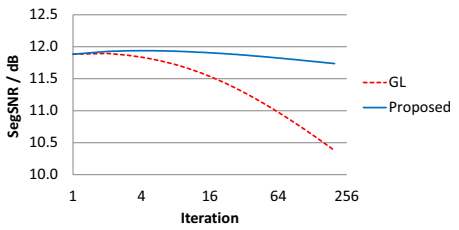

Figure 5: *A comparison of iterative performance of Griffin and Lim's and the proposes on SegSNR. X-axis is in logarithm scale to make the iterative performances clear.*

in the first few iterations and degraded slowly afterwards. When compared with [22], in short, the proposed method achieved very similar LSDs and showed a great advantage on SegSNRs.

A detailed comparison is given in Table 1, where iterative methods were measured after running 20 iterations. Starting with noise speech (indicated by 'N') it shows that when reconstructing the waveform using the noisy phase ('NP'), for example, on the third row of SNR at -5 dB, 0.60 dB was lost on LSD from the system with the DNN-predicted magnitude (system 'M'). Phase enhancement with 'KG' [17] made it even worse and lost an extra 0.14 dB, and phase recovery with 'GL' [22] got 0.55 dB back. 'GL' and the proposed 'RP' had a small 0.03 dB difference on average. On SegSNR, 'GL' got an average degradation of 0.42 dB from 'NP' and got a larger decrease at higher SNRs, while the proposed 'RP' was even slightly better than 'NP'. A reason could be that 'GL' has an issue of stagnation [24], while the proposed 'RP' masked some frequency bins' phase and reduced the stagnation effect. On PESQ, 'GL' is better than 'NP', 'KG', and about the same as 'RP'.

By taking advantage of the information stored in the spectral phase of noisy speech, the proposed method adds values to Griffin and Lim's method, with which the reconstructed signal will converge but may not converge to clean speech. On the other hand, traditional phase enhancement method cannot beat the iterative phase recovery method in our experiments. We believe it is because the state-of-the-art DNN based spectral magnitude enhancement algorithm has an excellent estimation of the clean spectral magnitude, and thus phase enhancement cannot further remove some residual noises neither could it solves the in-frame inconsistency issue. Furthermore, the proposed method required an estimation of IRM which is a byproduct of multi-objective learning [30].

5. Conclusion and Future Work

In this paper, an iterative phase recovery framework for waveform reconstruction in speech enhancement is proposed. It modifies the classical Griffin and Lim's algorithm [22], and attempts to resolve the problem mentioned in [11]. By removing the inconsistency in phases between the overlapped frames, the proposed mask-based framework brings out the potential advantages of DNN based enhancement on performances measured in LSD and SegSNR. We would continue to work on using phase recovery in different application areas, such as bandwidth expansion, speech separation, voice conversion, etc. On the other hand, embedding phase enhancement like [19] into the magnitude enhancement framework and learning masks from phase instead of magnitude could also be a good direction.

6. References

[1] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 873–902.

[2] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Proc. ICASSP*. IEEE, 2015, pp. 4375–4379.

[3] C. Weng, D. Yu, S. Watanabe, and B.-H. F. Juang, "Recurrent deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2014, pp. 5532–5536.

[4] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1415–1426, 2009.

[5] Y. Xu, J. Du, L. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, pp. 65–68, 2014.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 1, pp. 7–19, 2015.

[7] Y. Tu, J. Du, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *9th Int. Symp. Chinese Spoken Lang. Process. (ISCSLP)*. IEEE, 2014, pp. 250–254.

[8] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. ICASSP*, 2015.

[9] 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, and Lang. Process.*, vol. 23, no. 6, pp. 982–992, 2015.

[10] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE Signal Process. Lett.*, 2016, submitted.

[11] J. R. Fienup, "Reconstruction of an object from the modulus of its fourier transform," *Optics letters*, vol. 3, no. 1, pp. 27–29, 1978.

[12] S. H. Nawab, T. F. Quatieri, and J. S. Lim, "Signal reconstruction from short-time Fourier transform magnitude," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 4, pp. 986–998, 1983.

[13] J. Miao, D. Sayre, and H. N. Chapman, "Phase retrieval from the magnitude of the Fourier transforms of nonperiodic objects," *JOSA A*, vol. 15, no. 6, pp. 1662–1669, 1998.

[14] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*. Cambridge university press, 1952.

[15] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin, "Painless reconstruction from magnitudes of frame coefficients," *Journal of Fourier Analysis and Applications*, vol. 15, no. 4, pp. 488–501, 2009.

[16] R. Balan, P. Casazza, and D. Edidin, "On signal reconstruction without phase," *Applied and Computational Harmonic Analysis*, vol. 20, no. 3, pp. 345–356, 2006.

[17] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 12, pp. 1931–1940, 2014.

[18] P. Mowlaee and J. Kulmer, "Phase estimation in single-channel speech enhancement: limits-potential," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 8, pp. 1283–1294, 2015.

[19] 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, no. 2, pp. 55–66, 2015.

[20] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Process. Lett.*, vol. 20, no. 2, pp. 129–132, 2013.

[21] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, in press.

[22] D. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, 1984.

[23] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," in *Proc. Int. Conf. Digital Audio Effects DAFx*, vol. 10, 2010.

[24] N. Sturmel and L. Daudet, "Signal reconstruction from STFT magnitude: a state of the art," in *Proc. Int. Conf. Digital Audio Effects DAFx*, vol. 2012, 2011, pp. 375–386.

[25] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice Hall Englewood Cliffs, NJ, 1988.

[26] J. Fienup and C. Wackerman, "Phase-retrieval stagnation problems and solutions," *JOSA A*, vol. 3, no. 11, pp. 1897–1907, 1986.

[27] 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 *Proc. ICASSP*, vol. 2. IEEE, 2001, pp. 749–752.

[28] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions." in *Proc. INTERSPEECH*, 2008, pp. 569–572.

[29] D. M. Allen, "Mean square error of prediction as a criterion for selecting variables," *Technometrics*, vol. 13, no. 3, pp. 469–475, 1971.

[30] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," in *Proc. INTERSPEECH*, 2015.

[31] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*. IEEE, 2013, pp. 7092–7096.

[32] Y. Li and D. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Communication*, vol. 51, no. 3, pp. 230–239, 2009.

[33] J. S. Garofolo *et al.*, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, 1988.

[34] J. B. Allen and L. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[35] G. Hu. (2004). [Online]. Available: http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html

[36] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011, pp. 1–4.

[37] ICSI QuickNet toolbox. Newbob approach is implemented in the toolbox. [Online]. Available: http://www1.icsi.berkeley.edu/Speech/qn.html

[38] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," Dept. Comput. Sci., Univ. Toronto, Tech. Rep. UTML TR 2010–003, 2010.

[39] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, *The HTK book*. Entropic Cambridge Research Laboratory Cambridge, 1997, vol. 2.