NOISE-TO-NOISE SPEECH ENHANCEMENT: SPEECH DENOISING WITHOUT CLEAN SPEECH

Yen-Yu Chang *Jing-Cheng Chang* *Hung-Yi Lee*

Department of Electrical Engineering, National Taiwan University
{b03901138, b04901138, hungyilee}@ntu.edu.tw

ABSTRACT

Speech enhancement is an important technique in audio signal processing and is often used as preprocessor in speech-related applications. In this paper, we propose a Noise-to-Noise method in speech denoising. With statistical arguments, we show that theoretically even without using the uncorrupted speech data, it is possible to transform the corrupted speech into its clean counterpart. In experiments, we evaluate our work with signal-to-noise ratio (SNR), mean squared error (MSE), short-time objective intelligibility (STOI), and perceptual evaluation of speech quality (PESQ) metrics. The experimental results indicate that the proposed method outperforms Wiener filtering, and it is also comparable to the typical supervised-learning-based model (Noise-to-Clean) in several conditions. To our best knowledge, this is the first paper investigating Noise-to-Noise speech enhancement.

Index Terms— Speech Enhancement, Noise-to-Noise

1. INTRODUCTION

Over the past few years, machine learning has achieved significant success and has benefited numerous real-world applications. Speech enhancement [1] is especially the important technique in audio signal processing and is often used as preprocessor in speech-related applications, such as automatic speech recognition (ASR), hearing aids, and communication systems. The goal of speech enhancement aims to improve the quality and intelligibility of target speech signals corrupted by background interference. In the past few decades, many classic noise reduction and speech enhancement algorithms such as spectral subtraction [2], minimum mean-square error short-time spectral amplitude (MMSE STSA) estimator [3], Wiener filtering [4], etc., were been proposed. Most of the aforementioned works focused on exploring the statistical difference between speech and noise and required some a priori conditions to be satisfied. However, due to these impractical assumptions, most of these statistical-based approaches often fail to construct good estimators which can generalize to the complex scenarios in the real world.

More recently, neural-network-based deep learning approaches [5, 6, 7] have been proposed and shown considerable improvements in performance. In these works, they utilized the nonlinearity characteristic of neural networks to simulate the mapping from noise data to its clean version. With a large number of pairs of noisy inputs and clean targets, the network learns to restore the clean speech given the noisy speech. In addition, convolutional neural networks (CNN) have also been introduced in this task [8, 9, 10]. CNN can model the spacial and temporal correlations and pay more attentions to neighboring regions. Also, due to its weight-sharing property, CNN can capture the local temporal nature of signal spectrogram more efficiently. To further intensify the performance and ability of speech denoising, adversarial learning has recently captured more attention. Based on generative adversarial networks (GAN) [11], [12] and [13] built in an encoder-decoder structure, which learned a mapping from the noisy speech spectrogram to the enhanced counterpart.

However, these above models trained in a supervised method always need large amounts of clean and unclean data pairs during the training phase. But in reality, sometimes speech may be interfered by environment noise during recording, and hence may not be able to collect the clean data. Inspired by [14], in this study, we propose Noise-to-Noise speech enhancement. With simple statistical arguments, we show that it is possible to recover the corrupted speech signals given only pairs of noisy speech. Moreover, we also do not need the signal priors or likelihood of the corrupted speech. In experiments, we give the quantitative results of our proposed approach. The experiments show that at large signal-to-noise ratio, the performance of Noise-to-Noise model is very comparable or even better than the Noise-to-clean model under the conditions of Gaussian noise and real-world noise data in terms of some evaluation metrics. The idea of Noise-to-Noise has been studied on image processing, but it has not been explored on speech yet. To our best knowledge, this is the first paper investigating Noise-to-Noise speech enhancement.

The paper is organized as follows. We give a formal theoretical background in Section 2. The network structure is depicted in Section 3. The description of experimental setup is presented in Section 4. The results and discussion are presented in Section 5, and in Section 6, the conclusion is given.

2. NOISE-TO-NOISE SPEECH ENHANCEMENT

In typical deep learning based speech enhancement, the network learns to minimize the following objective function $\mathcal{L}_{n2c}$.

$$\mathcal{L}_{n2c} = \mathbb{E}_{(y,x)}[L(f_\theta(y), x)], \tag{1}$$

where f_θ represents the network for speech enhancement, and θ is its parameters. Here x is clean speech, while y is its noisy counterpart. f_θ takes y as input and clean its noise to make $f_\theta(y)$ as close to x as possible. The function $L(.,.)$ in (1) is used to evaluate the difference between $f_\theta(y)$ and x.

In Noise-to-Noise speech enhancement, clean speech x is not available during training, and only noisy speech is available. In such condition, the network is learned to minimize the following objective function $\mathcal{L}_{n2n}$.

$$\mathcal{L}_{n2n} = \mathbb{E}_{(y,y')}[L(f_\theta(y), y')]. \tag{2}$$

Both y and y' are noisy speech corresponding to the same clean speech x which is latent in the task. f_θ learns to output noisy speech y' given another noisy speech y. In this paper, we are going to show that the model learns from (2) can still achieve speech enhancement to some good extent.

Noise-to-Noise speech enhancement does not make sense at the first glance. However, it has been verified to be useful on image enhancement [14]. Assume that y and y' is obtained from x by adding noises. That is, $y = x + n$ and $y' = x + n'$, where n and n' are random variables sampled from noise distributions $\mathcal{N}$ and $\mathcal{N}'$ respectively ($\mathcal{N}$ and $\mathcal{N}'$ do not have to be the same). The mean of noise n' has to be zero, that is, $\mathbb{E}[n'] = 0$ ($\mathbb{E}[n]$ does not have to be zero). With the assumption that $L(.,.)$ is L2 loss which is commonly used in speech enhancement, we have the following derivation.

$$\mathcal{L}_{n2n} = \mathbb{E}_{(y,y')}[(f_\theta(y) - y')^2] \tag{3}$$

$$= \mathbb{E}_{(y,x),n' \sim \mathcal{N}'}[(f_\theta(y) - (x + n'))^2] \tag{4}$$

$$= \mathbb{E}_{(y,x),n' \sim \mathcal{N}'}[((f_\theta(y) - x) - n')^2] \tag{5}$$

$$= \mathbb{E}_{(y,x),n' \sim \mathcal{N}'}[((f_\theta(y) - x)^2 - 2(f_\theta(y) - x)n' + (n')^2)] \tag{6}$$

$$= \mathbb{E}_{(y,x)}[(f_\theta(y) - x)^2] + \mathbb{E}_{n' \sim \mathcal{N}'}[(n')^2] \tag{7}$$

$$= \mathcal{L}_{n2c} + \mathbb{E}_{n' \sim \mathcal{N}'}[(n')^2]. \tag{8}$$

Due to the assumption $\mathbb{E}[n'] = 0$, the term $2(f_\theta(y) - x)n'$ in (6) can be removed. In (8), because n' is independent to network parameter θ, theoretically, $\mathcal{L}_{n2n}$ and $\mathcal{L}_{n2c}$ have the same solution. This implies that we can corrupt the clean training targets with zero-mean noise n' without changing what the neural network learns, so Noise-to-Noise speech enhancement is feasible.

Name	N_{out}	Function
INPUT	1	
CL$_{64,5}$-0	64	Convolution 5 × 5
CBL$_{128,5}$-1	128	Convolution 5 × 5
CBL$_{256,5}$-2	256	Convolution 5 × 5
CBL$_{512,5}$-3	512	Convolution 5 × 5
CBL$_{512,5}$-4	512	Convolution 5 × 5
CBL$_{512,5}$-5	512	Convolution 5 × 5
CBL$_{512,5}$-6	512	Convolution 5 × 5
CBR$_{512,5}$-7	512	Convolution 5 × 5
DCDR$_{512,5}$-7	512	DeConvolution 5 × 5
CONCAT-7	1024	Concatenate output of CBL$_{512,5}$-6
DCDR$_{512,5}$-6	512	DeConvolution 5 × 5
CONCAT-6	1024	Concatenate output of CBL$_{512,5}$-5
DCDR$_{512,5}$-5	512	DeConvolution 5 × 5
CONCAT-5	1024	Concatenate output of CBL$_{512,5}$-4
DCR$_{512,5}$-4	512	DeConvolution 5 × 5
CONCAT-4	1024	Concatenate output of CBL$_{512,5}$-3
DCR$_{256,5}$-3	256	DeConvolution 5 × 5
CONCAT-3	512	Concatenate output of CBL$_{256,5}$-2
DCR$_{128,5}$-2	128	DeConvolution 5 × 5
CONCAT-2	256	Concatenate output of CBL$_{128,5}$-1
DCR$_{64,5}$-1	64	DeConvolution 5 × 5
CONCAT-1	128	Concatenate output of CL$_{64,5}$-0
DCT$_{1,5}$-0	1	DeConvolution 5 × 5

Table 1: Network Architecture

3. IMPLEMENTATION

In this study, we use U-net [15] as our network. U-net is an encoder-decoder structure. Each layer in the encoder downsamples its input to the next layer until there is a bottleneck, and each layer in the decoder operates in the reverse process to upsample its input until it returns to the original shape. Assume N is the number of layers, we use the skip connections to concatenate feature maps of layer i in the encoder to the layer $N - i$ in the decoder.

The detail of our U-net is shown in table 1. Following [16], the details of our network are as follow. CBL$_{l,s}$ denotes a Convolution-BatchNorm-Leaky-ReLU layer with slope=0.2, where l is the number of the filters and $s \times s$ is the filter size. CBR$_{l,s}$ represents the same architecture with ReLU, and CL$_{l,s}$ has the same architecture without BatchNorm. DCDR$_{l,s}$ denotes the DeConvolution-BatchNorm-Dropout-ReLU with 50% dropout rate, and DCR$_{l,s}$ denotes the DeConvolution-BatchNorm-ReLU. We use CONCAT to denote concatenation layer. DCT$_{l,s}$ denotes DeConvolution-tanh. The U-net architecture which combines direct and skip connections is illustrated in Fig 1.

4. EXPERIMENTAL SETUP

4.1. Data Set

We set up the experiments by following the previous work[17, 12]. The speech data was provided by the Voice Bank corpus

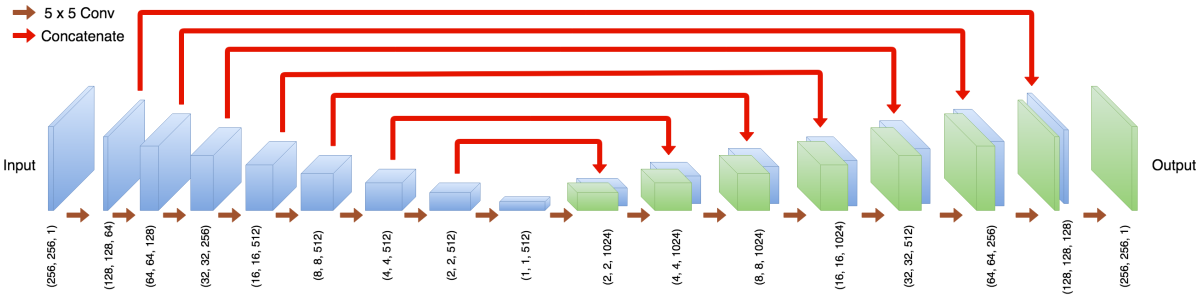


Fig. 1: U-net Architecture.

Dataset		White Gaussian Noise				Pink Gaussian Noise				Brown Gaussian Noise				DEMAND			
		SNR				SNR				SNR				SNR			
Metrics	Model	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB
SNR	NS	0.00	5.00	10.00	15.00	0.00	5.00	10.00	15.00	0.00	5.00	10.00	15.00	0.00	5.00	10.00	15.00
	Wiener	3.22	8.30	13.49	18.39	0.50	5.52	10.48	15.27	0.00	4.97	9.91	14.73	0.25	5.24	10.17	14.98
	N2C	8.44	11.58	14.91	17.39	5.07	8.47	12.13	16.25	5.28	7.98	12.07	15.94	4.49	7.88	11.64	15.65
	N2N	3.25	8.06	12.83	15.90	2.60	7.24	11.04	15.77	2.94	6.58	11.46	15.89	2.98	6.96	11.18	15.16
MSE	NS	0.0667	0.0375	0.0211	0.0119	0.0667	0.0375	0.0211	0.0119	0.0667	0.0375	0.0211	0.0119	0.0493	0.0292	0.0162	0.0092
	Wiener	0.0461	0.0257	0.0142	0.0080	0.0629	0.0353	0.200	0.0115	0.0668	0.0376	0.0213	0.0122	0.0649	0.0365	0.0207	0.0119
	N2C	0.0253	0.0177	0.0120	0.0090	0.0373	0.0290	0.0166	0.0103	0.0363	0.0267	0.0167	0.0106	0.0279	0.0199	0.0128	0.0081
	N2N	0.0459	0.0265	0.0153	0.0107	0.0495	0.0252	0.0187	0.0109	0.0475	0.0313	0.0179	0.0107	0.0341	0.0227	0.0136	0.0085
STOI	NS	0.7038	0.7676	0.8219	0.8617	0.7045	0.7821	0.8379	0.8719	0.8873	0.8976	0.9039	0.9080	0.8365	0.8810	0.9188	0.9429
	Wiener	0.7080	0.7745	0.8297	0.8664	0.7130	0.7905	0.8420	0.8719	0.8852	0.8951	0.9011	0.9048	0.8109	0.8806	0.9152	0.9358
	N2C	0.7222	0.7834	0.8300	0.8626	0.7478	0.8031	0.8459	0.8734	0.8532	0.8814	0.8998	0.9061	0.8109	0.8669	0.9093	0.9393
	N2N	0.6945	0.7723	0.8281	0.8615	0.6906	0.7890	0.8419	0.8747	0.8277	0.8645	0.8936	0.9050	0.8011	0.8691	0.9121	0.9404
PESQ	NS	1.44	1.73	2.08	2.41	1.55	1.90	2.26	2.59	3.04	3.33	3.62	3.91	1.54	1.57	1.53	1.55
	Wiener	1.48	1.79	2.15	2.51	1.61	1.98	2.34	2.67	3.07	3.35	3.63	3.91	1.04	1.00	1.02	1.00
	N2C	1.97	2.30	2.58	2.71	2.07	2.39	2.67	2.95	2.73	3.12	3.53	3.82	1.55	1.56	1.59	1.58
	N2N	1.45	1.79	2.24	2.54	1.54	1.99	2.37	2.70	2.49	2.96	3.23	3.55	1.58	1.57	1.61	1.63

Table 2: SNR, MSE, STOI, and PESQ performance for Noisy Speech (NS), Wiener filter, Noise-to-Clean (N2C), and Noise-to-Noise (N2N).

[18]. We use the subset of the Voice Bank corpus with 30 native English speakers with around 400 sentences each, 28 speakers are used for training and the other 2 speakers are use for testing. The recordings were operated at 48kHz, and we subsampled them to 16kHz.

For training and testing, we used four different noise types: white Gaussian noise, pink Gaussian noise, brown Gaussian noise, and real environment noise. Noise which were added to speech data with four different signal-to-noise ratio (SNR) values: 0, 5, 10, and 15. The environmentally noisy data were supplied by the Diverse Environments Multichannel Acoustic Noise Database (DEMAND) [19]. DEMAND provides recordings with 18 different environmental conditions like kitchen, bus, and park. DEMAND was operated under a 16 channels array at 48kHz, and we merged all the channels and subsampled them to 16kHz.

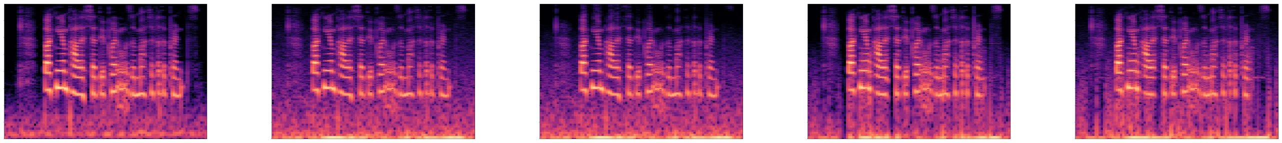
4.2. Preprocessing and Training

We used the log-spectral image of the speech signals as inputs of our work. We computed the time-frequency representation by frame length of 512, with a Hamming window size of 25ms and hop size of 10ms. Only the STFT magnitude was considered. We used the inputs phase for time-domain signal reconstruction. The data were divided into groups of 256 time bins and 256 frequency bins, so the network accepts $256 \times 256 \times 1$ input. Also, before fed into the network, the data are are normalized to be in the range of $[-1, 1]$.

The networks are trained for 30 epochs with the Adam optimizer [20] with the learning rate of 0.0002 and batch size of 16. We use the mean squared error (MSE) as the objective function. The best models are selected by the best performances on the validation set.

4.3. Baseline Methods

We compare the result of our Noise-to-Noise U-net with other two methods we consider as baselines: Wiener filtering and standard Noise-to-Clean U-net. Wiener filtering is a signal processing method, and it is widely used in speech denoising and source separation. The Noise-to-Clean U-net has the same network architecture as in table 1 and Fig 1, but uses noisy speech spectrograms as inputs and maps to clean speech spectrograms.



(a) Clean speech (b) Noisy speech (c) Wiener filtering (d) Noise-to-Clean (e) Noise-to-Noise

Fig. 2: From left to right: clean spectrogram, noisy spetrogram (brown Gaussian noise at 0 dB SNR), spectrogram of the signal enhanced with Wiener filtering, spectorgram of the signal enhanced with Noise-to-Clean, spectrogram of the signal enhanced with Noise-to-Noise.

5. RESULTS AND DISCUSSION

In this study, we compute the following objective metrics to evaluate the quality of the enhanced speech.

- SNR: signal-to-noise ratio (from 0 to ∞).
- MSE: mean squared error in time domain.
- STOI: short-time objective intelligibility [21] (from 0 to 1).
- PESQ: perceptual evaluation of speech quality [22] (from -0.5 to 4.5).

For SNR, STOI, and PESQ, higher value is better, while lower value is better for MSE.

Experimental results are shown in Table 2. In terms of SNR and MSE, the Noise-to-Noise model always outperforms noisy speech and Wiener filtering with pink noise, brown noise, and real environment noises (DEMAND). It is observed that the SNR and MSE scores of Noise-to-Clean model are always better than Noise-to-Noise model regardless of the noise type. However, at 10 dB and 15 dB, the performance of Noise-to-Noise model is very comparable with the performance of Noise-to-Clean model. At 15 dB, our approach only loses to the Noise-to-Clean with average 3.6% in SNR, and only loses with 0.32% in SNR of Brown Gaussian Noise.

In terms of STOI and PESQ, the performance of Noise-to-Clean model is better than Noise-to-Noise model in most cases. But in the high SNR noise condition, Noise-to-Noise performs very comparably to Noise-to-Clean model. Moreover, the performance of Noise-to-Noise is even better than the performance of Noise-to-Clean under real environment noises (DEMAND) with only one exception (STOI, 0dB). The reason why Noise-to-Noise can surpass Noise-to-Clean in some cases is still under investigation.

Figure 2 shows the spectrograms of a noisy utterance (Brown Gaussian noise at 0 dB SNR), together with its clean and enhanced versions with Wiener filtering, Noise-to-Clean, and Noise-to-Noise. It is observed that both the spectrogram enhanced by the Noise-to-Clean approach and Noise-to-Noise approach preserve the structure of the original signal. At the same time, it is observed that Noise-to-Noise model can effectively remove noise components from the noisy utterance better than Wiener filtering.

6. CONCLUSION AND FUTURE WORK

In this work, we investigate the Noise-to-Noise training method which is inspired from image processing study for speech enhancement. By utilizing some statistical arguments, we can corrupt the data with zero-mean noise without changing the neural networks ability, and hence it is possible to train the model in the absence of clean target data. Furthermore, we adapt U-net as our model framework. With only the convolutional and de-convolutional layers, we can effectively reduce the training parameters of the model and also get the promising results. At the experimental phase, we demonstrate our work in comparison with Wiener filtering and the Noise-to-Clean method. It is observed that the result of Noise-to-Noise is comparable to the former and the latter, respectively. Confirm the efficacy of Noise-to-Noise for speech enhancement.

Although the Noise-to-Noise approach does not need clean speech, it needs the same speech corrupted by different noises. Nevertheless, it can be applied on the scenario with multiple microphones or sensors, in which each microphone or sensor records simultaneously, but the recorded speech has different corruption due to different positions of the devices. We are looking for more applications of the Noise-to-Noise model in the real world.

7. REFERENCES

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

[2] 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, April 1979.

[3] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, December 1984.

[4] P. Scalart and J. V. Filho, “Speech enhancement based on a priori signal to noise estimation,” in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, May 1996, vol. 2, pp. 629–632 vol. 2.

[5] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori, “Speech enhancement based on deep denoising autoencoder,” in *INTERSPEECH*, 2013.

[6] Y. Xu, J. Du, L. Dai, and C. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan 2015.

[7] Y. Xu, J. Du, L. Dai, and C. Lee, “An experimental study on speech enhancement based on deep neural networks,” *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Jan 2014.

[8] E. M. Grais and M. D. Plumbley, “Single channel audio source separation using convolutional denoising autoencoders,” in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Nov 2017, pp. 1265–1269.

[9] Szu-Wei Fu, Yu Tsao, and Xugang Lu, “Snr-aware convolutional neural network modeling for speech enhancement.,” in *Interspeech*, 2016, pp. 3768–3772.

[10] Han Zhao, Shuayb Zarar, Ivan Tashev, and Chin-Hui Lee, “Convolutional-recurrent neural networks for speech enhancement,” *CoRR*, vol. abs/1805.00579, 2018.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[12] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, “SEGAN: speech enhancement generative adversarial network,” *CoRR*, vol. abs/1703.09452, 2017.

[13] Daniel Michelsanti and Zheng-Hua Tan, “Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification,” *arXiv preprint arXiv:1709.01703*, 2017.

[14] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila, “Noise2noise: Learning image restoration without clean data,” *CoRR*, vol. abs/1803.04189, 2018.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[16] Ori Ernst, Shlomo E Chazan, Sharon Gannot, and Jacob Goldberger, “Speech dereverberation using fully convolutional networks,” *arXiv preprint arXiv:1803.08243*, 2018.

[17] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi, “Investigating rnn-based speech enhancement methods for noise-robust text-to-speech,” in *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*. 2016, pp. 146–152, ISCA.

[18] Christophe Veaux, Junichi Yamagishi, and Simon King, “The voice bank corpus: Design, collection and data analysis of a large regional accent speech database,” in *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*. IEEE, 2013, pp. 1–4.

[19] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, “The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings,” *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[20] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[21] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4214–4217.

[22] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP’01). 2001 IEEE International Conference on*. IEEE, 2001, vol. 2, pp. 749–752.