ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747656

LOW RESOURCES ONLINE SINGLE-MICROPHONE SPEECH ENHANCEMENT WITH HARMONIC EMPHASIS

Nir Raviv[⋆] *Ofer Schwartz*[⋆] *Sharon Gannot*[†]

[⋆] CEVA-DSP, Israel
[†] Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel
{Nir.Raviv,Ofer.Schwartz}@ceva-dsp.com, Sharon.Gannot@biu.ac.il

ABSTRACT

In this paper, we propose a deep neural network (DNN)-based single-microphone speech enhancement algorithm characterized by a short latency and low computational resources. Many speech enhancement algorithms suffer from low noise reduction capabilities between pitch harmonics, and in severe cases, the harmonic structure may even be lost. Recognizing this drawback, we propose a new weighted loss that emphasizes pitch-dominated frequency bands. For that, we propose a method, applied only at the training stage, to detect these frequency bands. The proposed method is applied to speech signals contaminated by several noise types, and in particular, typical domestic noise drawn from ESC-50 and DEMAND databases, demonstrating its applicability to 'stay-at-home' scenarios.

Index Terms— Single-microphone speech enhancement, DNN, Speech harmonics presence detection, Ideal ratio mask

1. INTRODUCTION

A plethora of approaches to solve the problem of speech enhancement using a single microphone can be found in the literature (see e.g. [1]). Although microphone array algorithms are nowadays widely used, there are still applications in which only a single microphone is available.

The performance of current single microphone solutions is not always satisfactory. Classical model-based algorithms, such as the optimally modified log-spectral amplitude (OMLSA) estimator with the improved minima controlled recursive averaging (IMCRA) noise estimator [2, 3], were introduced to enhance speech signals contaminated by non-stationary noise signals using a virtue of spectral filtering. The most difficult task of filtering-based speech enhancement is signal to noise ratio (SNR) estimation at each time-frequency (TF) bin. When the noisy input exhibits rapid changes in the noise statistics, this task becomes more difficult.

In recent years, DNN-based algorithms were proposed to enhance noisy speech. A comprehensive survey of common approaches can be found in [4]. Recent contributions in the field can be found in [5–8]. These DNN-based approaches necessitate, in general, high computational and memory resources and are hence not applicable to battery-powered edge devices, as often used in home environments, such as true wireless stereo (TWS) earphones. The Deep Noise Suppression (DNS) challenge [9, 10] was specifically focusing on low-resource speech enhancement algorithms. Many DNN-based methods are estimating a mask, e.g. an ideal ratio mask (IRM) [11, 12], that can be applied to the noisy signal to obtain an enhanced spectrogram.

The estimation of such masks is closely related to the classical task of SNR estimation. Unfortunately, it also inherits the performance drop in rapidly changing noise statistics. This is specifically true for voiced-speech segments that are attributed by a pronounced harmonic structure and consequently by rapid spectral changes across the frequency axis. It is therefore expected that the enhanced signal will exhibit low spectral resolution and will fail to preserve the harmonic structure of the clean signal.

Following approaches for fundamental pitch estimation [13, 14], speech bands dominated by harmonic structure can be detected by measuring the periodicity of the signal in the time domain. In this work, we therefore introduce a new weighted loss to train a DNN-based speech enhancement algorithm that emphasizes frequency bands that are dominated by harmonic structure typical to human voice. For that, we first propose a method to detect frequency bands with pronounced harmonic structure using the auto-correlation of the clean input signal. The speech segment periodicity can be evaluated by the ratio between the auto-correlation of the speech signal, calculated in lags typical to the speech periods, and its variance. Then, the contribution of these detected frequency-bands to the overall loss function is emphasized by increasing their corresponding weights. The proposed DNN-based algorithm is capable of reducing noise even between the harmonics of voiced speech segments and of preserving the harmonic structure. The proposed method is applied to speech signals contaminated by several noise types, and in particular, typical domestic background noises.

2. PROBLEM FORMULATION

Let $x(t) = s(t) + v(t)$ denote the observed noisy signal at discrete-time t, where $s(t)$ denotes the clean speech signal and $v(t)$ an additive noise signal. The short-time Fourier transform (STFT) of $x(t)$ with frame-length L is denoted by $\bar{x}(k,n)$, where n is the frame-index and $k = 0, 1, \ldots, L-1$ denotes the frequency index. Similarly, $\bar{s}(k,n)$ and $\bar{v}(k,n)$ denote the STFT of the clean speech and the noise-only signals, respectively.

Different speech activation masks were proposed [4, 11], and the most commonly used mask is the ideal ratio mask (IRM). The IRM of a single frame is defined as follows:

$$\rho(k,n) = \left(\frac{|\bar{s}(k,n)|^2}{|\bar{s}(k,n)|^2 + |\bar{v}(k,n)|^2} \right)^{\gamma}, \tag{1}$$

where γ is commonly set to $\gamma = 0.5$.

We can cast the speech enhancement problem as estimating the IRM mask $\rho(k,n) \in [0,1]$ using only noisy speech utterances. Therefore, the task of the DNN is to infer the mask $\boldsymbol{\rho}$, where $\boldsymbol{\rho}_{k,n} = \rho(k,n)$. Explicitly, given the noisy signal $\bar{\mathbf{x}}$, the goal is to

estimate the IRM $\boldsymbol{\rho}$. Once it is computed, the enhanced signal $\bar{\mathbf{s}}$ can be estimated by $\hat{\bar{\mathbf{s}}} = \bar{\mathbf{x}} \odot \boldsymbol{\rho}$, where $\odot$ is the Hadamard product. In this work, as proposed in [12], we use a softer version of the enhancement task:

$$\hat{\bar{\mathbf{s}}} = \bar{\mathbf{x}} \odot \exp\{-(1-\boldsymbol{\rho}) \cdot \beta\} \tag{2}$$

to potentially alleviate *musical noise artifacts* [15,16] by limiting the attenuation factor to $\exp\{-\beta\}$ in noise-dominant bins.

3. SPEECH HARMONICS WEIGHTED LOSS

In this section, a time-frequency mask is derived by minimizing a special loss function that weights frequency-bands that are dominated by speech harmonics. This mask is used only during the training phase to emphasize the speech-harmonic bands, and consequently improving the training speed of the DNN and the overall performance of the enhancement algorithm.

3.1. Speech Harmonics Presence Detection

Speech-harmonics segments are characterized by periodicity in the time-domain with typical period times. Voiced human speech is attributed by a fundamental frequency in the range 80-250 Hz. In samples, typical speech periods are in the range $\tau \in \left[\lfloor \frac{f_s}{250} \rceil, \ldots, \lfloor \frac{f_s}{80} \rceil\right]$, where f_s is the sampling frequency of the acquisition device and $\lfloor \cdot \rceil$ is a rounding operation. Accordingly, speech segments dominated by harmonic structure can be detected by evaluating the periodicity of the signal. The level of periodicity can be measured by comparing the auto-correlation of the signal evaluated in candidate lag values, $\mathcal{R}(\tau)$, with its variance, $\mathcal{R}(0)$. The auto-correlation can be estimated by the inverse Fourier transform applied to the signals' auto-spectrum, $\mathcal{R}(\tau) = \mathcal{F}^{-1}\left\{\mathcal{S}(e^{j\omega})\right\}$, where ω is the angular frequency. Practically, the STFT of the signal can be used by substituting $\omega = \frac{2\pi k}{L}$. For each frame, an estimate of the auto-spectrum can be obtained by periodogram smoothing:

$$\mathcal{S}(k,n) = \alpha \mathcal{S}(k,n-1) + (1-\alpha)|\bar{s}_k(n)|^2, \tag{3}$$

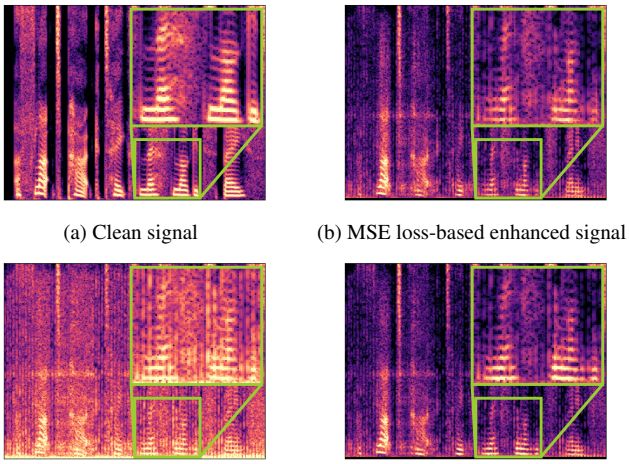
where α is a smoothing factor. Using the auto-spectrum, the auto-correlation can be estimated by $\mathcal{R}(\tau,n) = \frac{1}{L}\sum_{k=0}^{L} \mathcal{S}(k,n) e^{j\frac{2\pi k}{L}\tau}$. Alternatively, the auto-correlation can be estimated for each frequency by summing only a band of frequencies around the center frequency k,

$$\hat{\mathcal{R}}(\tau,k,n) = \mathfrak{Re}\left\{\frac{1}{2\mathcal{K}+1}\sum_{\tilde{k}=k-\mathcal{K}}^{k+\mathcal{K}} \mathcal{S}(\tilde{k},n) e^{j\frac{2\pi\tilde{k}}{L}\tau}\right\}, \tag{4}$$

where $2\mathcal{K}+1$ is the width of each band. In presence of dominant speech harmonics and $\mathfrak{Re}\{\cdot\}$ is the real function. $\hat{\mathcal{R}}(\tau,k,n)$ will have relatively high value for τ that corresponds to the speech period. Therefore, detection of the presence of speech harmonics will be obtained by examining the maximum value of the ratio between $\hat{\mathcal{R}}(\tau,k,n)$ and $\hat{\mathcal{R}}(0,k,n)$ for τ in the typical range of speech periodicity. Consequently, we define speech harmonics presence level $\mathcal{M}$ as follows:

$$\mathcal{M}(k,n) = \max_{\tau} \frac{\hat{\mathcal{R}}(\tau,k,n)}{\hat{\mathcal{R}}(0,k,n)}, \tag{5}$$

The speech harmonics presence level is confined to the range $0 \le \mathcal{M}(k,n) \le 1$, with $\mathcal{M}(k,n) = 1$ implies significant presence of harmonic structure in the speech signal, and $\mathcal{M}(k,n) = 0$ implies the absence of such harmonic structure.



(a) Clean signal (b) MSE loss-based enhanced signal

(c) Noisy signal (d) Our loss-based enhanced signal

Fig. 1: Example for the impact of our proposed speech harmonics weighted loss.

3.2. Objective

Common loss functions, such as mean squared error (MSE), equally weight all time-frequency bins. Therefore, in these cases the DNN training is solely data-driven. To incorporate important domain-knowledge regrading the speech harmonics, we present in this paper a novel weighted loss function.

To motivate our proposed method, we first examine a DNN-based algorithm (with equivalent architecture to the proposed method) that uses a uniform loss function across all time-frequency bins. It is clearly depicted in the highlighted area in Fig. 1b that the output signal suffer from low frequency resolution and that the algorithm fails to preserve the exact harmonic structure and hence fail to suppress the noise between the speech harmonics.

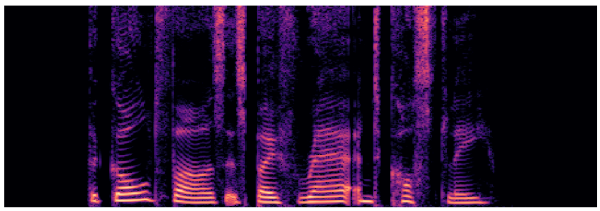
To overcome this pitfall, we propose to define a weighted loss. Let $\ell(k,n)$ be a loss function. In the current work we chose the mean square error (MSE) loss function $\ell = \|\boldsymbol{\rho} - \hat{\boldsymbol{\rho}}\|_{\mathbf{F}}^{\mathbf{2}}$, where $\|\cdot\|_F$ is the Frobenius norm. Then the weighted loss is defined as $\mathcal{L} = \sum_{k,n} w(k,n)\ell(k,n)$. The weights are calculated as follows:

$$w(k,n) = \begin{cases} \lambda, & \mathcal{M}(k,n) > \theta \\ 1, & \mathcal{M}(k,n) \le \theta \end{cases} \tag{6}$$

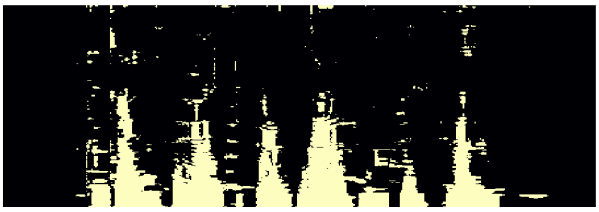
where, $\theta \in [0,1]$ is a threshold parameter, and λ is the weighting hyper-parameter. To emphasize bands dominated by harmonic structure, we set $\lambda > 1$. According to (5) and (6), the weight is controlled by the speech harmonics presence indicator, defined as $\mathbb{1}_{\mathcal{M}(k,n)>\theta}$. A sample speech utterance and the corresponding indicator are depicted in Fig. 2. It can be easily verified that the indicator equals '1' only in frequency bands that are dominated by speech-harmonics.

3.3. System Architecture

Low-resource real-time speech enhancement applications require low complexity, low latency, and low memory footprint. All these requirements mandate a relatively simple neural network. To reduce latency and support real-time applications, we decided to include a recurrent DNN (RNN) layer as part of our model. For these reasons, we designed a network comprising of two GRU layers followed by



(a) Clean signal STFT magnitude (dB)



(b) Harmonic indicator mask with $\theta = 0.4$

Fig. 2: Example of clean signal STFT and its harmonic indicator mask $\mathbb{1}_{\mathcal{M}(k,n)>\theta}$.

two fully-connected (FC) layers and a sigmoid activation to generate the IRM output. As input features to the IRM estimation network, we use the log-spectrum of the noisy signal at a single time-frame n.

3.4. Algorithm Summary

In this section, we introduced a procedure to detect frequency bands with strong harmonic structure, or equivalently speech segments with pronounced periodicity, by calculating the auto-spectrum and auto-correlation of each frequency band, using (3) and (4). Then, a speech harmonic presence level $\mathcal{M}$ is obtained by examining the ratio between the auto-correlation in peak values and the variance of the speech signal using (5). Finally, the weights are calculated using (6) and a desired objective function is applied in the training phase. The enhanced amplitude is obtained by applying (2). The time-domain signal is obtained by appending the noisy phase to the enhanced amplitude and then applying the inverse STFT.

4. EXPERIMENTAL STUDY

In this section we will report the relevant information to reproduce our results and the different experiments we have carried out.

Datasets: The training dataset was generated from the data in the DNS-Challenge [10], for better comparison to relevant competing methods. The speech data was mixed with background noise with SNR ranging from -5 dB to 25 dB, in 1 dB resolution. In addition, the speech data was also convolved with different room impulse responses (RIR) with RT60 ranging from 100 ms to 1000 ms, simulated with the gpuRIR library [17]. All other parameters were set as recommended by the challenge organizers.

The test dataset was generated to simulate different home environments. Speech data was drawn from Harvard dataset [18] mixed with background noise sampled from DEMAND [19] and ESC-50 [20] datasets. From DEMAND we only used 'kitchen', 'living room', and 'washing machine' noises, while from ESC-50 we used all available noise samples. We evaluated our method with SNR $\in \{0, 5, 10, 15\}$ dB with and without reverberation. For reverberant scenarios RT60 was set in the range from 100 ms to 1000 ms.

Training Setup: The algorithm is applied to a single speech signal sampled at 16kHz. The window function is Hann with 32ms frame-size and 8ms shift. The FFT size is 512, equal to the frame length. The GRU layers and the FC have 128 units and are trained in a stateful manner. During training, a 25% dropout is applied between the GRU layers. The Adam optimizer is used with initial learning rate of 0.001 and a gradient clipping of 3. The learning rate is reduced by 10% on plateau for 5 consecutive epochs. The model is trained on a batch-size of 64, and each sample is 2 seconds long.

Competing Methods: We first compare our method with different objective functions, trained with similar architecture. The first objective function is the unweighted MSE loss. The second objective function is the spectral convergence and log-STFT magnitude loss, proposed in [21] (will be referred to as STFT-loss) and is widely used in recent papers [8, 22]:

$$\mathcal{L}_{\text{STFT}} = \mathcal{L}_{\text{sc}} + \mathcal{L}_{\text{mag}} = \frac{\left\| |\bar{\mathbf{s}}| - |\hat{\bar{\mathbf{s}}}| \right\|_F}{\left\| |\bar{\mathbf{s}}| \right\|_F} + \frac{1}{L} \left\| \log |\bar{\mathbf{s}}| - \log |\hat{\bar{\mathbf{s}}}| \right\|_1 \quad (7)$$

where $\|\cdot\|_1$ stands for the L_1 norm.

We then compare our proposed method with different open-source noise suppression networks which comply with our real-time and low resource requirements. The first baseline is the DNS-Challenge NSNet[1] [23, 24] model, provided by the challenge organizers, that was optimized with MSE speech distortion objective function. It consists 3 gated recurrent units (GRU) layers with 257 hidden units and a FC layer with sigmoid activation function for mask estimation. In addition, our method is compared to the Dual-Signal Transformation LSTM (DTLN)[2] [25], optimized with the scale-sensitive negative SNR in the time-domain. This network consists of two consecutive blocks, each with two LSTM layers followed by a fully-connected layer and a sigmoid activation function. The first DTLN block produces a TF mask in the STFT domain, while the second produce a mask in a learnable 1D-Conv layer analysis domain and output a time-domain enhanced signal with a learnable 1D-Conv synthesis layer. These models have 1.26 M and 989 K parameters, respectively, and similar number of multiply–accumulate (MAC) operations. Both networks consume much more memory and require higher computational resources than our proposed network which has 297 K parameters and MACs.

Performance Evaluation: We evaluated the performance of the proposed method using several well-known objective measurements: (i) Perceptual Evaluation of Speech Quality (PESQ), using the wide-band version in ITU-T P.862.2 [26] (from -0.5 to 4.5), (ii) Short-Time Objective Intelligibility (STOI) [27] (from 0 to 100), and (iii) Scale Invariant Signal-to-Distortion Interference Ratio (SI-SDR) [28] (higher is better).

Results: The results of our evaluations are presented in Table 1 and Table 2. One can observe that our method is at the top-2 scores for all the presented scenarios and measures.

Table 1 summarizes the results for the chosen test set in different SNR values and with/without reverberation. Results suggest that our proposed objective function outperforms the original MSE loss in every scenario as well as the STFT loss in the average performance by approximately $0.1, 1.1$, and 0.5 for PESQ, STOI and SI-SDR, respectively. Moreover, it can be validated from Fig. 1d that the speech harmonics weighted loss indeed preserves the speech

[1] https://github.com/microsoft/DNS-Challenge

[2] https://github.com/breizhn/DTLN

Table 1: A comparison of the PESQ, STOI, and SI-SDR for speech signals mixed with different SNRs. Best results are indicated in boldface and second are underlined.

	PESQ					STOI					SI-SDR				
SNR (dB)	0	5	10	15	Avg.	0	5	10	15	Avg.	0	5	10	15	Avg.
—WITHOUT REVERBERATION—															
Noisy	1.14	1.23	1.35	1.62	1.33	79.3	87.2	92.9	96.5	89.0	-8.43	-3.59	1.44	6.43	-1.03
NSNet-2	**1.50**	**1.78**	2.11	2.50	1.97	**86.3**	**91.6**	**95.0**	97.1	**92.5**	**8.69**	**12.05**	15.02	17.48	13.30
DTLN	1.42	1.64	1.91	2.28	1.81	80.3	87.2	92.6	96.2	89.0	5.30	10.06	14.13	17.65	11.78
STFT loss	1.37	1.63	1.97	2.42	1.91	82.5	89.3	93.7	96.7	91.0	6.83	11.22	15.41	**19.62**	12.15
MSE loss	1.35	1.65	2.06	2.57	1.85	83.3	89.9	94.1	96.8	90.6	6.28	10.38	14.11	17.82	13.27
Ours	1.42	1.73	**2.14**	**2.62**	**1.98**	84.9	91.0	**95.0**	**97.2**	92.0	7.28	11.52	**15.46**	19.30	**13.40**
—WITH REVERBERATION—															
Noisy	1.18	1.27	1.42	1.74	1.40	73.9	83.4	90.1	95.0	85.6	-8.37	-3.54	1.23	6.29	-1.09
NSNet-2	1.44	1.65	1.93	2.24	1.82	**81.4**	87.5	91.3	94.0	88.6	**7.61**	10.62	12.97	14.78	11.50
DTLN	1.38	1.55	1.77	2.08	1.69	73.9	81.7	87.8	92.7	84.0	4.22	8.50	12.18	15.34	10.07
STFT loss	1.42	1.67	2.03	2.52	1.91	78.2	86.1	91.5	95.3	87.8	6.48	10.80	**14.83**	**18.98**	12.75
MSE loss	1.42	1.71	2.13	2.68	1.98	79.2	87.0	92.0	95.6	88.5	6.00	10.08	13.63	17.23	11.74
Ours	**1.48**	**1.80**	**2.23**	**2.75**	**2.07**	80.4	**88.0**	**92.7**	**96.0**	**89.3**	6.75	**11.04**	14.78	18.54	**12.78**

Table 2: PESQ results for speech signals mixed with different domestic background noises drawn from DEMAND dataset and convolved with different RIR simulated with gpuRIR. Best results are indicated in boldface.

	Kitchen					Living room					Washing machine				
SNR (dB)	0	5	10	15	Avg.	0	5	10	15	Avg.	0	5	10	15	Avg.
Noisy	1.08	1.13	1.26	1.57	1.26	1.05	1.11	1.26	1.62	1.26	1.03	1.05	1.13	1.40	1.15
NSNet-2	**1.52**	1.76	2.13	2.46	1.97	1.20	1.43	1.71	2.04	1.59	1.16	1.34	1.59	1.95	1.51
DTLN	1.23	1.40	1.67	1.99	1.58	1.23	1.48	1.73	2.10	1.64	1.07	1.09	1.18	1.46	1.20
Ours	**1.52**	**1.87**	**2.37**	**2.97**	**2.18**	**1.25**	**1.59**	**2.08**	**2.69**	**1.90**	**1.18**	**1.51**	**1.96**	**2.55**	**1.80**

harmonics and suppress the background noises between them, as explained in Sec. 3.

To compare the performance of the proposed method with other relevant models we have carried out two experiments. First, as depicted in Table 1, we show that our method in non-reverberant environment outperforms DTLN and is in-par with NSNet-2, which have $\times 3$ and $\times 4$, respectively, more computations and memory resources than our proposed model. In addition, in reverberant environments, our method achieves better performance than both methods. Second, since our main application is speech enhancement in domestic scenarios, which are characterized by reverberation and typical noise types, we compare the proposed model to DTLN and NSNet-2 for the three types of background noise for various SNR levels. All three background noise signals were drawn from the DEMAND dataset. As clearly indicated in Table 2, our method outperforms the other methods for all SNR values and all background noise types. We improve the overall PESQ at ('kitchen', 'living room', 'washing machine') by $(0.21,\ 0.31,\ 0.29)$ and $(0.6,\ 0.26,\ 0.6)$ for NSNet-2 and DTLN, respectively. It can be further noticed that our method performs better in the presence of a background noise with an harmonic structure, such as washing machine, and thus learns to better distinguish harmonics patterns, typical to human voice. Note that we only present the PESQ scores due to space constraints. STOI and SI-SDR exhibit similar trends.

5. CONCLUSION

This study introduces a *speech-harmonics weighted loss* for low-resources, online, single microphone speech enhancement task. Applying this method during training improves the overall performance by preserving the speech harmonics and suppressing the background noise between them. To weight the desired objective function we proposed a procedure that first detects the harmonics-dominated bands by seeking for the maximum value of the auto-correlation for each band to obtain a speech-harmonic presence indicator, which is then used to weight the loss function. A comprehensive set of experiments verified that the proposed algorithm outperforms other competing method in three objective quality and intelligibility measures, specifically in a domestic environment, demonstrating its applicability to tiny edge devices and 'stay-at-home' scenarios.

6. REFERENCES

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

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

[3] 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.

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

[5] J. Lee, Y. Jung, M. Jung, and H. Kim, "Dynamic noise embedding: Noise aware training and adaptation for speech enhancement," *arXiv preprint arXiv:2008.11920*, 2020.

[6] M. Kolbæk, Z.-H. 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.

[7] T. Lan, Y. Lyu, W. Ye, G. Hui, Z. Xu, and Q. Liu, "Combining multi-perspective attention mechanism with convolutional networks for monaural speech enhancement," *IEEE Access*, vol. 8, pp. 78979–78991, 2020.

[8] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[9] Chandan K.A. Reddy, Ebrahim Beyrami, Harishchandra Dubey, Vishak Gopal, Roger Cheng, Ross Cutler, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, Puneet Rana, Sriram Srinivasan, and Johannes Gehrke, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework," in *INTERSPEECH*, 2020.

[10] Chandan KA Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "INTERSPEECH 2021 deep noise suppression challenge," in *INTERSPEECH*, 2021.

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

[12] Shlomo E Chazan, Jacob Goldberger, and Sharon Gannot, "Speech enhancement with mixture of deep experts with clean clustering pre-training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 716–720.

[13] David Talkin and W Bastiaan Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.

[14] Adrian Von Dem Knesebeck and Udo Zölzer, "Comparison of pitch trackers for real-time guitar effects," in *Proc. of the 13th Int. Conference on Digital Audio Effects*, 2010.

[15] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[16] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[17] David Diaz-Guerra, Antonio Miguel, and Jose R Beltran, "gpuRIR: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.

[18] Institute of Electrical and Electronics Engineers, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[19] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," in *Proc. Meetings Acoust*, 2013, pp. 1–6.

[20] Karol J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. 2015, pp. 1015–1018, ACM Press.

[21] Sercan Ö Arık, Heewoo Jun, and Gregory Diamos, "Fast spectrogram inversion using multi-head convolutional neural networks," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 94–98, 2018.

[22] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.

[23] Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 871–875.

[24] Sebastian Braun and Ivan Tashev, "Data augmentation and loss normalization for deep noise suppression," in *International Conference on Speech and Computer*. Springer, 2020, pp. 79–86.

[25] Nils L Westhausen and Bernd T Meyer, "Dual-signal transformation LSTM network for real-time noise suppression," *arXiv preprint arXiv:2005.07551*, 2020.

[26] ITU-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.

[27] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper 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.

[28] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "SDR–half-baked or well done?," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.