A Cascaded Speech Enhancement for Hearing Aids in Noisy-Reverberant Conditions

Xi Chen[1,2*], *Yupeng Shi*[2], *Wei Xiao*[2], *Meng Wang*[2], *Tingzhao Wu*[2], *Shidong Shang*[2]
Qinglin Meng[3], *Nengheng Zheng*[1]

[1]The Guangdong Key Laboratory of Intelligent Information Processing, College of Electronic and Information Engineering, Shenzhen University, Shenzhen, China
[2]Tencent Ethereal Audio Lab, Shenzhen, China
[3]Acoustics Lab, South China University of Technology, Guangzhou, China

1900432053@email.szu.edu.cn
{yupengshi, denniswxiao, markuswang, tingzhaowu, simeonshang}@tencent.com
mengqinglin@scut.edu.cn, nhzheng@szu.edu.cn

Abstract

Hearing aid users suffer from poor listening experiences under noise and reverberation. This paper introduces a cascaded speech enhancement system to improve the intelligibility and perception of hearing impairments in noisy-reverberant environments. The system consists of three main parts: a deep learning-based noise reduction, a weighted prediction error-based dereverberation, and a personalized dynamic equalization. The proposed enhancement method is in cooperation with a hearing aid simulator for objective and subjective evaluations. In terms of modified binaural short-time objective intelligibility (MBSTOI), the proposed method outperforms the baseline on the test dataset in different noisy and reverberant conditions. The subjective listening test shows that our scheme obtains a lower word recognition rate under noise and speech interference than other teams.

Index Terms: noisy-reverberant speech, speech intelligibility, hearing aid, hearing loss, deep learning

1. Introduction

Over 20 percent of the global population have hearing loss (HL) problems, according to World Report on Hearing in 2021 [1]. Hearing aid (HA) is the primary clinical intervention for hearing impairments (HIs). However, the most common complaint from HA wearers is that they struggle to understand speech in the presence of noise and reverberation [2]. Therefore, it is worthwhile exploring denoising and de-reverberant algorithms to improve speech clarity in numerous environments for HA devices [3].

Over the past decades, many classical speech enhancement (SE) approaches including Wiener filter [4], Karhunen-Loève transform [5, 6], and et al., have been implemented for HA. Such unsupervised SEs significantly improve the output signal-to-noise ratio (SNR) and intelligibility of the HA [7, 8]. However, due to the diverse acoustic conditions in real applications, these conventional methods usually fail due to their inability to deal with non-stationary noise and severe reverberations [9].

Recent advances in deep learning (DL) have demonstrated its potential in the SE for hearing aids. For example, SEs based on the ideal binary masking (IBM) or ideal ratio masking (IRM), in which deep neural networks (DNN) were built for masking gains estimation, have been developed to improve the speech quality and intelligibility for HIs [10, 11]. For example, a two-stage DNN structure is utilized to leverage spectro-temporal information in [10]. The subjective tests demonstrate its significant intelligibility improvements for HI listeners. Thanks to the powerful learning ability of the DNN, the superiority of DNN-based SEs over the classical unsupervised ones is evident in objective and subjective evaluations, especially for non-stationary noises. Nevertheless, considering the diversity of acoustic interferences, in particular the reverberation, more innovative approaches are needed to fulfill the technical challenges in practical environments.

This paper proposes a cascaded speech enhancement method for tackling denoising and dereverberation problems defined in the First Clarity Enhancement Challenge (CEC1) [12]. A modified deep complex convolution recurrent network (MDCCRN) [13]is employed to suppress noise, followed by a multi-channel weighted prediction error (WPE) [14] for dereverberation. Besides, the personalized dynamic equalization scheme is utilized to compensate for individual HL. For the objective evaluation, the enhanced signals are processed by a HA simulator and an HL module to compute the modified binaural short-time objective intelligibility (MBSTOI) [15]. In the subjective listening test stage, the equalized audios were evaluated by mean word identification rate of each HI subject.

2. Proposed System

In this section, we will discuss the signal model, neural network-based denoising system and its training, multi-channel dereverberation, and dynamic equalization used in our system.

2.1. Signal model

As shown in Figure 1, let $s(t)$ and $n(t)$ denote the target and point source interfere signal, $h_1(t)$ and $h_2(t)$ represent the room impulse responses (RIRs) of target and interfere respectively, $y(t)$ denotes the monaural noisy and reverberant signal which the listener receives:

$$y(t) = s(t) * h_1(t) + n(t) * h_2(t) \tag{1}$$

where $*$ denotes convolution. More scenario details could be found in [12].

*This work was done when the author worked as intern at Tencent.

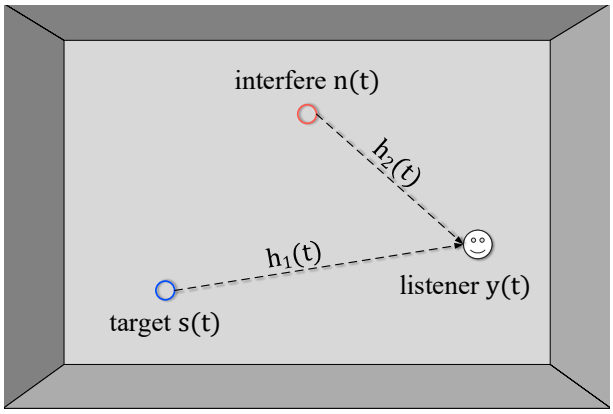


Figure 1: *Scenario diagram (top view).*

2.2. System overview

As illustrated in Figure 2, a cascaded speech enhancement system including MDCCRN and WPE is implemented to remove the noise and reverberant components in tandem.

The CEC1 proposed two stages to evaluate the objective and subjective intelligibility of submitted entries, respectively. In the objective stage, the enhanced signal is processed by the HA simulator and the HL simulator to compute the MBSTOI scores. Besides, dynamic equalization is utilized for listening perception improvement. The equalized gain of each frequency band is calculated on individual binaural pure-tone air-conduction audiograms.

2.3. NN-based denoiser

The NN-based denoiser is a built-in MDCCRN that introduces a complex encoder-decoder architecture with long-short term memory (LSTM) layers as the bottleneck layer. As depicted in Figure 3, MDCCRN consists of six complex conv2D/deconv2D layers. Besides, the complex LSTM layer is implemented to capture the temporal dependencies of the encoder outputs. The complex convolution $(S * W)$ illustrated in the dotted box of Figure 3 can be formulated as:

$$F_{out} = (S_r * W_r - S_i * W_i) + j(S_r * W_i + S_i * W_r) \quad (2)$$

where W_r and W_i are the real and imaginary kernel weights of the complex convolutional layers, respectively. Similarly, the complex operations in LSTM and FC layer are the same as Eq. (2).

Furthermore, skip connections (concatenation) and batch normalization (BN) are used to stabilize the model training. In this work, a casual DCCRN model configuration is designed for real-time applications. Compared with the original DCCRN, the proposed MDCCRN employs unidirectional LSTM layers and smaller convolutional filter depths for conv2D/deconv2D layers with causal padding.

The training target of MDCCRN is a complex ratio mask (CRM) which is optimized by signal approximation. Given the complex-valued spectrogram of the noisy input Y, the estimated output X can be computed as:

$$X = Y_{mag} \cdot M_{mag} \cdot e^{Y_{phase}+M_{phase}} \quad (3)$$

where Y_{mag} and Y_{phase} denote the magnitude and phase of Y respectively. Similarly, the Polar coordinate representation of the CRM is $M = M_{mag} \cdot e^{M_{phase}}$.

To further enhance the denoised output, we apply a post-filter to the estimated CRM magnitude part M_{mag} [16]. The postfilter introduces a wrapped gain as:

$$\widetilde{M}_{mag} = M_{mag} \times sin(\frac{\pi}{2} M_{mag}) \quad (4)$$

where $\widetilde{M}_{mag}$ approaches 0 (noisier band) denotes de-emphasize noise-domain frequency bins. Besides, a global gain compensation is applied to avoid over-attenuating the enhanced output signal:

$$G = \sqrt{\frac{(1+\beta)\frac{E_0}{E_1}}{1+\beta(\frac{E_0}{E_1})^2}} \quad (5)$$

where E_0 is the energy of the enhanced signal using M_{mag} and E_1 is the enhanced signal energy using the wrapped gain $\widetilde{M}_{mag}$. In this paper, β is set to be 0.02 as in [16]. Therefore, the final output could be formulated as:

$$X = G \cdot Y_{mag} \cdot \widetilde{M}_{mag} \cdot e^{Y_{phase}+M_{phase}} \quad (6)$$

2.4. WPE based dereverberation

This method is a delayed linear prediction-based technique, which only models the late reverberation into an auto-regressive (AR) process and leaves early reflections of the speech signal in the prediction residual. To account for the time-varying characteristics of speech, the statistical model-based approach [17] iteratively estimates the time-varying speech variance and normalizes the linear prediction with this speech variance.

2.5. Post-processing module

The post-processing module is designed to simulate the listening perception of HA users. This module includes the HA simulator and the baseline HL simulator to simulate the signal processing in hearing aids and impaired ears. Two options of HA simulators are optimized to improve objective intelligibility and subjective listening perception, respectively:

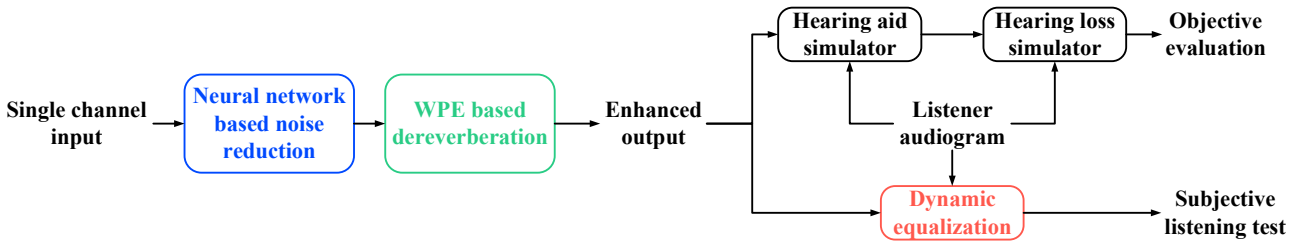


Figure 2: *Schematic diagram of the proposed system.*

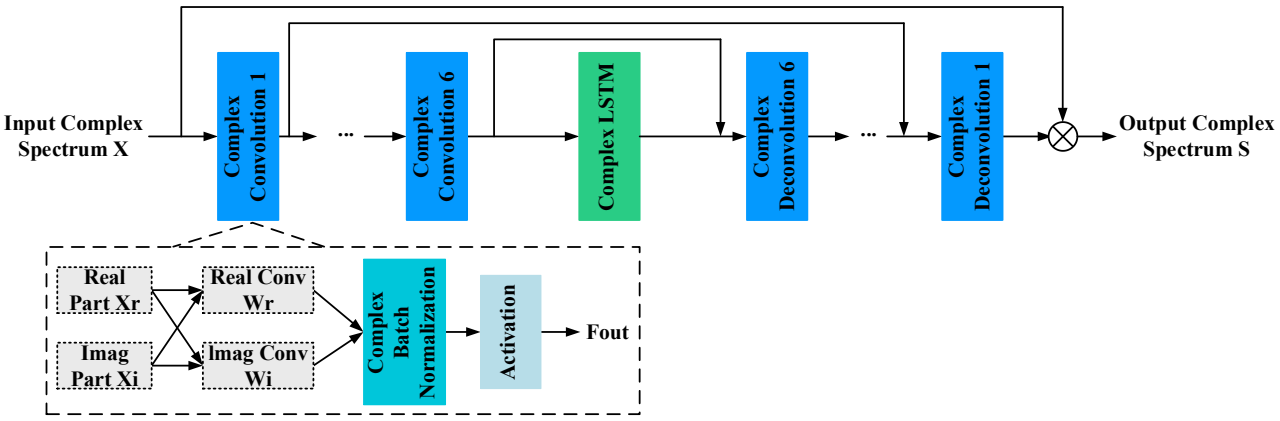


Figure 3: *Block diagram of MDCCRN-based denoising.*

Table 1: *Fitting strategy for dynamic equalization.*

SPL(dB)	HLT(dB)	Gain(dB)
$\leq 40dB$	$0dB \sim 20dB$	0
	$20dB \sim 60dB$	$HLT - 20$
	$\geq 60dB$	$0.5 \times HLT + 10$
$40dB \sim 65dB$	$0dB \sim 20dB$	0
	$20dB \sim 60dB$	$0.6 \times (HLT - 20)$
	$\geq 60dB$	$0.8 \times (HLT - 23)$
$65dB \sim 90dB$	$0dB \sim 40dB$	0
	$\geq 40dB$	$0.1 \times (HLT - 10)^{1.4}$

SPL: sound pressure level; HLT: hearing loss threshold

Option-1: the baseline HA algorithm offered by CEC1. The baseline fitting strategy is based on the Camfit compressive algorithm [18], which calculates compression ratios for eight subbands with center frequencies at [177, 297, 500, 841, 1414, 2378, 4000, 6727] Hz. The gains differ from each other according to their varying bilateral pure-tone audiograms (PTA). The HA module involves the openMHA configurations [19] which could fit the HIs' dynamic range and take advantage of the spatial cues from two microphones of HA devices.

Option-2: a dynamic equalization modifying the spectrum of the enhanced speech. This scheme is derived from the strategy in Figure 6 of [20], which employs a loudness normalization rationale. As shown in Table 1, the equalization gains are calculated based on the sound pressure level and the hearing loss threshold at each frequency band defined at Camfit compressive algorithm.

To simulate impaired hearing, the baseline HL module is developed by the Auditory Perception Group at the University of Cambridge [20]. For subjects with HI, the main symptoms are their reduced dynamic range and low resolution on temporal and frequency information. The module simulated those attenuated mechanisms with loudness recruitment and spectral smearing processes.

3. Experiments

3.1. Dataset

In this experiment, the target signals are from the British National Corpus recorded by 40 speakers [21], while speech interfere data come from Open-source Multi-speaker Corpora of the English Accents in the British Isles [22]. The noise interferences are a collection mainly from Freedsound [23] database. The binaural room impulse responses (BRIRs) are created in Real-time acoustic simulation software [24]. All the samples are stereo utterances with the sampling rate at 44.1 kHz and SNRs ranging from -6 to 12 dB. All the audios in the training dataset are downsampled to 16 kHz to extract input features for neural network training because the MBSTOI metric only focuses on the envelope below 5 kHz. Specifically, 2000 scenes in the development dataset are randomly selected for validation. Finally, mean MBSTOI scores are computed using the rest 500 scenes (3 listeners per scene) from the development dataset.

3.2. Settings

For all the schemes, the window length and hop size are 32 ms and 20 ms, and the FFT length is 512. The convolution channels of the encoder and decoder for the casual MDCCRN are {16-32-64-128-128-128} and {128-128-128-64-32-16}. The kernel size, stride, and padding are modified to (5,1), (2,1), and (0,0) for casual operations. The complex LSTM between the encoder and decoder consists of two unidirectional LSTM layers with 128 hidden units. Except for the final output layer, BN and complex LeakyReLU are implemented to all the hidden layers. AdaBelief optimizer [25] with $\{lr = 0.002, \varepsilon = 1 \times 10^{-12}, \beta = (0.9, 0.999)\}$ and CosineAnnealingLR [26] with $\{T_{max} = 25, \eta_{min} = 4 \times 10^{-8}\}$ are used to optimize the model for minimizing the objective loss function:

$$L = L_{SI-SNR} + L_{STOI} \tag{7}$$

where L_{SI-SNR} is the SI-SNR loss in [13] and L_{STOI} is the STOI loss obtained using torch-stoi [27].

3.3. Evaluation

The MBSTOI, a binaural intelligibility metric based on short-time objective intelligibility (STOI) [28], is employed for objective evaluation. Three candidate schemes are evaluated.

1) NN: a DNN-based speech enhancement system is applied to mixture input for denoising and de-reverberation.

2) NN-WPE: a DNN-based model followed by WPE to enhance the noisy-reverberant input. The DNN model first suppresses the noise, and then WPE removes the late reverberation to obtain the anechoic output.

3) WPE-NN: WPE followed by a DNN-based model, in

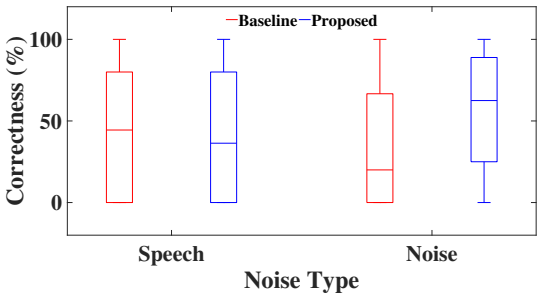


Figure 4: *Word correctness in different noise types of the proposed system, compared with the CEC1 baseline in subjective listening test.*

which the WPE and DNN are the same as that in 2). That is, the DNN-based model removes the noise from the de-reverberant output processed by WPE.

Subjective assessment is based on mean words identified correctness rate (i.e., correctness) of 27 HIs who completed the listening test conducted by the CEC1 organization.

3.4. Results and discussion

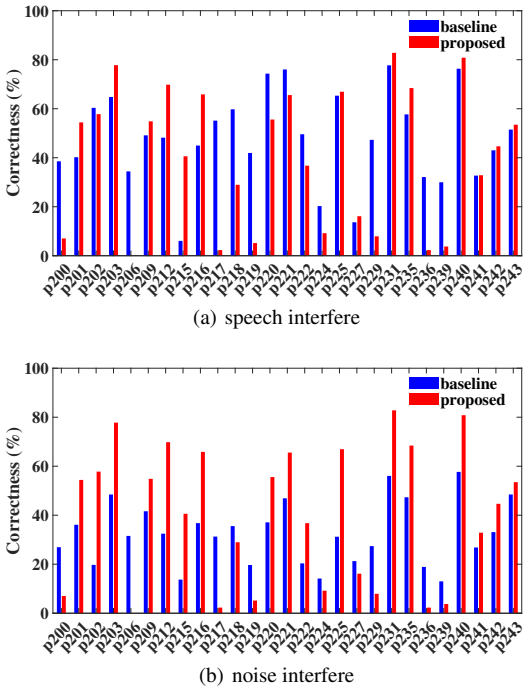
Table 2: *Mean Speech intelligibility results on the dev set. The bold value indicates the best-performing algorithm.*

Methods	MBSTOI
Baseline	0.53
NN	0.54
NN-WPE	**0.60**
WPE-NN	0.55
Clean	0.71

Table 2 gives the MBSTOI scores for different methods., where the 'Clean' represents the target signals, the 'Baseline' represents enhanced speech processed by the system baseline (provided by the CEC1 organizer). As shown, the 'NN-WPE' is the best (score 0.60) one in terms of the MBSTOI score. Therefore, the 'NN-WPE' enhanced signals are submitted to the stage of objective evaluation ($E004$). The casual model is trained for 80 epochs and the well-trained model with the best validating results is used for evaluation. Besides, the computational complexity is about 39 MFLOPs. The one-frame processing time of our PyTorch implementation of MDCCRN is approximately 1ms tested empirically on an Intel i7-9750U PC.

Based on the 'NN-WPE' enhancement module, two options of HA schemes are compared with objective metrics. The mean score of equalized signals is slightly lower (0.57) than those processed with the baseline HA module. However, it sounds more natural and steadier. Therefore, the equalized signals are submitted to the subjective evaluation stage ($E018$).

Figure 4 shows the median score and distribution of the proposed system between speech and noise interference compared with the baseline. The results show that the proposed method is better than the baseline in background noise conditions but worse in competing speaker scenes. The reason is that the DCCRN incorporated in our system focuses on noise reduction but ignores speech separation. Considering the hearing difference between individuals, Figure 5 shows the correctness among 27 HI listeners under two noise types, i.e., speech and noise interference. Concerning the speech interference, 12 listeners perform better with the proposed enhancement algorithms; the correctness of 17 listeners is higher than the baseline



(a) speech interfere

(b) noise interfere

Figure 5: *Word correctness comparison of each listener in different noise types.*

in background noise. These results show that the enhancement and equalization could be effective in specific circumstances for HIs. However, there are still 8 participants who could not benefit from our system under both noise-related conditions. The reason may be their HL severity, as the correctness is below 50% in the baseline scene.

4. Conclusions

This paper proposed a cascaded speech enhancement scheme for improving speech perception in noisy and reverberant environments. The scheme involves deep learning-based denoising, dereverberation utilized in WPE, and dynamic equalization-based hearing compensation. Both objective and subjective evaluation show that our system could improve speech intelligibility for HIs under noisy-reverberant conditions. However, the relatively poor performance under speech interfere shows the essence of speech separation. Further work includes the investigations of novel approaches for speech enhancement in noise-reverberant conditions (including speech separations) and more effective HA algorithms to compensate for moderate to severe HL.

5. ACKNOWLEDGE

This work is jointly supported by National Natural Science Foundation of China (61771320), Guangdong Key Area R&D Project (No. 2018B030338001). Qinglin Meng and Nengheng Zheng are the corresponding authors.

6. References

[1] World Health Organization, "World report on hearing," 2021.

[2] H. Levitt, "Noise reduction in hearing aids: A review," *Journal of rehabilitation research development*, vol. 38, no. 1, pp. 111–122, 2001.

[3] Y.-H. Lai and W.-Z. Zheng, "Multi-objective learning based speech enhancement method to increase speech quality and intelligibility for hearing aid device users," *Biomedical Signal Processing Control*, vol. 48, pp. 35–45, 2019.

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

[5] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Transactions on Speech Audio Processing*, vol. 9, no. 2, pp. 87–95, 2001.

[6] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on speech audio processing*, vol. 11, no. 4, pp. 334–341, 2003.

[7] Y. H. Lai, Y. Tsao, and F. Chen, "A study of adaptive WDRC in hearing aids under noisy conditions," *Int. J. Speech Lang. Pathol. Audiol*, vol. 1, no. 2, pp. 43–51, 2013.

[8] F. Chen, Y. Hu, and M. Yuan, "Evaluation of noise reduction methods for sentence recognition by mandarin-speaking cochlear implant listeners," *Ear and Hearing*, vol. 36, no. 1, pp. 61–71, 2015.

[9] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," *arXiv preprint arXiv:1703.07172*, 2017.

[10] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 3029–3038, 2013.

[11] E. W. Healy, S. E. Yoho, Y. Wang, F. Apoux, and D. Wang, "Speech-cue transmission by an algorithm to increase consonant recognition in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 136, no. 6, pp. 3325–3336, 2014.

[12] S. Graetzer, M. Akeroyd, J. P. Barker, T. J. Cox, J. F. Culling, G. Naylor, E. Porter, and R. V. Munoz, "Clarity: Machine Learning Challenges to Revolutionise Hearing Device Processing," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2021, Conference Proceedings.

[13] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv preprint arXiv:2008.00264*, 2020.

[14] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, Conference Proceedings, pp. 1–5.

[15] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Communication*, vol. 102, pp. 1–13, 2018.

[16] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy, "A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech," *arXiv preprint arXiv:.04259*, 2020.

[17] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[18] B. Moore, J. Alcántara, M. Stone, and B. Glasberg, "Use of a loudness model for hearing aid fitting: II. Hearing aids with multi-channel compression," *British journal of audiology*, vol. 33, no. 3, pp. 157–170, 1999.

[19] H. Kayser, T. Herzke, P. Maanen, C. Pavlovic, and V. Hohmann, "Open Master Hearing Aid (openMHA)—An integrated platform for hearing aid research," *The Journal of the Acoustical Society of America*, vol. 146, no. 4, pp. 2879–2879, 2019.

[20] Y. Nejime and B. C. Moore, "Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 603–615, 1997.

[21] BNC Consortium, "British national corpus, version 3 (bnc xml edition)," http://www.natcorp.ox.ac.uk/, 2007, distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. Accessed: 2021-03-01.

[22] I. Demirsahin, O. Kjartansson, A. Gutkin, and C. E. Rivera, "Open-source Multi-speaker Corpora of the English Accents in the British Isles," 2020.

[23] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, Conference Proceedings, pp. 411–412.

[24] F. Denk, S. M. Ernst, J. Heeren, S. D. Ewert, and B. Kollmeier, "The Oldenburg Hearing Device (OlHeaD) HRTF Database," Technical report, Report, 2018.

[25] J. Zhuang, T. Tang, Y. Ding, S. C. Tatikonda, N. Dvornek, X. Papademetris, and J. Duncan, "Adabelief optimizer: Adapting stepsizes by the belief in observed gradients," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[26] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, Conference Proceedings, pp. 4214–4217.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. 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.