On the Use of Absolute Threshold of Hearing-based Loss for Full-band Speech Enhancement

Rohith Mars and Rohan Kumar Das

Fortemedia Singapore, Singapore

{rohithmars, rohankd}@fortemedia.com

Abstract

In this paper, we investigate the use of a perceptually motivated loss function for training single-channel full-band speech enhancement models. Specifically, we modify the conventional squared error loss function by incorporating the use of a frequency-importance based weighting scheme utilizing absolute threshold of hearing (ATH). We placed more emphasis on the perceptually relevant frequency bins of the speech spectrogram by applying larger weights to train the speech enhancement model targeting for a higher perceptual quality. We compare the models trained using both the conventional loss and the loss utilizing the proposed ATH-based weighting scheme on the VCTK and 4^{th} DNS challenge datasets. The results demonstrate that the proposed loss using ATH-based weighting scheme achieves better performance than the conventional loss in terms of multiple objective speech quality metrics.

Index Terms: speech enhancement, deep neural networks, absolute threshold of hearing.

1. Introduction

For several decades, single-channel speech enhancement has remained one of the most important and challenging topics in audio signal processing. Given a noisy speech as input, the objective of a speech enhancement system is to process the noisy speech such that the listening quality and intelligibility of the speech is enhanced [1, 2]. Speech enhancement modules find many applications in automatic speech recognition (ASR) systems, audio/video communication and assistive listening devices to improve their performance under noisy and reverberant environments. In practical scenarios, the task of speech enhancement under low signal-to-noise ratio (SNR) conditions, room reverberations and in presence of non-stationary/transient noise is very challenging.

In the past, several approaches have been employed to address single-channel speech enhancement. The classical methods include the approaches that make use of some form of noise spectral estimation such as spectral subtraction [3], minimum mean square error short-time spectral amplitude (MMSE-STSA) [4] and sub-space methods [5]. With the advent of data-driven approaches, the deep learning techniques are now pre-dominantly applied for speech enhancement. A majority of such methods treat speech enhancement as a supervised learning problem [6]. Together with the development of modern neural network model architectures, these approaches have significantly helped to improve the performance of speech enhancement algorithms.

It is well-established that speech enhancement models can be trained either in the time-domain or in the time-frequency (TF) domain. Processing in time-domain has the advantage that it eliminates the requirement for explicit phase estimation. In contrast, phase estimation of the enhanced speech along with the magnitude spectrum is necessary for processing in TF domain. To begin with, magnitude spectrum of the enhanced speech or the TF mask to be applied on the noisy magnitude spectrum was estimated with the phase taken directly from the noisy signal for signal reconstruction [7, 8]. However, in [9–11], it was shown that phase reconstruction plays a crucial part in speech intelligibility, which subsequently led to the development and use of phase sensitive mask (PSM) and complex ratio mask (CRM) [12, 13].

Generally, deep learning-based models for speech enhancement consist of a neural network model trained on a given loss function. The performance of such models often scale with the model size. However, a prohibitively large model size has limited application for real-time inference. In contrast, for a given model size, performance gain can be achieved by choosing the optimal loss function used for the model training. In addition, the model inference time is independent of the choice of loss function since it is computed only during the training phase. As such, there has been several studies on the choice of loss function used for training speech enhancement models both in the time-domain and TF domain [14–16].

One of the most commonly used loss function for supervised speech enhancement in the TF domain is the squared error between the enhanced and the clean speech spectrum. Applying a conventional squared error loss would treat all the frequency bins of the spectrum equally during model training. However, it is well-known from psychoacoustics that all the frequency bins are not equally perceptually important [17, 18]. This is particularly more important when training a full-band speech enhancement model due to the large bandwidth involved, with varying perceptual relevance. Hence, it would be advantageous if the model emphasizes more on the perceptually relevant frequency bins by replacing the squared loss with a weighted squared loss which incorporates frequency importance.

In [19], the use of weighted squared loss based on absolute threshold of hearing (ATH) [20] for speech enhancement using deep neural networks was explored. It utilized simple multi-layer feed-forward neural networks to perform enhancement on wide-band speech signals. It was also reported that use of such ATH-weighted loss only provided better performance under low SNR conditions. In this work, we re-examine and further explore the use of ATH-weighted loss for training modern neural network architectures utilized for speech enhancement. In addition, we extend the use of ATH weighting from wide-band to full-band speech signals along with ablation studies to investigate the performance achieved by using ATH-weighted loss under a wide range of SNR conditions.

The rest of the paper is organized as follows. Section 2 discusses the signal model. In Section 3, we present the ATH-based loss function. The details of the experiments and the results are discussed in Section 4 and Section 5, respectively. Finally, the paper is concluded in Section 6.

979-8-3503-9796-3/22/$31.00

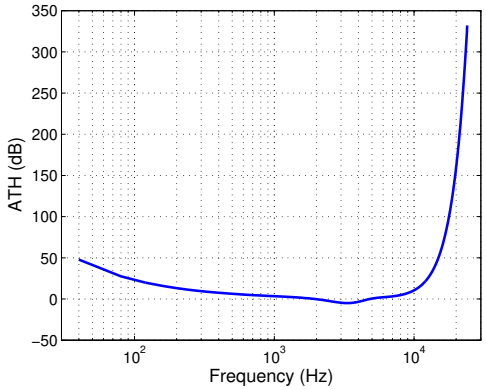

Figure 1: *ATH corresponding to frequencies upto 24000 Hz.*

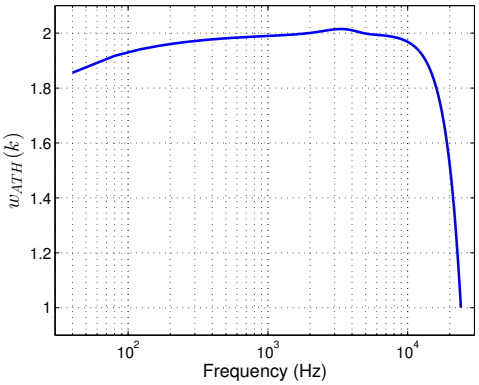

Figure 2: *ATH-based frequency-importance weight $w_{ATH}(k)$ corresponding to frequencies upto 24000 Hz.*

2. Signal Model

The time-domain noisy speech signal $x(t)$ received by a microphone can be expressed as

$$x(t) = s(t) * h(t) + n(t), \tag{1}$$

where t represents the sample index, $s(t)$ denotes the clean speech, $h(t)$ corresponds to the impulse response from the speech source to the microphone in the presence of the additive background noise $n(t)$ with $*$ denoting the convolution operation. Using the short-time Fourier transform (STFT), these signals can be converted to their corresponding TF representation as

$$X(k,\tau) = S(k,\tau)H(k,\tau) + N(k,\tau), \tag{2}$$

where k denotes the frequency bin index and τ denotes the time-frame index.

In order to recover the clean speech from the noisy speech signal, a TF mask needs to be estimated and applied on the noisy signal. For joint magnitude and phase estimation, a complex-ratio mask (CRM) $M(k,\tau)$ can be estimated and applied on the noisy speech $X(k,\tau)$ to obtain the enhanced speech $\tilde{S}(k,\tau)$ as

$$\tilde{S}(k,\tau) = X(k,\tau) \odot M(k,\tau), \tag{3}$$

where $\odot$ represents the element-wise complex-valued multiplication. The enhanced speech in the time-domain, $\tilde{s}(t)$ is then obtained by performing an inverse STFT (ISTFT) operation.

For training the deep neural network with focus on optimum reconstruction of speech, signal approximation (SA) approach [21] can be used. It utilizes the loss between the clean/target speech and the enhanced speech as

$$\mathcal{L} = Loss(S(k,\tau), \tilde{S}(k,\tau)), \tag{4}$$

which is minimized so that the enhanced speech $\tilde{S}(k,\tau)$ matches as close as possible to the target speech $S(k,\tau)$.

3. Perceptually Weighted Loss

The squared error (SE) loss is one of the conventionally used loss functions for training a neural network model for speech enhancement. Specifically, for a given time-frame index with N number of frequency bins, it computes the difference between the estimated and target spectrum of the speech as

$$SE(S,\tilde{S}) = \sum_{k=1}^{N} (S(k) - \tilde{S}(k))^2. \tag{5}$$

In the above formulation, the error corresponding to each of the frequency bins are given equal importance. However, it is well-known from the domains of audio coding and psychoacoustic principles that all frequencies are not perceptually equally important. The model could be trained such that more emphasis is given to frequency bins that are perceptually more relevant. In order to incorporate such perceptual importance during model training, the SE loss function in Eq. (5) is modified to as a weighted squared error (WSE) loss as

$$WSE(S,\tilde{S}) = \sum_{k=1}^{N} w(k)(S(k) - \tilde{S}(k))^2, \tag{6}$$

where $w(k) > 0$ is the weight corresponding to each frequency bin.

In [19], it was proposed to utilize the ATH for defining the frequency-importance weight $w(k)$ for training speech enhancement models. We extend the study in few aspects. To begin with, we replace the simple multi-layer feed-forward neural networks used in [19] with one of the state-of-the-art neural network architectures used for speech enhancement. In addition, we extend the application of ATH from wide-band speech signals to full-band speech signals. The effect of frequency-importance is more emphasized for a full-band spectrum due to the increased bandwidth of operation.

3.1. ATH based frequency-importance weighting

The ATH is defined as the minimum sound pressure level (in dB) required in a pure tone that an average human ear with normal hearing can hear in a quiet environment in the absence of other sounds. The ATH is dependent on the frequency of the tone. In [22], the ATH as a function of tone frequency f is approximated as

$$ATH(f) = 3.64\left(\tfrac{f}{1000}\right)^{-0.8} - 6.5e^{-0.6\left(\frac{f}{1000}-3.3\right)^2} + 10^{-3}\left(\tfrac{f}{1000}\right)^4. \tag{7}$$

We plot the ATH function defined above in Figure 1. It can be seen that the ATH decreases from the low frequencies and has the lowest value around 3000 Hz. Thereafter, it increases and rises sharply for higher frequencies. It suggests that the sound pressure level required to hear very low frequencies as well as the higher frequencies is high and is less critical perceptually.

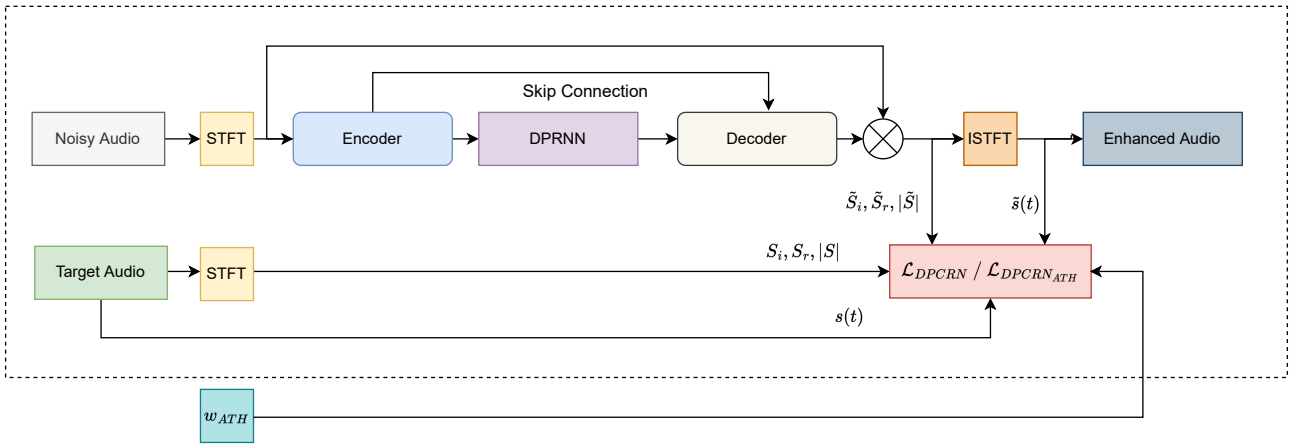

Figure 3: *Block diagram of the DPCRN model with the squared error loss $\mathcal{L}_{DPCRN}$ and the ATH-based weighted loss $\mathcal{L}_{DPCRN_{ATH}}$.*

As such, ATH can be used as a representative of frequency importance since a lower value of ATH implies that the corresponding frequency is easily audible by the humar ear and is more critical perceptually. Therefore, it is easy to note that the frequency-importance weights $w(k)$ have an inverse relationship with ATH. To estimate $w(k)$ using the ATH function defined in Eq. (7), we first compute the $ATH(f)$ corresponding to the center frequency of each frequency bin k for the bandwidth of interest. We then normalize the obtained $ATH(f)$ values such that the maximum corresponds to unity. Denoting the normalized values of $ATH(f)$ as $ATH_{norm}(f)$, the ATH-based frequency importance weight $w_{ATH}(k)$ is then obtained as

$$w_{ATH}(k) = 1 + (1 - ATH_{norm}), \tag{8}$$

where the weights are shifted such that the minimum weight corresponds to unity as shown in Figure 2. It can be seen that the perceptually less critical very low frequencies and the higher frequencies are given a comparatively lower weighting. Using $w_{ATH}(k)$, the weighted loss in Eq. (6) can re-written as

$$WSE_{ATH}(S, \tilde{S}) = \sum_{k=1}^{N} w_{ATH}(k)(S(k) - \tilde{S}(k))^2. \tag{9}$$

4. Experiments

4.1. Datasets

We perform experiments on two separate datasets, namely the VCTK dataset [23] and the 4$^{\text{th}}$ DNS Challenge dataset [24]. A brief description of those datasets are given in the following.

4.1.1. VCTK dataset

The VCTK dataset consists of a separate training and test dataset with utterances sampled at 48 kHz. The training set consists of 11,572 pairs of clean and noisy utterances from 28 speakers. It consists of 10 different noise-types mixed at 4 different SNR conditions, thus resulting in a total of 40 noisy conditions. The 10 different noise types include two synthetically generated (speech-shaped noise and babble) noise types and eight real noise types obtained from the DEMAND dataset [25]. The SNR levels for the training set are set as [0 dB, 5 dB, 10 dB, 15 dB]. This training set is 10 h in duration. We set aside 10% of the training utterances as validation set. The test set consists of 824 pairs of clean and noisy utterances from 2 unseen speakers with 5 different noise types and 4 different SNR conditions, i.e., [2.5 dB, 7.5 dB, 12.5 dB and 17.5 dB].

4.1.2. 4th DNS Challenge dataset (DNS-4)

This dataset consists of clean speech samples belonging to six different languages, including English, French, German, Italian, Russian and Spanish. It also consists of noise dataset which consists of over 62,000 clips belonging to 150 audio classes, totaling to 181 h of noise samples. Both the speech and noise clips are sampled at 48 kHz. From this dataset, we create 500 h of clean and noisy speech pairs by setting the SNR from [-5 dB, 5 dB]. Similar to the previous experiment, we set aside 10% of the training utterances as validation set. For the evaluation, we use DNS-4 blind test set provided by the challenge organizers which consists of 859 noisy speech utterances, each with 10 s duration. Out of these, 638 test clips are recorded using a mobile device while the rest are recorded using desktop PC/laptop.

4.2. Deep learning model for speech enhancement

We utilize one of the state-of-the-art speech enhancement models, namely the dual-path convolution recurrent network (DPCRN) [26] for training our models. It consists of an encoder-decoder architecture with dual-path recurrent neural network (DPRNN) block in between to model the frequency and temporal dependencies in the spectrum. Skip connections are used between the encoder and the decoder blocks. The DPCRN model is trained using SNR loss as the time domain loss, while a combination of squared error loss of the real spectrum, imaginary spectrum and magnitude spectrum is used as the TF domain loss. The DPCRN loss $\mathcal{L}_{DPCRN}$ can be expressed as

$$\mathcal{L}_{DPCRN} = f(s(t), \tilde{s}(t)) + \log(SE(S_r, \tilde{S}_r) + SE(S_i, \tilde{S}_i) + SE(|S|, |\tilde{S}|)) \tag{10}$$

where $f(s(t), \tilde{s}(t))$ denotes the SNR loss. We modify the squared loss of DPCRN to incorporate the ATH-based frequency importance weighting defined in Eq. (9) as

$$\mathcal{L}_{DPCRN_{ATH}} = f(s(t), \tilde{s}(t)) + \log(WSE_{ATH}(S_r, \tilde{S}_r) + WSE_{ATH}(S_i, \tilde{S}_i) + WSE_{ATH}(|S|, |\tilde{S}|)). \tag{11}$$

The block diagram of the DPCRN model with the loss functions $\mathcal{L}_{DPCRN}$ and $\mathcal{L}_{DPCRN_{ATH}}$ is shown in Figure 3.

Table 1: *Performance of DPCRN model trained with $\mathcal{L}_{DPCRN}$ and $\mathcal{L}_{DPCRN_{ATH}}$ loss on the VCTK test set.*

Model	PESQ-WB	CSIG	CBAK	COVL	STOI	SI-SDR
Noisy	1.97	3.32	2.43	2.61	0.92	8.44
DPCRN ($\mathcal{L}_{DPCRN}$)	2.57	3.78	3.23	3.17	**0.93**	**17.75**
DPCRN ($\mathcal{L}_{DPCRN_{ATH}}$)	**2.68**	**3.89**	**3.29**	**3.28**	**0.93**	**17.75**

Table 2: *Performance in PESQ-WB of DPCRN model trained with $\mathcal{L}_{DPCRN}$ and $\mathcal{L}_{DPCRN_{ATH}}$ loss at different SNR levels.*

Model	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
Noisy	1.09	1.14	1.23	1.40	1.66	2.02
DPCRN ($\mathcal{L}_{DPCRN}$)	1.19	1.37	1.63	1.92	2.23	2.55
DPCRN ($\mathcal{L}_{DPCRN_{ATH}}$)	**1.21**	**1.39**	**1.65**	**1.94**	**2.28**	**2.64**

4.3. Model parameters

We keep the DPCRN model architecture similar to the original architecture used in [26]. Since we use full-band speech, the FFT length is set as 1200 with 50% overlap. The input to the model is then the 601-dimensional complex spectrum. For the encoder block, the number of filters in the convolutional layers are set as (32,32,32,64,128). The kernel size and the stride corresponding to these convolutional layers are set as (5,2),(3,2),(3,2),(3,2),(3,2) and (2,1),(2,1),(1,1),(1,1),(1,1) in frequency and temporal dimension. Two DPRNN modules are stacked with BiLSTM layer to model frequency dimension and LSTM layer to model temporal dimension each with a hidden dimension of 128. For optimization, we use Adam optimizer [27] with a batch size of 8. The initial learning rate is set as 0.001 and it is reduced by half if the validation loss does not decrease after 5 epochs. Early stopping technique is used if the validation loss does not improve for 10 epochs.

4.4. Evaluation metrics

For the experiments on the VCTK dataset, we use the perceptual evaluation of speech quality (PESQ-WB) using the wideband version recommended in ITU-T P.862.2 as an objective quality measure [28]. We also use the composite measures, namely CSIG for signal distortion, CBAK for noise distortion evaluation, and COVL for overall quality evaluation [29]. In addition to these, we also employ the short-time objective intelligibility (STOI) [30] and the scale-invariant signal-to-distortion ratio (SI-SDR) [31] objective quality measures. For the experiments on the DNS-4 dataset, we use the local evaluation of DNSMOS [32] provided by the challenge organizers. The DNSMOS metric measures speech quality (SIG), background noise quality (BAK) and overall audio quality (OVRL).

5. Results and Analysis

We first consider the VCTK database for the studies. We evaluate the performance of the DPCRN model trained using $\mathcal{L}_{DPCRN}$ and $\mathcal{L}_{DPCRN_{ATH}}$ loss functions on VCTK test set with pre-mixed noisy utterances using PESQ-WB, CSIG, CBAK, COVL, STOI, SI-SDR metrics as shown in Table 1. It can be observed that the model trained using $\mathcal{L}_{DPCRN_{ATH}}$ loss achieves a PESQ-WB score of 2.68, which is higher than 2.57 obtained using $\mathcal{L}_{DPCRN}$. Similarly, it achieves CSIG, CBAK, COVL of 3.89, 3.29 and 3.28 compared to the scores of 3.78, 3.23 and 3.17 obtained using $\mathcal{L}_{DPCRN}$. In terms of the speech intelligibility and distortion metrics, STOI and SI-SDR, both models achieve comparable performance. The performance of the noisy signal without using any speech enhancement in terms of various metrics is also reported to show the impact of speech enhancement using both the methods, which is more evident when using ATH-based loss.

Table 3: *Performance on DNS-4 blind test set.*

Model	SIG	BAK	OVRL
Noisy	3.23	2.40	2.25
NSNet2	3.12	**3.84**	2.79
DPCRN ($\mathcal{L}_{DPCRN}$)	3.24	3.80	2.87
DPCRN ($\mathcal{L}_{DPCRN_{ATH}}$)	**3.25**	3.81	**2.88**

We then conduct another study using VTCK test set, by replacing the pre-mixed noisy utterance with synthetically generated noisy utterance. Specifically, in order to evaluate the performance of the trained models to unseen noise conditions, we utilize the noise dataset from the DNS-4 dataset and mix them with the clean speech utterances obtained from the VCTK test dataset under different SNR conditions. For a given clean speech clip, we randomly select a noise clip from the DNS-4 noise dataset and mix them at [-5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB] SNR. Therefore, for each SNR condition there are 824 test utterances. The performance of the models trained using the $\mathcal{L}_{DPCRN}$ and $\mathcal{L}_{DPCRN_{ATH}}$ loss functions are evaluated using the PESQ-WB metric and their results are reported in Table 2. It can be observed that for all the SNR conditions, the model trained using $\mathcal{L}_{DPCRN_{ATH}}$ produces a higher PESQ score. This result is in contrast with the results obtained in [19], where weighted loss showed improvement only under low SNR conditions. Thereby, it highlights the use of DPCRN based state-of-the-art model to utilize the ATH-based loss to increase robustness against a wide range of conditions.

Finally, we consider the DNS-4 dataset for the studies and the performance comparison of the DPCRN models trained using $\mathcal{L}_{DPCRN}$ and $\mathcal{L}_{DPCRN_{ATH}}$ loss is shown in Table 3. We also compare them with the performance of a reference model provided by the challenge organizers, named NSNet2 [33]. It can be observed that the DPCRN model with ATH-based loss, achieves higher SIG and OVRL scores using the DNSMOS metric indicating its effectiveness.

6. Conclusions

In this paper, we demonstrate the use of perceptually motivated loss for training full-band speech enhancement models. We reinvestigate the use of incorporating the ATH-based frequency-importance weight to the squared error loss. The studies on VCTK and DNS-4 dataset reveal that the proposed speech enchantment system using ATH-based loss performs better than the system with conventional loss. In addition, it is also robust to a varying range of SNR conditions showing applicability towards real-world systems.

7. References

[1] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2006.

[2] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[5] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Process.*, vol. 3, no. 4, pp. 251–266, 1995.

[6] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio Speech Lang Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio Speech Lang Process.*, vol. 23, no. 1, pp. 7–19, 2014.

[8] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*. IEEE, 2013, pp. 7092–7096.

[9] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, 2011.

[10] P. Mowlaee, R. Saiedi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," in *Proc. ICSLP*, 2012, pp. 1–4.

[11] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio Speech Lang Process.*, vol. 22, no. 12, pp. 1931–1940, 2014.

[12] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*. IEEE, 2015, pp. 708–712.

[13] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio Speech Lang Process.*, vol. 24, no. 3, pp. 483–492, 2015.

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

[15] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *Proc. 44th Int. Conf. Telecommun. Signal Process*. IEEE, 2021, pp. 72–76.

[16] S. Braun and H. Gamper, "Effect of noise suppression losses on speech distortion and ASR performance," in *Proc. ICASSP*. IEEE, 2022, pp. 996–1000.

[17] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*. Springer Science & Business Media, 2006, vol. 22.

[18] T. Painter and A. Spanias, "A review of algorithms for perceptual coding of digital audio signals," in *Proc. DSP*, vol. 1. IEEE, 1997, pp. 179–208.

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

[20] H. Fletcher, "Auditory patterns," *Reviews of modern physics*, vol. 12, no. 1, pp. 47–65, 1940.

[21] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. GlobalSIP*. IEEE, 2014, pp. 577–581.

[22] E. Terhardt, "Calculating virtual pitch," *Hearing research*, vol. 1, no. 2, pp. 155–182, 1979.

[23] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech." in *Proc. SSW*, 2016, pp. 146–152.

[24] H. Dubey, V. Gopal, R. Cutler, S. Matusevych, S. Braun, E. S. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "ICASSP 2022 deep noise suppression challenge," in *Proc. ICASSP*, 2022.

[25] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Int. Congr. Acoust*, vol. 19, no. 1. Acoustical Society of America, 2013.

[26] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-path convolution recurrent network for single channel speech enhancement," in *Proc. Interspeech*, 2021, pp. 2811–2815.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] I.-T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[29] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE/ACM Trans. Audio Speech Lang Process.*, vol. 16, no. 1, pp. 229–238, 2007.

[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*. IEEE, 2010, pp. 4214–4217.

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

[32] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: a non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*. IEEE, 2021, pp. 6493–6497.

[33] S. Braun and I. Tashev, "Data augmentation and loss normalization for deep noise suppression," in *Proc. Int. Conf. Speech Comput.* Springer, 2020, pp. 79–86.