PAN
ARCHIVES OF ACOUSTICS
Vol. **44**, No. 1, pp. 3–12 (2019)
DOI: 10.24425/aoa.2019.126347

Research Papers

Deep Neural Network for Supervised Single-Channel Speech Enhancement

Nasir SALEEM[(1), (2)*], Muhammad IRFAN KHATTAK[(2)]
Muhammad Yousaf ALI[(1)], Muhammad SHAFI[(3)]

[(1)] *Department of Electrical Engineering*
Gomal University
Dera Ismail Khan, 29050, Pakistan
*Corresponding Author e-mail: nasirsaleem@gu.edu.pk

[(2)] *Department of Electrical Engineering*
University of Engineering & Technology
Peshawar, Kohat Campus 26000, Pakistan

[(3)] *Department of Computer Science*
Air University
Islamabad, 26000, Pakistan

(*received May 17, 2018; accepted November 5, 2018*)

Speech enhancement is fundamental for various real time speech applications and it is a challenging task in the case of a single channel because practically only one data channel is available. We have proposed a supervised single channel speech enhancement algorithm in this paper based on a deep neural network (DNN) and less aggressive Wiener filtering as additional DNN layer. During the training stage the network learns and predicts the magnitude spectrums of the clean and noise signals from input noisy speech acoustic features. Relative spectral transform-perceptual linear prediction (RASTA-PLP) is used in the proposed method to extract the acoustic features at the frame level. Autoregressive moving average (ARMA) filter is applied to smooth the temporal curves of extracted features. The trained network predicts the coefficients to construct a ratio mask based on mean square error (MSE) objective cost function. The less aggressive Wiener filter is placed as an additional layer on the top of a DNN to produce an enhanced magnitude spectrum. Finally, the noisy speech phase is used to reconstruct the enhanced speech. The experimental results demonstrate that the proposed DNN framework with less aggressive Wiener filtering outperforms the competing speech enhancement methods in terms of the speech quality and intelligibility.

Keywords: deep neural network; intelligibility; speech enhancement; speech quality; supervised learning; Wiener filtering.

1. Introduction

Speech enhancement reduces the background noise and improves the quality and intelligibility of the degraded speech utterances in noisy conditions and has many applications, for example, hearing aid, mobile communication and automatic speech recognition (ASR). Single channel based speech enhancement is possibly the most desirable from application viewpoint. Compared to the multi channel, a single channel is less affected by the room reverberation and spatial sources. A number of single channel speech enhancement methods are available in literature, for example, minimum mean squared error (MMSE) (Ephraim, Malah, 1984); log MMSE (LMMSE) (Ephraim, Malah, 1985) estimation, spectral subtraction (Boll, 1979), Wiener filtering (WF) (Scalart, 1996), and many others, including (Saleem, Irfan, 2017; Kim *et al.*, 2012; Xu *et al.*, 2017; Kolbk *et al.*, 2017; Doire *et al.*, 2017; Sun *et al.*, 2016a; 2016b). The methods in (Ephraim, Malah, 1984; 1985; Boll, 1979; Scalart, 1996) are usually less efficient in strong noisy conditions and assume stationarity of the noise signals, therefore, they perform poorly in nonstationary noisy conditions. Statistical model based methods (Gerkmann, Hendriks, 2012; Schwerin, Paliwal, 2014) show improved results in good signal-to-noise ratio (SNR). In contrast, model based methods showed promising results in strong noisy conditions. Methods in (Hershey *et al.*, 2010; Narayanan, Wang, 2013) constructed probabilistic models which are based on the prior knowledge and showed considerable performance gain. The computational auditory scene analysis (CASA) based methods (Wang, Brown, 2006; Wang, Wang, 2013) replicate the process of human auditory system by exploiting the signal processing techniques and group them into an auditory stream using psycho-acoustic cues. In such methods, the speech enhancement is carried out as a binary classification problem to estimate the ideal binary mask (IBM) and ideal ratio mask (IRM) (Wang, 2005) and shows robustness to nonstationary noise sources in a wide range of acoustic conditions. IBM improves the speech quality and has been shown to improve speech intelligibility (Roman, Woodruff, 2013; Li, Loizou, 2008; Saleem *et al.*, 2015). Nonnegative matrix factorisation (NMF) is extensively used as a model based method for reducing of non-stationary noise signals (Mohammadiha *et al.*, 2013; Sun *et al.,* 2015). In NMF, nonnegative data matrices are estimated from the product of basis and encoding matrices containing nonnegative elements. However, NMF is considered as a linear model that cannot extract the complex features and is not effective as compared to non linear models. The empirical mode decomposition (EMD) gained enormous attention in speech enhancement. It is a pioneer work proposed by (Chatlani, Soraghan, 2012) where EMD is used as post filtering to reduce musical noise. A speech enhancement method is proposed by (Zao *et al.*, 2014) based on EMD and Hurst based intrinsic mode function (IMF) selection criterion. The Hurst exponent statistics is used to classify and choose IMFs which are notably affected by noise components. The restoration of enhanced speech is based on least degraded IMFs.

Recently, deep learning is used in speech enhancement (Xu *et al.*, 2014), which is a learning method with multiple layers representation. This representation is achieved by using a nonlinear module where each layer transforms the representation from a lower to a higher level. A regression based DNN is proposed by (Xu *et al.*, 2015) where the DNN framework is considered to estimate clean speech from the noisy speech. The huge training set of 104 noise sources is used to train a network. A DNN framework is proposed in (Wang *et al.*, 2014) to train diverse targets, including IBM, ideal ratio mask (IRM), FFT-mask, Gammatone frequency power spectrum (GFPS), and short-time Fourier transform (STFT) spectral magnitude for supervised speech separation. A joint optimisation of masking functions and deep recurrent neural networks (DRNN) for monaural source separation task is proposed by (Huang *et al.*, 2015). The joint optimisation of DRNN with an additional masking layer enforces a discriminative training approach for DNN to further improve the separation performance. Here we present a speech enhancement method which adds a less aggressive Wiener filter as an additional layer in a DNN framework to increase speech intelligibility and quality in adverse noisy conditions. A comparative performance study is carried out to access the performance of the proposed method in diverse noisy conditions in terms of the speech quality and intelligibility. Five competing speech enhancement methods are used for comparison purpose. The Matlab R2015b is used to develop the proposed method and all the simulations and evaluations are performed using this simulation platform. The remaining paper is organised as follows: an overview of our method is presented in Sec. 2; the experimental setup is presented in Sec. 3, while results and discussions are presented in Sec. 4. Finally, the concluding remarks are presented in Sec. 5.

2. Proposed speech enhancement overview

Consider a noisy speech $S_m(t)$ which can be expressed as a sum of the clean speech and noise as: $S_m(t) = x(t)+d(t)$, where $x(t)$ and $d(t)$ show the clean speech and additive noise signal, respectively. Usually, Wiener filtering is applied as a soft mask to the DNN outputs. However, in our method, we have combined a DNN and a less aggressive Wiener filter (LW) (Chen, Loizou, 2010) into a single structure. LW can synthesise spectral components more efficiently as compared to the conventional rigid Wiener filter, specifically the components with low *a priori* SNR without over attenuation. The proposed deep neural network framework is named as DNN+LW and shown in Fig. 1, where for the given input features, the DNN+LW computes the magnitude spectrums of the clean speech and noise signals. After acquiring magnitude spectrums, the LW is used as an additional layer. The DNN framework is trained by extracting acoustic features from the noisy and clean speech signals based on RASTA-PLP, as shown in Fig. 2. All acoustic features are extracted at the frame level and coupled with delta features. A second order ($k = 2$) ARMA filter is applied to smooth the temporal curves of extracted features, and is given by equation:

$$\widehat{F}(t) = \frac{\widehat{F}(t-k) + ... + F(t) + ... + F(t+k)}{2k+1}, \qquad (1)$$

where $F(t)$ shows the feature vector at time t, $\widehat{F}(t)$ are filtered feature vectors, and k shows the filter order. The dropout and back propagation techniques are used in the proposed method. Short time Fourier analysis is

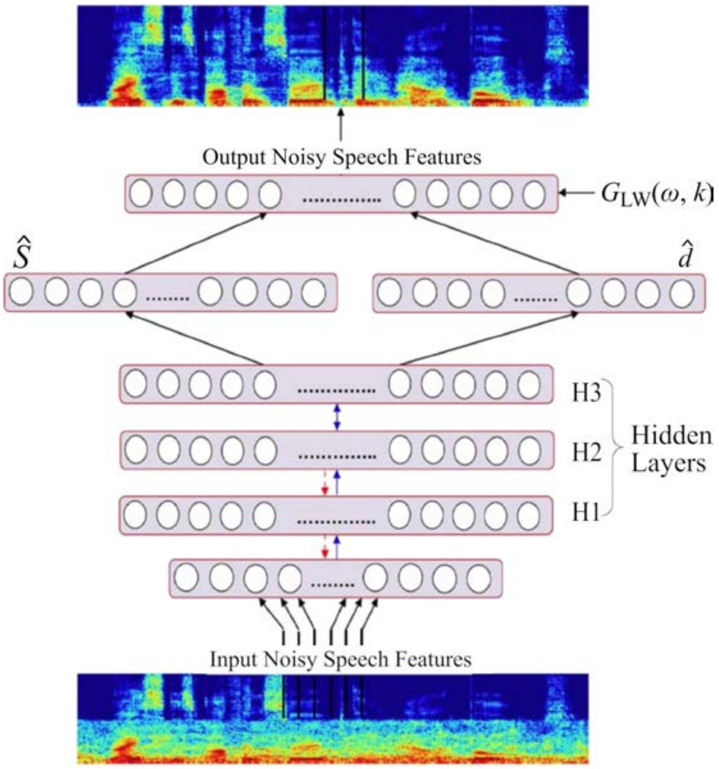
Output Noisy Speech Features
$G_{\mathrm{LW}}(\omega, k)$
$\hat{S}$
$\hat{d}$
H3
H2
H1
Hidden Layers
Input Noisy Speech Features

Fig. 1. Proposed DNN framework.

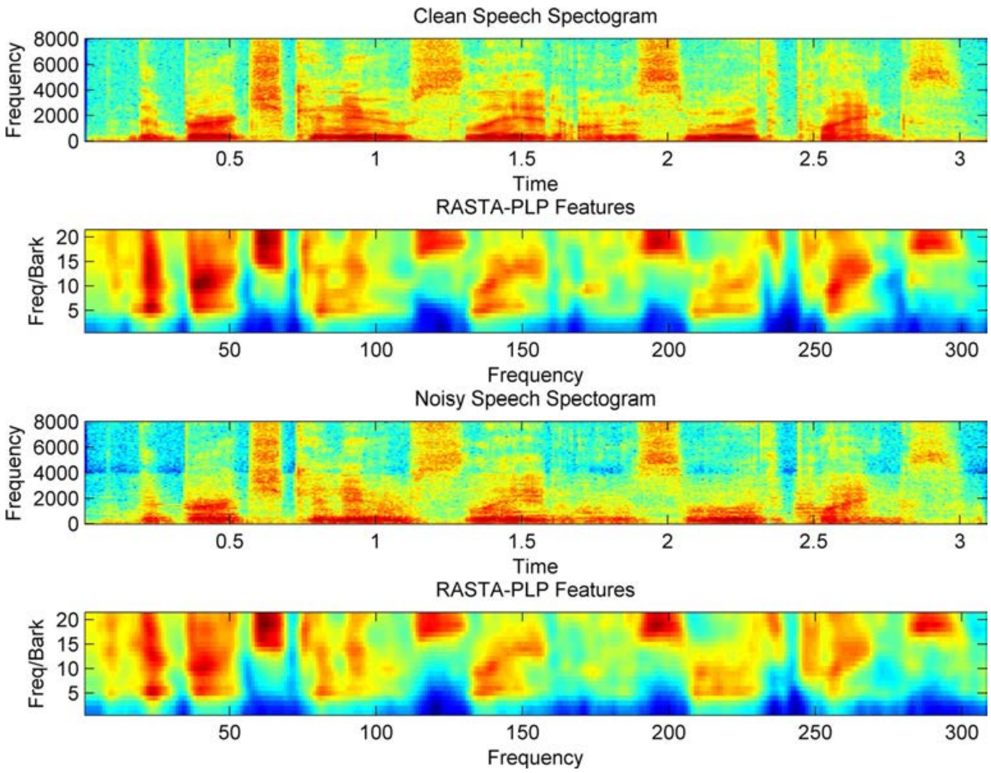

Fig. 2. RASTA-PLP extracted features for clean and noisy speech utterances.

applied to input signal and discrete Fourier transform (DFT) is computed for all overlapping frames. DNN predicts the magnitude spectrums of clean and noise signals and the less aggressive Wiener filtering can be seen as additional layer on the top of the DNN output layer. The filtering process is as follows:

$$\widetilde{X}(\omega,k) = \frac{|\widehat{x}(\omega,k)|^2}{|\widehat{x}(\omega,k)|^2 + |\widehat{d}(\omega,k)|^2} \odot Y_m(\omega,k), \tag{2}$$

$$\widetilde{X}(\omega,k) = \frac{E\left\{|\widehat{x}(\omega,k)|^2\right\}}{E\left\{|\widehat{x}(\omega,k)|^2\right\} + E\left\{|\widehat{d}(\omega,k)|^2\right\}} \odot Y_m(\omega,k). \tag{3}$$

The operator $\odot$ shows elementwise multiplication function and $E\{\cdot\}$ is expectation operator. In Eq. (3), outputs of DNN, that is, $\widehat{x}(\omega,k)$ and $\widehat{d}(\omega,k)$, are used to compute gain, and $Y_m(\omega,k)$ is magnitude spectrum of the noisy speech. The gain $G(\omega,k)$ is a function of *a priori* SNR $\xi(\omega,k)$ and *a posteriori* SNR $\gamma(\omega,k)$, respectively, and is given as:

$$\begin{aligned} \xi(\omega,k) &= \frac{E\left(|X(\omega,k)|^2\right)}{E\left(|D(\omega,k)|^2\right)}, \\ \gamma(\omega,k) &= \frac{|Y_m(\omega,k)|^2}{E\left(|D(\omega,k)|^2\right)}. \end{aligned} \tag{4}$$

The gain of conventional and LW filtering is given as:

$$\begin{aligned} G_C(\omega,k) &= \frac{\xi(\omega,k)}{\xi(\omega,k)+1}, \\ G_{\mathrm{LW}}(\omega,k) &= \frac{\sqrt{\xi(\omega,k)}}{\sqrt{\xi(\omega,k)+1}}. \end{aligned} \tag{5}$$

Here, the conventional rigid Wiener filter $G_C(\omega,k)$ is replaced with less aggressive Wiener filtering $G_{\mathrm{LW}}(\omega,k)$ because this gain can synthesise spectral components, specifically those with low *a priori* SNRs without any over-attenuation. The estimate of *a priori* SNR is computed by using a modified version of decision direct (DD) (Ephraim, Malah, 1984) approach given as:

$$\begin{aligned} \xi_{\mathrm{MDD}}(\omega,k) &= \beta \frac{\left|\widehat{S}(\omega,k-1)\right|^2}{\lambda_D(\omega,k-1)} + \eta(\omega,k) \\ &\quad + (1-\beta)\cdot\max\left[\frac{|Y(\omega,k)|^2}{\lambda_D(\omega,k)} - 1,\ 0\right], \end{aligned} \tag{6}$$

$$\eta(\omega,k) = \alpha[\xi(\omega,k-1) - \xi(\omega,k-2)],$$

$\xi_{\mathrm{MDD}}(\omega,k)$ is *a priori* SNR estimate using the modified decision direct method, β is the smoothing parameter ($\beta = 0.98$), α is the momentum parameter ($\alpha = 0.99$), $\eta(\omega,k)$ is momentum terms, and $\lambda_D(\omega,k)$ is the estimate of background noise variance. The magnitude spectrum of noisy speech $Y_m(\omega,k)$ is multiplied with $G_{\mathrm{LW}}(\omega,k)$ gain. During the enhancement stage, the trained DNN is fed with acoustic features to obtain the enhanced speech. The phase is appended from the noisy speech as it is not useful for human auditory perception. Finally, the enhanced speech is obtained by taking inverse DFT and overlap-and-add method.

3. Experimental setting

A dataset composed of 720 IEEE utterances (Rothauser, 1969) is used as training utterances, whereas the testing set consists of 250 utterances from unknown speakers of both genders. We used five noise sources from AURORA dataset (Hirsch, Pearce, 2000) in the training and testing process. The noise sources include: airport, babble, car, street, and train. The spectrograms of the noise sources are given in Fig. 3. All noise sources are considered nonstationary and the duration of each noise source is around 3.5 minutes.

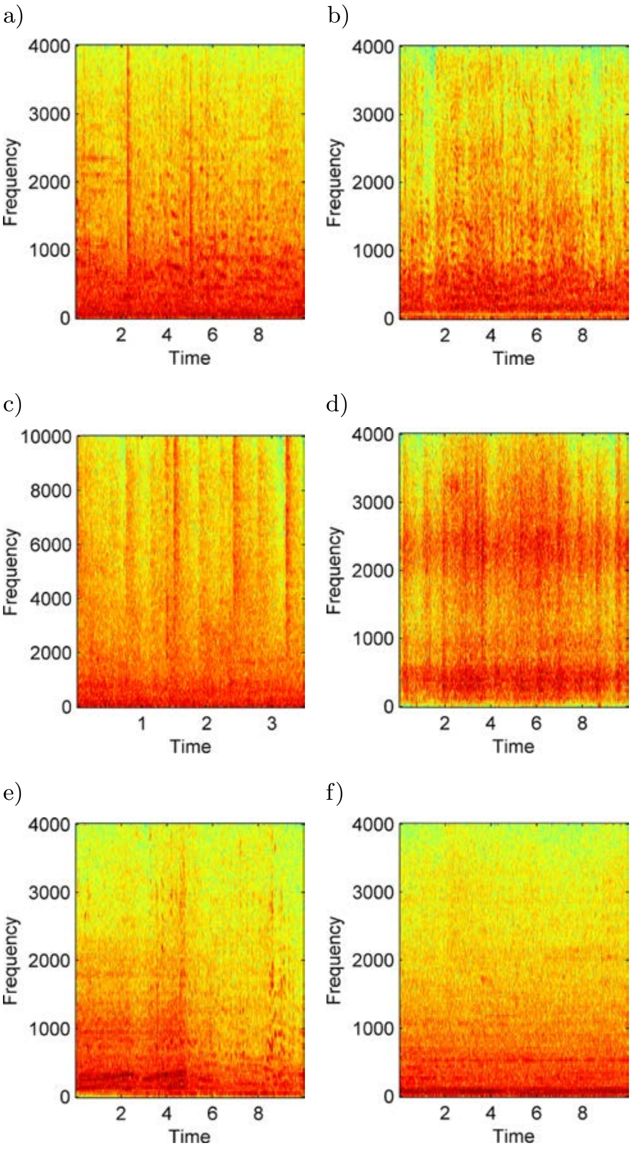

Fig. 3. Spectrograms of noise sources used in experiments.

In order to formulate the training set, random cuts from the first half of all noise sources are used and mixed with clean utterances at −10 dB, −6 dB, −2 dB, 2 dB, 6 dB, and 10 dB SNR, respectively. The testing mixtures are formulated by mixing random cuts from the second half of all noise sources. In our proposed method, DNN framework used three hidden layers. Each layer contains 1024 hidden units and the sigmoid function is used as an activation function. The standard back propagation method is used to train the network. To circumvent the mismatch between training and testing, the dropout regularisation method (Seltzer *et al.*, 2013) is adopted to improve the generalisation of DNN. The dropout rate is fixed at 0.2 and no unsupervised pretraining is used. The adaptive gradient descent (AGD) method (Duchi *et al.*, 2011) is tied with a momentum term ς to optimise the proposed DNN. For first 4 epochs, the momentum rate is set at 0.6, while the rate is increased and fixed at 0.8 in the remaining epochs. The mean square error (MSE) is used as the objective cost function to reduce errors during the training. For evaluation purpose, we have adopted the frequency weighted segmental SNR (FwSNRSeg) (Krishnamoorthy, 2011) and short time objective intelligibility (STOI) (Taal *et al.*, 2011) to measure objective speech intelligibility. STOI scores are computed by correlating the clean and enhanced speech signals and this measure has shown a strong connection to the human speech intelligibility. FwSNRSeg is preferred as an objective intelligibility measure since it showed a strong correlation to subjective speech intelligibility (Hu, Loizou, 2008). To measure objective speech quality, perceptual evaluation of speech quality (PESQ) is adopted (Rix *et al.*, 2001). The range of STOI scores is from 0 to 1, whereas for PESQ scores, the range is from −0.5 to 4.5. SNR based measure is used to evaluate the performance of the speech enhancement methods. However, standard SNR measure does not present a good correlation with the speech quality because averaging over the entire signal length can remove crucial speech contents. To handle this problem, the segmental SNR (SNRSeg) is used, which computes SNR in short segments. We considered this measure to examine the noise suppression in the synthesised speech. Four competing methods are chosen for performance comparison with the proposed method. The competing methods include spectral subtraction, Weiner filtering, LMMSE, and NMF.

4. Results and discussions

To evaluate the performance at all input SNRs, we have given mean and the best performance values of DNN, MMSE, LMMSE, NMF, WF, and DNN+LW.

4.1. Comparison with competing methods

Table 1 shows comparison of DNN+LW and competing methods in terms of PESQ, used to examine the overall quality of the synthesised speech. We have noticed that DNN+LW achieved consistent higher PESQ

Table 1. The objective quality analysis using PESQ.

Methods	−10 dB	−6 dB	−2 dB	2 dB	6 dB	10 dB	−10 dB	−6 dB	−2 dB	2 dB	6 dB	10 dB
Airport noise							Babble noise					
MMSE	1.19	1.42	1.73	2.01	2.31	2.41	1.17	1.40	1.58	1.79	2.11	2.34
LMMSE	1.31	1.66	1.91	2.19	2.44	2.76	1.23	1.41	1.65	1.83	2.19	2.43
NMF	1.29	1.61	1.88	2.18	2.38	2.68	1.29	1.38	1.81	1.92	2.20	2.54
WF	1.23	1.51	1.78	2.02	2.21	2.55	1.21	1.43	1.66	1.86	2.18	2.44
DNN	1.58	1.82	2.12	2.38	2.65	2.91	1.42	1.68	1.92	2.22	2.52	2.75
DNN+LW	1.69	1.91	2.18	2.45	2.78	2.96	1.58	1.87	2.15	2.35	2.70	2.94
Factory Noise							Street Noise					
MMSE	1.21	1.48	1.78	2.09	2.21	2.57	1.18	1.47	1.78	2.03	2.21	2.54
LMMSE	1.27	1.57	1.83	2.16	2.34	2.68	1.23	1.67	1.91	2.01	2.33	2.67
NMF	1.30	1.41	1.88	2.02	2.14	2.47	1.42	1.65	1.88	2.11	2.32	2.47
WF	1.22	1.47	1.75	2.06	2.24	2.58	1.28	1.78	1.94	2.02	2.21	2.65
DNN	1.41	1.79	1.90	2.19	2.46	2.76	1.47	1.82	2.01	2.23	2.52	2.81
DNN+LW	1.47	1.83	2.02	2.30	2.53	2.87	1.55	1.91	2.13	2.31	2.59	2.88
Subway Noise							Train Noise					
MMSE	1.16	1.39	1.76	2.04	2.21	2.44	1.66	1.99	2.22	2.45	2.66	2.82
LMMSE	1.21	1.52	1.88	2.11	2.35	2.51	1.72	2.08	2.33	2.58	2.70	2.86
NMF	1.18	1.48	1.71	2.09	2.21	2.45	1.65	1.89	2.29	2.55	2.49	2.85
WF	1.02	1.31	1.54	1.82	2.09	2.35	1.66	1.92	2.16	2.43	2.63	2.85
DNN	1.63	1.77	2.06	2.26	2.46	2.68	1.91	2.17	2.44	2.75	2.95	3.18
DNN+LW	1.66	1.88	2.15	2.37	2.65	2.71	2.07	2.32	2.58	2.96	3.01	3.27

scores at all input SNRs, however, a less significant PESQ scores are achieved in some noisy conditions at high SNRs. Considerable PESQ scores are attained by the DNN+LW at low SNRs, i.e., −10 dB, −6 dB, and −2 dB, in all noisy conditions. MMSE, LMMSE, NMF, and WF methods have lost considerable speech quality as compared to DNN+LW. For instance, the average predicted PESQ scores in airport noise are improved from 2.06 with LMMSE to 2.32 with DNN+LW. In the same way, for train noise, the average scores are improved from 2.27 with WF to 2.70 with DNN+LW. Also, the average PESQ scores in babble noise are improved from 2.08 with competing DNN to 2.26 with DNN+LW.

The overall average PESQ scores across all noise sources are improved from 1.28 with unprocessed noisy speech to 1.85 with DNN+LW at −10 dB and −6 dB. The highest and lowest PESQ improvements are achieved at −10 dB train noise and 10 dB babble noise, i.e., ΔPESQ = 0.63 and ΔPESQ = 0.54, respectively. Table 2 shows the comparison of DNN+LW and competing methods in terms of SNRSeg, used to examine noise suppression in synthesised speech. The results indicate that DNN+LW significantly reduced the background noise and achieved better SNRSeg scores in all noisy conditions and at all SNRs consistently. High values of SNRSeg measure at low SNRs (−10 dB and −6 dB) indicate that the DNN+LW based speech enhancement has the great capacity to reduce noise in adverse noisy conditions. In terms of SNRSeg, the DNN+LW performed better in train noisy condition and achieved the highest scores as compared to other competing methods. The average predicted SNRSeg scores in factory, subway, and street noise are improved from 3.38, 3.29, and 3.57 with WF to 4.65, 4.68, and 4.82, respectively, with DNN+LW. Similarly, average SNRSeg scores in five noise sources are improved from 4.28 with DNN to 4.51 with DNN+LW. Within-competing methods comparison, in terms of SNRSeg, the WF performed equally well for most SNR conditions and five noise sources, except for a few SNR conditions. For instance, the average SNRSeg scores in factory noise are improved from 2.25, 3.39, and 2.80 with MMSE, LMMSE, and NMF to 3.43 with WF. Speech enhancement mainly involves noise suppressing, so that to improve the quality of the noisy speech. But, in speech recognition, the intelligibility is the key attribute. The intelligibility scores for all methods are predicted using STOI measure, presented in Fig. 4. The highest and lowest STOI scores for DNN+LW are achieved at train and babble noise, respectively. All five noise sources led to the high intelligibility scores (STOI ≥ 80%) for SNR = 10 dB. However, large differences in STOI scores are found at low SNRs. DNN+LW outperformed the competing

Table 2. The objective quality analysis using SNRSeg.

Methods	−10 dB	−6 dB	−2 dB	2 dB	6 dB	10 dB	−10 dB	−6 dB	−2 dB	2 dB	6 dB	10 dB
Airport noise							Babble noise					
MMSE	0.26	0.36	1.00	2.49	3.86	5.65	0.31	0.70	1.38	2.35	3.87	5.76
LMMSE	0.31	0.86	1.81	3.12	5.04	7.03	0.17	0.47	1.24	2.75	4.77	6.93
NMF	0.37	1.21	1.92	3.22	4.32	5.97	0.20	0.56	1.41	2.30	3.07	4.95
WF	0.40	0.81	1.65	3.42	5.11	6.98	0.22	0.62	1.49	3.08	4.64	6.78
DNN	1.52	2.56	3.54	4.92	6.43	8.01	1.48	2.01	2.98	4.46	5.96	7.50
DNN+LW	1.63	2.74	3.72	5.04	6.54	8.25	1.57	2.07	3.11	4.60	6.05	7.67
Factory Noise							Street Noise					
MMSE	0.14	0.42	1.12	2.23	3.94	5.68	0.12	0.44	1.15	2.21	3.83	5.56
LMMSE	0.42	1.07	2.07	4.02	5.75	7.03	0.23	1.03	1.63	3.34	4.97	6.98
NMF	0.43	0.98	2.01	3.08	4.11	6.24	0.24	0.93	1.54	3.14	3.94	5.92
WF	0.45	1.22	2.23	3.86	5.70	7.12	0.44	1.28	3.17	3.84	5.18	7.51
DNN	1.43	1.59	2.83	4.28	5.59	7.21	1.77	2.59	3.69	4.65	6.61	8.09
DNN+LW	1.57	1.97	2.97	4.52	5.98	7.92	1.84	2.67	3.92	5.44	6.74	8.32
Subway Noise							Train Noise					
MMSE	0.11	0.19	0.35	0.64	1.15	2.08	0.22	0.57	1.21	2.43	3.87	5.98
LMMSE	0.52	1.14	2.11	3.76	5.35	7.33	0.79	1.81	3.31	4.62	6.75	8.37
NMF	0.17	0.57	1.98	2.19	3.12	4.98	0.57	1.29	2.97	4.22	5.43	7.11
WF	0.21	1.09	2.08	3.69	5.32	7.39	0.67	1.95	3.36	4.60	6.32	8.23
DNN	1.43	2.44	3.65	5.10	6.13	7.42	2.07	3.18	4.40	6.25	7.28	8.69
DNN+LW	1.45	2.45	3.71	5.43	6.64	8.24	2.11	3.28	4.70	6.28	7.43	8.98

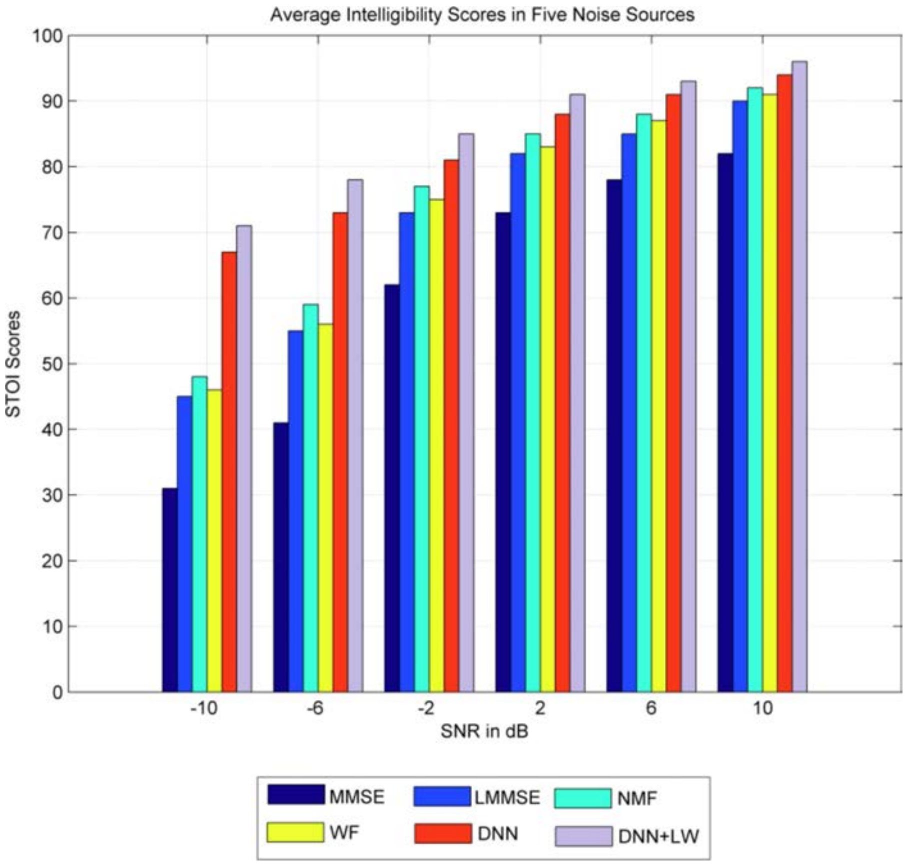

Fig. 4. Speech intelligibility scores using STOI.

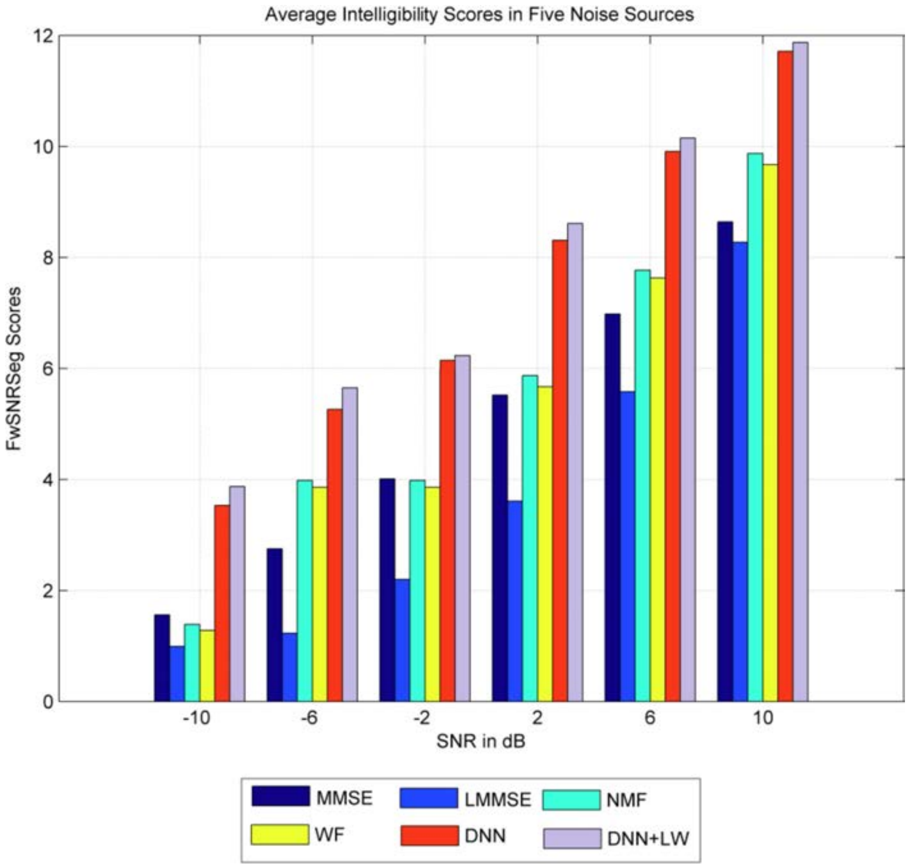

Fig. 5. Speech intelligibility scores using FwSNRSeg.

methods at all SNRs and led to the best overall average prediction rate: 88.51%. Note from the Fig. 4, as compared to MMSE, LMMSE, NMF, and WF, that the DNN+LW achieved the best average STOI scores in five nonstationary noisy conditions. For instance, the predicted STOI scores are improved from 61.51% with MMSE to 83.16% with DNN+LW. The average STOI scores for LMMSE, NMF, WF, and DNN are 66.16%, 73.1%, 74.16%, and 80.60%, respectively. Finally, the FwSNRSeg is used, which showed a high correlation with subjective speech intelligibility. Large FwSNRSeg scores describe improved intelligibility. Figure 5 gives FwSNRSeg scores obtained with DNN+LW and competing methods. The average comparison shows that DNN+LW achieved the best scores at all input SNRs for five nonstationary noise sources as compared to competing methods. When compared to the baseline DNN, DNN+LW achieved the best FwSNRSeg scores at all noise sources. Average scores demonstrate that DNN+LW performed very well at all SNRs in terms of STOI and FwSNRSeg. The experimental outcomes verify the superiority of DNN+LW-based speech enhancement in terms of speech intelligibility.

4.2. Spectrogram analysis

For obtaining further understanding about the residual noise we have examined the time varying spectrograms. Figure 6 shows the spectrograms of DNN+LW and competing methods. A speech utterance is contaminated by airport noise at +2 dB SNR, and has PESQ = 1.98 and SNRSeg = 2.42 dB. By examining spectrograms in Fig. 6c–h harmonic spectrums of the vowels are retained. Hence, the DNN+LW and

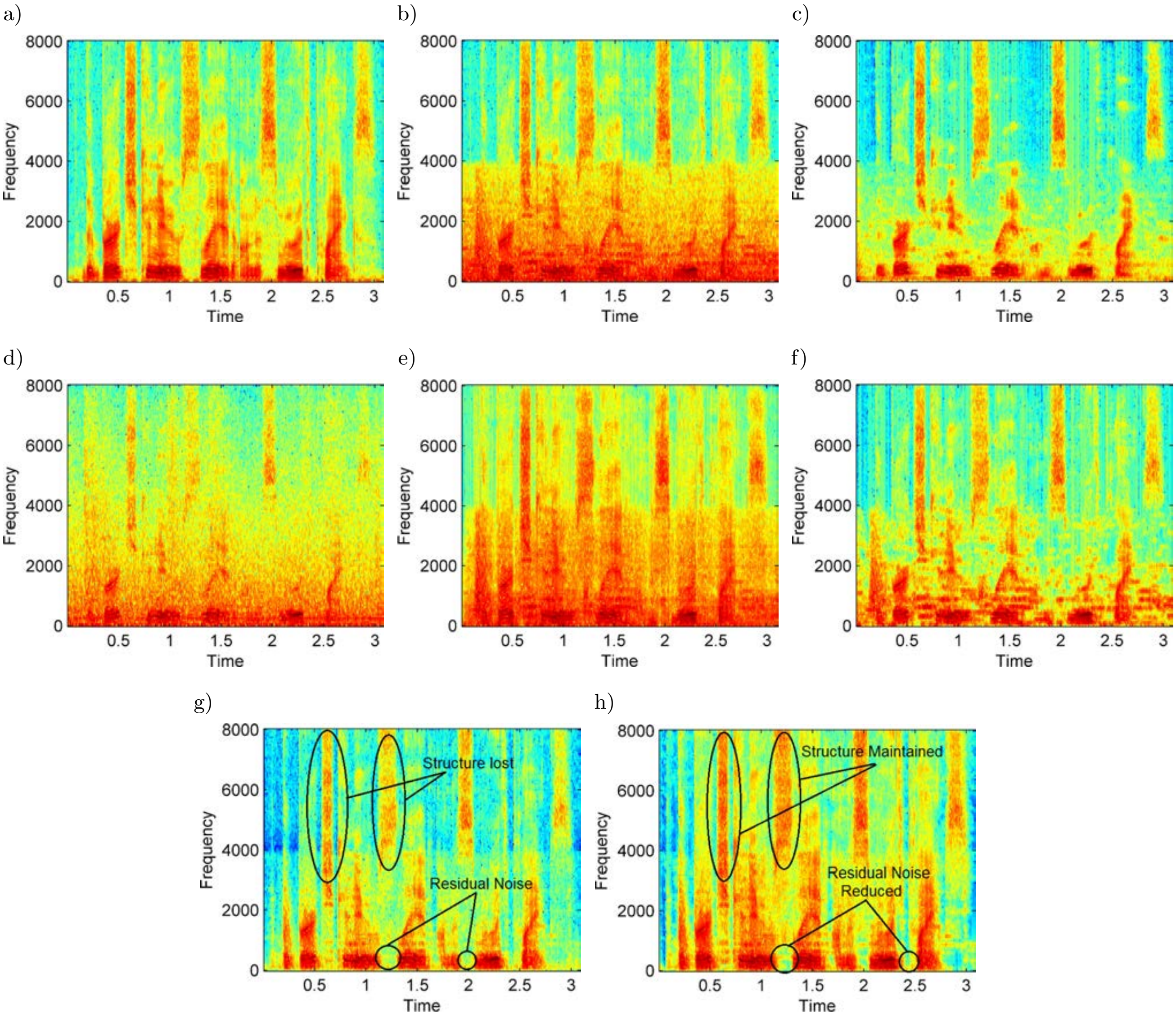

Fig. 6. Spectrogram analysis: a) clean speech, b) noisy speech, c) speech processed by LMMSE, d) speech processed by MMSE, e) speech processed by NMF, f) speech processed by Weiner filtering, g) speech processed by DNN, and h) speech processed by DNN+LW.

competing methods did not suffer badly from over-attenuation. However, large residual noise is evident in spectrograms of MMSE, LMMSE, NMF, and WF. During speech-pause areas, the DNN+LW are adequate in removing background noise as compared to other methods. Weak harmonic structures in high frequency subbands are retained by DNN+LW. That is why, perceptual quality of DNN+LW speech is better than that of the competing methods. The residual noise is evident in the spectrogram showing the output speech of DNN, shown in Fig. 6g, which is considerably reduced in the spectrogram showing the output speech of DNN+LW, shown in Fig. 6h. The weak energy contents are well preserved by DNN+LW, resulting in less speech distortion.

5. Conclusion

A supervised single-channel speech enhancement algorithm based on deep neural network (DNN) and less aggressive Weiner filtering gain is proposed in this paper. Usually, Wiener filtering is used as a soft mask to DNN outputs, but we have used DNN and less aggressive Wiener filter in a single framework. Four performance metrics and five competing methods are used in experiments to evaluate the performance of the proposed speech enhancement method. To summarise, the DNN+LW outperformed by yielding a high speech quality and the background noise is excellently reduced with less residual noise. Also, DNN+LW showed the highest intelligibility scores in all noise sources.

References

1. Boll S. (1979), *Suppression of acoustic noise in speech using spectral subtraction.* IEEE Transactions on Acoustics, Speech, and Signal Processing, **27**, 2, 113–120.
2. Chatlani N., Soraghan J.J. (2012), *EMD-based filtering (EMDF) of low-frequency noise for speech enhancement*, IEEE Transactions on Audio, Speech, and Language Processing, **20**, 4, 1158–1166.
3. Chen F., Loizou P.C. (2010), *Speech enhancement using a frequency-specific composite Wiener function*, 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 4726–4729.
4. Doire C.S., Brookes M., Naylor P.A., Hicks C.M., Betts D., Dmour M.A., Jensen S.H. (2017), *Single-channel online enhancement of speech corrupted by reverberation and noise*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, **25**, 3, 572–587.
5. Duchi J., Hazan E., Singer Y. (2011), *Adaptive subgradient methods for online learning and stochastic optimization*, Journal of Machine Learning Research, **12**, 2121–2159.
6. Ephraim Y., Malah D. (1984), *Speech enhancement using a minimum–mean square error short-time spectral amplitude estimator*, IEEE Transactions on Acoustics, Speech, and Signal Processing, **32**, 6, 1109–1121.
7. Ephraim Y., Malah D. (1985), *Speech enhancement using a minimum mean-square error log-spectral amplitude estimator*, IEEE Transactions on Acoustics, Speech, and Signal Processing, **33**, 2, 443–445.
8. Gerkmann T., Hendriks R.C. (2012), *Unbiased MMSE-based noise power estimation with low complexity and low tracking delay*, IEEE Transactions on Audio, Speech, and Language Processing, **20**, 4, 1383–1393.
9. Hershey J.R., Rennie S.J., Olsen P.A., Kristjansson T.T. (2010), *Super-human multi-talker speech recognition: a graphical modeling approach*, Computer Speech & Language, **24**, 1, 45–66.
10. Hirsch H.-G., Pearce D. (2000), *The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions*, ASR2000-Automatic Speech Recognition: Challenges for the New Millenium ISCA Tutorial and Research Workshop (ITRW).
11. Huang P.-S., Kim M., Hasegawa-Johnson M., Smaragdis P. (2015), *Joint optimization of masks and deep recurrent neural networks for monaural source separation*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, **23**, 12, 2136–2147.
12. Hu Y., Loizou P.C. (2008), *Evaluation of objective quality measures for speech enhancement*, IEEE Transactions on Audio, Speech, and Language Processing, **16**, 1, 229–238.
13. Kim H.-G., Jang G.-J., Park J.-S., Kim J.-H., Oh Y.-H. (2012), *Speech segregation based on pitch track correction and music-speech classification*, Advances in Electrical and Computer Engineering, **12**, 2, 15–20.
14. Kolbk M., Tan Z.-H., Jensen J. (2017), *Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems*, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), **25**, 1, 153–167.
15. Krishnamoorthy P. (2011), *An overview of subjective and objective quality measures for noisy speech enhancement algorithms*, IETE Technical Review, **28**, 4, 292–301.
16. Li N., Loizou P.C. (2008), *Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction*, The Journal of the Acoustical Society of America, **123**, 3, 1673–1682.
17. Mohammadiha N., Smaragdis P., Leijon A. (2013), *Supervised and unsupervised speech enhancement using nonnegative matrix factorization*, IEEE Transactions on Audio, Speech, and Language Processing, **21**, 10, 2140–2151.
18. Narayanan A., Wang D. (2013), *The role of binary mask patterns in automatic speech recognition in background noise*, The Journal of the Acoustical Society of America, **133**, 5, 3083–3093.

19. Rix A.W., Beerends J.G., Hollier M.P., Hekstra A.P. (2001), *Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs*, ICASSP '01 Proceedings of the Acoustics, Speech, and Signal Processing on Conference Proceedings, 2001 IEEE International Conference, **2**, 749–752.

20. Roman N., Woodruff J. (2013), *Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold*, The Journal of the Acoustical Society of America, **133**, 3, 1707–1717.

21. Rothauser E. (1969), *IEEE recommended practice for speech quality measurements*, IEEE Transactions on Audio and Electroacoustics, **17**, 225–246.

22. Saleem N., Irfan M. (2017), *Noise reduction based on soft masks by incorporating SNR uncertainty in frequency domain*, Circuits, Systems, and Signal Processing, 1–22.

23. Saleem N., Shafi M., Mustafa E., Nawaz A. (2015), *A novel binary mask estimation based on spectral subtraction gain-induced distortions for improved speech intelligibility and quality*, University of Engineering and Technology Taxila, Technical Journal, **20**, 4, 36.

24. Scalart P. (1996), *Speech enhancement based on a priori signal to noise estimation*, ICASSP '96 Proceedings of the Acoustics, Speech, and Signal Processing on Conference Proceedings, 1996 IEEE International Conference, **2**, 629–632.

25. Schwerin B., Paliwal K. (2014), *Using STFT real and imaginary parts of modulation signals for MMSE-based speech enhancement*, Speech Communication, **58**, 49–68.

26. Seltzer M.L., Yu D., Wang Y. (2013), *An investigation of deep neural networks for noise robust speech recognition*, 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7398–7402.

27. Sun C., Xie J., Leng Y. (2016a), *A signal subspace speech enhancement approach based on joint low-rank and sparse matrix decomposition*, Archives of Acoustics, **41**, 2, 245–254

28. Sun M., Zhang X., Zheng T.F. (2016b), *Unseen noise estimation using separable deep auto encoder for speech enhancement*, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), **24**, 1, 93–104.

29. Sun M., Li Y., Gemmeke J.F., Zhang X. (2015), *Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback-Leibler divergence*, IEEE Transactions on Audio, Speech, and Language Processing, **23**, 7, 1233–1242.

30. Taal C.H., Hendriks R.C., Heusdens R., Jensen J. (2011), *An algorithm for intelligibility prediction of time-frequency weighted noisy speech*, IEEE Transactions on Audio, Speech, and Language Processing, **19**, 7, 2125–2136.

31. Wang D. (2005), *On ideal binary mask as the computational goal of auditory scene analysis*, [in:] Divenyi P. [Ed.], *Speech separation by humans and machines*, Springer, Boston, pp. 181–197.

32. Wang D., Brown G.J. (2006), *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley-IEEE Press, Hoboken, NJ.

33. Wang Y., Narayanan A., Wang D. (2014), *On training targets for supervised speech separation*, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), **22**, 12, 1849–1858.

34. Wang Y., Wang D. (2013), *Towards scaling up classification-based speech separation*, IEEE Transactions on Audio, Speech, and Language Processing, **21**, 7, 1381–1390.

35. Xu Y. *et al.* (2017), *Unsupervised feature learning based on deep models for environmental audio tagging*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, **25**, 6, 1230–1241.

36. Xu Y., Du J., Dai L.-R., Lee C.-H. (2014), *An experimental study on speech enhancement based on deep neural networks*, IEEE Signal processing letters, **21**, 1, 65–68.

37. Xu Y., Du J., Dai L.-R., Lee C.-H. (2015), *A regression approach to speech enhancement based on deep neural networks*, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), **23**, 1, 7–19.

38. Zao L., Coelho R., Flandrin P. (2014), *Speech enhancement with EMD and Hurst-based mode selection*, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), **22**, 5, 899–911.