INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Constrained Ratio Mask for Speech Enhancement Using DNN

Hongjiang Yu[1], *Wei-Ping Zhu*[1] *and Yuhong Yang*[2]

[1]Department of Electrical and Computer Engineering, Concordia University, Canada
[2]National Engineering Research Center for Multimedia Software, Wuhan University, China

ho_yu@encs.concordia.ca, weiping@ece.concordia.ca, yangyuhong@whu.edu.cn

Abstract

Speech enhancement has found many applications concerning robust speech processing. A masking based algorithm, as an important method of speech enhancement, aims to retain the speech dominant components and suppress the noise dominant parts of the noisy speech. In this paper, we derive a new type of mask: constrained ratio mask (CRM), which can better control the trade-off between speech distortion and residual noise in the enhanced speech. A deep neural network (DNN) is then employed for CRM estimation in noisy conditions. The estimated CRM is finally applied to the noisy speech for denoising. Experimental results show that the enhanced speech from the new masking scheme yields an improved speech quality over three existing masks under various noisy conditions.

Index Terms: speech enhancement, constrained ratio mask, deep neural network

1. Introduction

Speech enhancement, which aims to obtain the clean speech estimate under noisy environment, has been widely adopted in robust speech processing related applications. Various speech enhancement methods have been proposed during the past decades, among which the masking based algorithms have received much attention and achieved a series of progresses [1]. Inspired by the masking effect of the human auditory mechanism, the goal of this kind of methods is to estimate a mask, which can be applied into the noisy speech to retain the speech dominant regions and suppress the noise dominant regions. An appropriate and accurate mask is of great importance to the enhancing performance. To this end, researchers have made large efforts from two aspects: finding an optimal mask and developing reliable mask estimation algorithms.

The ideal binary mask (IBM) [2] is one of the pioneering masks investigated in the literature. Given the clean speech on a time-frequency (T-F) representation, the mask value of a T-F unit is set to 1 if the local signal-to-noise ratio (SNR) is greater than a preset threshold, otherwise it is set to 0. This simplifies speech enhancement to a binary classification. However, speech enhancement using IBM has some limitations such as introducing the residual musical noise. As a result, the ideal ratio mask (IRM) [3, 4], which can be viewed as a smoothed form of IBM, is proposed. The IRM is obtained by computing the ratio between the energy of clean speech and that of noisy speech for each T-F unit. Denoising with IRM is actually assigning large ratios to the T-F units with higher local SNR and small ratios to those with lower local SNR. Another mask with similar concept is the spectral magnitude mask or ideal amplitude mask (IAM) [5], which computes the ratio of the magnitude of clean speech to that of noisy speech. Note that IRM and IAM are motivated by the frequency response of the Wiener filter, which achieves the optimal signal-to-noise ratio (SNR) gain for stationary signals. However, speech signals and many real-world noises are nonstationary. To overcome this problem, the optimal ratio mask (ORM) is proposed in [6], by considering the correlation between the desired speech and noise, leading to an improved SNR of the enhanced speech. The above mentioned masks only focus on enhancing the magnitude spectrogram. More recently, the phase information has been considered in masking techniques, such as the phase sensitive mask (PSM) [7] and the complex IRM (cIRM) [8], to better recover the complex speech spectrogram.

Probably, supervised learning algorithms, such as Gaussian mixture model (GMM) [9] or support vector machine (SVM) [10], are the most popular mask estimation methods in early works. In recent years, the deep learning based methods have made a great progress. A feed-forward deep neural network is adopted in [4] to learn the mapping between noisy acoustic features and IRM. Other achitectures can be found in [11, 12], where recurrent neural network and convolutional neural network are, respectively, employed as the estimation model. The deep structure and powerful learning capability enable DNN to better explore the non-linear relationship between noisy features and masks, leading to a better estimation result.

Although the DNN estimated masks have achieved good performance in improving speech intelligibility, none of these works further investigates the trade-off between the speech distortion and residual noise in the enhanced speech. Several traditional methods have been proposed to denoise with less speech distortion and remove residual noise as much as possible. In [13], the authors proposed a new weighting rule using masking properties, which calculates the weighting coefficients to keep the perceived noise to be equal to a pre-defined level. However, the speech distortion is not explicitly considered in their processing. In [14], a spectral constraint is applied into the short-time spectral amplitude (STSA) estimator, which adaptively suppresses the noise dominant regions and reduces the speech distortion in the speech dominant regions.

In this paper, we propose a constrained ratio mask (CRM) for speech enhancement, which is derived to minimize the speech distortion while suppressing the residual noise such that it falls below a threshold level. A DNN is then trained for CRM estimation, which learns a mapping from the noisy features to the CRM. Finally, the enhanced speech is obtained by applying the estimated CRM to the noisy speech. Compared with the previous mask based algorithms, which mainly focus on retaining the speech information, our CRM based system is the first one to consider both speech distortion and the residual noise level in the enhanced speech, by adaptively adjusting the value of the CRM for different T-F units according to their local SNRs. Experimental results have shown that our proposed system yields

The authors acknowledge the support from China Scholarships Council (CSC No.201606270200) and the NSERC of Canada under a CRD project sponsored by Microchip in Ottawa, Canada.

http://dx.doi.org/10.21437/Interspeech.2020-1920

better speech quality and similar speech intelligibility as compared to several previous mask based algorithms.

2. Constrained ratio mask

We consider a noisy speech $y(t)$ as the addition of clean speech $x(t)$ and background noise $n(t)$, with t denoting the time index. The time domain noisy speech can be transformed into a spectro-temporal spectrogram using short-time Fourier transform (STFT), namely,

$$Y(k,l) = X(k,l) + N(k,l) \tag{1}$$

where $Y(k,l)$, $X(k,l)$ and $N(k,l)$ denote the STFT spectrograms of the noisy speech, clean speech and noise, respectively, with k and l indicating the frequency bin and frame index. We denote the ratio mask as $M(k,l) \in [0,1]$, which will be applied into the magnitude of the noisy speech to get the enhanced magnitude, i.e., $|\hat{X}| = M \cdot |Y|$. For simplicity, we have omitted k and l without loss of generality. By directly using the noisy phase for speech reconstruction, we obtain the STFT of the enhanced speech,

$$\hat{X} = (M \cdot |Y|)\, e^{j\phi_y} = M \cdot Y \tag{2}$$

To derive a CRM, we first introduce the mathematical definitions of the speech and noise distortions as given in [13, 15]. The estimation error e of $\hat{X}$ can be decomposed into two items as follows,

$$\begin{aligned} e &= \hat{X} - X = M \cdot (X + N) - X \\ &= (M-1) \cdot X + M \cdot N \\ &= e_x + e_n \end{aligned} \tag{3}$$

The power spectrums of e_x and e_n can be written as,

$$\begin{aligned} d_x = \mathrm{E}\left[{e_x}^2\right] &= (M-1)^2 \cdot P_x \\ d_n = \mathrm{E}\left[{e_n}^2\right] &= M^2 \cdot P_n \end{aligned} \tag{4}$$

where P_x and P_n are the power spectrums of the clean speech and noise, respectively; d_x denotes the speech distortion and d_n the noise distortion. We regard d_x as the distortion to the original clean speech introduced by the enhancement algorithm, while d_n is the distortion caused by the residual noise. The above mentioned two distortion terms with respect to the value of mask M are plotted in Fig.1 for three different values of input SNR $\xi = P_x/P_n$.

Ideally, we prefer a large value of M to yield a small speech distortion d_x under the circumstance of the clean speech being much stronger than the noise (i.e., $\xi \gg 1$). Conversely, if the signal is weaker than the noise (i.e., $\xi \ll 1$), a small value of M is required so that the residual noise or d_n will be small. However, we found that when employing a DNN to estimate the IRM (tested under four different noises and input SNRs), the values of estimated IRM are $0.15\% \sim 5.67\%$ higher than those of reference IRM, which results in a larger d_n, indicating that the enhanced speech suffers more residual noise.

To better remove the residual noise with no noticeable speech distortion, we need to derive a mask to minimize the speech distortion while constraining the noise distortion below a threshold. To this end, we establish the following constrained optimization problem,

$$\begin{aligned} &\min_{M} \; d_x \\ &\text{subject to}\, d_n \le \delta \end{aligned} \tag{5}$$

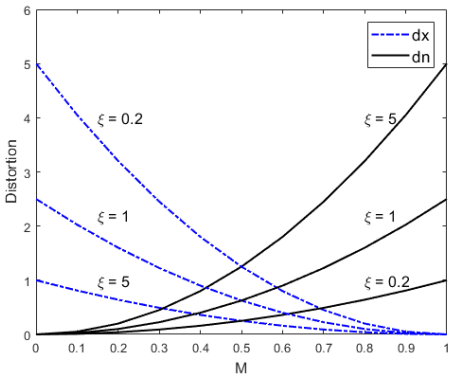

Figure 1: *The relationship between M and distortions*

where δ is a preset threshold. It has been shown in [14] that the optimal M for (5) satisfies the following equation:

$$(M-1)\, P_x + \mu M P_n = 0 \tag{6}$$

where μ ($\mu \ge 0$) is the Lagrange multiplier (also named as the controlling factor in our paper). From (6), the CRM can be expressed as

$$M = \frac{P_x}{P_x + \mu P_n} = \frac{\xi}{\xi + \mu} \tag{7}$$

It should be noted that due to the unknown SNR ξ, the controlling factor μ is to be determined. In other words, μ can be viewed as a function of the local SNR or ξ in dB. Using (7) into (4), the speech and noise distortions can be rewritten as,

$$\begin{aligned} d_x &= (M-1)^2 \cdot P_x = \left(\frac{\mu}{\xi+\mu}\right)^2 \cdot P_x \\ d_n &= M^2 \cdot P_n = \left(\frac{\xi}{\xi+\mu}\right)^2 \cdot P_n \end{aligned} \tag{8}$$

By adaptively adjusting the value of μ for each T-F unit, our CRM can balance the trade-off between the speech and noise distortions. For example, we would like to set a small value of μ for the speech dominant unit, in order to minimize the speech distortion and conserve the speech information; while for the noise dominant unit, a large value of μ is chosen to remove the noise as much as possible. As such, we propose the following empirical expression for μ:

$$\mu = \begin{cases} \mu_0 - \mathrm{SNR}/s & , \quad S_l \le \mathrm{SNR} \le S_u \\ \mu_{\min} & , \quad \mathrm{SNR} > S_u \\ \mu_{\max} & , \quad \mathrm{SNR} < S_l \end{cases} \tag{9}$$

where $\mathrm{SNR} = 10\log_{10}\xi$, $\mu_{\max}$ and $\mu_{\min}$ are the maximum and minimum values of μ, respectively, S_l and S_u are the lower and upper bounds of the local SNR, respectively, and μ_0 and s are constants related to $\mu_{\max}$ and $\mu_{\min}$.

3. Proposed speech enhancement system

The proposed system is made of two stages: the off-line training stage and the on-line enhancement stage. In the training stage, a DNN is employed to learn the mapping between the noisy acoustic features and the reference CRM computed from speech databases. In the enhancement stage, given a new noisy speech, its features are extracted and sent to the well-trained

DNN to obtain the CRM estimate, which is then applied to obtain the enhanced magnitude. Finally, the enhanced speech is reconstructed with the enhanced magnitude and noisy phase. The main steps involved in the proposed speech enhancement system are explained below.

3.1. Input features

Choosing appropriate input features for DNN plays an important role in the deep learning algorithms. In order to obtain good estimation performance, different acoustic features have been investigated in [16]. Generally speaking, four kinds of acoustic features are widely adopted in most works as input for DNN based mask estimation. They are the amplitude modulation spectrum (AMS); the relative spectral transform and perceptual linear prediction (RASTA-PLP); the Mel-frequency cepstral coefficients (MFCC) and their deltas; the Gammatone filterbank energies (GF) and their deltas. Since these features lie in different ranges, normalization is required to scale the input features for achieving better results. Moreover, to make use of the temporal information of the speech, the features of two adjacent time frames are incorporated with the current frame to form a input feature set.

3.2. Network structure

The architecture adopted in our method is a fully-connected feed-forward neural network. The DNN has a total of five layers that includes an input layer, an output layer and three hidden layers with 1024 units in each layer. We employ the rectified linear unit (ReLU) as activation function in the hidden layers, and employ the linear function in the output layer.

To learn the weights and biases in the network, the famous back propagation is adopted to update the parameters in the training process. Ideally, the model parameters are trained to minimize the cost function J, which is defined as the mean square error between the reference and the estimated CRM.

$$J = \frac{1}{2L}\sum_{l=1}^{L}\sum_{k=1}^{K}\left(\hat{M}(k,l) - M(k,l)\right)^2 \tag{10}$$

where L denotes the total number of speech frames.

3.3. Waveform reconstruction

In the enhancement stage, the estimated CRM is firstly output by the well-trained DNN. Afterwards, we apply the estimated CRM to the noisy magnitude spectrum to obtain the estimated magnitude. The enhanced speech spectrum is then reconstructed with the estimated magnitude $\hat{X}(k,l)$ and the noisy phase $\phi_y(k,l)$ as $\hat{X}(k,l) = |\hat{X}(k,l)|e^{j\phi_y(k,l)}$. Finally, the enhanced speech $\hat{x}(t)$ is obtained by performing the inverse STFT of $\hat{X}(k,l)$.

4. Experimental results

4.1. Experimental setup

The clean speech database used in our experiment is the TIMIT corpus [17], in which 731 utterances from different female and male speakers are used for the training and 87 utterances used for testing. several types of noises are picked from the NOISEX-92 corpus [18], in which four types (babble, white, buccaneer1, factory) are regarded as seen noises, and the other four (pink, buccaneer2, street, hfchannel) as unseen noises. In the training stage, we mix the clean training speeches with seen noises at four levels (-3dB, 0dB, 3dB, 6dB) of signal-to-noise rates (SNRs) to obtain 11696 noisy speeches. In the enhancement stage, both seen noises and unseen noises are mixed with clean testing speeches at the above four SNR levels. The number of noisy utterances used in enhancement stage is 1392 for both seen noises and unseen noises. The sampling rate of all speech utterances and noises is set to 16 kHz. Hamming window is used in framing and the window size of STFT is 320 with 50% overlap.

To assess the enhancement performance, three objective metrics are adopted in our experiment: the perceptual evaluation of speech quality (PESQ) [19], the short-time objective intelligibility (STOI) [20] and the signal-to-distortion ratio (SDR) [21]. PESQ has a high correlation with the subjective scores of the perceptual speech quality, while STOI focuses on the evaluation of the speech intelligibility. SDR is a measurement that is widely-used in speech separation and source enhancement, which is given by,

$$\text{SDR} = 10\log_{10}\frac{\|x(t)\|^2}{\|\hat{x}(t) - x(t)\|^2} \tag{11}$$

where $\|\cdot\|^2$ takes the power of the signal. For all metrics, a larger score indicates a better performance.

4.2. Controlling factor μ

In this section, we investigate the enhancement performance when setting different values for the controlling factor μ. Firstly, we consider the following three types of lower and upper bounds for the local SNR as given in Table 1. Moreover, we set $\mu_{\min} = 1$, $\mu_{\max} = 10$ and $s = 25/(\mu_{\max} - \mu_{\min})$. A T-F unit will be treated as a noise dominant unit when the local SNR is under S_l. In contrast, a T-F unit is regarded as a speech dominant unit when the local SNR is above S_u.

Table 1: *Different settings of lower and upper bounds*

Type	S_l (dB)	S_u (dB)	μ_0
#1	-15	10	$(3\mu_{\min} + 2\mu_{\max})/5$
#2	-10	15	$(2\mu_{\min} + 3\mu_{\max})/5$
#3	-5	20	$(\mu_{\min} + 4\mu_{\max})/5$
#4	0	25	$\mu_{\max}$

As shown in Table 2, if S_l and S_u are very small, the noise dominant unit would be falsely classified to speech dominant unit, the residual noise would be fully removed and thus the scores of SDR and PESQ will decrease. On the contrary, if the S_l and S_u are too large, the speech unit with low local SNR will be mistakenly considered as noise, which could suppress the speech information leading to a decrease of STOI score. In terms of all metrics, the optimal case is type 3. The corresponding improvement of PESQ and SDR scores is significant, while the STOI score has no obvious degradation, which indicates that it removes the background noise quite well without extra speech distortion. For type 4, although it has the best PESQ and SDR scores, this setting is not perfect as the decrease of STOI score shows that the speech information is damaged compared with other settings. Secondly, we also investigate the different settings for $\mu_{\min}$ (varies from 0.5 to 1.5) and $\mu_{\max}$ (varies from 5 to 15). However, the objective results of the enhanced speech do not change significantly.

Table 2: *Objective results with different controlling factors (on seen noise)*

		-3dB	0dB	3dB	6dB
PESQ	Noisy	1.35	1.54	1.75	1.97
	Type1	1.98	2.25	2.51	2.78
	Type2	2.00	2.27	2.53	2.79
	Type3	2.01	2.29	2.55	**2.81**
	Type4	**2.02**	**2.30**	**2.56**	**2.81**
STOI	Noisy	0.60	0.67	0.74	0.81
	Type1	0.75	0.82	0.87	0.90
	Type2	0.75	0.82	0.87	0.90
	Type3	0.75	0.82	0.87	0.90
	Type4	*0.74*	*0.81*	*0.86*	0.90
SDR	Noisy	-2.85	0.11	3.08	6.07
	Type1	7.08	9.29	11.33	13.46
	Type2	7.29	9.49	11.52	13.62
	Type3	7.44	9.63	11.64	13.73
	Type4	**7.54**	**9.71**	**11.72**	**13.78**

4.3. Performance comparison

To evaluate the enhancement performance, we compare our CRM with three existing masks as shown in Table 3. We adopted type 3 as the setting of the controlling factor. For fair comparison, all masks are estimated by the DNN with the same input features and configurations. As the difference between any two comparison methods is that the DNNs use different masks as their outputs. Hence, we use the mask's name to represent each method in this section.

Table 3: *Comparison masks and definitions*

Mask	Definition
IRM [4]	$\sqrt{P_x/(P_x+P_n)}$
IAM [5]	$\lvert X\rvert/\lvert Y\rvert$
OPM [6]	$(P_y+P_x-P_n)/2P_y$

a) Seen noise: Table 4 gives the average objective score on seen noise. Obviously, the enhanced speech from CRM reaches the highest score under all metrics in most cases. More specifically, the proposed method has a large improvement on SDR scores, which means our enhanced speech has a higher SNR. The improvement indicates that our system strengthens the suppression of noise in noise dominant units using a large controlling factor. Our enhanced speech also obtains the best PESQ score, which reflects a good perceptual speech quality. The improvement of SDR and PESQ demonstrates that our CRM is better at noise suppression, especially in the noise dominant regions. Compared with the SDR and PESQ, the improvement of STOI is not that obvious. This is because the STOI algorithm mainly focuses on evaluating the intelligibility of the high-energy speech frames, and our CRM employs a small value of μ in speech dominant units. In this case, the value of our CRM is similar to those of other masks and thus the STOI scores are close for all tested methods. In terms of three metrics, we can conclude that our proposed CRM removes more residual noise while minimizing the speech distortion.

b) Unseen noise: Table 5 shows the average objective scores on unseen noise. In general, our CRM still outperforms the other reference methods even under the unseen noise environment. However, compared to seen noise, the improvements of the objective scores on the enhanced speech decrease a bit, due to the increasing prediction error of masks. This result is not surprising since the mismatch in the types of the noises between the enhancement stage and the training stage makes the estimation of CRM with DNN more difficult.

Table 4: *Results of different algorithms on seen noise*

		-3dB	0dB	3dB	6dB
PESQ	Noisy	1.35	1.54	1.75	1.97
	IRM	1.97	2.23	2.47	2.70
	IAM	1.98	2.24	2.47	2.71
	OPM	2.00	2.27	2.50	2.74
	CRM	**2.01**	**2.29**	**2.55**	**2.81**
STOI	Noisy	0.60	0.67	0.74	0.81
	IRM	0.74	0.81	0.86	0.90
	IAM	**0.75**	0.81	0.86	0.90
	OPM	**0.75**	0.81	0.86	0.90
	CRM	**0.75**	**0.82**	**0.87**	**0.91**
SDR	Noisy	-2.85	0.11	3.08	6.07
	IRM	5.72	8.03	10.19	12.45
	IAM	5.89	8.18	10.33	12.55
	OPM	5.71	8.01	10.15	12.43
	CRM	**7.44**	**9.63**	**11.64**	**13.73**

Table 5: *Results of different algorithms on unseen noise*

		-3dB	0dB	3dB	6dB
PESQ	Noisy	1.34	1.52	1.73	1.94
	IRM	1.73	1.97	2.19	2.43
	IAM	1.74	1.98	2.20	2.44
	OPM	1.76	1.99	2.21	2.45
	CRM	**1.77**	**2.01**	**2.23**	**2.47**
STOI	Noisy	0.63	0.69	0.76	0.82
	IRM	0.71	0.77	0.83	0.88
	IAM	**0.72**	0.78	**0.84**	**0.89**
	OPM	0.71	0.77	0.83	0.88
	CRM	**0.72**	**0.79**	**0.84**	**0.89**
SDR	Noisy	-2.84	0.10	3.07	6.06
	IRM	2.54	5.31	8.03	10.77
	IAM	2.68	5.49	8.23	10.94
	OPM	2.42	5.20	7.93	10.70
	CRM	**3.92**	**6.72**	**9.39**	**12.01**

5. Conclusion

In this paper, a new type of mask named constrained ratio mask (CRM) is proposed for speech enhancement. Compared with traditional masks, employing CRM is able to control the speech and noise distortions by adaptively adjusting the value of the controlling factor based on the local SNR, which minimizes the speech distortion while removing the residual noise as much as possible. The CRM is predicted by the well-known DNN model and is applied to the noisy speech to obtain the desired one. Experimental results show that our enhanced speech achieves better speech quality compared to the others resulting from existing masks under both seen and unseen noises. Future work could be a possible extension of the CRM to jointly enhance the magnitude and phase of the noisy speech.

6. References

[1] G. J. Brown and D. Wang, "Separation of speech by computational auditory scene analysis," in *Speech Enhancement*. Springer, 2005, pp. 371–402.

[2] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*. Springer, 2005, pp. 181–197.

[3] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[4] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7092–7096, 2013.

[5] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing (TASLP)*, vol. 22, no. 12, pp. 1849–1858, 2014.

[6] S. Liang, W. Liu, W. Jiang, and W. Xue, "The optimal ratio time-frequency mask for speech separation in terms of the signal-to-noise ratio," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. EL452–EL458, 2013.

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

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

[9] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[10] K. Han and D. Wang, "A classification based approach to speech segregation," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475–3483, 2012.

[11] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 23, no. 12, pp. 2136–2147, 2015.

[12] S. Chakrabarty, D. Wang, and E. A. Habets, "Time-frequency masking based online speech enhancement with multi-channel data using convolutional neural networks," *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 476–480, 2018.

[13] S. Gustafsson, P. Jax, and P. Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 397–400, 1998.

[14] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 59–67, 2004.

[15] P. C. Loizou, "Contrained Wiener filtering," in *Speech enhancement: theory and practice*. CRC press, 2013.

[16] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 21, no. 2, pp. 270–279, 2012.

[17] 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*, vol. 93, 1993.

[18] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[19] ITU-R, "Perceptual evaluation of speech quality (PESQ) an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Recommendation P.862*, 2001.

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

[21] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 14, no. 4, pp. 1462–1469, 2006.