A consolidated view of loss functions for supervised deep learning-based speech enhancement

Sebastian Braun, Ivan Tashev
Microsoft Research, Redmond, WA, USA
{sebastian.braun, ivantash}@microsoft.com

***Abstract*—Deep learning-based speech enhancement for real-time applications recently made large advancements. Due to the lack of a tractable perceptual optimization target, many myths around training losses emerged, whereas the contribution to success of the loss functions in many cases has not been investigated isolated from other factors such as network architecture, features, or training procedures. In this work, we investigate a wide variety of loss spectral functions for a recurrent neural network architecture suitable to operate in online frame-by-frame processing. We relate magnitude-only with phase-aware losses, ratios, correlation metrics, and compressed metrics. Our results reveal that combining magnitude-only with phase-aware objectives always leads to improvements, even when the phase is not enhanced. Furthermore, using compressed spectral values also yields a significant improvement. On the other hand, phase-sensitive improvement is best achieved by linear domain losses such as mean absolute error.**

***Keywords*—speech enhancement, noise reduction, recurrent neural network, loss functions**

I. Introduction

Speech enhancement using neural networks has seen large attention and success in the recent years [1]. While classic single-channel statistical model-driven speech enhancement techniques used in practical systems often only leverage signal models for quasi-stationary noise [2], neural networks can potentially learn more complex speech characteristics, which also allows reduction of highly non-stationary, transient noise, and non-speech sound sources.

Unfortunately, state-of-the-art deep learning (DL) based noise reduction performance is currently only achieved by architectures requiring large look-ahead, large amounts of temporal context data input [3]–[5], or computationally expensive network architectures [3], [6]–[9]. As the performance seems to scale with the network size this often prohibits the use in real-time speech communication systems such as live-messengers or mobile communication devices.

However, the training loss function is independent of the inference complexity, and has therefore potential to improve performance at no cost. Although the most popular choice for regression-based DL is the mean-squared error (MSE), this might arguably be not the optimal choice for speech enhancement. Loss functions and training targets for speech enhancement have shifted from the MSE between several versions of enhancement filters or masks [3], [10] to signal-based metrics, such as spectral magnitude-based MSE, phase-sensitive MSE [11] and finally the complex spectral MSE [6]. Approaches originating from a source separation background often use the time-domain MSE or signal-to-distortion ratio (SDR) loss [5], [12].

While recent attempts were made integrating perceptually motivated metrics in the loss function [13], [14], optimizing on perceptual metrics alone is often insufficient, and is therefore combined again with lower-level criteria such as the spectral magnitude MSE. It is often observed that optimization on some objective metrics like perceptual evaluation of speech quality (PESQ) or short-time objective intelligibility (STOI) improves the test results for the optimized metric, but fails to outperform other baselines in terms of other metrics [13], [14]. While the log-energy sigmoid weighting proposed in [15] does not generalize as it is highly heuristic and signal level dependent, we also could not verify improvements using a noise shaping weighting as proposed in [14] for our tested networks and data. Therefore, we take a step back and investigate different basic signal distance metrics as optimization criteria, which does not exclude the possibility to add perceptually motivated weightings.

As in the last years a large variety of speech enhancement loss functions have been proposed, it is impossible to quantify their individual contribution to success due to the use of different enhancement systems and datasets. The study in [16] compares a selection of loss functions for a convolutional time-domain network. These results may differ greatly from our study due to a complex network architecture with larger delay, an inference complexity more than 30 times larger than our network, and training/evaluation on non-reverberant speech, which is rarely encountered in practice. In this work, we compel an overview and comparison of different frequency-domain optimization criteria using a small recurrent neural network suitable for on-the-edge real-time inference. We classify the losses based on their distance metric in spectral magnitude and complex losses, propose some new losses closing gaps in this systematic search, and point out interesting relations. We show that the best performing of the tested loss functions are the compressed MSE, closely followed by the mean absolute error (MAE), which can be attributed to a better match to the signal distributions. We furthermore show that linear combination of magnitude and complex losses leads to improvement in all cases. Another interesting finding is that our results on a reverberant speech dataset did not confirm advantages of the recently proposed speech distortion-weighted (SDW) [17] and noise shaping losses [14].

II. Signal Model

In a pure noise reduction task, we assume that the observed signal is an additive mixture of the desired speech and noise. We denote the observed signal $X(k,n)$ directly in the short-time Fourier transform (STFT) domain, where k and n are the frequency and time frame indices as

$$X(k,n) = S(k,n) + N(k,n), \tag{1}$$

where $S(k,n)$ is the potentially reverberant speech, and $N(k,n)$ is the disturbing noise signal. The objective is to recover a speech signal estimate $\widehat{S}(k,n)$ by

$$\widehat{S}(k,n) = G(k,n)\, X(k,n). \tag{2}$$

where $G(k,n)$ is a filter that can be either a real-valued suppression gain, or a complex-valued filter. In this work, we consider only a suppression gain.

III. Loss functions

In this section, we review and introduce a wide range of training loss functions targeting recovery of the speech signal $S(k,n)$. All considered speech enhancement loss functions are distance metrics between the enhanced and target spectral representations. We can classify the loss functions summarized in Table I in magnitude distances and complex spectral distances, which also incorporate phase information. The operator $\langle Y(k,n)\rangle = \frac{1}{KN}\sum_{k,n} Y(k,n)$ denotes the arithmetic average over frequency and time indices, k and n, per sequence. Newly proposed loss functions are marked with a †. In the following, we introduce and discuss the loss functions in Table I.

A. Linear spectral distance norms

The most straightforward choice is the L2-norm or squared error between estimated and target signals. While this loss is often only magnitude based as in (3) [18], [19], its complex counterpart (4) is usually only used in direct spectral mapping approaches [6], [20], but has strangely never been used in filter prediction networks so far.

An actually better distance metric for the complex error is the L1-norm or MAE, as the distribution of STFT bins follow a more Laplacian distribution rather than Gaussian, as can be observed in Fig. 1 by the blue curves. The L1-norm of the magnitude and complex signal error are given by (5) and (6), respectively, where we define the L1 norm of a complex number as $\|x_R + jx_I\|_1 = |x_R| + |x_I|$. The complex L1-norm loss (6) has been termed *RI* loss in [21].

B. Logarithmic spectral distance

To account for the logarithmic perceptual nature of the human ear, the log spectral distance (LSD) given by (7) can be used, which was a standard in traditional model-based speech enhancement for decades [22]. Note that so far, the LSD has only been proposed in methods directly predicting the log power spectrum instead of a filter [13], [23], while we use it to predict a filter. The log compression creates a Gaussian-like distribution as shown in Fig. 1 by the yellow line.

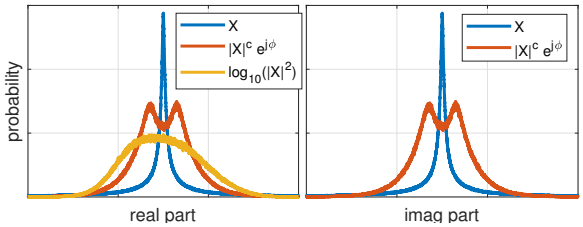

Fig. 1. Distributions of linear complex, compressed complex and log spectral signals for 5 min noisy speech.

To extend the LSD (7) with a phase-error term, we propose the phase-aware logarithmic spectrum distance (PLSD) by (8), where φ_S and $\varphi_{\widehat{S}}$ are the phase angles of $S(k,n)$ and $\widehat{S}(k,n)$, respectively. The first term in (8), the magnitude error, is identical to (7). The second term, the phase error, is connected to the magnitude error by bin-wise multiplication, which naturally decreases the phase error at bins with small magnitude error. The constant 2 ensures that the phase error lies within the range of $[1,3]$, preventing vanishing magnitude error at zero phase error. Note that the cosine phase difference can be calculated by $\cos(\varphi_{\widehat{S}} - \varphi_S) = \Re\left\{\frac{\widehat{S}S^*}{|\widehat{S}S^*|}\right\}$.

C. Weighted logarithmic loss

Due to the logarithmic compression, the standard LSD suffers from the problem of producing large errors also at low energy bins, which are perceptually less relevant. As limiting the log mitigates this problem only suboptimally, we propose to apply a bin-wise weighting based on the target speech signal in (9) with

$$W_{\text{lsd}}(k,n) = |\widehat{S}(k,n) + \gamma X(k,n)|^{0.3}, \tag{20}$$

where we chose $\gamma = 0.1$ to blend in the noisy signal to prevent applying zero weights where high noise reduction is achieved, and apply a compression exponent of 0.3. The same weighting can also be applied to the PLSD as given by (10).

D. Power-law compressed spectral distance

A similar dynamic compression as the logarithm can be achieved using power-law compression [24] applied to the magnitudes by (11) with a compression exponent $0 < c < 1$.

A phase-aware compressed loss can be obtained by multiplying the phase terms to the compressed magnitudes as given by (12), which was proposed in [4], [25]. A commonly used compression exponent is $c = 0.3$. In contrast to the logarithm, which has to be lower bounded to prevent undefined values, the compression produces well-behaved positive semi-definite values. We can observe in Fig. 1 by the red lines that the compression broadens the distributions complex compressed spectra, while values closer to zero occur less frequent.

E. Signal ratio losses

Commonly used ratios in speech enhancement are the signal-to-noise ratio (SNR) and SDR. The time-domain SDR has already been successfully used in DL based speech enhancement [26]–[28]. However, this metric is not restricted

TABLE I. LOSS FUNCTIONS APPLYING VARIOUS DISTANCE METRICS ON MAGNITUDES OR COMPLEX SPECTRA. NOTATION: $S = A\,e^{j\varphi_S}$.

metric	magnitude			complex		
L2	magMSE	$\left\langle \lvert\widehat{A} - A\rvert^2 \right\rangle$	(3)	cMSE	$\left\langle \lvert\widehat{S} - S\rvert^2 \right\rangle$	(4)
L1	magMAE	$\left\langle \lvert\widehat{A} - A\rvert \right\rangle$	(5)	cMAE	$\left\langle \lvert\widehat{S} - S\rvert \right\rangle$	(6)
log MSE	LSD	$\left\langle \lvert \log_{10}\widehat{A} - \log_{10} A\rvert^2 \right\rangle$	(7)	PLSD†	$\left\langle \lvert \log_{10} \lvert\frac{\widehat{S}}{S}\rvert \times \left(2 - \Re\left\{\frac{\widehat{S}S^*}{\lvert\widehat{S}S^*\rvert}\right\}\right) \right\rangle$	(8)
weighted log MSE	wLSD†	$\left\langle W_{\text{lsd}} \lvert \log_{10}\widehat{A} - \log_{10} A\rvert^2 \right\rangle$	(9)	wPLSD†	$\langle W_{\text{lsd}}\ (-\text{"}-)\ \rangle$	(10)
compressed	magComp	$\left\langle \lvert\widehat{A}^c - A^c\rvert^2 \right\rangle$	(11)	cComp	$\left\langle \lvert\widehat{A}^c e^{j\varphi_{\widehat{S}}} - A^c e^{j\varphi_S}\rvert^2 \right\rangle$	(12)
ratios	SNR†	$-\log_{10}\frac{\langle A^2\rangle}{\langle\lvert\widehat{A}-A\rvert^2\rangle}$	(13)	SDR	$-\log_{10}\frac{\langle\lvert S\rvert^2\rangle}{\langle\lvert\widehat{S}-S\rvert^2\rangle}$	(14)
correlation	magCorr†	$-\frac{\langle\widehat{A}A\rangle^2}{\langle\widehat{A}^2\rangle\langle A^2\rangle}$	(15)	cCorr†	$-\frac{\Re\{\langle\widehat{S}S^*\rangle\}}{\sqrt{\langle\lvert\widehat{S}\rvert^2\rangle\langle\lvert S\rvert^2\rangle}}$	(16)
speech dist. weight	SDW	$\lambda\left\langle\lvert S-GS\rvert^2\right\rangle + (1-\lambda)\left\langle\lvert GN\rvert^2\right\rangle$	(17)		–	
weighted L2	MSE-AMR	$\left\langle W_{\text{AMR}}\lvert\widehat{A} - A\rvert^2 \right\rangle$	(18)	cMSE-AMR	$\left\langle W_{\text{AMR}}\lvert\widehat{S} - S\rvert^2 \right\rangle$	(19)

to the time-domain, and can be equivalently computed in the frequency domain. We employ here the scale-variant SDR given by (14), as we believe a scaled output signal as in the scale-invariant SDR [28] is undesired.

In analogy, computing this ratio from magnitudes is more commonly termed the SNR, given by (13). Note that the SNR and SDR losses, (13) and (14), are simply related to the MSE losses (3) and (4), normalized by the speech power, as was also pointed out in [29].

F. Correlation based losses

The speech intelligibility index and related objective metrics [30] are based on signal envelope correlation. Motivated by this fact, we introduce the magnitude correlation loss given by (15).

The complex equivalent, the complex correlation coefficient given by (16), is better known as the *coherence*. While the range of (15) is $[0, 1]$, the range of (16) is $[-1, 1]$. Note that in [31], the coherence loss (16) has been termed source-to-distortion ratio. Special properties of the ratio and correlation based losses is that they are signal-level independent.

G. Speech distortion weighted loss

By using the signal components of speech and noise separately, the SDW loss [17], [32] given by (17) provides a trade-off parameter $0 < \lambda < 1$ between speech distortion and noise reduction. Note that while (17) does not explicitly use only magnitudes, the decomposed nature and absence of the noisy signal $X(k, n)$ implies that $G(k, n)$ as zero-phase filter is optimal. Therefore, the loss is categorized as magnitude loss. Drawbacks of the SDW loss are that the optimal weight λ is data dependent, and finding optimal adjustments of λ e.g. depending on the SNR, are heuristic and difficult to determine.

H. Weighted and combined losses

In [14], a weighting for the MSE based on the AMR codec is proposed to spectrally shape the noise error. We include this loss given by (18) and (19), while also other weightings can be applied to most distance metrics.

Several works have proposed combined losses using linear combinations of magnitude-only and phase-aware metrics [4], [8], [24] as

$$\mathcal{L}_{\text{mix}} = (1 - \beta)\mathcal{L}_{\text{mag}} + \beta\mathcal{L}_{\text{complex}}, \tag{21}$$

where $\mathcal{L}_{\text{mag}}$ is a magnitude-based loss, $\mathcal{L}_{\text{complex}}$ is a complex signal based loss, and $0 \leq \beta \leq 1$ is the mixing factor. We investigate all useful combinations per row in Table I.

IV. NETWORK AND TRAINING

We use a recurrent network architecture based on gated recurrent units (GRUs) [33] and feed forward (FF) layers, similar to the core architecture of [8], to estimate the enhancement filter $G(k, n)$. The architecture was chosen to maintain real-time constraints without delay and moderate complexity.

The network input is the logarithmic power spectrum $P = \log_{10}(|X(k, n)|^2 + \epsilon)$ with online mean and variance normalization [17]. We use a STFT size of 512 with 32 ms square-root Hann windows and 16 ms frame shift, but feed only the relevant 255 frequency bins into the network, omitting 0th and highest (Nyquist) bins, which do not carry useful information. The network consists of a FF embedding layer, two GRUs, and three FF layers with rectified linear unit (ReLU) activations and an output layer with *Sigmoid* activation. The enhancement system and network architecture with layer sizes is shown in Fig. 2, and has 2.8 M trainable parameters. The network satisfies real-time constraints on typical CPU platforms with a processing time of the ONNX runtime of 6 ms per second of audio on a Intel© Core™i7 QuadCore at 3.5 GHz.

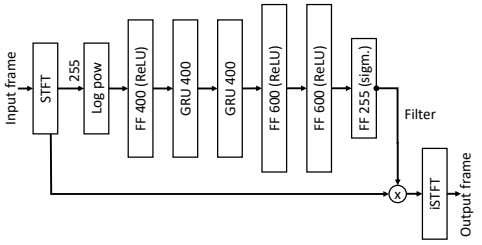

Fig. 2. Network architecture and enhancement system.

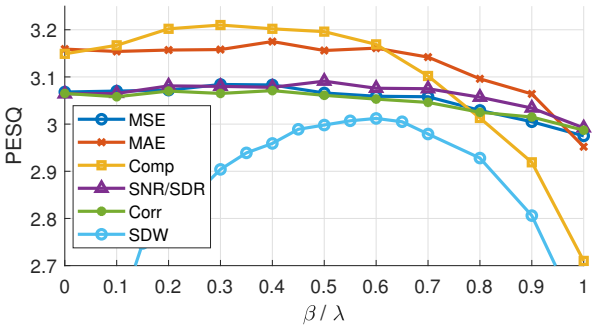

Fig. 3. Optimization of magnitude vs. complex loss weight β and speech distortion weight λ (see equations (17) and (21)) on validation set.

The network was trained using the AdamW optimizer [34] with a learning rate of 10^{-4}. The training was monitored every 10 epochs using a validation subset. The best model was chosen based on the highest PESQ [35] on the validation set. Also the optimal weighting factors for β, λ etc. were optimized by a grid search and choosing the best performing parameter for PESQ on the validation set.

V. Experiments

A. Dataset and evaluation metrics

We used the Chime-2 WSJ-20k dataset [36], which is, despite only of medium size, a realistic self-contained public dataset including matching reverberant speech and noise conditions. The dataset contains 7138, 2418, and 1998 utterances for training, validation and testing, respectively. The target speech signals are binaural and reverberant, and the mixtures contain noise recorded in the same rooms. Validation and test sets are mixed with SNRs in from -6 to 9 dB. For testing, we used only the left channel. We evaluate the speech enhancement performance in terms of PESQ [35] as an indicator for noise reduction and speech quality, and scale-invariant signal-to-distortion ratio (SI-SDR) [28] as a phase-sensitive metric.

B. Results and discussion

Each magnitude and complex loss per row in Table I was combined by linear mixing (21). The LSD losses were omitted as the PLSD is already a combined metric. The mixing factors were determined on the development set. The PESQ results for the parameter sweeps of β are shown in Fig. 3. We can observe that the combination of magnitude and complex loss leads to an improvement for all distance metrics. We can also see that the MAE and compressed losses outperform the other distance metrics significantly at the optimal weight β. Furthermore it is interesting, that the combined compressed loss of (11), (12) achieves the highest performance with $\beta = 0.3$, while for magnitude loss only ($\beta = 0$), compressed and MAE are similar, but for fully complex loss ($\beta = 1$), the compressed loss shows a significant performance drop. Although we experimented with "out-of-metric" combinations, in particular combining magComp with a better complex loss, e.g. cMAE, this did not lead to an improvement.

TABLE II. PESQ (SI-SDR) on test set.

loss	magnitude	complex	combined (21)
noisy		2.29 (1.92)	
MSE	3.16 (9.57)	3.10 (9.58)	3.17 (9.58)
MAE	**3.25** (**9.73**)	3.08 (**9.68**)	**3.25** (**9.75**)
LSD	3.04 (8.59)	3.03 (8.31)	–
wLSD	3.19 (9.12)	**3.21** (8.88)	–
Comp	**3.25** (9.45)	2.88 (9.21)	**3.31** (9.42)
SNR / SDR	3.15 (9.54)	3.11 (9.62)	3.19 (**9.66**)
Corr	3.16 (9.56)	3.11 (9.60)	3.16 (9.58)
SDW	3.12 (9.61)	–	–
MSE-AMR	3.01 (9.39)	2.98 (9.45)	–

The PESQ and SI-SDR results for all losses on the test set are shown in Table II, where the combined losses in the right column use the PESQ-optimal weightings. The best performers are highlighted in bold font. The PESQ results align well with the development set in Fig. 3, namely that MAE and compressed loss are good performers. While the pure LSD is even slightly worse than the MSE, the signal power-weighted wLSD outperforms the linear MSE. While the PLSD shows no advantage over the LSD, the wPLSD gives a slight advantage over the magnitude-based wLSD, which confirms the importance of attributing low weights to unimportant frequency bins for the LSD. It is not surprising that the SNR and SDR perform on par with the L2 norm, as they are merely normalized versions. The correlation-based losses are in the same range as well. It is surprising that on this reverberant dataset, the SDW loss performs significantly worse than the magMSE or cMSE, which has been shown differently on non-reverberant datasets in [17], [32]. This highlights also the data dependency of the speech distortion weight λ, which varies from 0.3 in [17], 0.5 in [32], and 0.6 in our case. Furthermore, on the reverberant Chime2 dataset, we also could not confirm the effectiveness of perceptually motivated weightings, such as the AMR weighting proposed in [14], which performed significantly worse than the unweighted MSEs. While the SI-SDR is less correlated with speech quality than PESQ, it shows the best results mostly for linear losses such as MAE and SDR. Overall we can say that magnitude compression and carefully chosen distance metrics according to the spectral domain's distribution can lead to more suitable loss functions.

VI. Conclusions

We have classified several signal-based frequency domain loss functions for speech enhancement and exploited relations and performance differences on the reverberant Chime2 dataset. Our experiments showed that for such realistic data, compressed losses are beneficial and that combined magnitude and complex losses improve the objective speech quality. We also showed different findings for weighted losses with reverberant speech than for anechoic data. Future work has to be done especially on improved phase-aware losses to further improve the quality.

References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," vol. 26, no. 10, Oct 2018.

[2] M. S. Kavalekalam, J. K. Nielsen, M. G. Christensen, and J. B. Boldt, "A study of noise PSD estimators for single channel speech enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5464–5468.

[3] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1492–1501, July 2017.

[4] 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. Graph.*, vol. 37, no. 4, Jul. 2018.

[5] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug 2019.

[6] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Separated noise suppression and speech restoration: LSTM-based speech enhancement in two stages," in *WASPAA*, Oct 2019.

[7] G. Wichern and A. Lukin, "Low-latency approximation of bidirectional recurrent networks for speech denoising," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2017, pp. 66–70.

[8] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, "Differentiable consistency constraints for improved deep speech enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 900–904.

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

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

[11] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," vol. 25, no. 10, Oct 2017.

[12] F. Bahmaninezhad, J. Wu, R. Gu, S.-X. Zhang, Y. Xu, M. Yu, and D. Yu, "A Comprehensive Study of Speech Separation: Spectrogram vs Waveform Separation," in *Proc. Interspeech 2019*, 2019, pp. 4574–4578.

[13] 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 Process. Lett.*, vol. 25, no. 11, pp. 1680–1684, Nov 2018.

[14] Z. Zhao, S. Elshamy, and T. Fingscheidt, "A perceptual weighting filter loss for DNN training in speech enhancement," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2019, pp. 229–233.

[15] 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. European Signal Processing Conf. (EUSIPCO)*, Aug 2017, pp. 1270–1274.

[16] M. Kolbæk, Z. Tan, S. H. Jensen, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 825–838, 2020.

[17] R. Xia, S. Braun, C. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

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

[19] H. Zhao, S. Zarar, I. Tashev, and C. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2401–2405.

[20] S. Fu, T. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *IEEE Intl. Workshop on Machine Learning for Signal Processing (MLSP)*, 2017, pp. 1–6.

[21] Z. Wang and D. Wang, "Multi-microphone complex spectral mapping for speech dereverberation," 2020.

[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," vol. 33, no. 2, 1985.

[23] Y.-H. Tu, I. Tashev, S. Zarar, and C. Lee, "A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 2531–2535.

[24] J. Lee, J. Skoglund, T. Shabestary, and H. Kang, "Phase-sensitive joint learning algorithms for deep learning-based speech enhancement," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1276–1280, 2018.

[25] K. Wilson, M. Chinen, J. Thorpe, B. Patton, J. Hershey, R. A. Saurous, J. Skoglund, and R. F. Lyon, "Exploring tradeoffs in models for low-latency speech enhancement," in *IWAENC*, Sep. 2018, pp. 366–370.

[26] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," vol. 14, no. 4, July 2006.

[27] H.-S. Choi, J. Kim, J. Hur, A. Kim, J.-W. Ha, and K. Lee., "Phase-aware speech enhancement with deep complex U-Net," in *Intl. Conf. on Learning Representations (ICLR)*, 2019.

[28] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - half-baked or well done?" May 2019.

[29] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying tasnet: A dissecting approach," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6359–6363.

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

[31] S. Venkataramani, J. Casebeer, and P. Smaragdis, "Adaptive front-ends for end-to-end source separation," in *Conference on Neural Information Processing Systems (NIPS)*, 2017.

[32] Z. Xu, S. Elshamy, and T. Fingscheidt, "Using separate losses for speech and noise in mask-based speech enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7519–7523.

[33] K. Cho, B. V. Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proc. 8th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*, 2014.

[34] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[35] ITU-T, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, International Telecommunications Union (ITU-T) Recommendation P.862, Feb. 2001.

[36] E. Vincent, J. Barker, S. Watanabe, and F. Nesta, "The second 'CHIME' speech separation and recognition challenge: datadata, tasks and baselines," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, June 2012.