WEIGHTED MAGNITUDE-PHASE LOSS FOR SPEECH DEREVERBERATION

Jingshu Zhang, Mark D. Plumbley, Wenwu Wang

Centre for Vision, Speech and Signal Processing, University of Surrey

ABSTRACT

In real rooms, recorded speech usually contains reverberation, which degrades the quality and intelligibility of the speech. It has proven effective to use neural networks to estimate complex ideal ratio masks (cIRMs) using mean square error (MSE) loss for speech dereverberation. However, in some cases, when using MSE loss to estimate complex-valued masks, phase may have a disproportionate effect compared to magnitude. We propose a new weighted magnitude-phase loss function, which is divided into a magnitude component and a phase component, to train a neural network to estimate complex ideal ratio masks. A weight parameter is introduced to adjust the relative contribution of magnitude and phase to the overall loss. We find that our proposed loss function outperforms the regular MSE loss function for speech dereverberation.

Index Terms— speech dereverberation, complex ideal ratio mask, deep neural network, loss function

1. INTRODUCTION

Speech captured in indoor environments usually contains reverberation components. It has been shown that reverberation has a detrimental effect on speech quality, and can degrade speech intelligibility for both hearing impaired listeners and normal hearing listeners [1]. Therefore, speech dereverberation plays an important role in processing speech signals produced in reverberant environments.

Due to recent advances in deep learning, approaches based on deep neural networks (DNNs) have been widely introduced in speech processing. For instance, regression models [2], feed forward neural networks [3], and recurrent neural networks (RNNs) [4] have been proposed for speech enhancement and dereverberation. These and many other approaches often enhance only the magnitude spectrogram and use the unprocessed noisy phase when reconstructing estimated speech [5]. However, the importance of phase has recently been discussed in [6], and phase processing in speech processing has been gaining increasing attention. The Griffin-Lim Algorithm [7] is a widely-adopted phase processing algorithm that tries to find the closest consistent spectrogram to a target. Following this idea, various phase processing approaches have been proposed [8]. Apart from these methods that focus on estimating phase, Williamson and Wang [9] proposed to estimate complex ideal ratio masks (cIRMs) to jointly modify magnitude and phase. The complex ratio mask has proven to be effective in speech denoising and dereverberation [9] [10], and has also been adopted in other speech enhancement approaches [11] [12].

Williamson and Wang [9] adopted a feed forward neural network with fully connected layers to estimate cIRMs with a mean square error (MSE) loss function. When estimating masks in the complex domain, MSE measures the Euclidean distance in the complex plane between the targets and the estimates. When computing the Euclidean distance, phase may in some case have more impact than magnitude, which can lead to estimates with a small Euclidean distance but a large magnitude difference compared to the target. However, it is generally considered that most of the insight about the structure of the speech is obtained from the magnitude [5], so we assume that the magnitude is generally more important than the phase. If this is true, estimates with a small magnitude difference should be favored, and the magnitude should weigh more than the phase when measuring the overall loss.

In this paper, we propose a loss function designed to measure the loss in magnitude and phase separately, and the relative contribution of phase can be adjusted by introducing a weight coefficient in the loss function. We train a convolutional recurrent neural network (CRNN) [13][14] to estimate cIRMs using our proposed loss function, and investigate the performance of our loss function as well as the impact of magnitude and phase in this context.

This paper is organized as follows. The cIRMs and our proposed loss function are introduced in Section 2 and Section 3 respectively. In Section 4, the experiments and the results are presented, and Section 5 concludes this paper.

2. COMPLEX IDEAL RATIO MASK FOR SPEECH DEREVERBERATION

The complex ideal ratio mask $M(t, f)$ at frame *t* and frequency bin *f* is defined as [15]

$$M(t,f) = \frac{S(t,f)}{Y(t,f)} = \frac{|S(t,f)|}{|Y(t,f)|} e^{j(\theta_s(t,f) - \theta_y(t,f))} \qquad (1)$$

where $S(t, f)$ and $Y(t, f)$ are the complex spectrograms of the desired signal and the observed signal at time frame *t* and frequency bin f, and θ_s and θ_y are their respective phases. This mask is defined in the complex domain, and can modify both magnitude and phase of the spectrogram, and hence can theoretically fully recover the desired signals.

The MSE loss function used to estimate cIRMs is defined as follows

$$\begin{aligned}\mathcal{L}_{\mathrm{MSE}} =& \frac{1}{2N}\sum_t\sum_f[(M_{\mathrm{R}}(t,f)-\hat{M}_{\mathrm{R}}(t,f))^2 \\ &+(M_{\mathrm{I}}(t,f)-\hat{M}_{\mathrm{I}}(t,f))^2]\end{aligned} \tag{2}$$

where $M_{\mathrm{R}}(t,f)$ and $M_{\mathrm{I}}(t,f)$ are real and imaginary components of target masks, and $\hat{M}_{\mathrm{R}}(t,f), \hat{M}_{\mathrm{I}}(t,f)$ are their estimates. The MSE loss measures the Euclidean distance between the target vector and its estimate in the complex domain.

Let us consider the situation shown in Figure 1(a), where $\mathbf{x}_a$ and $\mathbf{x}_b$ are estimates of the same target $\mathbf{s}$. The estimate $\mathbf{x}_b$ successfully estimates the magnitude $|\mathbf{s}|$, but is a poor estimate of the phase. In comparison, the phase of $\mathbf{x}_a$ is close to the target, but the difference in magnitude is rather large. The Euclidean distance between $\mathbf{s}$ and $\mathbf{x}_a$ is much smaller than that between $\mathbf{s}$ and $\mathbf{x}_b$, so a neural network trained using MSE loss is more likely to favor an estimate with a small Euclidean distance but a large difference in magnitude such as $\mathbf{x}_a$.

However, it is generally considered that a large proportion of the insight about the structure of the speech signal can be obtained from the magnitude spectrogram [5]. Therefore, we assume the magnitude is more important than phase in this context, and estimates with smaller magnitude difference, such as $\mathbf{x}_b$, may be preferred. Neural networks trained with the MSE loss would not consider magnitude errors more important than phase errors, and would possibly degrade the overall performance.

3. PROPOSED APPROACH

3.1. Weighted magnitude-phase loss function

To overcome this limitation, we look at this problem from a geometry point of view. As illustrated in Figure 1(b), the loss in magnitude and phase can be represented by the lengths of BB' and AB' respectively. The estimate $\mathbf{x}$ could approach the target $\mathbf{s}$ in two steps, closing the gap in magnitude and rotating to reduce the angle θ.

To create a weighted magnitude-phase (WMP) loss function, we divide the loss function into two parts: a magnitude part BB', and a phase part AB'. We would like the magnitude error BB' to be considered more important than the phase error AB', so a weight parameter is introduced to reduce the importance of phase error. This loss function can be formulated as follows

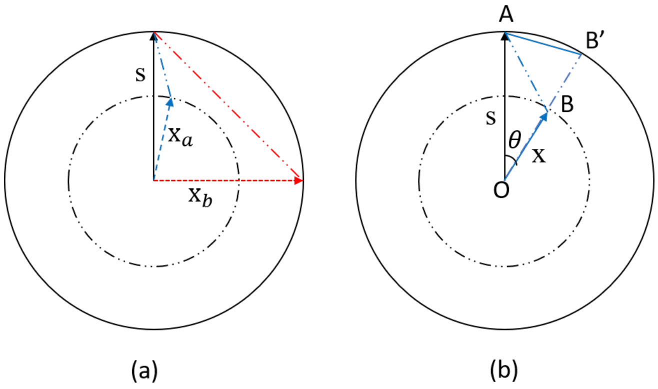


Fig. 1: (a) Illustration of MSE loss function. The black arrow is the target vector $\mathbf{s}$ while the blue dash arrow and the red dot arrows are two estimates $\mathbf{x}_a$ and $\mathbf{x}_b$ respectively. Two dot-dashed lines indicate the values of MSE loss of corresponding estimates. (b) Interpretation of proposed loss function. Vector $\mathbf{s}$ is the target and $\mathbf{x}$ is the estimate. Line AB' and BB' indicate the value of phase part and magnitude part in loss function respectively.

$$\begin{aligned}\mathcal{L}_{\mathrm{WMP}} =& \frac{1}{2N}\sum_t\sum_f(|M(t,f)|-|\hat{M}(t,f)|)^2 \\ &+\alpha\left[|M(t,f)|\sin\left(\frac{\theta_M(t,f)-\hat{\theta}_M(t,f)}{2}\right)\right]^2\end{aligned} \tag{3}$$

where M and $\hat{M}$ are the complex-valued target mask and its estimate, and θ_M, $\hat{\theta}_M$ are their corresponding phases. The parameter α is introduced as a weight parameter to adjust the contribution that the phase part makes to the the overall loss. The phase difference $\theta = \theta_M - \hat{\theta}_M$ can be wrapped into $[-\pi, \pi]$. When the wrapped phase difference is π or $-\pi$, the value of $\sin^2\left(\frac{\theta_M(t,f)-\hat{\theta}_M(t,f)}{2}\right)$ is maximized at 1, and gradually decreases to 0 when the wrapped phase difference approaches 0. We also include the magnitude of target signal $|M(t,f)|$ in the second term of Equation (3) so that the phases in time-frequency (T-F) units where the magnitude is large weigh more than that in other T-F units.

We illustrate how the value of the loss function changes with the different estimate in Figure 2, using a weight $\alpha = 0.1$. The lowest point (red spot in Figure 2) indicates the perfect estimate, where the value of the loss is 0. We can find a ring-shaped area (red ring in Figure 2) that is lower than the surrounding areas, where the magnitude of estimated vectors is the same with the target. Therefore, when trained with this loss function, a neural network is more likely to converge to a point in this ring-shaped area where a better estimate of the magnitude can be obtained.

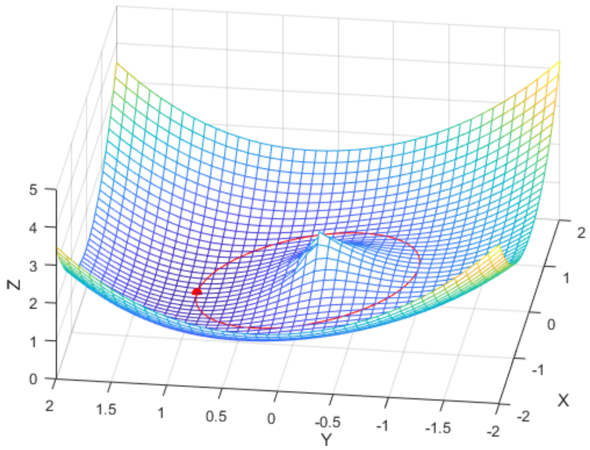


Fig. 2: The value of loss function on a complex plain. The X and Y represent real and imaginary parts, while the Z axis is the value of the loss function. The red dot is the target, and the red circle represents the estimates with the same magnitude with the target.

3.2. cIRM estimation using proposed loss function

In this study, we train a convolutional recurrent neural network (CRNN) to estimate cIRMs from reverberant speech features.

The input features are complex spectrograms of reverberant speech. The spectrograms are computed using a 512-point short-time Fourier transform (STFT), which are then reshaped into real-valued tensors with size of $(2 \times T \times 257)$, where T is the number of time frames. The targets are compressed cIRMs, which are obtained by using the compression function adopted in [9] to map the value of real and imaginary components of masks into the range of $[-K, K]$, after the masks are computed using Equation (1).

The architecture of our model is shown in Table 1. The hyperparameters of the convolution kernels in each layers are given in *kernelSize, strides, outChannels* format. The input size and the output size of each layer are specified in *featureMaps* × *timeSteps* × *frequencyBins* format. The number of feature maps in each decoder layer is doubled by skip connections, which are added between corresponding conv2d and deconv2d layers. After each convolution layer and deconvolution layer, PReLU activation and batchnorm are applied, while tanh activation is adopted after the output layer (deconv2d 1). After the cIRMs are obtained, the estimated clean speech is computed by multiplying the masks with the reverberant speech spectrograms followed by the inverse short time Fourier transform (iSTFT).

Table 1: Architecture of our adopted CRNN

layer name	input size	hyperparameters	output size
conv2d 1	$2 \times T \times 257$	$5 \times 5, (1, 2), 16$	$16 \times T \times 129$
conv2d 2	$16 \times T \times 129$	$5 \times 5, (2, 2), 32$	$32 \times T/2 \times 65$
conv2d 3	$32 \times T/2 \times 65$	$5 \times 5, (1, 2), 64$	$64 \times T/2 \times 33$
conv2d 4	$64 \times T/2 \times 33$	$5 \times 5, (2, 2), 128$	$128 \times T/4 \times 17$
conv2d 5	$128 \times T/4 \times 17$	$5 \times 5, (1, 2), 256$	$256 \times T/4 \times 9$
reshape	$256 \times T/4 \times 9$	-	$T/4 \times 2304$
blstm	$T/4 \times 2304$	2304	$2 \times T/4 \times 2304$
conv2d	$2 \times T/4 \times 2304$	$1 \times 1, (1, 1), 1$	$T/4 \times 2304$
reshape	$T/4 \times 2304$	-	$256 \times T/4 \times 9$
deconv2d 5	$512 \times T/4 \times 9$	$5 \times 5, (1, 2), 128$	$128 \times T \times 17$
deconv2d 4	$256 \times T/4 \times 17$	$5 \times 5, (2, 2), 64$	$64 \times T \times 33$
deconv2d 3	$128 \times T/2 \times 33$	$5 \times 5, (1, 2), 32$	$32 \times T \times 65$
deconv2d 2	$64 \times T/2 \times 65$	$5 \times 5, (2, 2), 16$	$16 \times T \times 129$
deconv2d 1	$32 \times T \times 129$	$5 \times 5, (1, 2), 2$	$2 \times T \times 257$

4. EXPERIMENT

4.1. Experiment setting

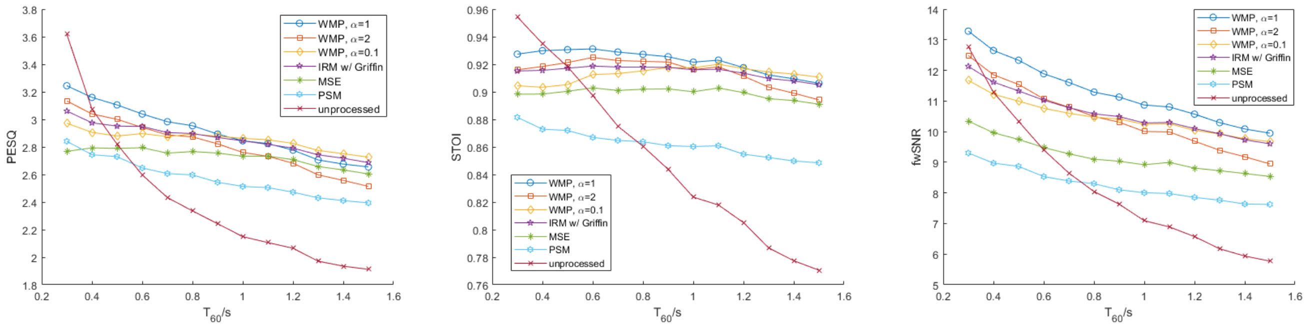
We use speech signals with a sample rate of 16kHz. When performing the STFT, the speech signals are divided into frames of 32 ms with an 8 ms frame shift (i.e. 75% overlap), and a Hanning window is applied to the frames.

The DNN is trained with simulated room impulse responses (RIRs) generated using a RIR generator[1] based on the image method. We use a similar setting to [4] and [16] to generate our RIRs. The simulated RIRs are generated for five simulated rooms with size of $9m \times 8m \times 7m$, $10m \times 7m \times 3m$, $6m \times 6m \times 10m$, $8m \times 10m \times 4m$, and $7m \times 7m \times 8m$. A sound source and a microphone are randomly placed in each simulated room with a fixed distance of 1m between the source and the microphone. The reverberation time (T_{60}) is set to increase from 0.3s to 1.5s with a step of 0.1s. Twenty different RIRs are generated for each T_{60} and for each room, which results in $5 \times 13 \times 20 = 1300$ RIRs in total. The RIRs from first three rooms (780 RIRs) are used for training, while the RIRs from the fourth room (260 RIRs) are used for validation and those from the fifth room (260 RIRs) are used for testing.

We use speech utterances from the ARU speech corpus [17], which comprises single channel recordings of IEEE (Harvard) sentences spoken by native British English speakers. Following [3] and [9], we use 500 utterances from a single speaker for training. In the training stage, each RIR from the training set is combined with 50 random utterances chosen from the 500 training utterances. A validation set is formed by convolving 10 different utterances from the same speaker with validation RIRs, and another 10 sentences by the same speaker are convolved with testing RIRs to form a test set. Before convolving RIRs with utterances, initial delays in RIRs are removed to improve time alignment.

The model is trained using our proposed WMP loss function with different weights α, and compared with the MSE

[1] https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator



(a) PESQ score (b) STOI score (c) frequency-weighted SNR

Fig. 3: PESQ score, STOI score and frequency-weighted SNR of unprocessed speech and speech processed by different methods under different T_{60}

loss function used by Williamson and Wang [9]. For comparison, we also train models to estimate phase-sensitive masks (PSMs) [18] and ideal ratio masks (IRMs) [19] using the MSE loss. The phases of IRMs are modified using the Griffin-Lim algorithm. The performance is evaluated using perceptual evaluation of speech quality (PESQ) [20], short-time objective intelligibility (STOI) [21] and frequency-weighted segmental signal-to-noise ratio ($\mathrm{SNR_{fw}}$) [22]. The source code is available at Github.[2].

4.2. Experiment results

The results of the experiments are shown in Figure 3. It can be seen that our approach leads to improved speech quality except for the case when the reverberation time is small ($T_{60} \leq 0.4s$), where the quality of unprocessed speech is already high. The performance of the models trained with our proposed WMP loss function surpasses the model trained using the MSE loss function in most cases. In terms of the PESQ score, the model trained using the WMP loss function with a weight of $\alpha = 1$ gives the best results when reverberation is not too high ($T_{60} < 1s$), while a smaller weight of $\alpha = 0.1$ gives a better result in high reverberation conditions ($T_{60} > 1s$). A similar trend can be seen for the STOI score: a weight of $\alpha = 1$ leads to the best STOI score in low and moderate reverberation conditions, while a smaller weight of $\alpha = 0.1$ performs better when the reverberation time is very large ($T_{60} > 1.2s$). For frequency-weighted segmental SNR, the weight of $\alpha = 1$ achieves the best performance. We also notice that the performance of WMP loss with a weight of $\alpha = 0.1$ is close to that of the IRM with the Griffin-Lim algorithm. The reason might be that WMP with a small weight focuses more on estimating the magnitude.

We also measure the average squared difference of magnitude between unprocessed or estimated speech and clean speech (Δmagnitude) as well as the average difference of phase (Δphase) measured in radians. The results are shown in Table 2. Our proposed loss function leads to estimates with smaller magnitude difference compared to MSE, PSM and IRM. The weight of $\alpha = 1$ gives the best estimates of magnitude and phase, corresponding to the best results in speech quality as shown in Figure 3.

Table 2: Average difference of magnitude and phase between unprocess/estimated speech and clean speech

	Δmagnitude	Δphase/rad
unprocessed	0.0509	2.083
MSE	0.0349	1.232
WMP, α=2	0.0280	1.235
WMP, α=1	**0.0258**	**1.215**
WMP, α=0.1	0.0260	1.263
IRM w/ Griffin	0.0270	2.081
PSM	0.0411	1.589

5. CONCLUSION

In this paper, we propose a weighted magnitude-phase loss function where magnitude and phase are measured separately. A weight parameter α is introduced to adjust the proportion of magnitude and phase in loss function. Experiments are conducted to test its performance in speech dereverberation. The result shows that our loss function outperforms the conventional MSE loss function, as well as PSM and IRM.

6. ACKNOWLEDGEMENTS

This research was supported by a PhD scholarship from the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey.

[2] https://github.com/ZhangJingshu/WMP-loss-for-dereverb

7. REFERENCES

[1] J. Xia, B. Xu, Sh. Pentony, J. Xu, and J. Swaminathan, "Effects of reverberation and noise on speech intelligibility in normal-hearing and aided hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1523–1533, Mar. 2018.

[2] B. Wu, K. Li, M. Yang, and C. H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 102–111, Jan. 2017.

[3] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, June 2015.

[4] J. F. Santos and T. H. Falk, "Speech dereverberation with context-aware recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1236–1246, July 2018.

[5] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, Mar. 2015.

[6] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, Apr. 2011.

[7] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[8] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Deep Griffin–Lim iteration," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, May 2019, pp. 61–65.

[9] D. S. Williamson and D. Wang, "Speech dereverberation and denoising using complex ratio masks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, Mar. 2017, pp. 5590–5594.

[10] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1492–1501, July 2017.

[11] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "PHASEN: A phase-and-harmonics-aware speech enhancement network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 9458–9465, 04 2020.

[12] J. Lee and H. G. Kang, "A joint learning algorithm for complex-valued T-F masks in deep learning-based single-channel speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1098–1108, June 2019.

[13] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Interspeech 2018*. Sept. 2018, pp. 3229–3233, ISCA.

[14] T. Grzywalski and S. Drgas, "Application of recurrent U-net architecture to speech enhancement," in *2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, Poznan, Sept. 2018, pp. 82–87, IEEE.

[15] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[16] Y. Li and D. S. Williamson, "A return to dereverberation in the frequency domain using a joint learning approach," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 7549–7553.

[17] C. Hopkins, S. Graetzer, and G. Seiffert, "ARU adult British English speaker corpus of IEEE sentences (ARU speech corpus) version 1.0," 2019. DOI: 10.17638/datacat.liverpool.ac.uk/681.

[18] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, Apr. 2015, pp. 708–712.

[19] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[20] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, UT, USA, 2001, vol. 2, pp. 749–752.

[21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sept. 2011.

[22] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387, 2009.