A Perceptually-Weighted Deep Neural Network for Monaural Speech Enhancement in Various Background Noise Conditions

Qingju Liu, Wenwu Wang, Philip JB Jackson
Centre for Vision, Speech and Signal Processing
University of Surrey, UK

Yan Tang
Acoustics Research Centre
University of Salford, UK

***Abstract*—Deep neural networks (DNN) have recently been shown to give state-of-the-art performance in monaural speech enhancement. However in the DNN training process, the perceptual difference between different components of the DNN output is not fully exploited, where equal importance is often assumed. To address this limitation, we have proposed a new perceptually-weighted objective function within a feedforward DNN framework, aiming to minimize the perceptual difference between the enhanced speech and the target speech. A perceptual weight is integrated into the proposed objective function, and has been tested on two types of output features: spectra and ideal ratio masks. Objective evaluations for both speech quality and speech intelligibility have been performed. Integration of our perceptual weight shows consistent improvement on several noise levels and a variety of different noise types.**

I. Introduction

Recent advances in the speech processing field have witnessed the deep neural network (DNN) [1] as a versatile and effective tool in many applications, e.g. speech recognition [2] and speech synthesis [3]. More recently, the DNN has been applied to speech separation [4]–[7] and enhancement/denoising [8]–[10], particularly for monaural recordings [4]–[6], [8]–[10]. When processing mixtures of target speech signals and competing noise, speech separation may be considered as speech enhancement.

In order to recover the underlying target speech embedded in noise, most of the deep neural networks, either recurrent [4], [5], [10] or feedforward [4], [6], [8], [9], [11], are trained to optimize some objective functions such as the mean squared error (MSE) between the true and predicted outputs. The inputs to the DNN are often (hybrid) features such as time-frequency (TF) domain spectral features [4]–[6], [8]–[10] and filterbank features [4], [5], [11]; while the output can be the TF unit level features that can be used to recover the speech source, such as ideal binary/ratio masks (IBM/IRM) [4]–[6], [11], direct magnitude spectra [9], [10] or their transforms such as log power (LP) spectra [8].

However, existing methods employing the prevailing energy minimization scheme have an essential limitation, that the perceptual importance of each predicted component is not considered, where each output unit often bears the same importance in the DNN learning. Take the IBM output for example, suppose there are two TF units both dominated by the target speech: one is perceptually audible with high energy and the other one is inaudible with very low energy. The listener's perception on the target speech may not be considerably affected even if the inaudible unit is further suppressed. However, if both units are mis-classified as noise-dominated, the DNN optimization in the back-propagation process will then use a gradient, where contributions from both units are equally weighted. As representative components possessing high energy are often more important to the listener's perception than those with lower energy [12]–[14], the current back-propagation process does not correctly reflect the psychoacoustic findings on human listeners. To address this issue, we attempt to integrate a novel perceptually-weighted objective function into a feedforward regression DNN model. The proposed weighting method takes both the groundtruth and estimated speech signals into account. It intends to maintain the perceptually important TF units in the groundtruth, while suppressing existing distortions in the estimated signal.

The remainder of the paper is organized as follows. Section II introduces the overall proposed scheme, followed by experimental results and analysis in Section III. Conclusions and insights for future work are given in Section IV.

II. The Proposed Method

Considering an additive model which assumes the microphone picks up the signals from both the target speech and the noise sources:

$$\begin{cases} Z(t,f) & = S(t,f) + N(t,f), \\ \mathbf{z}(t) & = \mathbf{s}(t) + \mathbf{n}(t), \end{cases} \tag{1}$$

where Z, S and N are respectively the spectra of the mixture, the target and the noise after applying short time Fourier transform (STFT) to the time-domain signals, indexed by the TF location (t, f); $\mathbf{z}$, $\mathbf{s}$ and $\mathbf{n}$ are the spectra vectors at each time frame. In order to recover the target speech, a five-layer ($L = 5$, with three hidden layers) feedforward regression DNN is utilized in our proposed system. The dimension, i.e. the number of neurons, at the l-th layer is denoted as D_l. The regression model has shown good performance in speech enhancement [8], using LP features $Z^{\text{LP}}(t,f) = \log(|Z(t,f)|^2)$ where $|\cdot|$ is the modulus operator. At each time frame t, the LP vectors associated with the mixture and target speech are

ISBN 978-0-9928626-7-1

denoted as $\mathbf{z}^{\text{LP}}(t)$ and $\mathbf{s}^{\text{LP}}(t)$ respectively, both $\in \mathbb{R}^{N_{\text{fft}}/2+1}$, with N_{fft} being the FFT size.

At the input layer, we concatenate LP spectra in the $2M+1$ neighboring frames as the input feature vector $\mathbf{x}(t) = [\mathbf{z}^{\text{LP}}(t-M)^T, .., \mathbf{z}^{\text{LP}}(t+M)^T]^T$ where the superscript T denotes transpose, such that the strong temporal correlation in speech signals can be exploited [8]. At the hidden layers, rectified linear units (ReLU) are employed, due to its simplicity in gradient calculation and quick convergence in the training process [15]. At the output layer, we have considered two types of features as the output vector $\mathbf{y}(t)$: the LP spectra $\mathbf{s}^{\text{LP}}(t)$ as well as IRM of the target speech $\mathbf{m}(t) = \frac{\mathbf{s}^2(t)}{\mathbf{s}^2(t)+\mathbf{n}^2(t)}$. Linear units are used in the output layer if the target is the LP spectra vector $\mathbf{y}(t) = \mathbf{s}^{\text{LP}}(t)$, and sigmoid units are employed instead if the output is the IRM $\mathbf{y}(t) = \mathbf{m}(t)$, such that the mask value is confined in the range of $(0, 1)$.

Hereafter, we omit the time index in these feature vectors and denote the i-th element in $\mathbf{y}$ and $\hat{\mathbf{y}}$ as y_i and $\hat{y}_i$ respectively. This notation also generalizes to other feature vectors.

Between each two neighboring layers l and $l+1, l = 1, ..., L-1$, there exist a transition matrix $\mathbf{W}^{(l)} \in \mathbb{R}^{D_l \times D_{l+1}}$ and a bias vector $\mathbf{b}^{(l)} \in \mathbb{R}^{D_{l+1}}$. The parameter set $\Theta = \{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}\}_l, l = 1, ..., L-1$, together with these neurons, compose the mapping process $\mathbf{f}(\mathbf{x}) := \mathbf{x} \in \mathbb{R}^{D_1} \rightarrow \hat{\mathbf{y}} \in \mathbb{R}^{D_L}$. In order to gain good speech quality, the DNN training process needs to find the optimal parameter set such that the predicted signal $\hat{\mathbf{y}} = \mathbf{f}(\mathbf{x})$ is as close to $\mathbf{y}$ as possible. As introduced earlier, minimization of the MSE term $\frac{1}{D_L}\sum_{i=1}^{D_L}(\hat{y}_i - y_i)^2$ is very popular in conventional methods in which the distortions caused by each element contribute equally to the DNN convergence. However, the relationship between the predicted speech quality and the MSE term is not a simple linear mapping, thus a lower MSE does not necessarily lead to a better quality. Speech quality is not clearly defined, and many factors can affect speech quality [16]. Yet, from the psychoacoustic point of view, auditory perceptual models have been utilized in several audio quality evaluation metrics [12]–[14], where signal components with high energy often play more important roles than low-energy components. These perceptual evaluation metrics spark us to investigate and incorporate the energy-dependent perceptual importance into the DNN learning. Therefore, we present a new objective function:

$$\mathbf{f}_{\text{opt}} = \underset{\mathbf{f}(\cdot)}{\operatorname{argmin}} \left(\begin{array}{c} \frac{1}{QD_L}\sum_{t=1}^{Q}\sum_{i=1}^{D_L} w_{\text{ei}}(\hat{s}_i^{\text{LP}}, s_i^{\text{LP}})(\hat{s}_i^{\text{LP}} - s_i^{\text{LP}})^2 \\ + \\ \frac{1}{L-1}\sum_{l=1}^{L-1}\frac{\lambda_1\|\mathbf{W}^{(l)}\|_1 + \lambda_2\|\mathbf{W}^{(l)}\|_2}{D_L D_{L+1}} \end{array} \right), \tag{2}$$

where Q is the total time frame number, $\hat{\mathbf{y}} = \mathbf{f}(\mathbf{x})$ and

$$\hat{s}_i^{\text{LP}} = \begin{cases} \hat{y}_i, & \text{if } \mathbf{y} = \mathbf{s}^{\text{LP}}, \\ 2\log(\hat{y}_i) + z_i^{\text{LP}}, & \text{if } \mathbf{y} = \mathbf{m}. \end{cases} \tag{3}$$

The top row in Eq. (2) is the perceptually-weighted squared error. The bottom row contains the penalty terms to mitigate the overfitting problem, where $\|\cdot\|_1$ and $\|\cdot\|_2$ are l_1 and l_2 norm for sparsity and energy regularization terms, respectively.

In some existing perceptual evaluation metrics, high-energy components play more important roles. For instance, in [12], units with high-energy distortion will greatly affect the perception; in [13], the quality measure is weighted by the internal representation energy of the degraded signal. Motivated by this mechanism, we proposed a novel perceptual weight $w_{\text{ei}}(\hat{s}_i^{\text{LP}}, s_i^{\text{LP}})$ model that gives more priority to high-energy components and distortions, based on the LP of both the original target speech and the predicted one:

$$w_{\text{ei}}(\hat{s}_i^{\text{LP}}, s_i^{\text{LP}}) = g(s_i^{\text{LP}}) + (1 - g(s_i^{\text{LP}}))g(\hat{s}_i^{\text{LP}}) \tag{4}$$

where $g(\cdot)$ is a sigmoid function with a translated and scaled argument

$$g(s) = \frac{1}{1 + \exp(-(s-\mu)/\sigma)}, \tag{5}$$

which aims to approximate the perceptual importance of a TF unit. Note that, since a defined mathematical formulation between the perceptual importance and signal energy does not exist, the perceptual importance function $g(\cdot)$ used here is an empirical balance between boosting high energy components and suppressing low energy components.

Our proposed weight model $w_{\text{ei}}(\hat{s}_i^{\text{LP}}, s_i^{\text{LP}})$ contains two parts, which are a trade-off between the following two extreme scenarios:

- When the target speech signal s_i^{LP} is perceptually important, i.e. $g(s_i^{\text{LP}}) \rightarrow 1$, we have $w_{\text{ei}}(\hat{s}_i^{\text{LP}}, s_i^{\text{LP}}) \approx g(s_i^{\text{LP}}) \rightarrow 1$, and any distortion between $\hat{s}_i^{\text{LP}}$ and s_i^{LP} will be taken into account in the DNN learning.
- When the target speech signal s_i^{LP} is perceptually unimportant, i.e. $g(s_i^{\text{LP}}) \rightarrow 0$, we have $w_{\text{ei}}(\hat{s}_i^{\text{LP}}, s_i^{\text{LP}}) \approx g(\hat{s}_i^{\text{LP}})$. If the predicted unit $g(\hat{s}_i^{\text{LP}}) \rightarrow 0$, which means the distortion is treated as if it does not affect perception, then the overall weight is suppressed. Otherwise, the distortion has caused perceptual change in the predicted data that we want to avoid, thus the overall weight is maintained.

III. Experiments

Here we evaluate our proposed algorithm on large-scale data, and analyze the experimental results. Considering either LP spectra or IRM as the DNN output, we have two baselines (denoted as "DNN-LP" and "DNN-IRM") and two algorithms with the proposed perceptual weight (denoted as "DNN-LP-w" and "DNN-IRM-w").

A. Data and setup

The Harvard sentences [17] uttered by 8 speakers (5 male, 3 female) were used to generate the noise-corrupted mixtures with a sampling rate of 16 kHz. A total of 1598 sentences were prepared, of which 80% were used for training and 20% for testing, respectively. As for the noise, we used the 100 Nonspeech Sounds [18], which were downsampled to 16 kHz. We considered three different signal to noise ratios

ISBN 978-0-9928626-7-1 © EURASIP 2017

(SNR), $[0, 5, 10]$ dB. At each SNR level, we randomly chose 10 different noise sequences for each training speech sentence to generate 10 additive mixtures. In total, approximately 40-hour training and 10-hour testing data were generated.

To extract training features, each mixture was normalized thus that its maximum magnitude was 1, while the associated target and noise were equally scaled with the same value. Then 512-point STFT ($N_{\text{fft}} = 512$) with half-overlapped Hamming window was applied. At each frequency bin, the LP features were further normalized with mean and variance calculated from all the mixtures. We chose $M = 5$ such that in total 11 frames covering around 200 ms were used to extract the input feature vector. The training data contained pairs of vectors $(\mathbf{x} \in \mathbb{R}^{2827}, \mathbf{y} \in \mathbb{R}^{257})$, such that the input and output layer dimensions were $D_1 = 2827$ and $D_5 = 257$ respectively. We set the three hidden layers with 3000 neurons for each layer. Of the training dataset, 80% was used for training (32 hours) and the remaining 20% for validation. In the objective function, we set $\lambda_1 = 100$ and $\lambda_2 = 1000$ for the regularization terms, and $\mu = -7$ and $\sigma = 0.5$ for the perceptual weight (Eq. (5)). Fig. 1 illustrates the proposed perceptual weight over the LP spectra associated with the dry speech signals. The cumulated weights, i.e., integral of the product of distribution and the weight, for band $[0\ 1]$ kHz, band $[7\ 8]$ kHz, and overall band are respectively 82%, 41% and 59%.

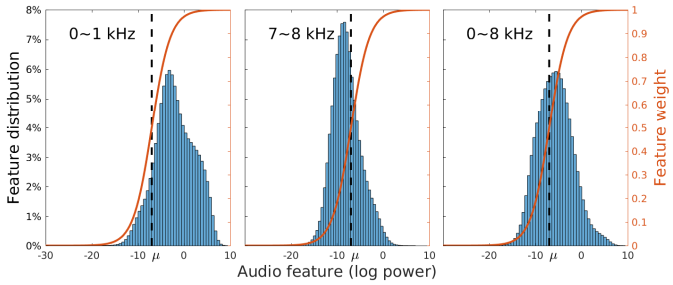

Fig. 1: Distributions of the LP features (left axis) from normalized clean speech signals and the proposed perceptual weight (right axis), at different bands. In the low-frequency band $[0\ 1]$ kHz, most features have high energy, and only 12% data has reduced weights lower than 0.5. However, the high frequency band $[7\ 8]$ kHz is dominated by low-energy features and thus 67% data has greatly suppressed weights. Overall, 38% data has suppressed weights lower than 0.5.

In the backpropagation of DNN training, we chose to use root mean square propagation optimization (RMSProp) [19], for its learning rate adaptation. The dropout was set to 0.5. Mini-batches spanning 4096 frames lasting about 1 minute were used for each update, and 50 mini-batches lasting about 1 hour were used for each iteration. The order of the training data were randomized after each epoch.

B. Results and analysis

We first show the convergence rate in Table I using the loss ignoring the regularization terms, i.e. the top-row value in Eq. (4). The losses were calculated on the validation dataset and were normalized such that losses at the 0-th iteration are with unit values.

TABLE I: Normalized loss over iterations, with (in gray) or without the perceptual weight.

Normalized loss	Iteration number					
	5	10	20	50	100	converged
DNN_LP_w	0.11	0.10	0.09	0.08	0.07	0.07
DNN_LP	0.13	0.12	0.11	0.11	0.10	0.09
DNN_IRM_w	0.26	0.22	0.20	0.17	0.15	0.12
DNN_IRM	0.38	0.35	0.35	0.28	0.26	0.24

From Table I we notice that LP-based DNN methods have a much faster convergence rate than IRM-based methods. Note that, the converged values for the LP-based DNN methods are smaller than these by the IRM-based DNN methods, which means LP features yield global minimum values that might introduce a higher gradient in the backpropagation, which is also proved by their quick convergence rate. However, it may not mean that LP features are better than IRM features for enhancement. Considering only LP features or only IRM features, we notice that methods with the perceptual weight converged to lower loss values, consistently for all iteration numbers. In other worlds, the converged model using a perceptual weight yields a much reduced error proportional to the initial weighted error. An example of using the above four algorithms is shown in Fig. 2. It can be observed that more details are maintained exploiting the perceptual weight (left bottom two). However using the conventional methods, some spectral components/regions with high energy are more suppressed (right bottom two).

Objective quality and intelligibility of the signals enhanced by the four approaches were evaluated respectively using signal to distortion ratio (SDR) and perceptual evaluation of speech quality (PESQ) [12], and short-time objective intelligibility (STOI) [20]. The same evaluations were performed on the input signals without processing as well, whose average results were denoted as "Input". For each SNR scenario, the average results from 3200 simulated mixtures were calculated, as shown in Fig. 3. We notice that overall IRM-based methods ("DNN_IRM_w" and "DNN_IRM") outperform LP-based methods in terms of SDR. However, "DNN_LP_w" gains the best performance in PESQ. One reason is that the above methods introduce different levels of distortion components such as artifact and interference, which have different impact on the perceptual evaluation [14]. The STOI results show a similar trend as the SDR results, with reduced difference between the two types of features. However, the two LP-based methods suffer in STOI as compared to the input signals without processing at 10 dB SNR noise scenarios. This is because the speech intelligibility is high in such low-noise conditions, and the LP-based DNN models have introduced extra artifacts that degrade speech intelligibility. Most importantly, our proposed perceptual weight shows advantages over the conventional methods over all the evaluation metrics, and consistent improvements can be observed over all SNR levels.

ISBN 978-0-9928626-7-1 © EURASIP 2017

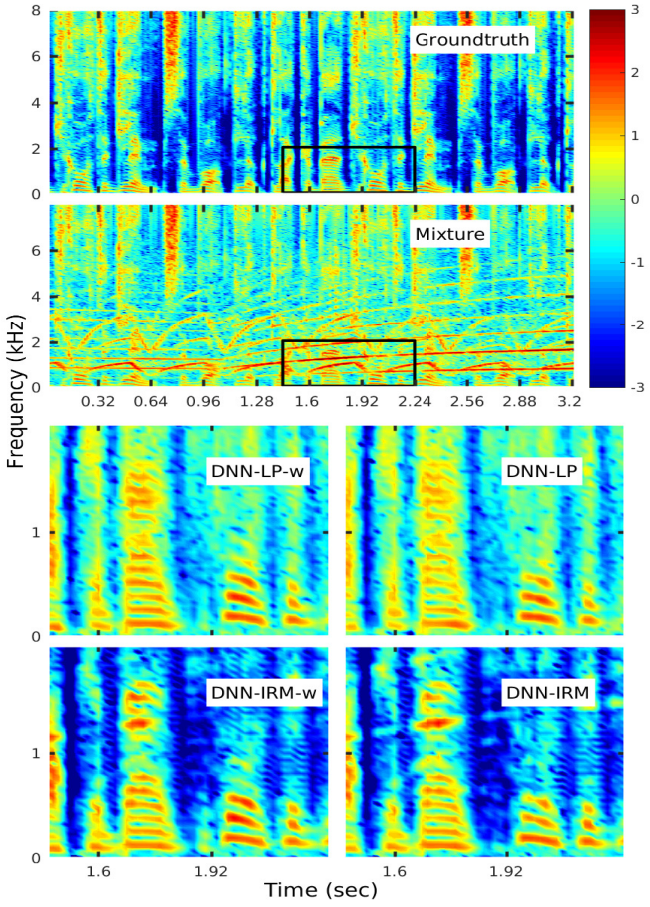

Fig. 2: An example of normalised LP spectrograms of the groundtruth speech (top) and its corrupted mixture by siren noise at 0 dB SNR (second row), and the enhanced LP by DNN models converged from the four approaches. Note that the groundtruth and enhanced LP are associated with the highlighted part in the groundtruth and mixture.

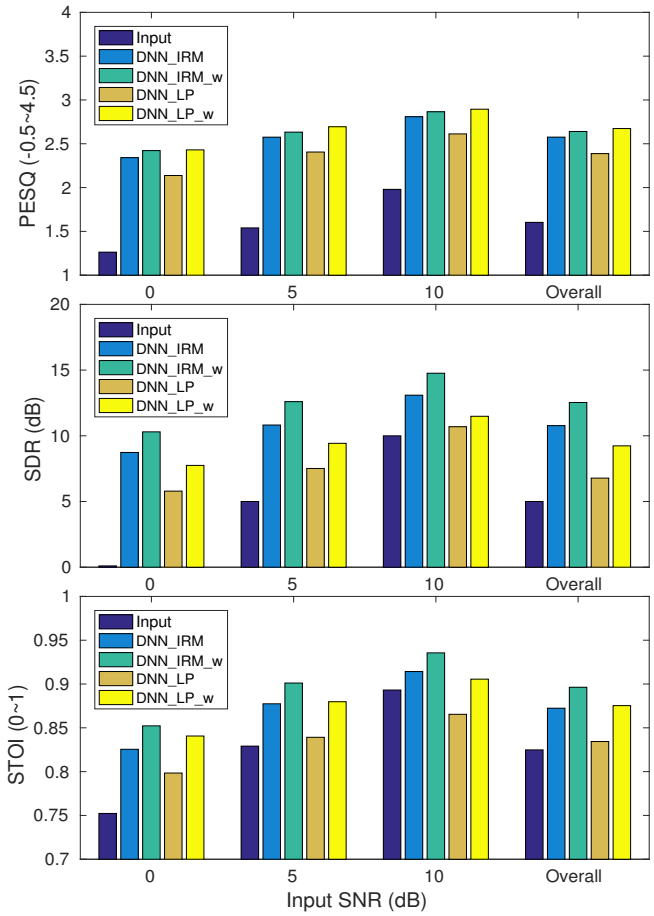

Fig. 3: Objective evaluations using PESQ, SDR and STOI at different SNR levels.

On average, 0.18 improvement in PESQ, 1.6 dB improvement in SDR and 0.03 improvement in STOI were obtained. This advantage is more significant when we consider LP as the DNN output. To test the statistical significance of our proposed scheme, we also ran a t-test as follows. For each of the three evaluation metrics (PESQ, SDR, STOI) and the two types of DNN output (LP, IRM), we performed the paired-sample t-test to 9600 (3200 samples, 3 SNR levels) pairs of evaluation results. For each of the above conditions, the p-value$< 10^{-100}$ was obtained. A p-value less than a threshold (e.g. 0.05) rejects the null hypothesis that, there is no performance difference with or without the perceptual weight. As a result, statistically significant results are justified and thus prove the effectiveness of our proposed weight model.

IV. Conclusions

We have proposed a new perceptual weight model that can be integrated into DNN frameworks for enhancing speech from monaural recordings. The perceptual weight model takes psychoacoustic characteristics into account. Having tested on a feedforward regression DNN, the proposed approach showed consistent improvement in both objective speech quality and intelligibility for the enhanced signals, as compared to the conventional methods with uniform weights. In the future, we plan to investigate the speech perception mechanisms at feature levels to further improve the objective function. Also, the optimal choice of the shift and scale parameters in the proposed weight model may be made more flexible. For instance, frequency-dependent parameters could be employed. In addition, we will consider generalization to other DNN structures such as recurrent neural networks.

Acknowledgment

The authors of the paper would like to acknowledge the support of the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership. We also thank Dr Yong Xu for helpful discussions concerning the implementation of the DNN framework used in this work.

ISBN 978-0-9928626-7-1 © EURASIP 2017

REFERENCES

[1] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Parallel distributed processing, vol. 1: Explorations in the microstructure of cognition: Foundations," chapter Learning Internal Representations by Error Propagation, pp. 318–362. MIT Press, Cambridge, MA, USA, 1986.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.

[3] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

[4] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1562–1566.

[5] 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*, vol. 23, no. 12, pp. 2136–2147, December 2015.

[6] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 3734–3738.

[7] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, September 2016.

[8] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, January 2014.

[9] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Interspeech*, 2014.

[10] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation*, 2015, pp. 91–99.

[11] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, July 2013.

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

[13] R. Huber and B. Kollmeier, "PEMO-Q –a new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, November 2006.

[14] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, September 2011.

[15] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2011, vol. 15, pp. 315–323.

[16] B. Barsties and M. De Bodt, "Assessment of voice quality: Current state-of-the-art," *Auris Nasus Larynx*, vol. 42, no. 3, pp. 183–188, 2015.

[17] E. H. Rothauser, W. D. Chapman, N. Guttman, H. R. Silbiger, M. H. L. Hecker, G. E. Urbanek, K. S. Nordby, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[18] G. Hu, "100 Nonspeech Sounds," online, retrieved in September 2016, http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html.

[19] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude.," COURSERA: Neural Networks for Machine Learning, 2012.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 4214–4217.

ISBN 978-0-9928626-7-1 © EURASIP 2017