WEIGHTED SPEECH DISTORTION LOSSES FOR
NEURAL-NETWORK-BASED REAL-TIME SPEECH ENHANCEMENT

Yangyang Xia[1], Sebastian Braun[2], Chandan K. A. Reddy[2],
Harishchandra Dubey[2], Ross Cutler[2], Ivan Tashev[2]

[1] Dept. of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA
[2] Microsoft Corporation, Redmond, WA, USA
`raymondxia@cmu.edu`, {`firstname.lastname`}`@microsoft.com`

ABSTRACT

This paper investigates several aspects of training a RNN (recurrent neural network) that impact the objective and subjective quality of enhanced speech for real-time single-channel speech enhancement. Specifically, we focus on a RNN that enhances short-time speech spectra on a single-frame-in, single-frame-out basis, a framework adopted by most classical signal processing methods. We propose two novel mean-squared-error-based learning objectives that enable separate control over the importance of speech distortion versus noise reduction. The proposed loss functions are evaluated by widely accepted objective quality and intelligibility measures and compared to other competitive online methods. In addition, we study the impact of feature normalization and varying batch sequence lengths on the objective quality of enhanced speech. Finally, we show subjective ratings for the proposed approach and a state-of-the-art real-time RNN-based method.

Index Terms— Real-time speech enhancement, recurrent neural networks, loss function, speech distortion, mean opinion score

1. INTRODUCTION

Speech enhancement (SE) algorithms aim at improving speech quality and intelligibility of speech signals degraded by additive noise [1], in order to improve human or machine interpretation of speech. Typical SE applications are hearing aids, automatic speech recognition, and audio/video communications in noisy environments. Most SE methods apply a spectral suppression gain or filter to the noisy speech signal in a time-frequency domain [2]. In recent supervised learning methods using deep neural networks (DNNs), a DNN is typically set up to estimate this time-varying gain function [3] from one or more sets of features derived from noisy speech.

Online processing capability is an attractive feature of a SE algorithm and required for real-time communication applications. Although most classical SE methods have to accommodate their approaches [4, 5, 6, 7] for fulfilling causality, many DNN-based methods in the literature [3, 8, 9] do not enforce this constraint. Several DNN-based approaches report high-quality enhancement using generous look-ahead [8, 9], but their performance for decreasing look-ahead is not well investigated. Nevertheless, DNN-based systems are preferred over classical methods for their ability to accurately suppress transient noise. In this work, we study real-time speech enhancement with recurrent neural network (RNN). Recent works involving RNNs demonstrated promising results [10], even at very low signal-to-noise ratio (SNR) scenarios [11, 12].

A key challenge in designing a SE algorithm for audio/video communication is to preserve perceived (subjective) speech quality to the best extent possible while suppressing the noise. In classical literature, optimizing such a compound global objective can be done by solving a constrained objective function [13]. Alternatively, one can optimize a simpler objective such as the (log) mean-squared error (MSE) [6, 14] and employ post-processing modules such as residual noise removal [7] and gain limiting [15]. By contrast, one major benefit of the deep learning framework is the relative ease to incorporate complex learning objectives one believes would drive the enhanced speech towards better quality and intelligibility. Methods along this line of thought include learning multiple objectives from heterogeneous features [16, 17, 18], jointly optimizing the final goal and its sub-targets (*e.g.* speech-presence probability) [10, 12], and directly optimizing towards an objective measure of speech quality or intelligibility [19, 20]. The latter seems a promising way to improve objective quality, although both models have to incorporate the standard MSE due to the band limitation of each objective measure. [21] reported that a simple perceptually weighted wide-band MSE alone does not improve objective speech quality or intelligibility, suggesting that the MSE is still a reliable learning objective for wide-band speech enhancement.

In this paper, we propose a DNN-based online speech enhancement system for real-time applications. First, we will discuss features and normalization techniques that would facilitate pattern learning with a RNN. We then describe a compact RNN that produces a gain function from a single noisy frame. Next, we introduce two simple MSE-based loss functions with separate control of speech distortion and noise reduction. During the evaluation, we thoroughly examine the effect of error weighting on the subjective and objective speech quality and intelligibility measures. Furthermore, we discuss how the objective metrics are affected by different feature normalization techniques and training strategies.

2. PROBLEM FORMULATION

We assume the microphone signals to be described in the short-time Fourier transform (STFT) domain by

$$X[t,k] = S[t,k] + N[t,k], \qquad (1)$$

where $X[t,k]$, $S[t,k]$, and $N[t,k]$ denote the STFT at time frame t and frequency bin k of the observed noisy speech, clean speech, and noise, respectively. Our system seeks for a time-varying gain function in the short-time Fourier transform magnitude (STFTM) domain, $G[t,k]$, that recovers $|S[t,k]|$ to the best extent possible.

$$|\hat{S}[t,k]| = G[t,k]|X[t,k]| \qquad (2)$$

ICASSP 2020

In real-time processing, $G[t,k]$ shall depend only on the past and present information of input, and is given by

$$G[t,k] = n(g(f(|X[l,k]|)); \Theta), l \leq t, \qquad (3)$$

where f is a transform function applied on the STFTM of the noisy signal, g is a normalization function, and n is a DNN whose adaptable parameters are together denoted by Θ. Finally, noisy phase of $X[t,k]$ is applied to $|\hat{S}[t,k]|$ to obtain the enhanced signal.

In the following sections, we will review the state-of-the-art methods, before discussing our choices for f and g, the architecture of n, two learning objectives for Θ, and further considerations in training that we believe will impact the quality of enhanced speech.

3. STATE-OF-THE-ART ONLINE NOISE REDUCTION

Classical online SE methods typically seek for the optimal gain function by optimizing some objective functions in a statistical sense. One of the most effective methods in this category assumes the clean and noise STFT are uncorrelated, complex Gaussian distributions, and solves for $G[t,k]$ by minimizing the MSE between clean and enhanced STFTM [6] or log-STFTM [14]. Although more advanced noise and speech-presence probability models could be incorporated to improve speech quality and prevent musical noise [4, 5], retaining speech quality while removing highly non-stationary noise is still a challenging task.

In recent DNN-based methods, statistical assumptions about the distribution of noisy and clean STFTM are typically dropped, while the minimum MSE (MMSE) objective becomes a loss function for which a DNN optimizes by stochastic gradient descent. One of the most popular loss functions has been the MSE between clean and enhanced STFTM

$$L(\vec{G}; \vec{S}, \vec{X}) = \mathrm{mean}(||\vec{S} - \vec{G} \odot \vec{X}||_2^2), \qquad (4)$$

where $\vec{A}$ denotes $|A[t,k]|$ in vector form, and $\odot$ is the element-wise product. A competitive method proposed recently [10] estimates the optimal gain function of a smoothed energy contour using a RNN and interpolates the spectral details by pitch filtering. Experiments [10, 22] report strong objective and subjective speech quality from enhanced speech produced by this RNNoise system.

4. PROPOSED METHODS

4.1. Feature representation

Selecting appropriate features and normalization is important for successfully training a DNN. We consider two basic features in STFTM and log-power spectra (LPS), and apply global, frequency-dependent (FD), and frequency-independent (FI) normalization, respectively, to train our network.

The STFTM used in all our systems is computed based on a 32 ms Hamming window with 75% overlap between frames and a 512-point discrete Fourier transform. The LPS is taken with the natural logarithm and floored at -120 dB, i. e.,

$$f_{LPS}(|X[t,k]|) = \log(\max(|X[t,k]|^2, 10^{-12})) \qquad (5)$$

We explore three types of normalization to be each individually combined with either STFTM or LPS mentioned above. First, we consider global normalization, in which case each frequency bin is standardized by its mean and standard deviation accumulated from a training set:

$$g_G(f(|X[t,k]|)) = \left[f(|X[t,k]|) - \mu_{f(x)}[k] \right] / \sigma_{f(x)}[k] \qquad (6)$$

Second, we consider online FD mean and variance normalization, in which case the running mean and variance are smoothed by a decaying exponential:

$$\mu_{f(x)}[t,k] = c\,\mu_{f(x)}[t-1,k] + (1-c)\,f(|X[t,k]|) \qquad (7)$$

$$\sigma_{f(x)}^2[t,k] = c\,\sigma_{f(x)}^2[t-1,k] + (1-c)\,f(|X[t,k]|)^2 \qquad (8)$$

$$g_{FD}(f(|X[t,k]|)) = \frac{f(|X[t,k]|) - \mu_{f(x)}[t,k]}{\sqrt{\sigma_{f(x)}^2[t,k] - \mu_{f(x)}^2[t,k]}} \qquad (9)$$

where $c = \exp(-\Delta t/\tau)$, Δt is the frame shift in seconds (8 milliseconds in our setting), and τ is a time constant that controls the adaptation speed. The idea is that the normalized spectra will facilitate the recurrent neural network learning long-term patterns. Finally, we also have the FI online normalization, in which case the mean and variance from each frequency are averaged and applied to all frequencies. This method retains the relative dynamics across frequency bins, but might pose a more challenging learning task to the learning machine. In all our experiments apart from the feature experiments, we use FD online normalization with $\tau = 3s$.

4.2. Learning machine

Our learning machine that takes in one frame of noisy speech spectra and outputs one frame of magnitude gain function is based on the gated recurrent unit (GRU) [23]. GRUs are preferred over long short-term memory (LSTM)[24] given their computational efficiency and superior performance in real-time SE tasks [22]. We stack three GRU layers followed by a fully-connected (FC) output layer with sigmoid activation to predict the gain function $G[t,k]$.

It is worth mentioning that we do not apply convolution layers as often done in other related work [11, 25] because of the relatively arbitrary process involved in choosing the amount of frequency span and filter taps. Previous studies [26] have shown that a naïve convolution layer applied on past and present input noisy frames did not improve objective quality of enhanced speech. Instead, we explore the temporal modeling capability of the network by training with sequences of different lengths, features, and loss functions.

4.3. Loss functions

We use three loss functions to train our system. First, we use the regular MSE between clean and enhanced STFTM in (4). To obtain a better control of the loss, we propose to separate the error into speech distortion and noise reduction terms

$$L_{\mathrm{speech}} = \mathrm{mean}(||\vec{S}_{\mathrm{SA}} - (\vec{G} \odot \vec{S})_{\mathrm{SA}}||_2^2) \qquad (10)$$

$$L_{\mathrm{noise}} = \mathrm{mean}(||\vec{G} \odot \vec{N}||_2^2) \qquad (11)$$

where subscript SA denotes a subset of frames where speech is active. In our experiments, we adopted a simple energy-based frame-level voice activity detector operating on the power spectra of clean utterances. The short-time speech energy is accumulated between 300 Hz and 5000 Hz and smoothed over 3 frames by a moving-average filter. Finally, a frame is decided to be voiced above a threshold of 30 dB below the peak energy of the whole utterance.

As the estimated gain approaches all-pass, the speech distortion error is minimized and the noise error is maximized, and vice versa. Therefore, we can control the relative importance of speech distortion to noise reduction with a *fixed-weighted* loss,

$$L(\vec{G}; \vec{S}_{\mathrm{SA}}, \vec{N}) = \alpha L_{\mathrm{speech}} + (1-\alpha)L_{\mathrm{noise}}, \qquad (12)$$
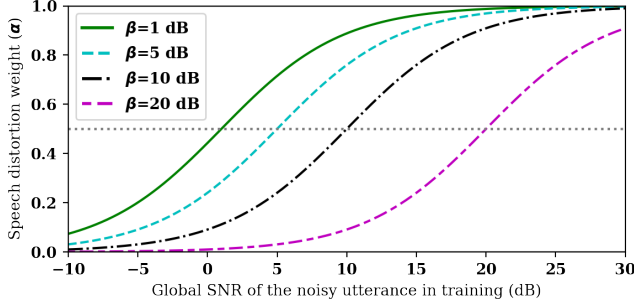
Fig. 1: Selected *SNR-weighted* speech distortion weighting. Horizontal line marks equal weighting of L_{speech} and L_{noise}.
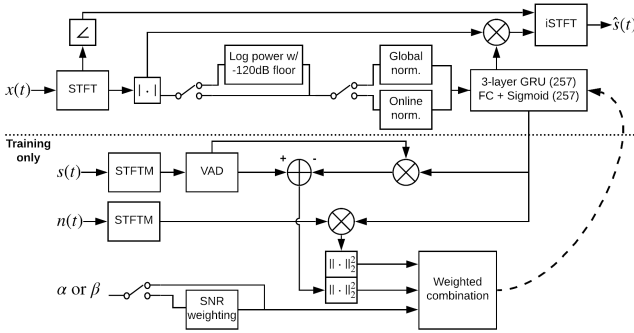
Fig. 2: Flow diagram of the proposed system.

where α is a constant in range [0, 1]. We notice a similar loss has been developed independently and termed two-component loss (2CL) [27]. Next, we discuss an extension to this fixed weighting.

In classical speech enhancement literature, the suppression rule is often adapted based on the SNR [15, 13]. Specifically, suppression should be limited at high SNR to avoid artifacts, and be aggressive at low SNR. Motivated by this principle, our second *SNR-weighted* loss adjusts α in (12) using the global SNR of each utterance

$$\alpha = \frac{\text{SNR}}{\text{SNR} + \beta}, \qquad (13)$$

where $\text{SNR} = ||\vec{S}||_2^2 / ||\vec{N}||_2^2$ and β is a constant. Note that $d\alpha/d[10\log_{10}(\text{SNR})]$ is maximized when $\text{SNR} = \beta$. In this way, β controls the global SNR at which a fixed amount of deviation would cause the maximum drift in speech distortion weighting. Furthermore, β also indicates the global SNR, where the two loss terms are equally weighted. We illustrate this in Fig. 1.

The proposed method is depicted in a flow diagram shown in Fig. 2. During training, both clean speech and noise are required for computing the weighted loss. The trained model enhances the noisy STFTM one frame at a time and utilizes the noisy phase for reconstructing the enhanced speech waveform.

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

5.1. Corpora & Experimental setup

We train and evaluate all DNN-based systems using a dataset synthesized from publicly available speech and noise corpus using the MS-SNSD dataset [22] and toolkit [1]. 14 diverse noise types are se-

---

[1] https://github.com/microsoft/MS-SNSD

Table 1: Effect of sequence lengths in a one-minute minibatch.

Length (s)	SI-SDR (dB)	CD	STOI (%)	PESQ (MOS)
1	13.7	3.78	90.1	2.58
2	13.7	3.80	90.3	2.57
5	14.1	3.72	90.5	2.59
10	14.1	3.73	90.7	2.64
20	14.0	3.73	90.6	2.64

lected for training, while samples from 9 noise types not included in the training set are used for evaluation. Our test set includes challenging and highly non-stationary noise types such as munching, multi-talker babble, keyboard typing, *etc*. All audio clips are resampled to 16 kHz. The training set consists of 84 hours each of clean speech and noise while 18 hours (5500 clips) of noisy speech constitute the evaluation set. All speech clips are level-normalized on a per-utterance basis, while each noise clip is scaled to have one of the five global SNRs from {40,30,20,10,0} dB. During the training of all DNN-based systems described below, we randomly select an excerpt of clean speech and noise, respectively, before mixing them to create the noisy utterance.

We performed a comparative study of proposed methods with three baselines based on several objective speech quality and intelligibility measures and subjective tests. Specifically, we include perceptual evaluation of speech quality (PESQ) [28], short-time objective intelligibility (STOI)[29], cepstral distance (CD), and scale-invariant signal-to-distortion ratio (SI-SDR) [30] for objective evaluation of enhanced speech in time, spectral, and cepstral domains. We conducted a subjective listening test using a web-based subjective framework presented in [22]. Each clip is rated with a discrete rating between 1 (very poor speech quality) and 5 (excellent speech quality) by 20 crowd-sourced listeners. Training and qualification are ensured before presenting test clips to these listeners. The mean of all 20 ratings is the mean opinion score (MOS) for that clip. We also removed obvious spammers who consistently selected the same rating throughout the MOS test. Our subjective test complements the other objective assessments, thus providing a balanced benchmark for evaluation of studied noise reduction algorithms.

We compare our proposed methods with three baseline methods. We used a classical enhancer, which is a slightly optimized implementation of the MMSE log-spectral amplitude (LSA) estimator [14] described in [31]. DNN-based baselines include the improved RNNoise (RNNoise$_l$) [22] and a RNN (RNNoise$_{257}$) that replicates the network architecture of RNNoise [10] but operates on 257-point spectra, is trained on (4), and does not have the originally proposed post-processing component. RNNoise$_{257}$ realizes a system with a comparable number of parameters as the proposed methods.

In the next section, we discuss the impact of feature normalization and training on various sequence lengths on the objective quality of enhanced speech. Then, we explore the optimal weighting for the proposed *fixed-weighted* and *SNR-weighted* loss functions. Finally, we compare the subjective and objective quality of enhanced speech produced by our systems to several competitive online methods.

5.2. Results & Discussions

We want to evaluate how training with long or short sequences affects temporal modeling in the RNN. Although long sequences are expected to help deal with long-term noise patterns, it might also potentially degrade speech that is only short-term stationary. Table 1 summarizes this impact of sequence lengths on objective speech quality. For each setting, we adjust the number of sequences in a minibatch so that one batch always contains one minute of noisy
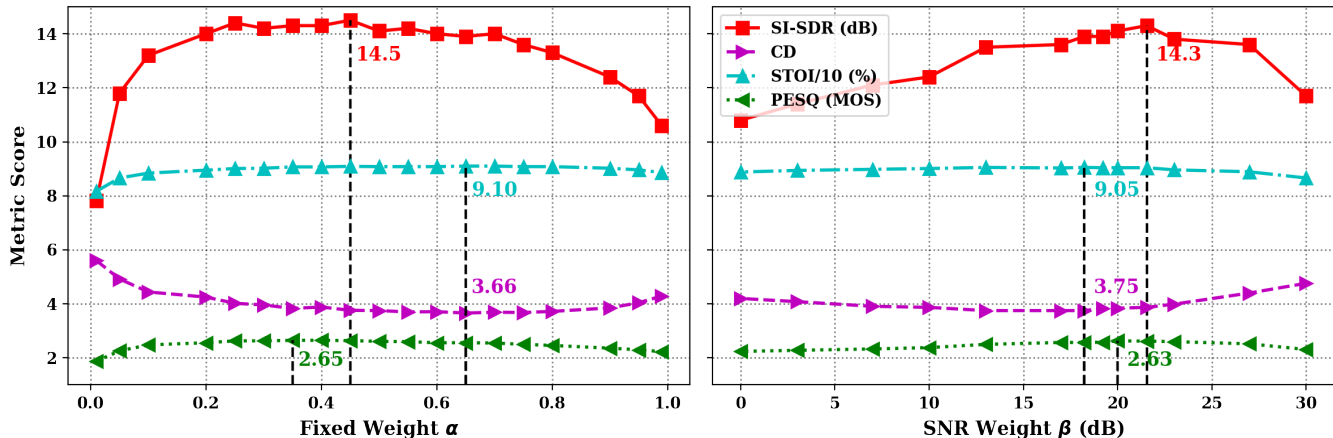
Fig. 3: Effect of fixed weighting and SNR weighting on objective speech quality and intelligibility measures. Black dashed vertical lines indicate the optimal coefficient for each metric. Note that the optimal points coincide for STOI and CD at $\alpha = 0.65$ and $\beta = 18.2$ dB.

Table 2: Subjective MOS from 5500 clips and 20 ratings per clip.

Method	MOS (mean $\pm$ std.)
Noisy	2.63 ± 0.03
RNNoise$_I$ [22]	3.26 ± 0.03
Proposed ($\alpha = 0.05$)	3.93 ± 0.03
Proposed ($\alpha = 0.1$)	3.92 ± 0.03
Proposed ($\alpha = 0.2$)	3.74 ± 0.03
Proposed ($\alpha = 0.3$)	3.65 ± 0.03

Table 3: Comparison of objective metrics with baseline online SE systems. Refer to text for details about each setup.

Method	# Param.	SI-SDR (dB)	CD	STOI (%)	PESQ (MOS)
Noisy	–	9.81	4.56	88.0	2.22
LSA [14, 31]	–	6.10	4.64	84.7	2.33
RNNoise$_I$ [22]	61.2 K	10.4	3.83	88.0	2.55
RNNoise$_{257}$	2.64 M	13.0	3.88	89.3	2.56
Proposed$_{0.35}$	1.26 M	**14.3**	**3.83**	**90.7**	**2.65**
Wiener	Oracle	20.5	2.13	98.1	3.82

speech. We observe a noticeable improvement in performance as each segment increases to 5 seconds, beyond which the improvement starts to diminish. We do not show the result for the feature test due to space limitation, but overall there is little difference between all normalized variants of STFTM and LPS features, while no normalization results in degradation. In general, we recommend FD online normalization due to its invariance to varying signal levels. We also suggest using segments that are no less than 5-second long each during training.

The effect of speech distortion weighting is shown in Fig. 3, where α or β are changed to search the optimal points for each objective measure. Curiously, only STOI and CD agree on the same coefficient in both cases, while both PESQ and SI-SDR suggest smaller weight on speech distortion. The optimal SNR weights for all metrics are concentrated around 20 dB, meaning that the speech distortion weight should only rapidly increase when the noisy signal is relatively clean. Overall, a fixed weighting is slightly better than SNR weighting in all metrics.

During experiments, we notice that even though our systems

trained on MSE (*e.g.* row 4 in Table 1) could achieve similar objective measures compared to those trained on the proposed weighted losses (12), the corresponding subjective quality of systems trained on the weighted loss is a lot better. The most noticeable improvement of systems trained on our loss functions, especially with small α, is that the estimated gain function is much more frequency-selective than systems trained on regular MSE, resulting in higher noise suppression, especially at high SNRs. To testify this, we present the result of the online subjective listening test in Table 2. Not only did all our selected systems significantly outperform the improved RN-Noise (RNNoise$_I$) trained on MSE presented in [22], we were surprised that the listening test subjects preferred a rather low setting for the speech distortion weight α. This trend is mispredicted by all objective measures as well as the authors' subjective preference of about $\alpha = 0.35$. We observed noticeable speech distortion as α goes below 0.35, while noise became more suppressed. It is evident that more detailed investigations are required in future work to shed more light on speech distortion and noise reduction preferences for different groups of listeners.

Finally, we report the objective evaluation from each baseline method, the noisy reference, and an oracle Wiener filter as upper bound in Table 3. The selected system from our method is trained using fixed speech distortion weighting with $\alpha = 0.35$, which we believe strikes a good balance between speech distortion and noise reduction. Although this setup might not be the most preferred for human listeners, it can be easily tuned to different applications. It is nevertheless important to show that it outperforms all tested classical or DNN-based methods in all objective metrics.

6. CONCLUSIONS

In this paper, we proposed and evaluated a real-time speech enhancement approach based on a compact recurrent neural network trained with a simple MSE-based speech distortion weighted loss function. We show the impact of various feature normalization techniques and sequence lengths on the objective quality of enhanced speech. We also demonstrate how to control the amount of speech distortion with fixed-weighted and SNR-weighted coefficients in the loss function. Both objective and subjective tests show that our method outperforms other competitive online methods. In the future, we will explore time-varying speech distortion weighting and its influence on subjective and objective speech quality.

7. REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[2] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, Springer, 2005.

[3] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[4] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE signal processing letters*, vol. 9, no. 1, pp. 12–15, 2002.

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

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[7] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[8] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation," *ACM Trans. on Graphics (TOG)*, vol. 37, no. 4, pp. 112, 2018.

[9] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *ISCA INTERSPEECH 2017*, 2017, pp. 3642–3646.

[10] J.-M. Valin, "A hybrid DSP/deep learning approach to real-time full-band speech enhancement," in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, 2018, pp. 1–5.

[11] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *ISCA INTERSPEECH*, 2018, pp. 3229–3233.

[12] Y. Xia and R. Stern, "A priori SNR estimation based on a recurrent neural network for robust speech enhancement," in *ISCA INTERSPEECH*, 2018, pp. 3274–3278.

[13] S. Braun, K. Kowalczyk, and E. Habets, "Residual noise control using a parametric multichannel wiener filter," in *IEEE ICASSP*, 2015, pp. 1–5.

[14] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[15] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement system," in *IEEE ICASSP*, 2009, pp. 1–5.

[16] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *IEEE Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 136–140.

[17] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," in *ISCA INTERSPEECH 2015*, pp. 1508–1512.

[18] F. G. Germain, Q. Chen, and V. Koltun, "Speech Denoising with Deep Feature Losses," in *Proc. Interspeech 2019*, 2019, pp. 2723–2727.

[19] J. M. Martín-Doñas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal processing letters*, vol. 25, no. 11, pp. 1680–1684, 2018.

[20] Y. Zhao, B. Xu, R. Giri, and T. Zhang, "Perceptually guided speech enhancement using deep neural networks," in *IEEE ICASSP*, 2018, pp. 5074–5078.

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

[22] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A Scalable Noisy Speech Dataset and Online Subjective Test Framework," in *ISCA INTERSPEECH 2019*, 2019, pp. 1816–1820.

[23] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *IEEE ICASSP*, 2018, pp. 2401–2405.

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

[27] Z. Xu, S. Elshamy, Z. Zhao, and T. Fingscheidt, "Components loss for neural networks in mask-based speech enhancement," *arXiv preprint arXiv:1908.05087*, 2019.

[28] 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 (Cat. No. 01CH37221)*, 2001, vol. 2, pp. 749–752.

[29] 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*, 2010, pp. 4214–4217.

[30] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *IEEE ICASSP*, 2019, pp. 626–630.

[31] I. J. Tashev, *Sound capture and processing: practical approaches*, John Wiley & Sons, 2009.