2022 International Workshop on Acoustic Signal Enhancement (IWAENC) | 978-1-6654-6867-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IWAENC53105.2022.9914801

TRAINING STRATEGIES FOR OWN VOICE RECONSTRUCTION IN HEARING PROTECTION DEVICES USING AN IN-EAR MICROPHONE

Mattes Ohlenbusch[1], *Christian Rollwage*[1], *Simon Doclo*[1,2]

[1]Fraunhofer Institute for Digital Media Technology IDMT,
Oldenburg Branch for Hearing, Speech and Audio Technology HSA, Germany
[2]Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all,
University of Oldenburg, Germany
Email: mattes.ohlenbusch@idmt.fraunhofer.de

ABSTRACT

In-ear microphones in hearing protection devices can be utilized to capture the own voice speech of the person wearing the devices in noisy environments. Since in-ear recordings of the own voice are typically band-limited, an own voice reconstruction system is required to recover clean broadband speech from the in-ear signals. However, the availability of speech data for this scenario is typically limited due to device-specific transfer characteristics and the need to collect data from in-situ measurements. In this paper, we apply a deep learning-based bandwidth-extension system to the own voice reconstruction task and investigate different training strategies in order to overcome the limited availability of training data. Experimental results indicate that the use of simulated training data based on recordings of several talkers in combination with a fine-tuning approach using real data is advantageous compared to directly training on a small real dataset.

Index Terms— Own voice reconstruction, in-ear microphone, training strategies, data augmentation, domain adaptation

1. INTRODUCTION

In noisy working environments, workers often rely on hearing protection devices. Since such devices do not only attenuate external noise, but also hinder direct speech communication, devices enabling radio communication may present an advantage [1]. One option for recording the own voice of the person wearing such a device is the use of a microphone placed inside of the occluded ear canal. However, the in-ear microphone picks up the own voice at a limited frequency range up to about 2 kHz with different transfer characteristics than a close-talking microphone due to occlusion and body-conduction effects and with body-produced noise (e.g., breathing, heartbeats) [2] . Hence an own voice reconstruction system is required to recover clean broadband speech.

The Oldenburg Branch for Hearing, Speech and Audio Technology HSA is funded in the program ≫Vorab≪ by the Lower Saxony Ministry of Science and Culture (MWK) and the Volkswagen Foundation for its further development. Part of this work was funded by the German Ministry of Science and Education BMBF FK 16SV8811. This work was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project ID 352015383 (SFB 1330 C1).

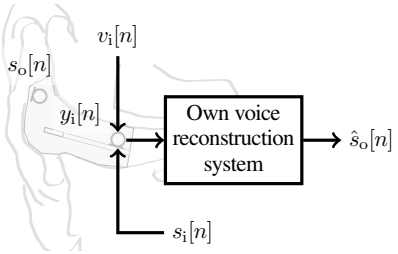

Fig. 1. Illustration of the considered hearing protection device and the own voice reconstruction task, aiming at estimating the clean broadband speech signal $s_\mathrm{o}[n]$ from the noisy and band-limited in-ear microphone signal $y_\mathrm{i}[n]$.

For this task, it has been proposed in [3] to first use an adaptive filter to reduce noise in the in-ear microphone and then apply a basic bandwidth extension (BWE) system to reconstruct high-frequency content. For a similar scenario, in [4] an autoencoder neural network has been applied to directly reconstruct broadband speech from recordings made with a bone-conduction microphone. In [5], a neural network is utilized to compute time-varying filters used to estimate broadband speech from in-ear recordings. Recently, in [6] a multi-modal approach has been investigated using both a bone- and an air-conduction microphone as input signals to a fully convolutional neural network. In [7] a fully convolutional neural network approach in the time-domain has been proposed to estimate clean broadband speech using two in-ear microphones. Similarly, in [8] it has been proposed to utilize a U-Net architecture to enhance bone-conducted signals in the short-time Fourier transform (STFT) domain.

In this paper, we consider the deep learning-based BWE approach proposed in [9], which reconstructs the high-frequency content using a time-domain U-Net, and adapt it to the own voice reconstruction task, i.e. not only reconstructing high-frequency content but also compensating for transfer characteristics and reducing body-produced noise. Since the transfer characteristics are device-specific, in-ear own voice recordings have to be made in-situ with a talker wearing the device, such that the amount of speech data available for training is typically limited. We propose to overcome data shortage by simulating artificial recordings for

use in data augmentation-based training strategies. The simulation framework relies on modeling the transfer characteristics between two device microphones using relative transfer functions (RTFs). We investigate the effects of single- and multi-talker transfer characteristics, the number of RTFs estimated per talker, the influence of body-produced noise, and the usage of real data fine-tuning. Results indicate that employing a training paradigm adopted from BWE is viable to the own voice reconstruction task. Experimental results show that training on simulated in-ear signals can be used to perform reconstruction on recordings of in-ear signals. In particular, pre-training the proposed system with simulated data and fine-tuning it with real data leads to the largest improvement in terms of objective metrics.

2. SIGNAL MODEL

We consider a scenario where a talker is wearing a hearing protection device equipped with a single microphone located at the inside of the occluded ear (see Fig. 1). Since a large component of the own voice speech is transmitted through bone and cartilage [2], the speech captured by the in-ear microphone exhibits different characteristics than speech captured by a microphone outside of the talkers' body (e.g., a close-talk microphone or a microphone placed at the outer side of the hearing protection device). Most prominently, high-frequency components are heavily attenuated, while low-frequency components are amplified. It is assumed here that the in-ear microphone does not pick up any external noise from outside of the device, but picks up body-produced noise such as breathing sounds. The considered scenario differs from BWE, since the transfer characteristics may vary based on hearing protection device, ear canal characteristics, and body-produced noise may need to be accounted for.

Fig. 1 illustrates the signal model for the own voice reconstruction task. In the absence of external noise, the signal $y_{\mathrm{i}}[n]$ recorded at the in-ear microphone (subscript i is given by

$$y_{\mathrm{i}}[n] = s_{\mathrm{i}}[n] + v_{\mathrm{i}}[n], \tag{1}$$

where n denotes the discrete time index, $s_{\mathrm{i}}[n]$ denotes the own voice speech and $v_{\mathrm{i}}[n]$ denotes the body-produced noise recorded at the in-ear microphone. The objective of own voice reconstruction is to estimate a clean broadband speech signal (as it would be captured by a microphone in front of the talkers mouth) from the band-limited and noisy microphone signal $y_{\mathrm{i}}[n]$. Although own voice captured at the outer microphone (subscript o) does not have the same long-term spectrum as speech recorded from a microphone in front of the talker's mouth, we still assume that they are similar such that in this paper we will aim at estimating the own voice captured at the outer microphone. In this paper, we will use the clean speech signal $s_{\mathrm{o}}[n]$ captured by a microphone at the outer side of the hearing protection device as the desired speech signal. It should be noted that the speech signal $s_{\mathrm{o}}[n]$ is only used for training and evaluation purposes in this paper, but is not available in practice. The speech signals $s_{\mathrm{i}}[n]$ and $s_{\mathrm{o}}[n]$ are related to each other by a linear, time-varying transfer characteristic (due to body transmission and mouth movements).

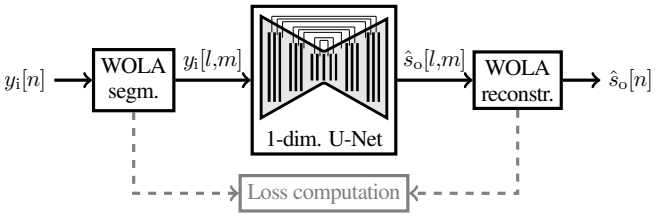

Fig. 2. Overview diagram of the considered own voice reconstruction system.

3. OWN VOICE RECONSTRUCTION SYSTEM

In this section, we propose a deep learning-based own voice reconstruction system based on the BWE system in [9]. In Section 3.1, we describe the used processing framework and network architecture. In Section 3.2, we describe a small dataset recorded using the target device. To overcome the limited availability of in-ear recordings, in Sections 3.3 and 3.4 we propose procedures to simulate in-ear recordings. In Section 3.5, we discuss training strategies for the proposed system.

3.1. Processing framework

The overall system is illustrated in Fig. 2. Following [9], it uses a weighted overlap-add (WOLA) framework, where the in-ear microphone signal $y_{\mathrm{i}}[n]$ is first segmented in segments $y_{\mathrm{i}}[l,m]$ of length P, where l and m denote the segment and segment time index. Each segment $y_{\mathrm{i}}[l,m]$ is then processed by the U-Net to estimate the clean speech segment $s_{\mathrm{o}}[l,m]$ at the outer microphone. The estimated segments $\hat{s}_{\mathrm{o}}[l,m]$ are then reconstructed to obtain $\hat{s}_{\mathrm{o}}[n]$. The WOLA segmentation and reconstruction is carried out using a sqrt-Hann window and an overlap of $\frac{P}{2}$ samples.

In this work, a U-Net architecture as described in [9] is utilized. This architecture has been used for BWE in [9], but has also previously been applied to noise reduction [10, 11]. Input frames consisting of $P = 2048$ samples are processed in the time domain, at a sampling frequency of $f_s = 16$kHz. The input convolutional layer in the encoder part increases the number of channels from 1 to 64, and subsequent layers decrease the sampling frequency towards the bottleneck while incrementally increasing the amount of filters. All filters have a length of 11 samples. The decoder part of the network after the bottleneck mirrors the encoder part. From each encoding layer, a skip-connection towards the corresponding decoder layer is added. The skip-connections are realized by concatenating the additional channels from the encoder to the output signals of the decoder layer. A parametric rectifier linear unit (PReLU) activation is utilized after each layer except for the last decoder layer, where a linear activation is employed. The U-Net has 10.2M parameters.

3.2. Real Dataset

A small dataset of own voice recordings was obtained with 14 different talkers (4 female, 10 male) wearing the closed-vent variant of the commercially available Hearpiece device [12] in each ear. The microphones used are the in-ear microphone and the outer microphone near the concha. Speech was recorded at

these microphones while the talkers are reading a German text out loud in a sound-proof listening booth. The talkers were seated during recording, so that body-produced sounds from movement are not expected. The overall size of the dataset is approximately thirty minutes, which apparently is not sufficiently large to train the proposed system and obtain satisfactory results (see Section 4.2). In addition, recordings of body-produced noise were gathered from each participant wearing the devices while being silent.

3.3. Simulated Dataset

Data augmentation strategies may help to overcome limitations imposed by small amounts of available training data [13]. We therefore propose to train on a larger speech dataset by simulating data. In order to easily perform data augmentation, we will approximate this transfer characteristic as time-invariant:

$$\tilde{s}_{\mathrm{i}}[n] = \tilde{h}[n] * s_{\mathrm{o}}[n], \tag{2}$$

where $*$ denotes the convolution operator, $\tilde{s}_{\mathrm{i}}[n]$ denotes the approximated own voice speech component at the in-ear microphone and $\tilde{h}[n]$ denotes the relative impulse response (ReIR) between the outer and in-ear microphone, which corresponds to the RTF in time domain.

In-ear own voice signals are simulated according to the signal model in (2) by convolving broadband speech with ReIRs estimated from the real dataset and adding body-produced noise $v_{\mathrm{i}}[n]$ recorded at the in-ear microphone from the same talkers as in the real dataset, i.e.

$$\tilde{y}_{\mathrm{i}}[n] = \tilde{h}[n] * s_{\mathrm{o}}[n] + \alpha \cdot \tilde{v}_{\mathrm{i}}[n], \tag{3}$$

where the scaling factor α determines the signal-to-noise-ratio (SNR) of the simulated in-ear signal.

The VCTK corpus containing approximately 44 hours of recordings is used as source material for broadband speech [14]. We investigate different options for the ReIR estimates $\tilde{h}[n]$ between the outer microphone and the in-ear microphone, which correspond to RTF estimates in the frequency domain, and the body-produced noise $\tilde{v}_{\mathrm{i}}[n]$ in (3):

- RTF estimated using recordings from a single talker (1T) vs. RTFs estimated using recordings from all talkers (14T)
- Estimation of a single RTF from a single utterance per talker (s-RTF) vs. estimation of RTFs from a multiple utterances per talker through temporal segmentation (m-RTF)
- additive randomly chosen body-produced noise segments, scaled to achieve an SNR randomly varied in [10, 60] dB between the in-ear speech and the body-produced noise, included vs. not included

3.4. RTF Estimation

First, the own voice recordings are divided into individual utterances using an energy threshold of -30 dB re. maximum peak value per recording for voice activity detection. Then, STFTs of the microphone signals are computed using a STFT frame size of $N = 256$ samples, a Hann window and an overlap of $\frac{N}{2}$ samples between frames. For each utterance, power spectral density (PSD) estimates are obtained using the Welch method [15] from the STFTs $Y_{\mathrm{i}}(k,l)$ and $S_{\mathrm{o}}(k,l)$ where k and l denote the STFT frequency and frame indices, L denotes the number of frames in an utterance, which varies between utterances, and $\cdot^{\dagger}$ denotes the complex conjugate:

$$\Phi_{\mathrm{i,o}}(k) = \frac{1}{L} \sum_{l=0}^{L-1} Y_{\mathrm{i}}(k,l) \cdot S_{\mathrm{o}}^{\dagger}(k,l) \tag{4}$$

$$\Phi_{\mathrm{o}}(k) = \frac{1}{L} \sum_{l=0}^{L-1} |S_{\mathrm{o}}(k,l)|^2. \tag{5}$$

Here, $\Phi_{\mathrm{i,o}}(k)$ is the cross-PSD between the in-ear and outer microphone signals, $\Phi_{\mathrm{o}}(k)$ is the PSD of the outer microphone signal. The relative transfer function $\tilde{H}(k)$ is then estimated as

$$\tilde{H}(k) = \frac{\Phi_{\mathrm{i,o}}(k)}{\Phi_{\mathrm{o}}(k)} \tag{6}$$

and the corresponding ReIR $\tilde{h}[n]$ used to generate simulated data is obtained by performing an inverse Fourier transform of the RTF.

However, due to changes in the speech excitation mechanism, it is highly likely that the transfer path changes over time. For this reason, speech RTFs are only estimated on individual utterances. In case of the s-RTF option, only the longest utterance is selected from which a single RTF is estimated. In case of the m-RTF option, all utterances with length over 1 second are used to estimate multiple RTFs.

3.5. Training Strategies

The U-Net (see Section 3.1) is trained using a batch size of 32 examples per batch, where each example is a single segment of P samples from an utterance randomly chosen from the dataset. Audio input and target signals are normalized to zero mean and unit variance for each recording. The U-Net is trained using the combined time- and phase constrained magnitude (T-PCM) loss function as proposed for BWE in [9]. The loss is computed between the output of the network (estimated clean speech) and either the outer microphone signal in case of the real dataset or the original corpus recording in case of training with the simulated dataset. The training is carried out using the Adam optimizer [16] with an initial learning rate of 10^{-4}, up to a maximum of 100 epochs. The learning rate is halved if the validation loss does not improve for 3 epochs, and early stopping is applied after 6 epochs without loss improvement. Dropout with a factor of 0.2 is performed after each three layers during training. The real dataset is split based on the device side: recordings and RTF estimates obtained from the left-side device are used in training and validation, whereas the right-side recordings are used as test subset. The training and validation set is further split into proportions of 0.88 and 0.12, respectively.

Aiming at investigating different training strategies, the U-Net is trained using differently composed datasets. First, we investigate suitability only using the real dataset for training. This dataset has the advantage of closely resembling the own voice reconstruction scenario, but has the drawback of limited data availability.

Second, we consider several variants of using the much larger simulated dataset for training. Since the signal model used to generate the simulated data is only an approximation, differences

between simulated and real data may lead to a lower performance than when training with the same amount of data from the real scenario. Finally, we perform a pre-training of the network on the simulated dataset, and then similarly to [17] fine-tune only the decoder weights on the real dataset using an initial learning rate of $5 \cdot 10^{-5}$. It is hypothesized that the encoding features required for own voice reconstruction may be learned from the simulated dataset, and the fine-tuning procedure enables the decoder to better approximate the clean own voice speech at the outer microphone, which is not available during inference.

4. EXPERIMENTAL EVALUATION

In this section, we compare the reconstruction performance of the proposed deep learning-based own voice reconstruction system using different training strategies. Additionally, we compare the results to our re-implementation of recently proposed single-channel sinc-dilated fully convolutional network (SDFCN) from [7], which is trained on our real multi-talker dataset instead of the single-talker dataset from the original publication.

4.1. Evaluation Procedure and Performance Metrics

For the experimental evaluation, we utilize the test subset of the real dataset. Speech recordings are cut to 10 seconds. To assess the own voice reconstruction performance, typical performance metrics used for bandwidth extension and speech enhancement tasks are considered. A metric which is often used to evaluate bandwidth extension systems is the log-spectral distance (LSD) [18]. Additionally, we consider the wideband (WB) setting of the perceptual evaluation of speech quality (PESQ) metric [19] and the short-time objective intelligibility index (STOI) [20]. We use the outer microphone signal, assuming to be only own voice speech, as the reference signal for the performance metrics. For all measures except LSD, a higher score indicates a better performance. LSD is computed on STFT spectra with a frame size of 2048 samples as in [21].

4.2. Reconstruction Performance

Table 1 shows the experimental results in terms of the considered objective metrics for the unprocessed in-ear microphone signal $y_{\mathrm{i}}[n]$ and the processed signal $\hat{s}_{\mathrm{o}}[n]$ using either the baseline SDFCN [7] or the U-Net for different training strategies.

Here, [R] denotes training on the real dataset, [S] denotes training on a simulated dataset without added body-produced noise, [S+] denotes training on a simulated dataset with added body-produced noise, and [S+R] indicates pre-training on simulated data with added body-produced noise and fine-tuning the encoder on real data. The RTF options are described in Sections 3.3 and 3.4.

First, it can be observed that both the SDFCN and the U-Net system trained on the real dataset yield improvements over the unprocessed input signal. Compared to the baseline SDFCN, the U-Net with a larger network size yields a higher PESQ score and a lower LSD score, but also a lower STOI score.

When the U-Net is trained with simulated data only, a performance decrease can be observed with respect to using real training data in terms of PESQ, and partly in terms of LSD and STOI. For the single-talker, single-RTF training condition, the results in terms of STOI and PESQ are actually worse than for the systems trained with real data. This can probably be attributed to the fact that in this case the U-Net only compensates the static transfer function of a single talker, which does not correspond to real recordings with different and time-varying transfer characteristics. When considering additive body-produced noise and multiple RTFs from one talker in the training dataset, the STOI and PESQ scores slightly improve, but the LSD score degrades compared to the 1T, s-RTF condition. The only training condition where all metrics are improved is the condition where the in-ear microphone signals are simulated using multiple RTFs from multiple talkers. However, it should be realized that even for the 14T, m-RTF training condition the STOI and PESQ scores are still worse than for the systems trained with real data, showing that the assumed signal model in (3) is probably not realistic enough. Finally, it can be observed that the training paradigm utilizing the simulated dataset for pre-training and the real dataset for fine-tuning yield large improvements in terms of all metrics compared to both the systems trained on only real data and the systems trained on only simulated data.

Table 1. Mean results for the unprocessed in-ear microphone signal, the baseline SDFCN system and the proposed U-Net system for different training strategies. Best performance is highlighted in bold.

System	Data	RTFs used	LSD / dB	PESQ	STOI
unproc.	-	-	2.51	1.31	0.79
SDFCN	[R]	-	1.53	1.47	0.74
U-Net	[R]	-	1.48	1.64	0.73
U-Net	[S]	1T, s-RTF	1.35	1.18	0.70
U-Net	[S+]	1T, s-RTF	1.54	1.19	0.69
U-Net	[S+]	1T, m-RTF	1.51	1.26	0.74
U-Net	[S+]	14T, m-RTF	1.24	1.36	0.72
U-Net	[S+R]	14T, m-RTF	**1.05**	**1.80**	**0.83**

Informal listening experiments confirm the signal quality predictions. It should however be noted that while the band-limitation of the in-ear signals appears to be accounted for by the pre-trained and fine-tuned system, there remains an audible difference between the target and the processed signals, probably since the proposed system is unable to account for individual differences in own voice transmission characteristics.

5. CONCLUSION

In this paper, we have investigated several training approaches for own voice reconstruction from band-limited noisy in-ear microphone recordings. We have proposed a method to simulate in-ear data by utilizing relative transfer functions between an outer and an in-ear microphone of a hearing device. Experimental results demonstrate a performance improvement from using simulated data in a pre-training approach. For pre-training, the device transfer characteristics seem to be best approximated using a multi-talker, multi-RTF simulation strategy. In future work, the influence of individual and device-specific own voice transmission factors and external noise will be further investigated.

6. REFERENCES

[1] S. Nordholm, A. Davis, P. C. Yong, and H. H. Dam, "Assistive listening headsets for high noise environments: Protection and communication," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Apr. 2015, pp. 5753–5757.

[2] R. E. Bouserhal, A. Bernier, and J. Voix, "An in-ear speech database in varying conditions of the audio-phonation loop," *The Journal of the Acoustical Society of America*, vol. 145, no. 2, pp. 1069–1077, Feb. 2019.

[3] R. E. Bouserhal, T. H. Falk, and J. Voix, "In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 1321–1331, Mar. 2017.

[4] H.-P. Liu, Y. Tsao, and C.-S. Fuh, "Bone-conducted speech enhancement using deep denoising autoencoder," *Speech Communication*, vol. 104, pp. 106–112, Nov. 2018.

[5] H. Park, Y.-S. Shin, and S.-H. Shin, "Speech Quality Enhancement for In-Ear Microphone Based on Neural Network," *IEICE Trans. on Information and Systems*, vol. 102, no. 8, pp. 1594–1597, 2019.

[6] C. Yu, K.-H. Hung, S.-S. Wang, Y. Tsao, and J.-W. Hung, "Time-Domain Multi-Modal Bone/Air Conducted Speech Enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 1035–1039, 2020.

[7] C.-L. Liu, S.-W. Fu, Y.-J. Li, J.-W. Huang, H.-M. Wang, and Y. Tsao, "Multichannel Speech Enhancement by Raw Waveform-Mapping Using Fully Convolutional Networks," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 28, pp. 1888–1900, 2020.

[8] Y. Li, Y. Wang, X. Liu, Y. Shi, and S.-F. Shih, "Enabling Real-time On-chip Audio Super Resolution for Bone Conduction Microphones," *arXiv preprint arXiv:2112.13156*, 2021.

[9] H. Wang and D. Wang, "Towards Robust Speech Super-Resolution," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 2058–2066, 2021.

[10] A. Pandey and D. Wang, "A New Framework for CNN-Based Speech Enhancement in the Time Domain," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, July 2019.

[11] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On Loss Functions for Supervised Monaural Time-Domain Speech Enhancement," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 28, pp. 825–838, 2020.

[12] F. Denk, M. Lettau, H. Schepker, S. Doclo, R. Roden, M. Blau, J.-H. Bach, J. Wellmann, and B. Kollmeier, "A one-size-fits-all earpiece with multiple microphones and drivers for hearing device research," in *Proc. AES International Conference on Headphone Technology*, San Francisco, USA, Aug. 2019, pp. 1–9.

[13] Q. Wen, L. Sun, F. Yang, X. Song, J. Gao, X. Wang, and H. Xu, "Time series data augmentation for deep learning: A survey," *arXiv preprint arXiv:2002.12478*, 2020.

[14] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, Gurgaon, India, Nov. 2013, pp. 1–4.

[15] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Trans. on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

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

[17] D. Cheng and E. Y. Lam, "Transfer Learning U-Net Deep Learning for Lung Ultrasound Segmentation," *arXiv preprint arXiv:2110.02196*, 2021.

[18] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.

[19] International Telecommunications Union (ITU), "ITU-T P.862, Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *International Telecommunications Union*, Feb. 2001, Geneva, Switzerland.

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

[21] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio Super-Resolution using Neural Networks," in *5th International Conference on Learning Representations (ICLR) 2017*, Toulon, France, 2017, pp. 1–8.