MONAURAL SPEECH ENHANCEMENT USING INTRA-SPECTRAL RECURRENT LAYERS IN THE MAGNITUDE AND PHASE RESPONSES

Khandokar Md. Nayem and Donald S. Williamson

Department of Computer Science, Indiana University, USA
knayem@iu.edu, *williads@indiana.edu*

ABSTRACT

Speech enhancement has greatly benefited from deep learning. Currently, the best performing deep architectures use long short-term memory (LSTM) recurrent neural networks (RNNs) to model short and long temporal dependencies. These approaches, however, underutilize or ignore spectral-level dependencies within the magnitude and phase responses, respectively. In this paper, we propose a deep learning architecture that leverages both temporal and spectral dependencies within the magnitude and phase responses. More specifically, we first train a LSTM network to predict both the spectral-magnitude response and group delay, where this model captures temporal correlations. We then introduce Markovian recurrent connections in the output layers to capture spectral dependencies within the magnitude and phase responses. We compare our approach with traditional enhancement approaches and approaches that consider spectral dependencies within a single time frame. The results show that considering the within-frame spectral dependencies leads to improvements.

Index Terms— speech enhancement, intra-spectral correlations, recurrent neural networks, long short-term memory

1. INTRODUCTION

Monaural speech enhancement is a challenging task that aims to remove unwanted background noise from a single audio channel, to improve perceptual speech intelligently and quality. Deep learning has resulted in improved performance, but additional improvement is needed in noisy environments.

Speech enhancement generally takes two forms, mask-based approaches or signal approximation. In mask-based approaches, a time-frequency (T-F) mask is approximated that acts as a filter to remove the noise. Various masks have been proposed, including the ideal binary mask (IBM) [1] and ideal ratio mask (IRM) [2]. In [3], it is shown that estimating the ideal ratio mask outperforms other T-F masking and signal approximation approaches. Since then, ratio masks have incorporated phase information, e.g., the phase-sensitive mask (PSM) [3], complex ideal ratio mask (cIRM) [4] and parametric complex-valued T-F mask [5]. Alternatively, signal approximation directly estimates the clean speech signal in either the time [6] or T-F domains [7, 8]. A variety of network architectures have been used, including, DNNs [3, 7, 8], autoencoders [9, 10], long short-term memory (LSTM) networks [7, 11], and convolutional neural networks (CNNs) [12]. More recent approaches use deep clustering (DC), which groups learned activations into classes (e.g. speech dominant or noise dominant), to form a binary mask (BM) [13]. In [14], an end-to-end model that uses an utterance-based objective function shows promising results and it preserves high and low-frequency spectral information.

In the above approaches, the approximated T-F output is based on prior network layers and prior (in time) outputs of that T-F unit. The output, however, is not based on the adjacent or nearby frequency points within the magnitude response. It is known, however, that speech has spectral dependencies along the frequency axis [15], but current architectures often ignore these correlations. Recently, two approaches have been developed to address frequency-level dependencies, but they have only been evaluated for automatic speech recognition [16] or audio restoration after coding [17]. Both approaches use dedicated LSTM modules to learn spectral dependencies, but this is either done at the subband frequency-level or overall time. Additionally, these approaches do not consider local spectral dependencies over short-time instances. Nevertheless, these approaches have shown that incorporating spectral dependencies offers noticeable improvements, but it is not clear if this will have the same impact on speech enhancement.

We propose an intra-spectral (e.g. across-frequency) recurrent layer that captures frequency dependencies within each time frame of a speech signal. Given a noisy speech input, a LSTM network with multiple target loss functions learns the temporal dependencies of speech. We then append the proposed intra-spectral recurrent layer to enforce spectral-level dependencies. Our preliminary work showed that incorporating spectral-level dependencies within the magnitude domains leads to noticeable improvements [18]. In this work, we also incorporate spectral-level dependencies within the phase response, by applying an intra-spectral recurrent layer to the group delay of the signal. This is done

because multiple studies have shown how phase is important to signal quality. To the best of our knowledge, magnitude and phase-level dependencies have not been investigated for monaural speech separation with deep learning.

The rest of the paper is organized as follows. Conventional deep learning-based speech enhancement is discussed in section 2. In section 3, we describe our proposed intra-spectral bi-directional recurrent (ISBR) layer. The experimental setup is provided in section 4 and results are discussed in section 5. We conclude in section 6.

2. MAGNITUDE AND PHASE INFORMATION FOR SPEECH ENHANCEMENT

Let's define $S_{t,k}$ as the T-F domain speech signal at time t and frequency k, which is computed using the short-time Fourier transform (STFT). Correspondingly, $S_{t,k}$ has a magnitude response, $|S_{t,k}|$, and a phase response, $\theta^S_{t,k}$, where $S_{t,k} = |S_{t,k}|e^{i\theta^S_{t,k}}$. Similarly, $N_{t,k}$ is the STFT of the noise signal with magnitude $|N_{t,k}|$ and phase response, $\theta^N_{t,k}$.

Speech enhancement systems often enhance the magnitude response of noisy speech, $|M_{t,k}|$, in order to produce an estimated clean version, $|\widehat{S}_{t,k}|$. They often find the best estimate of clean speech and noise by minimizing the following loss function [19]:

$$\mathcal{L}_{mag} = \sum_{t,k} \left(|\widehat{S}_{t,k}| - |S_{t,k}|\right)^2 + \left(|\widehat{N}_{t,k}| - |N_{t,k}|\right)^2 \quad (1)$$

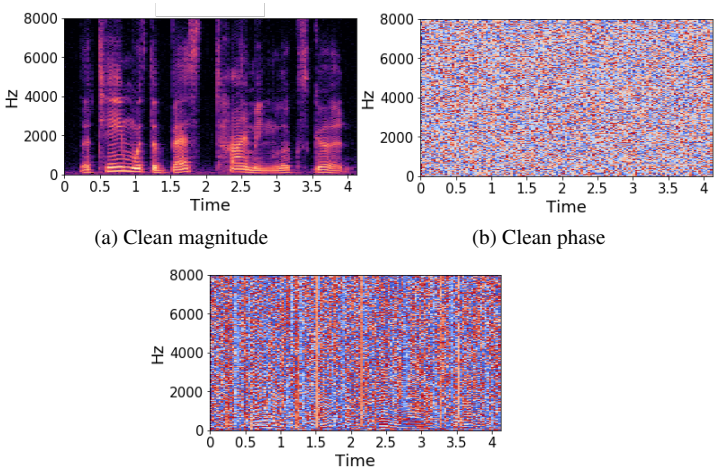
2.1. Group delay loss function

Phase enhancement has been shown to be important for speech quality [11, 4]. Unlike the magnitude response (Figure 1(a)), the phase of a speech does not show a clear structure (Figure 1(b)). On the other hand in Figure 1(c), group delay [20] of a speech shows a learn-able pattern in log-magnitude formulation. Therefore, instead of phase approximation, group delay (GD) approximation can show better success [19]. The group delay of signal $S_{t,k}$ is computed as $GD^S_{t,k} = \angle e^{i(\theta^S_{t,k+1} - \theta^S_{t,k})}$.

A mixture can be defined as a combination of two sound sources, which in our case are speech S and noise N. We can denote each of them as $\chi \in \{S, N\}$ where χ refers to a single sound source and $\neg\chi$ defines the other sound within the mixture signal. To incorporate group delay into signal approximation, a magnitude weighted cosine distance is used as the loss function:

$$\mathcal{L}_{gd} = \sum_{\chi\in\{S,N\}} \sum_{t,k} |\chi_{t,k+1}| \frac{(1 - cos(\widehat{GD}^{\chi}_{t,k} - GD^{\chi}_{t,k}))}{2} \quad (2)$$

where $GD^{\chi}_{t,k}$ and $\widehat{GD}^{\chi}_{t,k}$ are the unwrapped group delay and approximated group delay of a signal, either speech or noise.



(a) Clean magnitude (b) Clean phase

(c) Clean group delay

Fig. 1: Magnitude, phase and group delay of a clean signal.

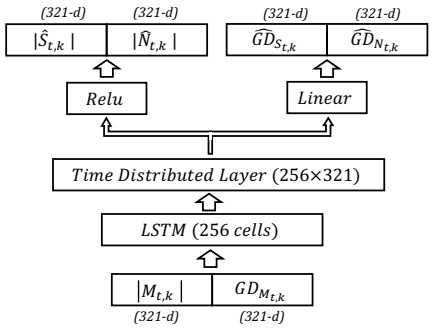


Fig. 2: Baseline LSTM architecture.

3. PROPOSED APPROACH

Signal-based speech enhancement needs to learn both temporal and spectral information to fully capture the structure (magnitude and phase) of the speech. To utilize the temporal information, a recurrent network that unrolls along the time axis is very effective. Since a LSTM is a recurrent deep network that can learn both short- and long-term temporal information, we use a LSTM network to learn the temporal information first. Then to learn the spectral dependencies among individual frequencies, we propose another recurrent architecture that captures the correlation among frequencies using a Markovian connection across the frequency axis of a single time frame. This recurrence is confined in the output layer of the pre-train LSTM network.

3.1. Long short-term memory (LSTM) architecture

In Figure 2, we show our proposed LSTM network which takes the magnitude of the mixture $|M_{t,k}|$ and the group delay of the mixture $GD^M_{t,k}$ as the inputs. The output layer of

the network is branched in two ways, one is for magnitude approximation $\widehat{\chi}_{t,k}$ and another is for group delay approximation $\widehat{GD}^{\chi}_{t,k}$ of both speech and noise. In other words, the outputs of the network are $\widehat{S}_{t,k}$ and $\widehat{GD}^{S}_{t,k}$ of enhanced clean speech, and $\widehat{N}_{t,k}$ and $\widehat{GD}^{N}_{t,k}$ of approximated noise. We reconstruct the phase of speech from the group delay using the trigonometric constraint, which gives an enhanced phase pair. We then estimate the phase difference between the enhanced speech and noise using:

$$\begin{aligned}\widehat{\delta}^{\chi}_{t,k} &= |\angle e^{i(\widehat{\theta}^{\chi}_{t,k} - \theta^{\neg\chi}_{t,k})}| \\ &= \arccos\left(\mathcal{T}\Big(\frac{|M_{t,k}|^2 + |\chi_{t,k}|^2 - |\neg\chi_{t,k}|^2}{2|M_{t,k}| \otimes |\chi_{t,k}|}\Big)\right) \end{aligned} \quad (3)$$

where $\chi \in \{S, N\}$, $\mathcal{T}(\cdot)$ truncates values to $[-1, 1]$ and $\otimes$ denotes element-wise multiplication. Then, the sign of each T-F unit, $\widehat{g}_{t,k} \in \{-1, 1\}$ is calculated by [19]:

$$\begin{aligned}\widehat{g}_{t,1}, \cdots, \widehat{g}_{t,K} = \underset{g_{t,1},\cdots,g_{t,K}}{\operatorname{argmax}} \sum_{k} \sum_{\chi\in\{S,N\}} cos\Big(\widehat{\theta}^{\chi}_{t,k+1}(g_{t,k+1}) \\ - \widehat{\theta}^{\chi}_{t,k}(g_{t,k}) - \widehat{GD}^{\chi}_{t,k}\Big) \end{aligned} \quad (4)$$

Here by the formulation of trigonometric property of group delay, $\widehat{\theta}^{\chi}_{t,k}(g_{t,k}) = \theta^{M}_{t,k} + \gamma g_{t,k}\widehat{\delta}^{\chi}_{t,k}$, with $\gamma = 1$ when $\chi = S$ and $\gamma = -1$ when $\chi = N$. Using dynamic programming within each frame, we can solve equation (4). The overall loss function of our LSTM network with hyper parameter λ is defined as:

$$\mathcal{L}_{mag+gd} = \lambda\mathcal{L}_{mag} + (1-\lambda)\mathcal{L}_{gd} \quad (5)$$

This LSTM network will learn the temporal mapping function between mixture to clean speech conditioning on both magnitude and phase information. However, this architecture does not consider the spectral axis connections to capture the dependencies among frequencies. For that reason, we use an intra-spectral layer over this LSTM network.

3.2. Intra-spectral Bi-directional Layer (ISBR)

Our proposed spectral recurrent layer captures intra-spectral correlations with a first-order Markov assumption. This layer models the frequency dependencies as a function of the adjacent spectral components. A Markov chain-like recurrent structure learns the spectral dependencies from low to high (increasing) and high to low (decreasing) frequencies. This is done along the entire frequency axis, and a depiction is shown in Fig. 3. This recurrent layer is denoted as an intra-spectral bi-directional recurrent (ISBR) layer. Each neuron of the layer represents a frequency bin of the signal.

The ISBR output layer uses the output of the prior LSTM network, $\boldsymbol{a}^{L-1}_{t}$, as input. The spectral output vector of the ISBR layer is denoted as $\boldsymbol{\psi}_t$. The individual spectral response at a corresponding frequency bin is denoted as $\psi_{k,t}$, where k indexes the frequency axis. Outputs from the ISBR layer are computed as follows:

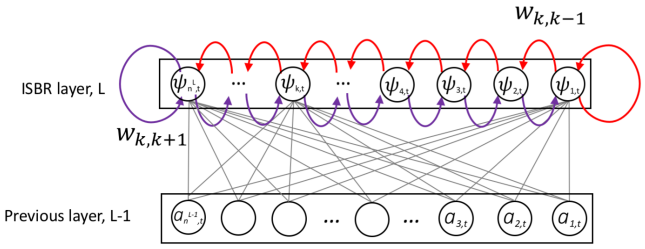


Fig. 3: Depiction of the proposed Intra-spectral bi-directional recurrent (ISBR) recurrent layer.

$$\boldsymbol{\Delta} = \sigma(\boldsymbol{R}^{L}\boldsymbol{a}^{L-1}_{t} + \boldsymbol{\beta}^{L}) \quad (6)$$

$$\begin{aligned}\psi_{1,t} &= \Delta_1 + \sigma_\psi(w_{1,2} \times \psi_{2,t}) \\ &\quad + \sigma_\psi(w_{1,1} \times \psi_{1,t-1}) \end{aligned} \quad (7)$$

$$\begin{aligned}\psi_{n^L,t} &= \Delta_{n^L} + \sigma_\psi(w_{n^L,n^L} \times \psi_{n^L,t-1}) \\ &\quad + \sigma_\psi(w_{n^L,n^L-1} \times \psi_{n^L-1,t}) \end{aligned} \quad (8)$$

$$\begin{aligned}\psi_{k,t} &= \Delta_k + \sigma_\psi(w_{k,k+1} \times \psi_{k+1,t}) \\ &\quad + \sigma_\psi(w_{k,k-1} \times \psi_{k-1,t}), \quad k \in [2, n^L - 1] \end{aligned} \quad (9)$$

where $\boldsymbol{\Delta}$ is the vector of activations, $\{\Delta_1, \ldots, \Delta_{n^L}\}$, based on inputs from the prior LSTM layer, $\boldsymbol{R}^{L} \in \mathbb{R}^{n^L \times n^{L-1}}$ is the weight matrix, and $\boldsymbol{\beta}^{L} \in \mathbb{R}^{n^L}$ is the bias vector. $\boldsymbol{w}_{k,k-1}$ represents the weight from the $(k-1)^{\text{st}}$ to k^{th} frequency component where as $\boldsymbol{w}_{k,k+1}$ represents the weight from the $(k+1)^{\text{st}}$ to k^{th} frequency component, within the recurrent output layer. σ and σ_ψ are the activation functions for the feed-forward and recurrent paths. Activation functions are applied separately to the feed-forward and recurrent paths, since this is similar to a logistic regression-based network, which has performed well for other tasks. The final output is therefore $\hat{\boldsymbol{y}}_t = \boldsymbol{\psi}^{L}_{t}$, which is the enhanced spectrum of the t^{th} time frame.

In our proposed LSTM network, we have 4 output layers in a parallel manner (see section 3.1). We first train our LSTM model with dense output layers because this will capture the temporal correlations in magnitude and phase structures. Then we replace the dense output layers with our proposed ISBR layer and retrain the network. This will learn the spectral relationships among the adjacent frequencies, which was absent in temporal training.

4. EXPERIMENTAL SETUP

We evaluate our proposed approach using the IEEE [21] and the TIMIT [22] speech corpora. The IEEE corpus consists of 720 utterances from a single male speaker and the TIMIT corpus has 6300 utterances from multiple male and female speakers. Three non-overlapping sets of 50, 11 and 18.3 hrs are developed for the training, cross-validation, and testing

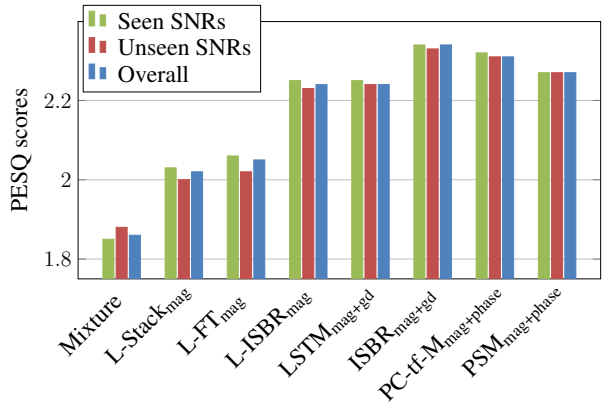


Fig. 4: PESQ scores for seen, unseen and overall SNR conditions for the IEEE corpus.

sets, respectively. The training and validation data is generated at -3, 0, and 3 dB signal-to-noise ratios (SNRs) using four noise types (speech-shaped noise, cafeteria, factory, and babble). We test with two additional SNRs (-6 and 6 dB), which are unseen by the model. All the signals are sampled at 16 kHz. The spectrogram is generated using a 640-point DFT with a Hann window of 40ms and a 20ms frameshift.

Our baseline LSTM network has a single LSTM layer of 256 cells, a time-distributed dense layer (321 units) and four separate output layers (one for each target) in parallel. The rectified linear (ReLU) function is used as the activation function for the two output layers that predict the clean and noise magnitude spectrograms. A linear activation function is used in the dense layer and the two output layers for predicting group-delay of the speech and noise signals. A sigmoid logistic function is used for the gate activation function, while hyperbolic tangent functions are used for the cell and hidden states. Batch normalization is performed between each layer. Adam optimization is used with momentum with the learning rate of 0.001. We use Xavier initialization to initialize the model. In the loss function, λ is set to 0.975. In our proposed approach, the output layer of the LSTM network is replaced by the ISBR layer (321 units), and the model is retrained. We denote our approach as ISBR$_{mag+gd}$.

We compare our approach against approaches that enhance only the magnitude of speech. In L-Stack$_{mag}$ model [17], prediction from a time LSTM (T-LSTM) and a frequency LSTM (F-LSTM) are stacked together to predict enhanced magnitude. L-FT$_{mag}$ [16] uses a F-LSTM to summarize frequency information in a super vector by scanning frequency sub-bands of a time frame. Then a T-LSTM layer uses this super vector to learn temporal dependencies. To evaluate the effectiveness of the ISBR layer, we compare our proposed ISBR output layer to our prior approach that uses a LSTM that is trained with the $\mathcal{L}_{mag}$ cost function (e.g., no group delay) [18]. This is denoted as L-ISBR$_{mag}$. Additionally, we define our baseline LSTM network with traditional dense layers as an output layer and denote this network as LSTM$_{mag+gd}$ (e.g., no ISBR layer). We compare two masking-based approaches [11, 5] because in general, the mask-based approaches have shown superior performance in speech enhancement, when compared to signal approximation approaches. A phase-sensitive filter is approximating in [11]. We denote this approach as PSM$_{mag+phase}$. Additionally, [5] proposes a parametric complex-valued T-F mask to estimate magnitude and phase through a joint learning network. We call this network PC-tf-M$_{mag+phase}$. All the approaches are evaluated with three commonly-used objective metrics, namely, the Perceptual Evaluation of Speech Quality (PESQ) [23], the short-time objective intelligibly (STOI) [24] and the scale-invariant speech distortion ratio (SI-SDR) [25, 26].

Table 1: Average scores for each approach. Best results are shown in **bold**.

	IEEE corpus			TIMIT corpus		
	PESQ	STOI	SI-SDR	PESQ	STOI	SI-SDR
Mixture	1.86	0.62	-1.47	1.58	0.51	-2.33
L-Stack$_{mag}$ [17]	2.02	0.59	-0.59	1.82	0.5	-0.84
L-FT$_{mag}$ [16]	2.05	0.6	-0.2	1.88	0.52	-0.26
L-ISBR$_{mag}$ [18]	2.24	0.64	0.22	1.93	0.52	-0.03
LSTM$_{mag+gd}$	2.24	0.64	0.12	1.97	0.53	0.1
ISBR$_{mag+gd}$	**2.34**	**0.67**	**0.92**	**2.04**	**0.58**	**0.84**
PC-tf-M$_{mag+phase}$[5]	2.31	**0.67**	0.85	**2.04**	**0.58**	0.72
PSM$_{mag+phase}$[11]	2.27	0.65	0.4	2	0.56	0.32

5. RESULTS

In Figure 4, we compare the performance scores of all the models across different SNR conditions. Evaluating the IEEE dataset with PESQ, our proposed ISBR$_{mag+gd}$ shows the best performance among all the models in seen (-3,0,3 dB), unseen (-6,6 dB) and overall SNRs on average. Additionally, L-ISBR$_{mag}$ shows big improvements among all the magnitude-based approaches. In Table 1, we compare the performance of the models in a single speaker (IEEE corpus) and multiple speakers (TIMIT corpus) scenario. ISBR$_{mag+gd}$ shows the best performance in all performance metrics. However, PC-tf-M$_{mag+phase}$ shows good performance and is tied with ISBR$_{mag+gd}$ multiple times, especially for the TIMIT corpus. It is important to note that our signal approximation approach outperforms the T-F masking approaches, which indicates that incorporating spectral-level magnitude and phase dependencies are beneficial.

6. CONCLUSION

Our proposed output layer with a base LSTM network successfully captures the temporal and spectral level dependencies in the magnitude and phase domains. The results show its superiority over traditional approaches and robustness on unseen noise and data. This model, however, considers only the first-order Markovian assumption. In the future, we will explore higher-order spectral dependencies along with sub-band spectral dependencies in a single time frame.

7. REFERENCES

[1] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *The Journal of the Acoustical Society of America*, vol. 123, pp. 1673–1682, 2008.

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

[3] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM TASLP*, vol. 22, pp. 1849–1858, 2014.

[4] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM TASLP*, vol. 24, pp. 483–492, 2016.

[5] J. Lee and H.-G. Kang, "A joint learning algorithm for complex-valued tf masks in deep learning-based single-channel speech enhancement systems," *IEEE/ACM TASLP*, vol. 27, pp. 1098–1109, 2019.

[6] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[7] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. GlobalSIP*, pp. 577–581, 2014.

[8] B. O. Odelowo and D. V. Anderson, "A study of training targets for deep neural network-based speech enhancement using noise prediction," in *Proc. ICASSP*, pp. 5409–5413, 2018.

[9] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM TASLP*, vol. 23, pp. 7–19, 2014.

[10] D. S. Williamson, "Monaural speech separation using a phase-aware deep denoising auto encoder," in *Proc. MLSP*, pp. 1–6, 2018.

[11] 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*, pp. 708–712, 2015.

[12] H. Zhao, S. Zarar, I. Tashev, , and C.-H. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *Proc. ICASSP*, pp. 2401–2405, 2018.

[13] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, pp. 31–35, 2016.

[14] S. W. Fu, T. W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM TASLP*, vol. 26, pp. 1570–1584, 2018.

[15] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Upper Saddle River, NJ: Prentice Hall, 1st ed., 2002.

[16] J. Li, A. Mohamed, G. Zweig, and Y. Gong, "LSTM time and frequency recurrence for automatic speech recognition," in *Proc. ASRU*, pp. 187–191, 2015.

[17] J. Deng, B. Schuller, F. Eyben, D. Schuller, Z. Zhang, H. Francois, and E. Oh, "Exploiting time-frequency patterns with lstm-rnns for low-bitrate audio restoration," *Neural Computing and Applications*, pp. 1–13, 2019.

[18] K. M. Nayem and D. S. Williamson, "Incorporating intra-spectral dependencies with a recurrent output layer for improved speech enhancement," in *Proc. MLSP*, 2019.

[19] Z.-Q. Wang, K. Tan, and D. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," in *Proc. ICASSP*, pp. 71–75, 2019.

[20] H. A. Murthy and B. Yegnanarayana, "Speech processing using group delay functions," *Signal Processing*, vol. 22, pp. 259–267, 1991.

[21] E. Rothauser, "IEEE recommended practice for speech quality measurements," *Proc. IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[22] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[23] 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, pp. 749–752, 2001.

[24] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM TASLP*, vol. 19, pp. 2125–2136, 2011.

[25] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM TASLP*, vol. 14, pp. 1462–1469, 2006.

[26] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. ICASSP*, pp. 626–630, 2019.