Interspeech 2018
2-6 September 2018, Hyderabad

Deep Noise Tracking Network: A Hybrid Signal Processing/Deep Learning Approach to Speech Enhancement

Shuai Nie[1,3], *Shan Liang*[1*], *Bin Liu*[1,3], *Yaping Zhang*[1,3], *Wenju Liu*[1] *and Jianhua Tao*[1,2,3]

[1.]National Laboratory of Pattern Recognition, Institute of Automation,Chinese Academy of Sciences
[2.]CAS Center for Excellence in Brain Science and Intelligence Technology
[3.]School of Artificial Intelligence, University of Chinese Academy of Sciences
{shuai.nie, sliang, bin.liu, yaping.zhang, lwj, jhtao}@nlpr.ia.ac.cn

Abstract

Noise statistics and speech spectrum characteristics are the essential information for the single channel speech enhancement. The signal processing-based methods mainly rely on noise statistics estimation. They perform very well for stationary noise, but have remained difficult to cope with non-stationary noise. While the deep learning-based methods mainly focus on the perception on the spectrum characteristics of speech and have a capacity in dealing with non-stationary noise. However, the performance would degrade dramatically for the unseen noise types, which could be due to the over-reliance on data and the ignorance to domain knowledge of signal process. Obviously, the hybrid signal processing/deep learning scheme may be a smart alternative. In this paper, we incorporate the powerful perceptual capabilities of deep learning in the conventional speech enhancement framework. Deep learning is used to estimate the speech presence probability and the update factor of noise statistics, which are then integrated into the Wiener filter-based speech enhancement structure to enhance the desired speech. All components are jointly optimized by a spectrum approximation objective. Systematic experiments on CHiME-4 and NOISEX-92 demonstrate the proposed hybrid signal processing/deep learning approach to noise suppression in noise-unmatched and noise-matched conditions.

Index Terms: speech enhancement, noise tracking, deep learning, signal processing.

1. Introduction

In real-world environments, the acquired speech signals are inevitably corrupted by various noises and reverberation. These degradations are known to significantly degrade the intelligibility and quality of speech [1, 2], and also deteriorate the performance of automatic speech recognition (ASR) [3–5]. To cope with such acoustic environments, it is essential to establish effective speech enhancement technologies. Various techniques including signal processing and deep learning have been applied to speech enhancement.

Speech signals are inherently sparse in the time-frequency domain, which allows for continuous tracking and reduction of background noise [6]. To implement the noise-reduction filters, noise statistics are usually required and need to be continuously estimated [7–11]. Temporal moving-average over time frames is a common way to the estimation of noise statistics, in which a voice activity detector (VAD) is usually used to decide to update/hold the noise statistics. While spotting time instants and frequency bins without/with active speech components based on speech presence probability (SPP) is a finer way, which can obtain more accurate noise power spectral density (PSD). However, these methods often assume that noise is stationary or slowly varying, which is hard to be met in real-world environments. To deal with this issue, Martin proposed a minimum statistic-based method to track the spectral minima of the noisy signal per frequency bin [8]. Cohen improved the minimum statistic approach and proposed a so-called minima controlled recursive averaging (MCRA) in which the noise estimate is obtained by SPP-based smoothing-average of PSD [10, 12], and the SPP is controlled by the principle of minimum statistics tracking. Hendriks proposed minimum mean-squared error (MMSE) based noise PSD tracking approaches and improved the performance for non-stationary noise sources [13, 14]. Although these approaches achieve reliable performance in non-stationary noise conditions, speech enhancement in real-world environments is still a challenging task and its performances are far from being satisfactory.

Owing to the powerful perceptual capabilities of deep learning to speech and noise, recently the deep learning-based speech enhancement has achieved remarkable performance improvements over conventional signal processing methods [15, 16], especially when noise is non-stationary or the signal-to-noise ratio (SNR) is low. A typical supervised speech enhancement system usually uses a trained deep network to directly cast noisy features into certain ideal masks or magnitude spectrograms of interest frame-by-frame. These approaches rely largely on the data-driven principle to perform noise-reduction and ignore the domain knowledge of signal process. When it comes to an unmatched acoustic environments, such as unseen noise types and SNRs, the performances would degrade dramatically due to the noise overestimation or underestimation. While these estimation errors could be "averaged out" by calculating the noise statistics as the conventional noise tracking framework does.

The combination of signal processing and deep learning techniques may be the more advisable strategy for speech enhancement. In this paper, we propose a novel deep noise tracking network (DNTN) that consists of a gated recurrent unit (GRU) [17] and a feed-forward network. It incorporates the powerful perceptual capabilities of deep learning in the conventional speech enhancement framework. The GRU is used to estimate the SPP from the noisy features, and the long-term state of GRU combined with the current noisy features feed into the feed-forward network to estimate the update factor of noise statistics. A proven temporal moving-average technique are then used to update the PSD of noise signals with the estimated SPP and update factor. Finally, a Wiener filter is established to extract the desired speech and attenuate the annoying noise. All components are jointly optimized by directly minimizing the spectral distance between

This work was supported by the National Key R&D Program of China (No. 2017YFB1002804) and the China National Nature Science Foundation (No. 61573357, No. 61503382, No. 61403370, No. 61273267, No. 91120303).

10.21437/Interspeech.2018-1020

the enhanced speech and the desired speech.

2. Signal Model and Problem Formulation

We consider a scenario where a single point-like speech source is captured by a single microphone in a reverberant room. Let $s(k)$ and $v(k)$ denote speech and uncorrelated additive noise signals, respectively, where k is a discrete-time index. The observed signal is then given by

$$x(k) = g(k) * s(k) + v(k) = y(k) + v(k), \tag{1}$$

where $*$ denotes a convolution operator, $g(k)$ is the channel impulse response. $y(k) = g(k) * s(k)$ is the noise-free speech component. Let us assume that all signals are zero-mean random process. In the short-time Fourier transform (STFT) domain, the Eq (1) can be written as

$$x(f,t) = y(f,t) + v(f,t), \tag{2}$$

where f and t are the frequency and time-frame indexes, respectively.

Technically, speech enhancement involves not only noise reduction but also dereverberation. However, here we only focus on noise reduction. So our aim is to recover speech signal $y(f,t)$ and reduce the noise signal $v(f,t)$ by applying a linear filter $h(f,t)$ to the observation $x(f,t)$. The Wiener filter can be considered as one of the most fundamental noise reduction approaches and many algorithms are closely connected to this technique [18]. Assuming that speech signals and noise signals are uncorrelated, the general form of the Wiener filter gain is written as

$$h(f,t) = \frac{\phi_{yy}(f,t)}{\phi_{yy}(f,t) + \phi_{vv}(f,t)}, \tag{3}$$

where $\phi_{vv}(f,t)$ are the PSD of noise signals defined as

$$\phi_{vv}(f,t) = E\left\{v(f,t)v^*(f,t)\right\}, \tag{4}$$

where * denotes a complex conjugate operator. Since noise signals and speech signals are assumed to be uncorrelated, the PSD of the desired speech can be calculated as

$$\phi_{yy}(f,t) = E\left\{y(f,t)y^*(f,t)\right\} = \phi_{xx}(f,t) - \phi_{vv}(f,t), \tag{5}$$

where $\phi_{xx}(f,t) = E\left\{x(f,t)x^*(f,t)\right\}$ is the PSD of the observed signals, and $E\{\cdot\}$ denotes a mathematical expectation operator. To meet real-time requirements in practice, temporal recursive smoothing is usually used to approximate the mathematical expectations involved in the previous PSDs [6]. In other words, at time frame t, the PSDs of the noise signals and the observed signals are updated recursively as

$$\begin{aligned}\hat{\phi}_{vv}(f,t) &= \tilde{\alpha}_v(f,t)\hat{\phi}_{vv}(f,t-1) + \left(1-\tilde{\alpha}_v(f,t)\right)x(f,t)x^*(f,t)\\ \hat{\phi}_{xx}(f,t) &= \alpha_x(f,t)\hat{\phi}_{xx}(f,t-1) + \left(1-\alpha_x(f,t)\right)x(f,t)x^*(f,t),\end{aligned} \tag{6}$$

where $0 \leqslant \alpha_x(f,t) \leqslant 1$ and $0 \leqslant \tilde{\alpha}_v(f,t) \leqslant 1$ are the smoothing factor of the PSDs of the observed signals and the noise signals, respectively. They are essential to correctly update the observed and noise signals PSDs. In practice, $\alpha_x(f,t)$ is usually set to an appropriate constant α_x to reach a compromise between smoothing the noise signals and tracking the speech signals. While $\tilde{\alpha}_v(f,t)$ is a time-varying frequency-dependent smoothing factor. It should be small enough when the speech is absent so that the noise changes can be quickly followed, but when the speech is present, it should be sufficiently large to avoid the overestimation of noise PSD. Obviously, $\tilde{\alpha}_v(f,t)$ is closely related to the detection of speech presence/absence. The SPP $p(f,t)$ is commonly utilized to adjust it as follows [6]

$$\tilde{\alpha}_v(f,t) \;=\; \alpha_v \;+ (1-\alpha_v)p(f,t), \tag{7}$$

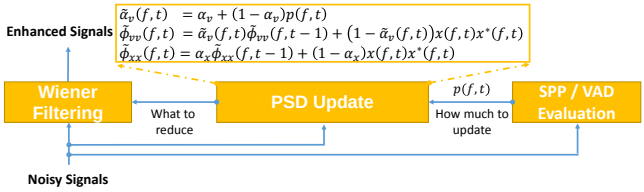

Figure 1: *The structure of conventional speech enhancement.*

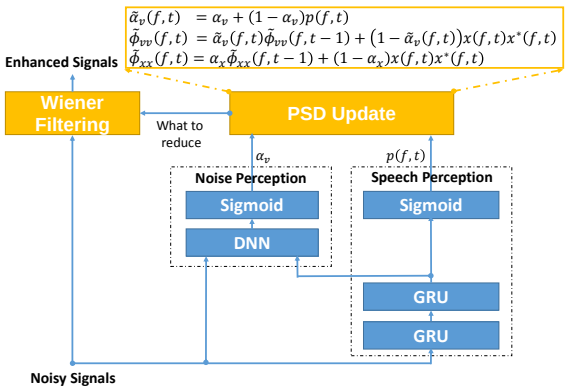

Figure 2: *The structure of the proposed DNTN.*

where $\alpha_v \in [0,1)$ determines the update factor of noise PSD when speech is absent. In conventional speech enhancement, it is commonly set to a constant based on the assumption that noise is stationary, which is hard to be met in practice.

3. A Hybrid Signal Processing/Deep Learning Approach

Figure 1 illustrates a block diagram of the general structure of most conventional speech enhancement algorithms [19]. In this scheme, the noise PSD estimation is one of the most important components, since it largely determines the amount of residual noise in the output of the Wiener filter [6]. A common approach to estimate the noise PSD is to average past spectral power values using a time-varying smoothing factor that is adjusted by the SPP [8, 9]. But when the SNR is low or noise is non-stationary, it is very hard to obtain the accurate SPP by conventional signal processing techniques, which restricts the tracking capability of the noise estimator in case of varying noise spectrum [20].

Deep learning has a powerful perceptual ability to speech and noise. To address the limitations of conventional signal processing methods, we propose a hybrid scheme, where we use deep learning to replace the SPP estimators that have traditionally been hard to correctly tune, while use basic signal processing building blocks for the typical PSD update and Wiener filtering. Deep learning and signal processing modules are combined into an organic whole and performed in an end-to-end manner instead of a separate or pipeline manner. In other words, the deep learning model is jointly optimized by the errors of the final Wiener filtering output rather than another separate optimization objective, such as ideal SPP. And there is no need to train the deep learning model separately.

Figure 2 illustrates the structure of the proposed hybrid scheme. The whole scheme closely follows the general structure of conventional speech enhancement algorithms as shown in Figure 1. We utilize a two-layers of GRU and a one-layer of

feed-forward network to construct a DNTN. The DNTN takes the responsibility of the SPP/VAD estimator in Figure 1. Specifically, a sigmoid output layer following the GRU generates a $1 \times F$ vector of $[0, 1]$ element, which will be used as the SPP $p(f, t)$ of the current frame to adjust the smoothing factor of noise PSD. The long-term states of the GRU combined with the current frame of noisy features feed into the feed-forward network to estimate α_v of $[0, 1]$ through a sigmoid output layer. After obtaining α_v and $p(f, t)$, we can calculate the smoothing factor $\tilde{\alpha}_v(f, t)$ of the noise PSD by Eq (7). Then the PSDs of the noise signals and observed signals can be updated according to Eq (6), and an optimal Wiener filter is established by Eq (3). Finally, we apply the established Wiener filter to the observed signals to obtain the desired speech signals as follows

$$\tilde{y}(f, t) = \frac{\phi_{xx}(f, t) - \phi_{vv}(f, t)}{\phi_{xx}(f, t)} x(f, t). \tag{8}$$

Speech enhancement is aimed at recovering speech signal $y(f, t)$ from the noise signals $v(f, t)$. Hence, we use an MSE-based magnitude approximation objective to jointly optimize all components.

$$J = \frac{1}{T} \sum_{t=0}^{T} \sum_{f=0}^{F} |\tilde{y}(f, t) - y(f, t)|^2, \tag{9}$$

where $|\cdot|$ denotes the absolute value operator in the complex domain, while F and T are the numbers of frequency bins and time frames, respectively.

4. Experiments

4.1. Dataset and Evaluation Metrics

We apply the proposed DNTN to single channel speech enhancement to examine its effectiveness and systematically evaluate its performances on the CHiME-4 [21] and NOISEX-92 [22] corpora. The CHiME-4 corpus consists of "real data" and "simulated data". The "real data" is recorded in 4 real noisy environments[1] and uttered by actual talkers. The "simulated data" has been generated by artificially mixing clean speech data with the real-world noise backgrounds, which means that the noise-free speech component in the noisy signals is known. Therefore, the "simulated data" can be employed to train the proposed DNTN. Although each utterance in the CHiME-4 corpus consists of 6 channels, we randomly choose a channel signal for the following experiments. The NOISEX-92 contains 15 common types of noise[2] in real-world environments, with a length of about 4 minutes for each. We mention that these noises are quite different from those in the CHiME-4 corpus. All audio data were sampled at 16 kHz and 16 bits.

The training set of CHiME-4 consists of 1,600 real and 7,138 simulated utterances in the 4 noisy environments. We choose the "simulated data" as our training set. Similarly, we choose the "simulated data" (410 (simulated) $\times$ 4 (environments)) in the development set of CHiME-4 as our development set. For testing, we randomly choose 1,000 clean speech utterances from the WSJ0 development part. They are randomly mixed with 15 types of noise from NOISEX-92 to generate 1,000 mixture utterances at a continuous SNR from 0dB to 10dB. These noises are unseen in the training set, which is used to test the generalization of the proposed DNTN to the unmatched noise. While mixing speech and noise, in order to ensure that the different parts of each noise utterance are mixed with the clean speech utterances, we randomly cut each noise utterance of NOISEX-92 into different parts according to the time length of a speech utterance.

We take the source to interference ratio (SIR), source to artifacts ratio (SAR), source to distortion ratio (SDR) [23] and perceptual evaluation of speech quality (PESQ $\in [-0.5, 4.5]$) [24] as evaluation metrics. SIR, SAR and SDR measure the ratios of source to interference, artifacts and distortion, respectively, and can be computed by the BSS Eval toolbox [23]. The PESQ score quantifies the objective speech quality. All evaluation metrics are the weighted means of all testing clips weighted by their lengths. Higher values mean the better performances.

4.2. Comparison Methods and Configurations

Deep learning-based supervised speech enhancement usually learns a mapping function from noisy features to certain ideal masks or magnitude spectrograms of interest. In this paper, we take mask-approximated and magnitude-approximated supervised speech enhancement methods as the comparisons. The mask-approximated approach uses a two-layers of GRU to estimate an ideal ratio mask (IRM) [25] from the noisy features, denoted as "GRU-IRM", while the magnitude-approximated approach uses a two-layers of GRU to estimate the magnitude spectrograms of target speech from the noisy features, denoted as "GRU-MAG". Since the IRM is bounded between 0 and 1, GRU-IRM uses a sigmoid function as the activation function of the output layer. Instead of directly predicting the magnitude spectrograms of target speech, GRU-MAG applies a sigmoid output to the mixture magnitude spectrograms as a masking operation to generate the magnitude spectrograms of the desired speech, which can be regarded as indirect masking [26]. The proposed DNTN, besides the similar GRU structure for speech perception, also has a feed-forward network with one hidden layer of 512 rectified linear units (ReLUs) [27] for noise perception. The additional feed-forward network has a sigmoid output of one unit, while the output layers of GRU networks of GRU-IRM, GRU-MAG and DNTN have 256 units according to the STFT points. Each GRU layer of GRU-IRM, GRU-MAG and DNTN has 512 cells. In addition, we apply batch normalization to the input-to-hidden transitions of each GRU layer, which can lead to a faster convergence of the training criterion [28].

In following experiments, we use the 256-dimension log power spectrum as the input feature, and each dimension of input features is normalized to have zero mean and unit variance over the training set. All networks are trained from a random initialization by the Adam optimizer [29] with a learning rate of 0.001. The maximum epoch is set to 25 and the size of mini-batch is set to 256. The spectral representation is extracted by applying a 512-point STFT to the mixture signals with 32-ms frame length that windowed by a 512-point Hamming window with 50% overlap. The 256-dimension log power spectrum is obtained by the log operator on the power magnitude spectrograms and cutting off the symmetrical parts.

4.3. Results and Discussions

Firstly, we systematically evaluate how the parameter α_x affects the performance of speech enhancement. α_x is a smoothing factor and determines the updating/holding speed of the observed signals PSD, as shown in Eq (6). Excessively large α_x restricts the update of the observed signals PSD and cannot timely track the speech signal and follow the varying acoustic environment. While excessively small α_x cannot smooth the noise signal in

[1] bus, cafe, pedestrian area, and street junction.

[2] babble, factory1, buccaneer1, destroyerengine, white, machinegun, pink, volvo, hfchannel, factory2, buccaneer2, destroyerops, f16, leopard, m109.

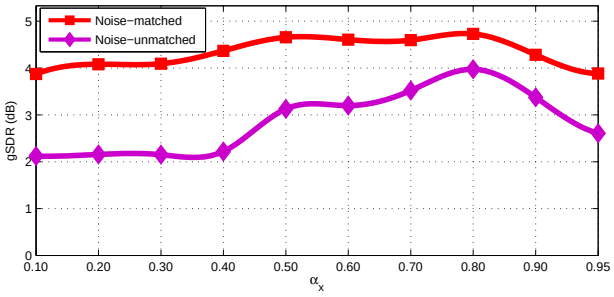

Figure 3: *The performances using various smoothing factors α_x in noise-matched and noise-unmatched conditions.*

the mixture signals. Therefore, its choice is essential to correctly update the observed signal PSD and should be appropriately set to reach a compromise between smoothing the noise signals and tracking the speech signals. In the experiments, we choose α_x from $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$. Figure 3 reports the average gains of SDR (gSDR) obtained by using different smoothing factors α_x in noise-matched (CHiME-4) and noise-unmatched (NOISEX-92) conditions, respectively. The gSDR can be computed as follows:

$$gSDR(\tilde{y}, y, x) = SDR(\tilde{y}, y) - SDR(x, y), \qquad (10)$$

where $\tilde{y}$ is the enhanced speech and the gSDR reflects the improvement of overall performance. It can be seen that the proposed speech enhancement system achieves best performance near $\alpha_x = 0.8$, in either noise-matched or noise-unmatched condition, which is very close to the calculated value in theory [9]. Hence, we set α_x to 0.8 in the following experiments.

Secondly, we systematically evaluate the performances of the proposed model (DNTN) and the comparisons (GRU-IRM and GRU-MAG). Table 1 reports the speech enhancement performances of different models in noise-matched and noise-unmatched conditions, respectively. We observe that all models have significantly enhanced speech in both noise-matched and noise-unmatched conditions, but GRU-MAG and DNTN consistently and significantly outperform GRU-IRM. This is because GRU-MAG and DNTN directly optimize an actual enhancement objective rather than ideal masks which is used as an intermediate target by GRU-IRM. It also suggests that the masking-based magnitude-approximated objective outperforms the direct mask-approximated objective. We also observe the proposed DNTN has achieved roughly equivalent performance with GRU-MAG in noise-matched condition, and even has slight improvements on SDR, SAR, gSDR and PESQ. However, in noise-unmatched condition, the proposed DNTN achieves best performances on various evaluation metrics and significantly outperforms GRU-MAG. It has indicated that DNTN has a better generalization ability to unseen noise, which is extremely important for practical applications. The better generalization is mainly attributable to the utilization of the domain knowledge of signal process. The proposed DNTN incorporates the powerful perceptual capabilities of deep learning to speech and noise in the conventional speech enhancement framework based on signal processing. The hybrid scheme not only exploits speech spectrum characteristics, but also utilizes noise statistics, which concentrate the strengths of both deep learning and signal processing techniques for speech enhancement. In fact, the recursive smoothing-based noise tracking component can be regarded as a regularization term of DNTN.

Table 1: *The speech enhancement performances of different models in noise-unmatched and noise-matched conditions.*

	Models	SDR	SIR	SAR	gSDR	PESQ
Unmatched	Mixture	5.03	5.03	–	–	1.26
	GRU-IRM	6.29	7.53	8.97	1.26	1.37
	GRU-MAG	8.00	10.67	10.28	2.98	1.36
	DNTN	**9.00**	**12.08**	**10.65**	**3.97**	**1.41**
Matched	Mixture	3.86	3.86	–	–	1.17
	GRU-IRM	6.95	8.75	8.67	3.10	1.34
	GRU-MAG	8.54	**11.63**	10.62	4.69	1.34
	DNTN	**8.58**	11.19	**11.25**	**4.73**	**1.36**

Finally, we exhibit some visible results of DNTN to further make each component of DNTN clear as shown in Figure 4. Although there is no ideal SPP as the supervised target, the GRU automatically learns the SPP from noisy signals. We also observe that the output of feed-forward network seems to reflect the smoothness of the noise. It indicates that the domain knowledge of signal processing has instructed neural networks what to learn.

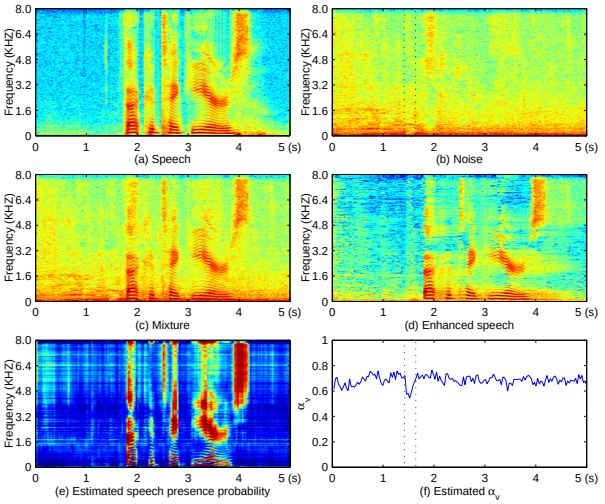

Figure 4: *The speech enhancement results of DNTN. (a-d) are the log power spectrums of clean speech, noise, mixture signals and enhanced speech, respectively. (e) is the output of the GRU, which looks like the speech presence probability. (f) is the output of the feed-forward network, which seems to reflect the smoothness of the noise.*

5. Conclusions

In this paper, we proposed a novel deep noise tracking network, which incorporates the powerful perceptual capabilities of deep learning in the mature speech enhancement framework based on signal processing. The deep learning and signal processing components are jointly optimized by a spectrum approximation objective. We demonstrate that the hybrid signal processing/deep learning approach significantly outperforms a pure deep learning-based approach and has a better generalization to unseen acoustic environments. We believe that this technique can be easily extended to multichannel speech enhancement, which will be explored in the future.

6. References

[1] Tashev and I. Jelev, "Sound capture and processing," *Wiley & Sons*, 2009.

[2] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–94, 2009.

[3] X. Huang and A. Acero, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, 2001.

[4] M. Wlfel and J. Mcdonough, "Distant speech recognition," vol. 130, no. 5, pp. 5106 – 5109, 2009.

[5] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The pascal chime speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013.

[6] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Transactions on Audio Speech & Language Processing*, vol. 19, no. 7, pp. 2159–2169, 2011.

[7] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, Inc., 2007.

[8] R. MARTIN, "Spectral subtraction based on minimum statistics," *Proceedings of the EUSIPCO94, Edinburgh*, vol. 2, pp. 1182–1185, 1994.

[9] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[10] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech & Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[11] ——, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113–116, 2002.

[12] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[13] R. C. Hendriks, R. Heusdens, and J. Jensen, "Mmse based noise psd tracking with low complexity," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 2010, pp. 4266–4269.

[14] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio Speech & Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[15] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[16] D. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview," *arXiv preprint arXiv:1708.07524*, 2017.

[17] K. Cho, B. Van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *Computer Science*, 2014.

[18] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Transactions on Audio Speech & Language Processing*, vol. 14, no. 4, pp. 1218–1234, 2006.

[19] M. Parchami, W.-P. Zhu, B. Champagne, and E. Plourde, "Recent developments in speech enhancement in the short-time fourier transform domain," *IEEE Circuits and Systems Magazine*, vol. 16, no. 3, pp. 45–77, 2016.

[20] M. Taseska and E. A. Habets, "Nonstationary noise psd matrix estimation for multichannel blind speech extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2223–2236, 2017.

[21] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, 2016.

[22] 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.

[23] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[24] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2001)*, vol. 2, 2001, pp. 749–752.

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

[26] 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 & Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[27] X. Glorot, A. Bordes, Y. Bengio, X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2012, pp. 315–323.

[28] C. Laurent, G. Pereyra, P. Brakel, Y. Zhang, and Y. Bengio, "Batch normalized recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 2657–2661.

[29] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.