Comparative Study of Deep Learning Based and Traditional Single-Channel Noise-Reduction Algorithms

Ningning Pan *, Jingdong Chen *, and Biing-Hwang (Fred) Juang†
* Center of Intelligent Acoustics and Immersive Communications,
Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
† School of Electrical and Computer Engineering,
Georgia Institute of Technology, Atlanta, GA, USA

Abstract—**Deep neural networks (DNN) have been applied to the problem of noise reduction and promising results have been reported widely, leading to the impression that the traditional techniques based on blind noise estimation may no longer be needed. However, there lacks comprehensive and rigorous evaluation and comparison between DNN based and traditional noise-reduction algorithms for their pros and cons. In this work, we attempt to evaluate some widely used DNN based noise-reduction algorithms and compare them to a traditional noise-reduction method. We also evaluate a method that straightforwardly combines a DNN based regression method with the optimal filtering technique. Through experiments, it is observed that: 1) DNN based methods have advantages over the traditional methods in scenarios with non-stationary noise and low signal-to-noise ratios (SNRs); 2) generalization remains a challenging issue with DNN based methods; for noise types unseen in the training data, which happen often in practical environments, DNN based methods do not show any advantage over the traditional technique; 3) combining DNN-based regression and the optimal filtering technique shows some potential in improving the noise-reduction performance as well as system generalization.**

I. Introduction

Single-channel noise reduction aims to recover a clean speech signal of interest from its microphone observation that is corrupted by additive noise [1]. Generally, the goal of noise reduction is to improve perceptual quality or/and intelligibility of the noisy speech signal. Extensive efforts have been made to address this problem and a large number of methods have been proposed, including the optimal filtering methods [2], [3], the spectral subtraction type of techniques [4], [5], [6], the statistical approaches [7], the subspace method [8], [9], and various recently developed DNN based algorithms.

Traditional methods are based on noise estimation with an assumption on the aggregate statistical behavior or model of the noise, while DNN based algorithms rely on paired speech signals (both noisy and clean speech, with the former containing a precise copy of the latter) to train the networks. As they obtain signal statistics and perform noise reduction quite distinctly, these two types of methods have their their own pros and cons. Many noise estimators have been developed to meet the need in the traditional approach, including algorithms based on the use of voice activity detection [10], minimum statistics based algorithms [11], [12], [13], the subspace methods [14], and the minimum-mean-square-error (MMSE) [15], nonnegative-matrix-factorization [16], and codebook-driven methods [17], etc. However, over- and under-estimation of noise with those methods is not uncommon in practical environments, leading to either insufficient noise reduction (noise still audible in the processed result) or excessive noise reduction so as to distort the speech part.

In comparison, DNN based methods use synthetic paired data or carefully recorded paired data to train the network parameters. Tremendous efforts have been devoted to training targets [18], [19], [20], feature extraction [21], DNN models [22], or different types of noises [23], [24]. The interested reader is referred to [25] for an overview. Most DNN-based speech enhancement methods use the time-frequency spectrum of the clean speech as the training targets [18], either mapping-based [19] or masking-based [18]. It is widely reported that DNN methods outperform all traditional noise reduction methods in terms of perceptual evaluation of the speech quality (PESQ) [32] and short-time objective intelligibility (STOI) [33], leading to the impression that collecting a big database and training DNN would become the dominant way to go for noise reduction. However, a recent study on subjective evaluations of DNN based methods [26] shows that evaluation reported in the literature may not be comprehensive enough for a solid conclusion.

This work is therefore organized to perform a comprehensive evaluation of DNN based noise-reduction algorithms and compare them with traditional noise-reduction techniques. Due to the space limit, we report in this paper the initial results on PESQ and STOI for three DNN based methods: ideal ratio mask (IRM) [18], concatenation of speech and noise amplitude spectrum (AMP), and concatenation of smoothed speech and noise power spectrum (POW), and a traditional method: the optimally-modified-log-spectral-amplitude (OMLSA) algorithm [27], which uses the improved-minima-controlled recursive-averaging (IMCRA) method for blind noise estimation. Also evaluated is a new method that combines DNN based regression method and the optimal filtering technique. More comprehensive results including both objective and subjective evaluation will be reported once all the experiments are completed.

II. DNN-based noise reduction algorithms

In this section, we introduce DNN-based noise-reduction algorithms for three different estimation targets in the training stage and the corresponding post-processing in the test stage.

A. Signal model

Consider the single-channel noise-reduction problem where the time-domain noisy observation signal is in the following form:

$$y(n) = x(n) + v(n), \tag{1}$$

with $x(n)$ and $v(n)$ being, respectively, the clean speech signal of interest and the additive noise, and n denoting the discrete-time index. $x(n)$ and $v(n)$ are assumed to be uncorrelated. Transforming $y(n)$ into the short-time-Fourier-transform (STFT) domain gives

$$Y(k,l) = X(k,l) + V(k,l), \tag{2}$$

where k is the frequency index, l denotes the frame index, and $Y(k,l)$, $X(k,l)$ and $V(k,l)$ are, respectively, the STFTs of the noisy speech, clean speech and noise signals. The objective of noise reduction is then to recover the clean speech $X(k,l)$ given $Y(k,l)$.

B. Training stage

A critical decision to be made before training a DNN is the representation vectors that serve as the input and the output of the networks, respectively. In this work, we use the logarithmic amplitude (log-amplitude) spectra of the noisy speech as the input of DNN. Mathematically, the lth frame log-amplitude spectrum of $Y(k,l)$ is defined as

$$\mathbf{Y}^{l}_{\log} = [Y_{\log}(0,l), \ldots, Y_{\log}(k,l), \ldots, Y_{\log}(N/2,l)]^T, \tag{3}$$

where N is the FFT length and $Y_{\log}(k,l) = \log(|Y(k,l)| + c)$ with c being set to 10^{-10}. The input of DNN, denoted as $\mathbf{Y}^{l-\tau:l+\tau}_{\log}$, is a concatenation of $2\tau+1$ frames centered at $\mathbf{Y}^{l}_{\log}$, i.e.,

$$\mathbf{Y}^{l-\tau:l+\tau}_{\log} = \left[(\mathbf{Y}^{l-\tau}_{\log})^T, \ldots, (\mathbf{Y}^{l}_{\log})^T, \ldots, (\mathbf{Y}^{l+\tau}_{\log})^T \right]^T, \tag{4}$$

where the superscript $l-\tau : l+\tau$ denotes a set of frame indices $\{l-\tau, l-\tau+1, \ldots, l+\tau\}$. In the traditional terminology, this represents a *block*-based (as opposed to frame-based) processing.

In this paper, we focus on the evaluation of three different training targets as output of the DNN, namely, the ideal ratio mask (IRM), the signal amplitude spectrum (AMP), and the signal power spectrum (POW).

- The IRM is defined as

$$G_{\mathrm{IRM}}(k,l) = \sqrt{\frac{|X(k,l)|^2}{|X(k,l)|^2 + |V(k,l)|^2}}. \tag{5}$$

  The corresponding lth frame of IRM is the target of DNN.
- The AMP is a concatenation of the lth frame of $\mathbf{X}^{l}_{\log}$ and $\mathbf{V}^{l}_{\log}$, which are defined the same way as $\mathbf{Y}^{l}_{\log}$ in (3) and (4).
- The speech power spectrum is computed in a recursive way

$$\phi^X(k,l) = \alpha\phi^X(k,l-1) + (1-\alpha)|X(k,l)|^2, \tag{6}$$

  where α is the smoothing parameter, which is set to 0.95 in this work. The noise power spectrum $\phi^V(k,l)$ is defined the same way. The lth frame speech log-power spectrum is defined as

$$\boldsymbol{\phi}^{X,l}_{\log} = [\phi^X_{\log}(0,l), \ldots, \phi^X_{\log}(k,l), \ldots, \phi^X_{\log}(N/2,l)]^T$$

  where $\phi^X_{\log}(k,l) = \log(\phi^X(k,l) + c)$. (The noise log-power spectrum $\boldsymbol{\phi}^{V,l}_{\log}$ is defined analogously to $\boldsymbol{\phi}^{X,l}_{\log}$. The third target, POW, is a concatenation of the lth frame $\boldsymbol{\phi}^{X,l}_{\log}$ and $\boldsymbol{\phi}^{V,l}_{\log}$.

C. Test stage

In the test stage, we need to obtain the enhanced speech based on the use of the noisy speech and the aforementioned three different DNN targets.

- To recover the clean speech using IRM, we have

$$\hat{X}_{\mathrm{IRM}}(k,l) = \hat{G}_{\mathrm{IRM}}(k,l)Y(k,l), \tag{7}$$

  where $\hat{G}_{\mathrm{IRM}}(k,l)$ is the estimation of IRM from DNN.
- To recover the clean speech using AMP, two different ways could be applied. The first one is to directly recover speech signal from output of DNN:

$$\hat{X}_{\mathrm{AMP}}(k,l) = \exp\left[\hat{X}_{\log}(k,l)\right]. \tag{8}$$

  The other way is to combine DNN output with optimal filtering technique. The speech and noise power spectra can be estimated from (8) as (6), which are denoted as $\hat{\phi}^X_{\mathrm{AMP}}(k,l)$ and $\hat{\phi}^V_{\mathrm{AMP}}(k,l)$ respectively. Then the traditional Wiener filter is computed as

$$\hat{G}_{\mathrm{W,AMP}}(k,l) = \frac{\hat{\phi}^X_{\mathrm{AMP}}(k,l)}{\hat{\phi}^X_{\mathrm{AMP}}(k,l) + \hat{\phi}^V_{\mathrm{AMP}}(k,l)}. \tag{9}$$

  We then have

$$\hat{X}_{\mathrm{AMP,W}}(k,l) = \hat{G}_{\mathrm{W,AMP}}(k,l)Y(k,l). \tag{10}$$

  Besides the classic Wiener filter, many different filters could be applied [2].
- To recover the clean speech using POW, we first compute the speech and noise power spectra from the DNN outputs as

$$\begin{aligned}\hat{\phi}^X_{\mathrm{POW}}(k,l) &= \exp\left[\hat{\phi}^X_{\log}(k,l)\right], \\ \hat{\phi}^V_{\mathrm{POW}}(k,l) &= \exp\left[\hat{\phi}^V_{\log}(k,l)\right].\end{aligned} \tag{11}$$

  Then, the Wiener filter defined in (9) can be computed as

$$\hat{G}_{\mathrm{W,POW}}(k,l) = \frac{\hat{\phi}^X_{\mathrm{POW}}(k,l)}{\hat{\phi}^X_{\mathrm{POW}}(k,l) + \hat{\phi}^V_{\mathrm{POW}}(k,l)}. \tag{12}$$

  Applying this filter to $Y(k,l)$ gives

$$\hat{X}_{\mathrm{POW,W}}(k,l) = \hat{G}_{\mathrm{W,POW}}Y(k,l). \tag{13}$$

Then the time-domain enhanced signal is obtained using the overlap-add technique.

III. Experimental Setup

A. Training Data

We construct the noisy training data by artificially adding noises to clean speech signals at a specified SNR level. The noise data set was collected from various sound packs published on https://www.freesound.org, which consists of 432 noise signals recorded in different acoustic environments and the total duration is approximately 18 hours. The packs we use are the same as those in [21]. This data set varies in amount, diversity and duration in comparison with the Hu noise database [29] and it is an extended version of [30]. As for the clean speech data set, 4620 clean utterances from the TIMIT database [28] were used, which were spoken by 462 female and male speakers. Each sentence was used 50 times, which gives us 4620×50 clean sentences in total. The time-domain energy level of the clean sentences, including the reused ones, was scaled to be between -22 dB and -3 dB in order to make DNN generalize to different energy levels. Noise was then added to the clean speech sentences, where noise was randomly selected from the constructed freesound noise dataset at a SNR level randomly chosen between -10 dB and 15 dB at a step of 1 dB. In total, more than 100 hours noisy training data were constructed. The generated training data were split into training and validation sets at a proportion of $9:1$.

B. Test data

We constructed three different test sets.

1) 15 different noise signals were randomly picked up from the training noise set, each of which is mixed with 100 sentences extracted from the TIMIT test set (with 50 from male speakers and the other 50 from female speakers) at five input SNR levels ranging from -5 dB to 15 dB at a step of 5 dB. Similar to the training set, the time-domain energy level is normalized to be between -22 dB and -3 dB.
2) Different from the first test set, 15 types of noise from the NOISEX-92 database are used, which is considered to be unseen noise types.
3) The third test set is recorded in an anechoic chamber. There are two loud speakers set up in the chamber, with one louder speaker playing pink noise and the other playing back a clip of clean speech in English read by a female speaker, which is not included in the TIMIT database. A single microphone is used to record the mixed speech.

C. Configurations

The frame length and frame shift are set to 32 milliseconds and 16 milliseconds respectively. The sampling rate is 16 kHz. STFT has 512 frequency bins, which corresponds to a 257-dimensional output for IRM, and 514-dimensional output for AMP and POW. The input frames τ is set to 2, leading to 5×257-dimensional input. The DNN has 4 hidden layers, each with 1024 hidden units. The learning rate is set to 0.0008 at the first epoch, and for the subsequent epoches, adjusted according to exponential decay as $LR = \max(0.008 \cdot 0.95^{E-1}, 0.0001)$, where E is the epoch. The batch size is set to 512. The dropout rate is set to 0.2. In the hidden layers, all nodes use a rectified linear activation function. In the output layer, all nodes are linear for AMP and POW, and sigmoid for IRM. The hyper parameter α in (6) is chosen to be 0.2 for AMP-Wiener in the test stage.

TABLE I
THE AVERAGE PESQ AND STOI SCORES OVER 15 TYPES OF NOISE FROM THE TRAINING NOISE SET.

SNR (dB)		-5	0	5	10	15
PESQ	Noisy	1.48	1.80	2.14	2.49	2.83
	IRM	**2.08**	**2.49**	**2.82**	**3.11**	**3.34**
	AMP	1.76	2.09	2.42	2.72	3.03
	AMP-Wiener	2.04	2.41	2.71	2.97	3.23
	POW-Wiener	1.75	2.08	2.41	2.72	3.02
	OMLSA	1.46	1.92	2.34	2.71	3.05
STOI	Noisy	0.63	0.73	0.83	0.90	0.95
	IRM	**0.73**	**0.83**	**0.90**	**0.94**	**0.97**
	AMP	0.69	0.78	0.83	0.87	0.90
	AMP-Wiener	**0.73**	**0.83**	0.89	0.93	0.96
	POW-Wiener	0.64	0.74	0.83	0.90	0.95
	OMLSA	0.63	0.73	0.83	0.90	0.95

IV. Experimental Results

We evaluated the performance using two metrics: PESQ and STOI. The PESQ is a metric for speech quality. Its score ranges from -0.5 to 4.5. The higher the PESQ score, the better the speech quality. The STOI is computed from the correlation of the temporal envelopes of the degraded speech signal and its clean reference. It has been shown empirically that STOI scores are strongly correlated with human speech intelligibility scores.

The aforementioned DNN based noise-reduction algorithms are compared to the OMLSA algorithm [27], which is a widely used traditional algorithm for noise reduction. In OMLSA, noise statistics are estimated with the IMCRA algorithm.

Table I lists the average PESQ and STOI scores of four DNN based algorithms, i.e., IRM, AMP, AMP-Wiener, POW and the traditional OMLSA algorithm on 15 types of noise taken from the training set. One can see that DNN based algorithms show great advantages in comparison with OMLSA. We can also observe that instead of estimating the speech spectrum directly with DNN (AMP), it is better to estimate speech power spectrum and apply speech enhancement filters to the noisy speech (AMP-Wiener), which give us relatively competitive PESQ and STOI scores to IRM.

Table II presents the average PESQ and STOI scores of DNN based algorithms and the traditional OMLSA algorithm on 15 types of noises from the NOISEX-92 data set, which is not included in the training set. Overall, the OMLSA algorithm outperformed DNN based algorithms in PESQ, while IRM and AMP-Wiener performs better on STOI. Compared with results in table I, the performances of DNN based algorithms degrade significantly in this scenario, which indicates the problem of DNN with generalization.

To gain more insights into the results in table II, we present in Fig. 1 the PESQ scores of noisy and enhanced speech signals for IRM and OMLSA for all types of noise from the NOISEX-92 data set with SNR set at 10 dB. Performances of both DNN based and traditional algorithms varied over noise

TABLE II
THE AVERAGE PESQ AND STOI SCORES OVER 15 NOISES FROM NOISEX-92 DATA SET.

SNR (dB)		−5	0	5	10	15
PESQ	Noisy	1.47	1.80	2.17	2.53	2.87
	IRM	**1.69**	2.09	2.50	2.87	3.19
	AMP	1.30	1.62	2.02	2.32	2.60
	AMP-Wiener	1.67	2.06	2.46	2.80	3.11
	POW-Wiener	1.60	1.98	2.36	2.70	3.03
	OMLSA	1.68	**2.15**	**2.59**	**2.95**	**3.28**
STOI	Noisy	0.62	0.72	0.82	0.89	0.95
	IRM	0.63	**0.75**	**0.85**	**0.92**	**0.96**
	AMP	0.59	0.60	0.78	0.85	0.89
	AMP-Wiener	**0.64**	**0.75**	0.84	0.91	**0.96**
	POW-Wiener	0.63	0.74	0.83	0.90	0.95
	OMLSA	0.60	0.71	0.82	0.89	0.95

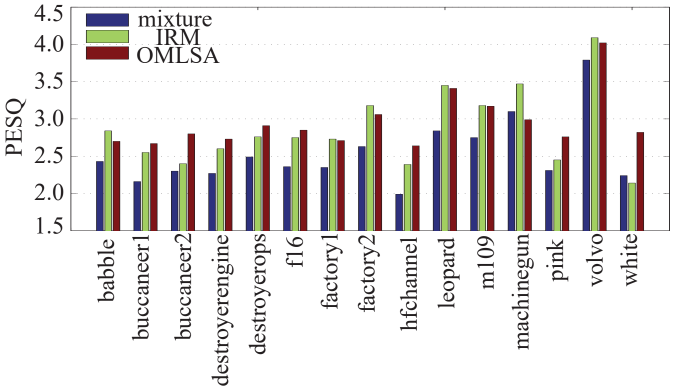

Fig. 1. PESQ scores of mixtures, IRM and OMLSA enhanced speeches corrupted by noises from NOISEX-92 database at an SNR of 10 dB.

types. From Fig. 1, one can see that DNN works better on non-stationary types of noise, like babble, factory, and machine-gun noise. However, it failed to generalize to white noise, which is not included in the training set. The traditional algorithm, OMLSA, works better on stationary noises such as white, pink, and hfchannel noise.

Figure 2 plots noisy and enhanced speech signals by the DNN based algorithm, IRM, and the traditional OMLSA method as well as their spectrograms on the third test set, i.e., signals recorded in an anechoic chamber. One can see that IRM removed almost all noise during speech silence. It also degrades speech, particularly at high frequencies, which are highlighted in red boxes. Besides IRM, the other three DNN based algorithms also have the tendency to remove high frequency speech components, which is plotted to same space. In comparison, OMLSA preserves more high frequency components, but produces less noise reduction in silence parts. It also seems to have more speech distortion at low frequencies as shown in blue boxes. This is mainly due to the low SNR of the recording, which results in over estimation of noise, thereby causing distortion to enhanced speech.

V. SUMMARY

This work presented a comparative study of widely used DNN based noise-reduction algorithms (IRM, AMP and POW) and the traditional OMSLA noise-reduction method. We also presented an algorithm that combines DNN results with Wiener filter, denoted as AMP-Wiener. The following conclusions are made through experimental results.

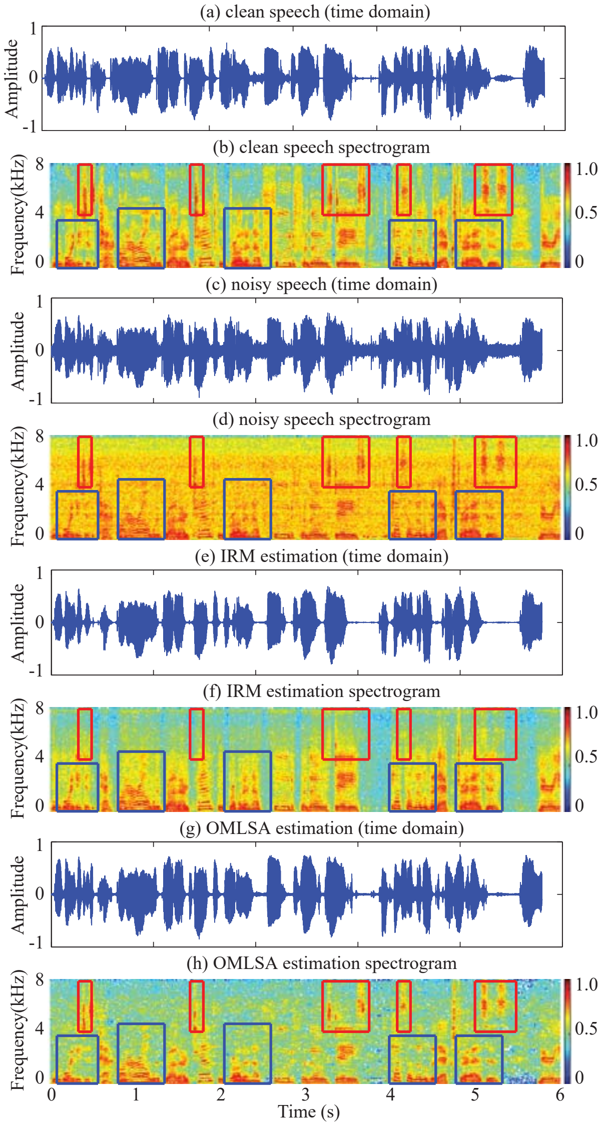

Fig. 2. A visualized comparison between OMLSA and IRM dealing with a recording corrupted by pink noise. (a) Clean speech in time domain. (b) The spectrogram of the clean speech. (c) Noisy speech (corrupted by pink noise) recorded in anechoic chamber. (d) The spectrogram of the noisy speech. (e) The enhanced speech by IRM. (f) The estimated speech spectrogram by IRM. (g) The enhanced speech by OMLSA. (h) The estimated speech spectrogram by OMLSA.

1) DNN based algorithms performs better in low-SNR and non-stationary noise cases than the traditional OMSLA method.
2) DNN based methods achieves more noise reduction in silence periods; however, they have tendency to add more speech distortion at high frequencies where the subband SNR is low.
3) Generalization is indeed a big issue for DNN based

algorithms. In the NOISEX-92 data set where different types of noise are not included in the training set, DNNs methods did not show superiority over the OMSLA method.

4) The AMP-Wiener method performes better than AMP in terms of quality, intelligibility and generalization, which gives us a hint that combining DNN and traditional methods could be promising.

Work is in progress to conduct more comprehensive evaluation for traditional and DNN based noise-reduction algorithms, including subjective tests.

References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice, Second edition*. Boca Raton Florida: CRC Press, 2013.

[2] J. Benesty, J. Chen, and E. Habets *Speech enhancement in the STFT domain*. Berlin: Springer-Verlag, 2011.

[3] G. Huang, J. Benesty, T. Long, and J. Chen, “A family of maximum SNR filters for noise reduction,” *IEEE/ACM Trans. Acoust., Speech, Signal Process.*, vol. 22, pp. 2034–2047, Dec. 2014.

[4] T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano, and K. Kondo, “Theoretical analysis of musical noise in generalized spectral subtraction based on higher order statistics,” *IEEE Trans. Audio, Speech, Lang. Process*, vol. 19, pp. 1770–1779, Aug. 2011.

[5] R. Miyazaki, H.Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, and K. Kondo, “Musical-noise-free speech enhancement based on optimized iterative spectral subtraction, ” *IEEE Trans. Audio, Speech, Lang. Process*, vol. 20, pp. 2080–2094, Sep. 2012.

[6] K. Hu and D. Wang, “Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction,” *IEEE Trans. Audio, Speech, Lang. Process*, vol. 19, pp. 1600–1609, Aug. 2011.

[7] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, “Codebook-based Bayesian speech enhancement for nonstationary environments,” *IEEE Trans. Audio, Speech, Lang. Process*, vol. 15, pp. 441–452, Feb. 2007.

[8] Y. Hu and P. C. Loizou, “A Generalized subspace approach for enhancing speech corrupted by colored noise”, *IEEE Trans. Speech Audio Process.,* vol. 11, pp. 334–341, Jul. 2003.

[9] N. Pan, J. Benesty, and J. Chen, “On single-channel noise reduction with rank-deficient noise correlation matrix” *ELSEVIER Appl. Acoust.,* vol. 126, pp. 126–135, May 2017.

[10] X. Zhang and D. Wang, “Boosting contextual information for deep neural network based voice activity detection,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, pp. 252–264, Feb. 2016.

[11] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 9, pp. 504–512, July 2001.

[12] I. Cohen, “Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging,” *IEEE Trans. Speech, Audio Process.*, vol. 11, pp. 466–475, Sep. 2003.

[13] S. Rangachari and P. C. Loizou, “A noise-estimation algorighm for highly non-stationary environments,” *ELSEVIER Speech Commun.*, vol. 48, pp. 220–231, Feb. 2006.

[14] R. C. Hendriks, J. Jensen, and R. Heusdens, “Noise tracking using DFT domain subspece deconpositions,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, pp. 541–553, Mar. 2008.

[15] Timo Gerkmann and R. C. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 1383–1393, May 2012.

[16] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, “Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback-Leibler divergence,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 1233–1242, July. 2015.

[17] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, ”Codebook-base bayesian speech enhancement for nonstationary environments,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*,vol. 15, pp. 441–452, Feb. 2007.

[18] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Trans Audio Speech Lang Process*, vol. 22, no. 12, pp. 1849–1858, 2014.

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

[20] B. Odelowo and D. Anderson, “ A study of training targets for deep neural network-based speech enhancement using noise prediction,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,*, pp. 5409–5413, May 2018.

[21] R. Rehr and T. Gerkmann , “ An analysis of noise-aware features in combination with the size and diversity of training data for DNN-based speech enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,*, pp. 601–605, May 2019.

[22] H. Zhao, S. Zarar, I. Tashev and C.-H. Lee “ Convolutional-recurrent neural networks for speech enhancement,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,*, pp. 2401–2405, May 2018.

[23] M. Kolbæk, Z.-H. Tan, and J. Jensen, “Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 153–167, 2017.

[24] 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 Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 102–111, 2017.

[25] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *arXiv preprint arXiv:1708.07524*, 2017.

[26] F. Gelderblom, T. Tronstad and E. Viggen, “Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 583–594, Nov. 2018.

[27] I. Cohen and B. Berdugo “Speech enhancement for non-stationary noise environments,” *ELSEVIER Signal Process.*, pp. 2403–2418, Feb. 2001.

[28] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “DARPA TIMIT acoustic phonetic continuous speech corpus,” 1993 [Online]. Available at: http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[29] G. Hu, “A corpus of nonspeech sounds, ”http://web.cse.ohiostate. edu/pnl/corpus/HuNonspeech/HuCorpus.html, 2005.

[30] Y. Xu, J. Du, Z. Huang, D. Rong, and C.-H. Lee, “Multi- objective learning and mask-based post-processing for deep neural network based speech enhancement,” in *Interspeech*, Dresden, Germany, Sep. 2015.

[31] A. Varga and H. J. M. Steeneken, “Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Commun.*, vol. 12, pp. 247–253, July 1993.

[32] *Mapping Function for Transforming Raw Results Scores to MOS-LQO,* ITU-T Rec. P. 862.1, 2003

[33] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 2125–2136, Sep. 2011.