This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2018.2871419, IEEE Signal Processing Letters

A Deep Learning Loss Function based on the Perceptual Evaluation of the Speech Quality

Juan M. Martín-Doñas, Angel M. Gomez, Jose A. Gonzalez, and Antonio M. Peinado, *Senior Member, IEEE*

***Abstract*—This paper proposes a perceptual metric for speech quality evaluation which is suitable, as a loss function, for training deep learning methods. This metric, derived from the perceptual evaluation of the speech quality (PESQ) algorithm, is computed in a per-frame basis and from the power spectra of the reference and processed speech signal. Thus, two disturbance terms, which account for distortion once auditory masking and threshold effects are factored in, amend the mean square error (MSE) loss function by introducing perceptual criteria based on human psychoacoustics. The proposed loss function is evaluated for noisy speech enhancement with deep neural networks. Experimental results show that our metric achieves significant gains in speech quality (evaluated using an objective metric and a listening test) when compared to using MSE or other perceptual-based loss functions from the literature.**

***Index Terms*—Deep Learning, Loss function, Speech Enhancement, PESQ, DNN.**

I. Introduction

SPEECH enhancement methods based on deep neural networks (DNNs) have recently attracted much research interest as they have shown very promising improvements over classical enhancement algorithms [1]–[3]. These methods aim to improve the quality and/or intelligibility of noisy speech signals. However, despite these signals are usually meant to be listened by humans, modern deep neural architectures are frequently trained by using the mean squared error (MSE) between clean and enhanced (log) power spectra [4]–[7] as optimization criterion. That is, none or very weak perceptual considerations are imposed during the training stage.

In the last years, several researchers have investigated the introduction of psychoacoustic criteria based on the human auditory perception during DNN training. Thus, Shivakumar *et al.* [8] introduced a constant penalty against signal removal in the loss function while Han *et al.* [9] proposed a joint DNN training and audible noise suppression framework. Applying a frequency-dependent weighting during the MSE loss computation is also a common strategy followed by other works, [10]–[14], which try to account for some perceptual features as the absolute threshold of hearing, the auditory masking or the perceptual relevance of each frequency band. Similarly, Chai *et al.* [15] proposed a maximum likelihood approach that models the errors between spectral features as Gaussian random variables, which finally yields a weighted MSE loss function.

Alternatively, Generative Adversarial Networks (GANs) [16] have recently been transposed from image processing to speech enhancement, achieving promising results [17], [18]. Nevertheless, instead of resorting to the training of a discriminative DNN for speech quality assessment, a more direct approach is incorporating already available and well-established objective speech quality metrics as criteria to the training loss function. That is the case of the short-time objective intelligibility (STOI) metric [19] which, once adapted and in combination with MSE, significantly improves the intelligibility of the enhanced speech [14], [20]–[22].

In this paper we adapt the perceptual evaluation of speech quality (PESQ) algorithm [23], which is one of the best known objective metrics for speech quality evaluation, as a loss function for DNN-based methods. To the best of our knowledge, a loss function based on this metric has not yet been proposed. To this end, the loudness-based disturbance terms described in the PESQ standard [23] are simplified and adapted for gradient-based training (i.e. differentiable), and computed in a per-frame basis from the power spectra of the target and enhanced signals. We show that our proposal achieves significant improvements in terms of objective perceptual quality when applied to a speech enhancement task, while subjective tests conducted also confirm that better perceived speech quality is obtained.

The rest of this paper is organized as follows. Section II describes the proposed perceptual function. The experimental framework used for evaluating our proposal as well as the results obtained are reported in Section III. Finally, conclusions are summarized in Section IV.

II. Perceptually-motivated speech-quality loss

Let us consider the MSE loss function commonly used for DNN training. In the log-power spectra (LPS) domain, after mean and variance normalization are considered, this metric can be expressed as,

$$\begin{aligned}\mathrm{MSE}_t &= \frac{1}{F}\sum_f \left(\frac{\log|X_{t,f}|^2-\mu_f}{\sigma_f} - \frac{\log|\widehat{X}_{t,f}|^2-\mu_f}{\sigma_f}\right)^2 \\ &= \frac{1}{F}\sum_f \frac{1}{\sigma_f^2}\left(\log\frac{|X_{t,f}|^2}{|\widehat{X}_{t,f}|^2}\right)^2,\end{aligned} \tag{1}$$

where $|X_{t,f}|^2$ and $|\widehat{X}_{t,f}|^2$ are, respectively, the target and enhanced power spectra obtained through the short-time Fourier

This work has been supported by the Spanish MINECO/FEDER Project TEC2016-80141-P and the Spanish Ministry of Education through the National Program FPU (grant reference FPU15/04161). We also acknowledge the support of NVIDIA Corporation with the donation of a Titan X GPU.

Juan M. Martín-Doñas, Angel M. Gomez and Antonio M. Peinado are with the Department of Signal Theory, Telematics and Communications, Universidad de Granada, Granada, Spain (e-mail: {mdjuamart,amgg,amp}@ugr.es).

Jose A. Gonzalez is with the Department of Languages and Computer Sciences, Universidad de Malaga, Malaga, Spain (e-mail: j.gonzalez@uma.es).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2018.2871419, IEEE Signal Processing Letters

transform (STFT), μ_f is the mean log-power spectrum and σ_f is its standard deviation. Indices t and f indicate, respectively, frame and frequency, while F is the number of frequency bins. As can be observed, this loss function essentially averages a weighted squared log ratio between the target and enhanced power spectra across frequency bands. This way, relevant perceptual considerations, as loudness difference, masking and threshold effects are completely neglected when optimizing the DNN parameters.

To take the perceptual features mentioned above into account, we modify the MSE loss by incorporating two disturbance terms inspired by the PESQ algorithm: a symmetrical and an asymmetrical disturbance, both computed in a frame-by-frame basis. The symmetrical disturbance, $D_t^{(s)}$, considers the absolute difference between the enhanced and clean loudness spectra when auditory masking effects are accounted for. On the other hand, the asymmetrical disturbance, $D_t^{(a)}$ is computed from the symmetrical disturbance but weighting the positive and negative loudness differences differently. This is because negative differences (omitted or attenuated spectral components) are perceived differently than positive ones (additive noise) owing to masking effects.

These terms are intended to meliorate the MSE loss function because, despite all the simplifications and adaptations we propose, the resulting PESQ criterion includes highly non-linear and non-fully differentiable (due to singular points) operators which can lead to gradient misguidance if applied alone. Thus, the final loss function is defined as,

$$J = \frac{1}{T}\sum_t \left(\text{MSE}_t + \alpha D_t^{(s)} + \beta D_t^{(a)}\right), \tag{2}$$

where α and β are weighting factors experimentally determined and T is the number of frames in the training batch. The previous equation can be seen as a multiobjective optimization function where not only the MSE error must be minimized but also, at the same time, two PESQ-based disturbance terms which introduce a perceptual criterion. In the next subsections we describe the procedure followed to compute these terms.

A. Perceptual domain transformation

The symmetrical and asymmetrical disturbances are computed in the loudness spectrum domain, which is perceptually closer to the human listening. In our proposal, the power spectra are vector transformed into a Bark frequency scale by means of a (pre-computed) Bark transformation matrix, $\mathbf{H}$, as follows,

$$\mathbf{b}_t = \mathbf{H}\cdot\mathbf{x}_t, \tag{3}$$

where $\mathbf{b}_t = \left[B_{t,0},\ldots,B_{t,Q-1}\right]^\top$ is the Bark spectrum with Q Bark bands and $\mathbf{x}_t = \left[|X_{t,0}|^2,\ldots,|X_{t,F-1}|^2\right]^\top$. Then, the Zwicker's law [24] is applied to transform each band of the Bark spectrum to a sone loudness scale as,

$$S_{t,q} = s_l \cdot \left(\frac{P_0(q)}{0.5}\right)^\gamma \cdot \left[\left(0.5 + 0.5\frac{B_{t,q}}{P_0(q)}\right)^\gamma - 1\right], \tag{4}$$

where, as in [23], s_l is a loudness scaling factor, $P_0(q)$ is the absolute hearing threshold for the q-th Bark band and γ is set to 0.23 (i.e. normal hearing). Finally, those bands with a loudness value lower than the absolute hearing threshold $P_0(q)$, are set to 0, as these cannot be perceived by humans.

These vector transformations are applied to both the target and enhanced power spectra, $\mathbf{x}_t$ and $\hat{\mathbf{x}}_t$, obtaining a target and enhanced loudness spectra $\mathbf{s}_t = \left[S_{t,0},\ldots,S_{t,Q-1}\right]^\top$ and $\hat{\mathbf{s}}_t = \left[\widehat{S}_{t,0},\ldots,\widehat{S}_{t,Q-1}\right]^\top$, respectively.

B. Symmetrical and asymmetrical disturbances computation

Here we simplify the computation of the symmetrical disturbance vector proposed in PESQ by applying a center-clipping operator over the absolute difference between the loudness spectra as,

$$\mathbf{d}_t^{(s)} = \max\left(|\hat{\mathbf{s}}_t - \mathbf{s}_t| - \mathbf{m}_t, \mathbf{0}\right), \tag{5}$$

with a clipping factor

$$\mathbf{m}_t = 0.25\cdot\min\left(\hat{\mathbf{s}}_t, \mathbf{s}_t\right), \tag{6}$$

where $|\cdot|$, $\min(\cdot)$ and $\max(\cdot)$ are applied element-wise and $\mathbf{0}$ is a zero-filled vector of length Q (note that, although non-derivable at singular points, previous operators allow to compute a sub-gradient for backpropagation at these points). This way, the psychoacoustic process by which small spectra differences are inaudible when loud signals are present is accounted for [23].

We obtain the asymmetrical disturbance vector as $\mathbf{d}_t^{(a)} = \mathbf{d}_t^{(s)} \odot \mathbf{r}_t$, where $\odot$ indicates an element-wise multiplication and $\mathbf{r}_t$ is a vector of asymmetry ratios whose components are computed from the Bark spectra as,

$$R_{t,q} = \left(\frac{\widehat{B}_{t,q} + \epsilon}{B_{t,q} + \epsilon}\right)^\lambda. \tag{7}$$

The asymmetry ratio accounts for positive ($\widehat{B}_{t,q} > B_{t,q}$) and negative ($\widehat{B}_{t,q} < B_{t,q}$) differences between the enhanced and the target spectra in the Bark domain by correspondingly applying a gain or an attenuation to the symmetrical disturbance. The constants ϵ and λ, set to 50 and 1.2 respectively (see next subsection), stabilize the ratio against very small Bark spectrum values and magnify the effect of the resulting ratio, respectively. Prior to the element-wise multiplication, asymmetry ratios $R_{t,q}$ are upper-bounded by a maximum value of 12, while those lower than 3 are set to 0, as in [23].

Finally, we can obtain the symmetrical and asymmetrical disturbance terms in a vectorized way, for each frame, as weighted norms of the symmetrical and asymmetrical vectors, respectively, as

$$D_t^{(s)} = ||\mathbf{w}||_1^{\frac{1}{2}} \cdot ||\mathbf{w}\odot\mathbf{d}_t^{(s)}||_2 \tag{8}$$

$$D_t^{(a)} = ||\mathbf{w}\odot\mathbf{d}_t^{(a)}||_1 = \mathbf{w}^\top\cdot\mathbf{d}_t^{(a)}, \tag{9}$$

where $\mathbf{w}$ is a vector filled with weights proportional to the width of the Bark bands, obtained from [23] (note that elements of $\mathbf{d}_t^{(a)}$ and $\mathbf{d}_t^{(s)}$ are always positive). As in the PESQ algorithm, a post-processing step is applied to the previous disturbance values, which are scaled by the frame audible power and upper-bounded [23].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2018.2871419, IEEE Signal Processing Letters

C. Spectral pre-processing and equalization

In order to reuse the already established perceptual constants and values described in the PESQ standard [23] (i.e. Bark transformation coefficients, s_l, $P_0(q)$, ϵ, λ and $\mathbf{w}$ values), a pre-processing step must be applied over $\mathbf{x}_t$ and $\hat{\mathbf{x}}_t$ in order to obtain PESQ-equivalent spectra. Thus, in the PESQ algorithm, prior to the Bark transformation, the level of both signals is equalized to a standard listening level. Gains are computed based on the estimated RMS values of the band-pass filtered (from 350 to 3250 Hz) speech signals. In our proposal, as time-domain signals are unavailable, this gain normalization is instead accomplished in the spectral domain as,

$$\bar{\mathbf{x}}_t = \mathbf{x}_t \cdot \frac{P_c}{\frac{1}{T}\sum_{\mathrm{t}} \left(\mathbf{g}^{\top} \cdot \mathbf{x}_{\mathrm{t}}\right)}, \tag{10}$$

where $\mathbf{g}$ is a spectral weighting mask which replicates the band-pass filtering and P_c is a power correction factor which accounts for the frame length, overlapping and windowing applied during the spectral computation (via STFT).

Additionally, the PESQ algorithm implements, in the Bark spectrum domain, a frequency equalization to compensate non-severe constant filtering effects, and a gain equalization to correct short-term gain variations, as listeners do not perceive these effects as degrading the quality. During frequency equalization, a per-band factor is computed as the ratio of the degraded Bark spectrum to the original Bark spectrum and applied over the latter one (limited to the range of $[-20, 20]$ dB) [23]. On the other hand, the gain equalization is applied over the degraded Bark spectra by computing a gain factor between the average power of the original and degraded spectra across the audible bands only (i.e. $S_{t,q} > P_0(q)$). This gain factor is bounded to the range of $[3{\cdot}10^{-4}, 5]$ and smoothed over time [23]. In our proposal, the frequency equalization is applied over the degraded (estimated) spectrum, $\hat{\mathbf{s}}_t$, instead of the reference (target) spectrum, $\mathbf{s}_t$, to prevent modifications over reference spectra which can deceive the loss function during training. Moreover, the gain factor is only bounded but not smoothed as our proposed disturbance terms are intended to be computed per frame.

III. Experimental Evaluation

We evaluated the performance of our proposed perceptual loss function in a DNN-based speech enhancement task. To this end, we used the VCTK corpus [25] downsampled to 8 kHz (as in narrowband PESQ). This database contains speech material from 108 native English speakers with various accents, each one reading about 400 sentences. A total of 72 speakers were considered for the training set, 18 were chosen for DNN validation and the remaining 18 speakers were saved for testing purposes. Speech utterances from the speakers were artificially contaminated by adding noise at several signal-to-noise ratio (SNR) levels. To this end, we recorded 5-minute length noises at eight different locations and grouped them in two different datasets: set A, comprising the noises babble, car, street and mall, and set B, comprising bus, cafe, pedestrian street and bus station noises. Six different SNR levels from -5 dB to 20 dB (with 5 dB increase step) are considered in both sets. The signals in the training and validation sets were distorted using only noises included in set A, whereas in the testing set, signals were distorted using noises from set B. This way, testing is performed with noises unseen during training.

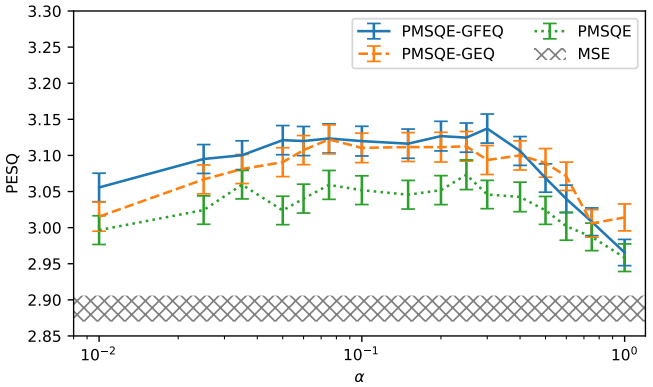

Fig. 1. Average PESQ scores with 95% confidence intervals obtained on AURORA2 test set by the proposed metric without (PMSQE) and with equalizations (PMSQE-GEQ, PMSQE-GFEQ), and several hyperparameter α values. Results from the MSE loss function (MSE) are also shown (95% confidence interval band).

Log-power spectra (LPS) feature vectors of 129 components were obtained from half power spectra after applying a log operator. Power spectra are computed via STFT with 256-sample frame length, 128-sample overlapping and Hanning windowing. For speech signal re-synthesis, the overlap-add method was applied using the enhanced magnitude spectrum and the noisy phase.

Although our proposal can be used by more complex network architectures, for simplicity sake, we chose a relatively simple feed-forward DNN regressor as in [4], [7]. Our loss function was implemented in TensorFlow [26] and it is available at [27]. Three hidden layers with 2048 rectifier linear units (ReLU) and a linear output layer of 129 units are considered. A temporal context of 4 previous and subsequent frames is applied in the input layer (i.e. 1161 components). LPS vectors are mean and variance normalized. To prevent overfitting, dropout is applied in the hidden layers with a deactivation probability factor of 0.1, as well as early stopping using the validation dataset with a patience of 20 iterations. Finally, DNN parameters are optimized by following the *ADAM* method with a learning rate of 10^{-4} [28].

A. Hyperparameter optimization

We used the AURORA2 noisy speech database [29] to optimize the weights α and β in (2) with a totally independent dataset. To reduce the search space, we first set the same relative weighting between symmetrical and asymmetrical disturbances as in the PESQ algorithm, i.e. $\beta = 0.309\alpha$. In addition, we also evaluated the effect of the two equalization steps adapted from the PESQ algorithm (see Section II-C), that is, the gain equalization and the frequency equalization when applied over the DNN output spectra.

Fig. 1 reports the average PESQ scores and 95% confidence intervals obtained on AURORA2 test set when different α values are considered. The variations of the loss function

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2018.2871419, IEEE Signal Processing Letters

TABLE I
PESQ SCORES OBTAINED FOR NOISY AND DNN ENHANCED SPEECH WITH DIFFERENT LOSS FUNCTIONS OVER THE TEST SET.

Method	**SNR (dB)**						**Avg.**
	-5	**0**	**5**	**10**	**15**	**20**	
Noisy	1.62	1.82	2.10	2.42	2.73	3.00	2.28
MSE	1.77	2.12	2.47	2.76	3.00	3.20	2.55
wMSE	1.77	2.14	2.50	2.80	3.04	3.25	2.58
wMSE-SVS	1.77	2.11	2.47	2.78	3.04	3.26	2.57
PMSQE-GFEQ	**1.89**	**2.27**	**2.62**	**2.89**	**3.13**	**3.34**	**2.69**

TABLE II
SDR VALUES (IN DB) OBTAINED FOR NOISY AND DNN ENHANCED SPEECH WITH DIFFERENT LOSS FUNCTIONS OVER THE TEST SET.

Method	**SNR (dB)**						**Avg.**
	-5	**0**	**5**	**10**	**15**	**20**	
Noisy	-	-	-	-	-	-	-
MSE	-2.62	3.03	7.17	10.15	12.03	12.91	7.11
wMSE	-2.74	2.89	7.22	10.47	12.59	13.62	7.34
wMSE-SVS	-3.06	2.80	7.46	**11.08**	**13.62**	**15.02**	**7.82**
PMSQE-GFEQ	**-1.53**	**4.14**	**7.85**	9.94	10.89	11.19	7.08

evaluated are our proposed perceptual metric for speech quality evaluation (PMSQE) without equalization, with gain equalization only (PMSQE-GEQ), and including both gain and frequency equalization (PMSQE-GFEQ). The results obtained with the bare MSE loss function are also shown as a reference. As can be observed, PMSQE-GFEQ performs the best in general, yielding a plateau in performance when α is around 0.1, so this value is selected for the rest of the evaluation.

B. Objective evaluation results

We first evaluated the performance of our proposal in terms of objective perceptual quality measured using the PESQ algorithm and the signal-to-distortion ratio (SDR) [30]. Tables I and II show the average results, for each SNR level, obtained for PESQ and SDR, respectively. As a reference, noisy speech (Noisy) scores and those from the same DNN regressor trained with the MSE loss are also reported. In addition, the results from a recently proposed perceptually oriented metric, the Mel-frequency weighted MSE loss (wMSE) and a variant including a regularization by spectral variation similarity (wMSE-SVS) [14], which can be applied in a per-frame basis in the spectral domain, are also included for comparison purposes.

As expected, our proposal achieves the best results in terms of PESQ, whereas wMSE and wMSE-SVS yield PESQ scores almost identical to MSE. On the other hand, these techniques outperform our proposal in terms of average SDR. Thus, it seems that the improvement on the perceptual quality that our technique achieves is at the cost of some speech distortion (or, at least, both cannot be improved at the same time at high SNRs). Nonetheless, average SDR reduction in comparison with MSE is small while significant improvements can be observed at low SNRs.

C. Subjective evaluation results

We also conducted a listening test to subjectively evaluate the perceived quality of the enhanced signals. We followed a *Comparative Mean Opinion Score* (CMOS) evaluation [31] in which listeners were asked to compare pairs of enhanced signals in terms of overall perceived quality using a Likert-style scale from -3 to 3 (-3: the 1st signal sounds much better than the second one, ..., 0: both signals sounds equally well, ..., 3: the 2nd signal sounds much better than the first one). The listening test was conducted in a quiet room using headphones and a web-based interface. Twenty-three listeners with normal hearing and no previous knowledge on speech processing participated in the listening test. Each listener evaluated a total of 20 randomly-chosen signal pairs from the test set (a pair for each noise and SNR condition, with the SNR range limited from 0 to 20 dB). A pair of stimuli was generated from each utterance by enhancing the noisy speech signal with the proposed PMSQE-GFEQ approach and with the wMSE-SVS metric. The signal pairs were presented to the listeners in a random order to control for order effect bias.

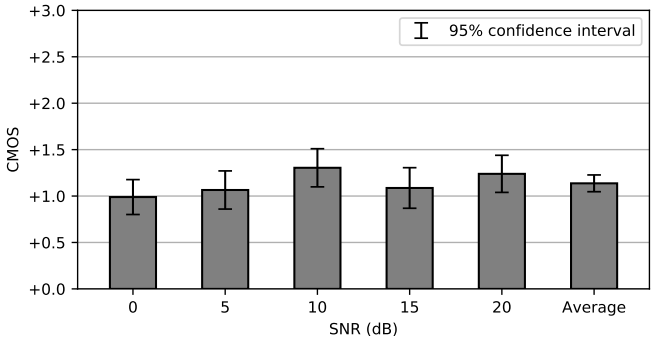

Fig. 2. CMOS test scores averaged per SNR level when comparing speech signal enhanced with PMSQE-GFEQ respect to wMSE-SVS.

Fig. 2 shows the average CMOS scores obtained by our proposal for each SNR level and in average. As can be observed, positive CMOS scores confirm that our proposal outperforms wMSE-SVS in terms of subjective quality with real listeners in each tested SNR condition.

IV. CONCLUSIONS

In this paper we proposed a novel perceptually-motivated loss function to objectively evaluate speech quality in deep learning based methods. The proposed approach improves the widely used MSE loss with two disturbance terms inspired by the well-known PESQ algorithm. In a speech enhancement task, the proposed loss function significantly improved the performance of DNN methods, yielding an absolute average increase of 0.12 points in PESQ score respect to other state-of-the-art metrics under unseen noise conditions. Subjective tests conducted with real listeners also confirmed a noticeable increase of quality (slightly above $+1.0$ CMOS score). Finally, it is worth to note that the proposed function can be used within other neural network architectures, could also be extended by including other metrics (e.g. STOI), and can be applied to other speech related tasks, as speech coding, frame loss concealment and robust speech recognition.

REFERENCES

[1] 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, 2014.

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

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

[4] 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 Speech and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[5] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement," in *Proc. Interspeech*, 2016, pp. 3768–3772.

[6] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proc. IEEE HSCMA*, 2017, pp. 136–140.

[7] J. M. Martín-Doñas, A. M. Gomez, I. López-Espejo, and A. M. Peinado, "Dual-channel DNN-based speech enhancement for smartphones," in *Proc. IEEE MMSP*, 2017, pp. 1–6.

[8] P. G. Shivakumar and P. Georgiou, "Perception optimized deep denoising autoencoders for speech enhancement," in *Proc. Interspeech*, 2016, pp. 3743–3747.

[9] W. Han, X. Zhang, G. Min, M. Sun, and J. Yang, "Joint optimization of audible noise suppression and deep neural networks for single-channel speech enhancement," in *Proc. IEEE ICME*, 2016, pp. 1–6.

[10] B. Xia and C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication*, vol. 60, pp. 13–29, 2014.

[11] W. Han, X. Zhang, G. Min, X. Zhou, and W. Zhang, "Perceptual weighting deep neural networks for single-channel speech enhancement," in *Proc. WCICA*, 2016, pp. 446–450.

[12] A. Kumar and D. Florencio, "Speech enhancement in multiple-noise conditions using deep neural networks," in *Proc. Interspeech*, 2016, pp. 3738–3742.

[13] Q. Liu, W. Wang, P. J. B. Jackson, and Y. Tang, "A perceptually-weighted deep neural network for monaural speech enhancement in various background noise conditions," in *Proc. Eusipco*, 2017, pp. 1270–1274.

[14] T. G. Kang, J. W. Shin, and N. S. Kim, "DNN-based monaural speech enhancement with temporal and spectral variations equalization," *Digital Signal Processing: A Review Journal*, vol. 74, pp. 102–110, 2018.

[15] L. Chai, J. Du, and Y. Wang, "Gaussian density guided deep neural network for single-channel speech enhancement," in *Proc. IEEE MLSP*, 2017, pp. 1–6.

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 3, 2014, pp. 2672–2680.

[17] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[18] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *Proc. Interspeech*, 2017, pp. 2008–2012.

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

[20] M. Kolbaek, Z.-H. Tan, and J. Jensen, "Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure," in *Proc. ICASSP*, 2018.

[21] S. Fu, T. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[22] Y. Zhao, B. Xu, R. Giri, and T. Zhang, "Perceptually guided speech enhancement using deep neural networks," in *Proc. ICASSP*, 2018.

[23] "ITU-T. Rec. P.862, Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assesment of narrow-band telephone networks and speech codecs," International Telecommunication Union-Telecommunication Standarisation Sector, Tech. Rep., 2001.

[24] E. Zwicker and R. Feldtkeller, *Das Ohr als Nachrichtenempfanger*. S. Hirtzel Verlag Stuttgart, 1967.

[25] J. Yamagishi. (2012) English multi-speaker corpus for CSTR voice cloning toolkit.

[26] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: http://download.tensorflow.org/paper/whitepaper2015.pdf

[27] J. M. Martín-Doñas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "Perceptual metric for speech quality evaluation (PMSQE): Source code and audio examples," 2018. [Online]. Available: http://sigmat.ugr.es/PMSQE

[28] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference for Learning Representations*, 2015, pp. 1–13.

[29] D. Pearce and H. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ICSLP*, 2000.

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

[31] "ITU-T. Rec. P.800, Methods for Subjective Determination of Transmission Quality - Series P: Telephone Transmission Quality; Methods for Objective and Subjective Assessment of Quality," International Telecommunication Union-Telecommunication Standarisation Sector, Tech. Rep., 1996.