Speech Denoising and Compensation for Hearing Aids Using an FTCRN-Based Metric GAN

Jiaming Cheng, Ruiyu Liang, *Member, IEEE*, Li Zhao, Chengwei Huang, and Björn W. Schuller, *Fellow, IEEE*

***Abstract*—Hearing aids aims to improve speech intelligibility for hearing impaired patients to levels comparable to those for normal hearing listeners. However, the interference of environmental noises greatly increase the difficulty of hearing loss compensation. Most related research only focuses on one aspect of noise reduction and hearing loss compensation. In this letter, we propose a metric generative adversarial framework based on a frequency-time convolution recurrent network for joint noise reduction and hearing loss compensation. The audiogram is extended along the frequency axis to form embedded features. A metric discriminator is introduced and the optimization of the generator is guided by an evaluation score related to hearing loss compensation. Additional perceptual-based losses are set to stabilize optimization. Experimental results show that the proposed method can better reduce noise and compensate for hearing loss compared with other algorithms.**

***Index Terms*—Hearing aid, noise reduction, deep learning, metric generative adversarial network.**

I. Introduction

AROUND 500 million people worldwide suffer from hearing loss [1]. Hearing aids are the most effective means of hearing rehabilitation for patients with hearing loss. However, the existing hearing aids are still difficult to compensate for hearing loss in a noisy environment [2], [3]. Hearing-impaired patients are often more sensitive to noises [4]. Therefore, noise reduction is necessary for hearing aids. Noise reduction methods of hearing aids mainly include non-spatial (single channel) time-frequency methods and spatial selectivity methods [5]. The single channel methods mainly use the time-frequency differences between the target speech and the interference, but the improvement of speech intelligibility is limited [6], [7]. Spatial selectivity (e. g., beamforming) methods keep the direction of the target speech more sensitive [8]. However, the noise may come from different directions and the orientation of the target speech and the interference may also change with time in real scenarios.

In recent years, breakthroughs have been made in noise reduction algorithms based on deep neural networks (DNNs) [9], [10], [11]. Deep learning-based algorithm is used in [12] to restore speech intelligibility for hearing aid users, establishing advantages over classic methods. Further, deep filtering is introduced in [13] to improve the effect of noise reduction for hearing aids while maintaining low latency compared to previous methods. However, these methods cannot properly combine noise reduction with hearing loss compensation. Actually, different hearing impaired patients have different gain requirements for specific frequencies. Hearing loss compensation will produce overall gain, while noise amplification will greatly affect the hearing quality. Therefore, noise reduction while compensation is a challenge.

Traditional algorithms have tried to combine noise reduction and dynamic range compression in hearing aids [14], [15], but this is rare in the algorithms based on DNNs. The recent clarity challenge [16] also targets both compensation and enhancement, which shows the importance of combining the two stages. In this letter, we explore a framework for simultaneous noise reduction and hearing loss compensation. Firstly, an embedding method is introduced to learn the correspondence between audiograms and frequency intervals. Secondly, inspired by the metric generative adversarial network [17], we introduce the metric generative adversarial strategy in hearing loss compensation. The metric discriminator is used to predict the hearing-aid speech quality index (HASQI) [18] score of the compensated speech and is alternately trained with the generator, guiding the latter with respect to the evaluation score. Finally, we strengthen noise reduction and stabilize optimization by setting additional perceptual-based losses.

II. Methodology

A. System Overview

In this letter, we propose a metric generative adversarial framework based on the frequency-time convolution recurrent network (FT-CRN). The overall framework is shown in Fig. 1. For the generator, the short-time Fourier transform (STFT) is first performed on the noisy speech to obtain the noisy complex

This work was supported in part by the National Key Re- search and Development Program of China under Grants 2020YFC2004002 and 2020YFC2004003, and in part by the National Natural Science Foundation of China under Grant 62001215. *(Corresponding author: Jiaming Chen.)*

Jiaming Cheng and Li Zhao are with the School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: 230198469@seu.edu.cn; zhaoli@seu.edu.cn).

Ruiyu Liang is with the School of Information Science and Engineering, Southeast University, Nanjing 210096, China, and also with the School of Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, China (e-mail: liangry@njit.edu.cn).

Chengwei Huang is with the Zhejiang Laboratory, Hangzhou 310000, China (e-mail: huangchengwei@zhejianglab.com).

Björn W. Schuller is with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany, and also with the GLAM – Group on Language, Audio, and Music, Imperial College London, SW7 2AZ London, U.K. (e-mail: bjoern.schuller@imperial.ac.uk).

Digital Object Identifier 10.1109/LSP.2023.3263788

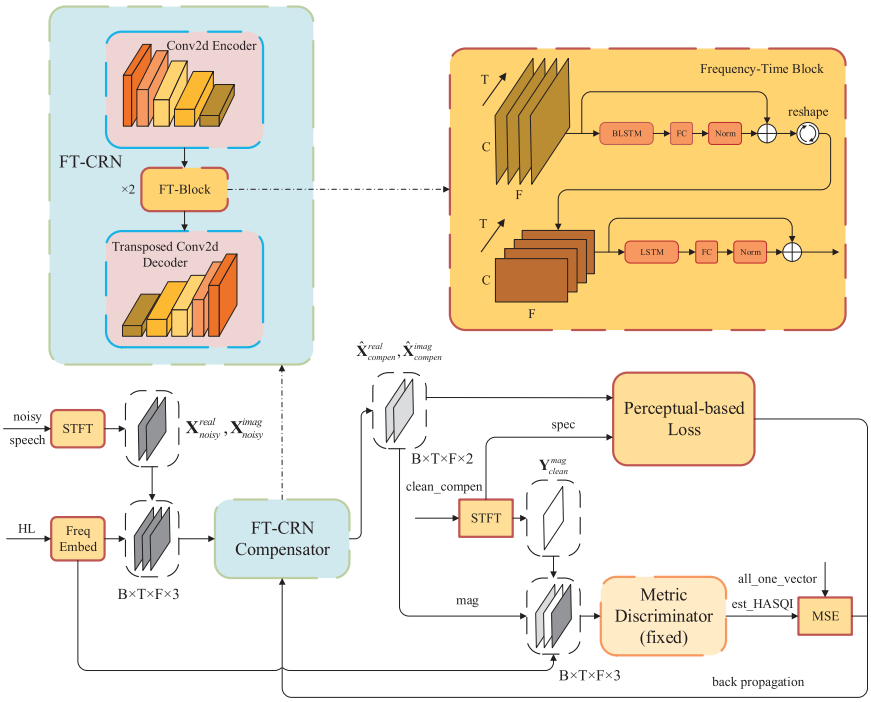

Fig. 1. An overview of the proposed architecture. The entire framework includes a FT-CRN compensator as the generator and a metric discriminator. The generator is trained with the target of compensated clean speech and optimized by perceptual-based losses and adversarial training with a metric discriminator. The hearing loss metric discriminator is used as a differentiable estimator for the HASQI [18] score.

spectrogram. Then, the audiogram is extended along the frequency axis, and the embedded audiogram is concatenated with the noisy complex spectrogram to form the input of the generator. The magnitude of clean compensated and estimated speech are concatenated with the embedding of the audiogram as the input of the metric discriminator. The discriminator is trained to learn how to evaluate the HASQI score of the estimated speech, and will be fixed during the training of the generator to become an instructor so that the predicted HASQI score is constantly approaching the optimal value.

TABLE I
THE CORRESPONDENCE BETWEEN AUDIOGRAMS AND FREQUENCY INTERVALS

threshold / dB	frequency bins	frequency interval / Hz
HL[1]	1∼8	0∼250
HL[2]	9∼16	250∼500
HL[3]	17∼32	500∼1000
HL[4]	33∼64	1000∼2000
HL[5]	65∼128	2000∼4000
HL[6]	129∼257	4000∼8000

B. FT-CRN Based Generator

1) FT-CRN Structure: The FT-CRN includes an encoder, a dual-path frequency-time module, and a decoder. For the encoder, five 2-D convolutional layers are used to extract local patterns and reduce feature resolution along the frequency axis, where each convolutional layer is accompanied by a batch normalization and a parametric rectified linear unit (PReLU) activation function. The decoder can be seen as a mirror of the encoder, with 2-D transposed convolutional layers to restore the compressed features to their original size. Skip connections are used between the encoder and the decoder to help the convergence of the model. The frequency-time module is arranged along dual paths. The input features are first sliced along the time axis, and the spectrum of a single time frame is modeled by a bidirectional long short-term memory (BLSTM) layer. The output is then sliced along the frequency axis, and the temporal correlation of each frequency bin is captured using an LSTM layer. A dense layer and a layer normalization are set on each path for post-processing.

2) Extended Embedding of Audiograms: The audiogram is the hearing loss threshold HL[1-6] collected at [250, 500, 1000, 2000, 4000, 8000] Hz, which characterizes the physiological condition of hearing. The dimensions of the audiogram and spectrogram features need to be matched for the concatenation of the two. Considering that if the audiogram is directly transformed for embedding, the correspondence between audiogram and speech spectrum cannot be used. In order to learn the correspondence information between the audiograms and the frequency intervals, we extend the audiogram along the frequency axis to obtain the alignment on the frequency intervals. The correspondence between audiograms and frequency intervals is shown in Table I (using 512-point FFT, 256-point frame shift, sampled at 16 kHz). That is to say, each audiogram threshold is normalized as a gain for each frequency interval. For the input of the generator and discriminator, the audiogram is extended to form a 257-dimensional gain embedding to achieve alignment with spectrogram features. Additionally, the audiogram is replicated

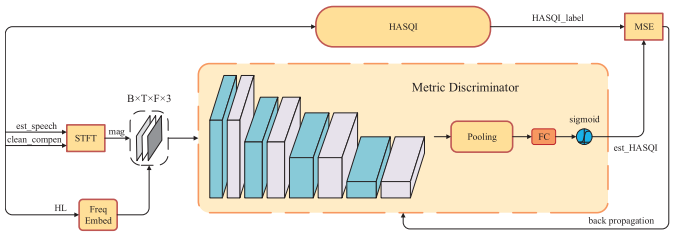

Fig. 2. Hearing loss compensation metric discriminator.

along the time axis so that it has the same contribution to each frame.

C. Metric Discriminator

1) Optimization Goal: HASQI [18] is often used to evaluate the effect of hearing loss compensation. It is based on a model of the auditory periphery, comprising a nonlinear term sensitive to noise and nonlinear distortion and a linear term sensitive to long-term spectral changes, which is used to measure changes in the signal envelope, temporal fine structure, and the spectrum caused by audio processing. The objective functions of existing compensation algorithms are often not directly related to the evaluation score. In this letter, improving the HASQI score is set as the optimization goal. Audios are compensated for different hearing-impaired patients based on the FIG6 formula [19]. FIG6 is based on the 'loudness normalization' criterion, as it aims to restore the loudness perception of the impaired listener to the same loudness perceived by normal hearing. The clean speech compensated by FIG6 is taken as the ideal target to guide the generator.

2) Metric Discriminator Network: Since the calculation of HASQI has non-differentiable parts, it cannot be directly used for backpropagation optimization. In this letter, a metric discriminator is introduced to mimic the HASQI score. The discriminator is trained alternately with the generator, guiding the latter with respect to the evaluation score. The structure is shown in Fig. 2. The magnitude of the estimated speech from the generator and the corresponding clean compensated speech are concatenated with the embedding of the audiogram as the input of the metric discriminator, and the output is the predicted evaluation score. The metric discriminator uses four convolutional blocks for feature extraction. Each block consists of a 2-D convolutional layer, an instance normalization layer, and a PReLU activation function. A global average pooling operation is performed after the convolutional blocks, with final processing by two dense layers and a sigmoid function.

D. Training Procedure

Throughout the training process, the generator and the discriminator are optimized alternately. In the training stage of the metric discriminator, the estimated speech generated by the generator is evaluated by the discriminator, and the score is compared with the ideal label computed by HASQI. In addition, the two inputs are both set to the clean speech compensated by FIG6, and the target is set to all-one vector as an ideal upper limit for estimation. During the training stage of the generator, the metric discriminator is frozen and the generator is optimized to approach a HASQI label of 1. Perceptual-based losses are additionally set to strengthen the model's ability to reduce noises. The joint losses for the whole framework are described as follows:

$$\begin{aligned}\mathcal{L}_G = &\mathbb{E}_{X_{clean},\hat{X},HL}\left[\left\|D\left(X_{clean},\hat{X},HL\right)-1\right\|^2\right]\\&+\lambda\cdot\mathbb{E}_{X_{clean},\hat{X}}\left[PMSQE\left(X_{clean},\hat{X}\right)\right]\\&+\mu\cdot\mathbb{E}_{X_{clean},\hat{X}}\left[PASE\left(X_{clean},\hat{X}\right)\right],\end{aligned}\tag{1}$$

$$\begin{aligned}\mathcal{L}_D = &\mathbb{E}_{X_{clean},X_{clean},HL}\left[\left\|D\left(X_{clean},X_{clean},HL\right)-1\right\|^2\right]\\&+\mathbb{E}_{X_{clean},\hat{X},HL}\left[\left\|D\left(X_{clean},\hat{X},HL\right)-Q_{HASQI}\right\|^2\right],\end{aligned}\tag{2}$$

where, X_{clean}, $\hat{X}$ and HL denote denote the clean compensated speech, estimated speech, and the corresponding audiogram sampled in each batch, respectively. D represents the discriminator. $PMSQE$ [20] is a perceptual metric for speech quality evaluation, and $PASE$ [21] uses the pre-trained speech feature encoder to measure the distance between the estimated and the clean compensated speech. λ and μ are the weight coefficients of two perceptual-based losses, which are set to 1 and 0.25 in this letter.

III. Experiment and Analysis

A. Experiment Settings

The Interspeech 2020 DNS Challenge dataset [22] is used to prepare experimental speech data. The DNS dataset contains 500 hours of clean clips from 2150 speakers and 65 000 noise clips in a total of 180 hours. We randomly split more than 60 000 utterances into a training set and a validation set with a ratio of 60:1.

The noisy utterances are generated by mixing randomly selected speech and noise at random SNR between -5 and 15 dB. Utterance segments of 3-second length are randomly sampled as the input signals during training. The official non-reverb test set is used for objective scoring comparison.

Audiograms in the experiments come from the National Health and Nutrition Examination Survey (NHANES) [23]. Part of the survey was to evaluate the hearing status of the subjects through pure-tone audiometry. The audiograms are obtained according to the pure-tone audiometry standard measurement. We use a total of 114 audiograms. An audiogram is randomly selected for each noisy utterance, so that the model can learn to adapt to different levels of hearing impairment.

All the utterances are sampled at 16 kHz. We provide codes and test set to ensure reproducibility.[1] All models are optimized using Adam with a learning rate of 0.0006, and a batch size of 32. The training will last for 20 epochs, and 59000 noisy utterances are used in each epoch to update the model parameters.

B. Experimental Results and Discussion

In this letter, we use three indicators: HASQI [18] (for evaluating the speech quality of hearing-impaired listeners), WB-PESQ [24] (for evaluating the perceptual quality of speech),

[1]The network codes, FIG6 codes and test set are available at https://github.com/JMCheng-SEU/FTCRN-based-Metric-GAN-for-Hearing-Aids.

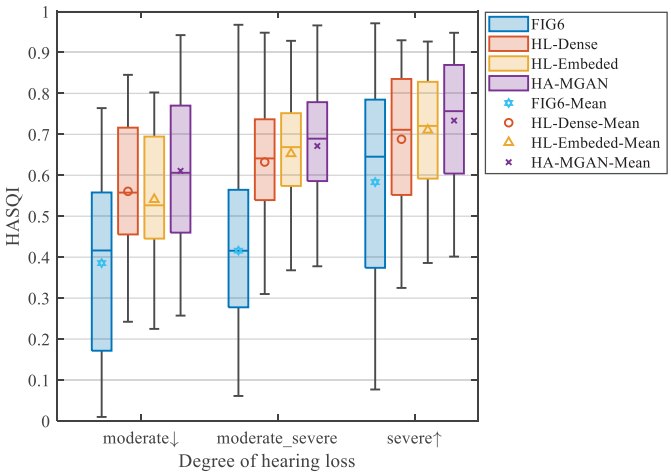

Fig. 3. Indicator comparison under different degrees of hearing loss.

TABLE II
COMPARISON OF TEST SET INDICATORS FOR DIFFERENT COMPENSATION ALGORITHMS

Compensation Algorithm	Param.(M)	HASQI	WB-PESQ	STOI
FIG6	-	0.465	1.540	0.863
RNNoise [26]	0.086	0.562	1.977	0.891
CRN [27]	17.579	0.487	2.027	0.914
NUNet-TLS [28]	2.830	0.572	2.420	0.932
DCCRN [29]	3.671	0.579	2.268	0.925
GCRN-complex [30]	8.146	0.593	2.186	0.922
DB-AIAT [31]	2.811	0.623	**2.689**	**0.944**
HL-Dense	0.709	0.624	2.354	0.930
HL-Embeded	0.707	0.644	2.457	0.934
HA-MGAN*	0.707	0.668	2.490	0.940
HA-MGAN	0.707	**0.675**	2.527	0.941

and STOI [25] (for evaluating speech intelligibility) to compare the effect of the algorithms. Since this letter aims to improve the effect of hearing loss compensation in noisy environments, we use the clean speech compensated by FIG6 [19] as the target speech for evaluation.

The effect of each introduced module is first tested under different levels of hearing impairment. Impairment classification is based on the average loss thresholds at 500, 1000, 2000, and 4000 Hz. In order to balance the number of audiograms under each category, we classify the audiograms with thresholds less than 50 dB as moderate or below (moderate ↓), while those greater than 50 dB and less than 65 dB are moderately severe (moderate-severe), and those greater than 65 dB are classified as severe or above (severe ↑).

FT-CRN is set as the baseline model, and three methods are used for comparison, including direct transformation of the audiogram through the fully connected layer (HL-Dense), the proposed extended embedding of the audiogram (HL-Embeded) and the introduction of a metric discriminator based on the extended embedding to conduct adversarial training (HA-MGAN). HASQI indicators of each method under different levels of hearing loss are shown in Fig. 3. All three methods are superior to direct FIG6 compensation for noisy speech at each level, which shows that joint noise reduction and hearing loss compensation can effectively improve the compensation quality in noisy environments. Note that, compared with the HL-Dense method, HL-Embeded has an advantage at moderate-severe level and above, but lower below the moderate level, which indicates that the extended embedding of the audiogram is more effective for patients with severe hearing loss. After the introduction of the metric discriminator for HASQI-oriented optimization, the evaluation scores have been further improved in each hearing loss level.

Next, an average test is performed on the noisy test set, in which audiograms are randomly selected for each speech. At this stage, we compare our model with other compensation methods. To the best of our knowledge, current approaches which jointly perform noise reduction and hearing loss compensation are rare to find. Therefore, the representative models in single channel speech enhancement are chosen for comparative experiments. RNNoise [26], CRN [27], and NUNet-TLS [28] are algorithms based on magnitude, while DCCRN [29], GCRN-complex [30], and DB-AIAT [31] are designed based on complex spectrum. For fair comparison, audiogram embeddings are added for each algorithm, and the rest of the settings remain the same as in the original paper. In addition, HA-MGAN* represents the proposed model trained without PMSQE and PASE terms. The input and target of all comparison systems are set to be the same, and all models are trained from the beginning under the same dataset. Table II shows the mean values of indicators for different algorithms.

We observed that despite using the highest number of parameters, the CRN model achieves the lowest HASQI score. The RNNoise model has the lowest model complexity due to mostly using gated recurrent units (GRUs), but its improvement on indicators related to perceptual quality is limited. The DB-AIAT model achieves the highest WB-PESQ and STOI scores, but its noncausal structure is not conducive to deployment on hearing aids. In terms of the models based on FT-CRN, the introduction of frequency extended embedding of the audiogram achieves an improvement in each indicator compared to the direct transformation. The guidance of the metric discriminator on this basis can achieve further advantages. Overall, the proposed method (HA-MGAN) achieves the highest HASQI score, second only to the DB-AIAT model on WB-PESQ and STOI, reflecting that the metric generative adversarial strategy is capable of reducing noises while compensating for audiograms. Additionally, the proposed model is fully causal and has a low amount of parameters (0.707 M), which is beneficial for deployment in real scenarios.

IV. CONCLUSION

In this letter, we proposed a metric generative adversarial framework for joint noise reduction and hearing loss compensation. In the proposed framework, we extended the audiogram along the frequency axis to let the network learn the distribution of hearing loss across frequency intervals. To the best of our knowledge, the metric generative adversarial strategy was applied for the first time to the hearing loss compensation task, guiding the generator with respect to the evaluation score. Experimental results confirmed that the proposed model achieves better performance than other models in terms of hearing loss compensation and noise reduction effect. In the future, we will further compress the model to accommodate the real-time computing needs of hearing aids.

REFERENCES

[1] B. S. Wilson, D. L. Tucci, M. H. Merson, and G. M. O'Donoghue, "Global hearing health care: New findings and perspectives," *Lancet*, vol. 390, no. 10111, pp. 2503–2515, 2017.

[2] D. Hartley, E. Rochtchina, P. Newall, M. Golding, and P. Mitchell, "Use of hearing aids and assistive listening devices in an older australian population," *J. Amer. Acad. Audiol.*, vol. 21, no. 10, pp. 642–653, 2010.

[3] E. M. Picou, "Marketrak 10 (mt10) survey results demonstrate high satisfaction with and benefits from hearing aids," in *Seminars in Hear.*, vol. 41, no. 01. New York, NY, USA: Thieme Med. Publishers, 2020, pp. 021–036.

[4] A. H. Andersen et al., "Creating clarity in noisy environments by using deep learning in hearing aids," in *Seminars in Hearing*, vol. 42, no. 03. New York, NY, USA: Thieme Med. Publishers, Inc., 2021, pp. 260–281.

[5] B. McPherson, "Innovative technology in hearing instruments: Matching needs in the developing world," *Trends Amplification*, vol. 15, no. 4, pp. 209–214, 2011.

[6] F. Y. Chong and L. M. Jenstad, "A critical review of hearing-aid single-microphone noise-reduction studies in adults and children," *Disabil. Rehabil.: Assistive Technol.*, vol. 13, no. 6, pp. 600–608, 2018.

[7] C. Völker, A. Warzybok, and S. M. Ernst, "Comparing binaural pre-processing strategies III: Speech intelligibility of normal-hearing and hearing-impaired listeners," *Trends Hear.*, vol. 19, 2015, Art. no. 2331216515618609.

[8] D. B. Hawkins and W. S. Yacullo, "Signal-to-noise ratio advantage of binaural hearing aids and directional microphones under different levels of reverberation," *J. Speech Hear. Disord.*, vol. 49, no. 3, pp. 278–286, 1984.

[9] S. Tamura, "An analysis of a noise reduction neural network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1989, pp. 2001–2004.

[10] F. Xie and D. Van Compernolle, "A family of MLP based nonlinear spectral estimators for noise reduction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1994, vol. 2, pp. II/53–II/56.

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

[12] T. Goehring, F. Bolner, J. J. Monaghan, B. Van Dijk, A. Zarowski, and S. Bleeck, "Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users," *Hear. Res.*, vol. 344, pp. 183–194, 2017.

[13] H. Schröter, T. Rosenkranz, A.-N. Escalante-B, and A. Maier, "Low latency speech enhancement for hearing aids using deep filtering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2716–2728, 2022.

[14] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. Holdt Jensen, "A combined multi-channel wiener filter-based noise reduction and dynamic range compression in hearing aids," *Signal Process.*, vol. 92, no. 2, pp. 417–426, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168411002635

[15] K. Ngo, S. Doclo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen, "An integrated approach for noise reduction and dynamic range compression in hearing aids," in *Proc. IEEE 16th Eur. Signal Process. Conf.*, 2008, pp. 1–5.

[16] S. Graetzer et al., "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Proc. Interspeech* 2021, pp. 686–690.

[17] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2031–2041.

[18] A. A. Kressner, D. V. Anderson, and C. J. Rozell, "Evaluating the generalization of the hearing aid speech quality index (HASQI)," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 407–415, Feb. 2013.

[19] M. C. Killion, "The 3 types of sensorineural hearing loss: Loudness and intelligibility considerations," *Hear. J.*, vol. 46, no. 11, pp. 31–36, 1993.

[20] J. M. Martin-Donas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal Process. Lett.*, vol. 25, no. 11, pp. 1680–1684, Nov. 2018.

[21] M. Ravanelli et al., "Multi-task self-supervised learning for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6989–6993.

[22] C. K. Reddy et al., "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech*, 2020, pp. 2492–2496.

[23] Centers for Disease Control and Prevention (CDC), "About the national health and nutrition examination survey," 2017. [Online]. Available: https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

[24] A. Takahashi, A. Kurashima, C. Morioka, and H. Yoshino, "Objective quality assessment of wideband speech by an extension of ITU-T recommendation P. 862," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005.

[25] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[26] J.-M. Valin, "A hybrid DSP/deep learning approach to real-time full-band speech enhancement," in *Proc. IEEE 20th Int. Workshop Multimedia Signal Process.*, 2018, pp. 1–5.

[27] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. Interspeech*, 2018, pp. 3229–3233.

[28] S. Hwang, Y. Park, and S. Park, "Monoaural speech enhancement using a nested U-net with two-level skip connections," in *Proc. Interspeech*, 2022, pp. 191–195.

[29] Y. Hu et al., "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.

[30] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 380–390, 2020.

[31] G. Yu, A. Li, C. Zheng, Y. Guo, Y. Wang, and H. Wang, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7847–7851.