A Non-Invasive Speech Quality Evaluation Algorithm for Hearing Aids With Multi-Head Self-Attention and Audiogram-Based Features

Ruiyu Liang, *Member, IEEE*, Yue Xie, *Member, IEEE*, Jiaming Cheng, Cong Pang, and Björn Schuller, *Fellow, IEEE*

***Abstract*—The speech quality delivered by hearing aids plays a crucial role in determining the acceptance and satisfaction of users. Compared with invasive speech quality evaluation methods that require pure signals as a reference, this paper proposes a non-invasive speech quality evaluation algorithm for hearing aids with multi-head self-attention and audiogram-based features. Initially, the audiogram of hearing-impaired individuals is extended along the frequency axis, enabling the speech quality evaluation model to learn the gain requirements specific to frequency bands for hearing-impaired individuals. Subsequently, the spectrogram is extracted from the speech signals to be evaluated. These features are combined with the transformed audiogram to create input features. To extract deep frame-level feature, a network employing multiple two-dimensional convolutional modules is utilized. Then, the temporal features are modeled using bidirectional long short-term memory networks (BiLSTM), while a multi-head self-attention mechanism is employed to integrate contextual information. This mechanism enables the model to focus on key frame information. Experimental results demonstrate that, compared to currently available advanced algorithms, the proposed network exhibits a higher correlation with the Hearing Aid Speech Quality Index (HASQI) and demonstrates robustness under various noise conditions.**

***Index Terms*—Audiogram, hearing aid, multi-head self-attention, speech quality evaluation.**

I. Introduction

ACCORDING to the World Health Organization's Global Hearing Report 2021, approximately one-fifth of the global population experiences varying degrees of hearing loss [1]. Wearing hearing aids is a common method to improve the hearing abilities of individuals with hearing impairment. However, achieving satisfactory results with hearing aids often requires expert fitting, which can be time-consuming and may not always yield optimal outcomes, especially for elderly patients. While the OTC Hearing Aid Act approved by the United States Congress in 2017 allows for the direct sale of hearing aids to mild to moderate hearing-impaired consumers without a prescription, evaluating the quality of speech signal processing in hearing aids without professional hearing experts remains a research topic.

The speech quality of hearing aids, which refers to the quality of speech after being processed by hearing aids, reflects the technical level of speech output and considerably influences people's acceptance and satisfaction with hearing aids [2]. Higher speech quality in hearing aids leads to more natural-sounding speech for users [3]. Similar to the evaluation of regular speech quality [4] for speech enhancement [5], [6], [7], the speech quality of hearing aids can also be subjectively assessed by hearing-impaired individuals. Due to the need to recruit hearing-impaired patients for evaluation, the speech quality evalution of hearing aids requires more time and economic resources compared to regular speech tests. Therefore, the research on speech quality evaluation algorithms for normal individuals has been relatively early, and many scholars have proposed many evaluation methods [8], [9], [10], [11], [12], [13], [14], [15], [16], among which the Perceptual Evaluation of Speech Quality (PESQ) [16] and Short-Time Objective Intelligibility (STOI) [17] are commonly used. In recent years, researchers have gradually been exploring objective indicators, similar to PESQ [16], to evaluate the speech quality of hearing aids.

Regarding speech evaluation indicators for hearing aids, the Hearing Aid Speech Quality Index (HASQI) [18] is a typical indicator that uses a simulated auditory structure related to the degree of hearing loss to process both the test and reference signals [19]. The final quality evaluation score is calculated by extracting short-term fine structure and long-term spectrum features. Experiments have shown that HASQI has shown good correlation with perceived hearing aid quality [20], [21]. Another speech quality evaluation model specifically designed for hearing-impaired individuals is the Perception Model-Quality

This work was supported in part by the Project of China Disabled Persons Federation under Grant 2023CDPFHS-02, in part by the Key Research and Development Program of Jiangsu Province under Grant BE2022059-3, and in part by the National Natural Science Foundation of China under Grant 62001215. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Juan Ignacio Godino- Llorente. *(Corresponding author: Yue Xie.)*

Ruiyu Liang and Yue Xie are with the School of Information and Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, China (e-mail: liangry@njit.edu.cn; xieyue0109@njit.edu.cn).

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

Björn Schuller is with the ZD.B Chair of Embedded Intelligence for Health Care, Wellbeing University of Augsburg, 86135 Augsburg, Germany, and also with the Group on Language, Audio, and Music, Imperial College London, SW7 2BX London, U.K. (e-mail: schuller@ieee.org).

Digital Object Identifier 10.1109/TASLP.2024.3378107

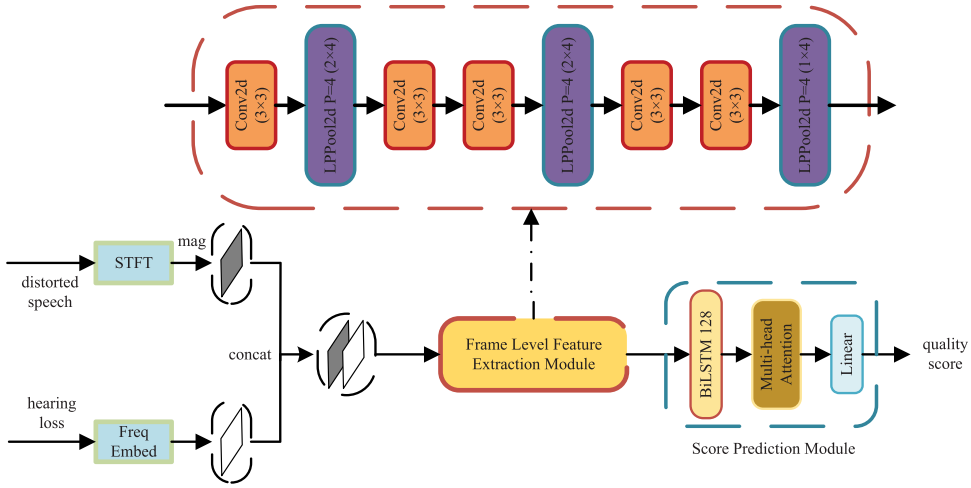

Fig. 1. Architecture of the speech quality evaluation network.

Extended for Hearing-Impaired (PEMO-Q-HI) [22]. Based on PEMO-Q [23], it modifies the auditory model by incorporating the patient's listening threshold into the center frequency of the outer filter group, considering both Inner Hair Cell Loss (IHC Loss) and Outer Hair Cell Loss (OHC Loss). The model adjusts the amplitude of the speech signal according to the loss of inner and outer hair cells. However, these indicators are invasive, requiring a clean reference speech with a similar frequency shape and time alignment to the speech being evaluated [24], which is often challenging to obtain in practical scenarios, limiting their applicability.

In comparison, non-invasive objective speech quality evaluation methods that do not require reference signals are more practical and suitable for real-time evaluation of speech quality in systems and devices [25]. However, there are relatively few related studies about the speech quality evaluation for hearing aids. In 2013, Suelzle et al. proposed the Speech to Reverberation Modulation Energy Ratio-Hearing Aid (SRMR-HA) metric, which incorporated a calculation model of cochlear hearing loss to evaluate hearing aid speech quality [24]. In 2015, Falk et al. examined different objective indicators for predicting the speech quality of hearing aids [26]. The survey results showed differences between invasive and non-invasive speech quality evaluation indicators in several databases, which further stimulated research on non-invasive speech quality evaluation indicators for hearing aids. In the same year, Salehi studied non-invasive speech evaluation indicators for hearing aids based on low-complexity quality assessment (LCQA-HA) [27]. This method expanded a large set of speech-specific features extracted through LCQA using the importance-weighted signal-to-noise ratio (iSNR) metric [28], and assimilated the most important features through regression functions to obtain predicted quality scores. Additionally, non-invasive hearing aid speech quality indicators based on support vector regression (SVR) were proposed [29].

The above models are implemented using traditional machine learning algorithms, deep learning networks are more suitable for predicting subjective [17], [30] or objective evaluation scores of speech quality [31], [32] and comprehensibility [32], [33] due to their ability to effectively model temporal information [34]. However, the application of deep learning networks in evaluating hearing aid speech quality is not as prevalent as in the telecommunications field. An end-to-end, non-invasive speech quality evaluation model called Quality-Net was proposed by Szu-Wei Fu et al. in 2018 [31]. This model, based on a Bidirectional Long Short-Term Memory (BiLSTM) structure, can effectively predict the PESQ scores for noisy or processed speech signals. The same research team later updated the model in 2020 for speech intelligibility evaluation, known as STOI-Net [33]. STOI-Net, based on a Convolutional Neural Network-Bidirectional Long Short-Term Memory (CNN-BiLSTM) structure, utilizes a multiplication attention mechanism to identify and weight important information. In 2019, Mittag et al. proposed a non-invasive speech quality assessment model (NISQA) for predicting the quality of ultra-wideband speech transmissions [35]. The model employs a CNN to predict frame-level quality and an LSTM network to aggregate values from each frame for overall speech quality estimation. In the same year, a quality assessment model for voice conversion (VC) systems called MOSNet was also introduced [36]. MOSNet, composed of a CNN-BiLSTM structure, effectively predicts the average opinion score of transformed speech. To address scoring bias caused by different evaluators' personal preferences, an improved version of MOSNet called MBNet was proposed, consisting of MeanNet and BiasNet, both based on the CNN-BiLSTM structure [37]. Furthermore, Cauchi et al. proposed a network that evaluates the quality of speech processed by different signal processing algorithms, combining modulation energy and LSTM units to consider the time dependence of the target signal, achieving high accuracy in evaluating speech quality [38]. In 2022, Reddy et al. have developed a non-intrusive speech quality metric called Deep Noise Suppression Mean Opinion Score (DNSMOS) based CNN using the scores from ITU-T Rec. P. 808 subjective evaluation [39]. In 2023, Jaiswal et al. proposed a deep neural network(DNN) where its input is combined features of the speech signal and its output is corresponding speech quality score [40]. The used features included multi-resolution auditory model, mel-frequency cepstral coefficients and line spectral frequencies. In addition, multi task learning strategies [32], [41], [42] have also begun to be applied to speech quality estimation.

In the field of hearing aids, research on non-invasive quality evaluation based on deep learning is limited. In 2021, Chiang et al. proposed HASA-Net [43], which combines BiLSTM and attention mechanisms to jointly predict HASQI and the Hearing Aid Speech Intelligibility Index (HASPI). To contribute to the research on non-invasive quality evaluation based on deep learning in the field of hearing aids and provide accurate evaluations of hearing aid speech quality under different degrees of hearing loss, this paper proposes a non-invasive speech quality evaluation network for hearing aids, called Speech Quality Index Net-Hearing Loss Level (SQINet-HLlevel). The network takes distorted speech spectrogram and audiograms of hearing-impaired patients as inputs and predicts the HASQI score as output. The network first extracts frame-level deep features from the input features using a CNN. These deep features are then processed by a BiLSTM to model temporal features, followed by a multi-head self-attention layer to integrate sequence context information, allowing the model to discern the importance of different speech frames. Finally, the weighted features are linearly mapped to quality scores.

The main contributions of this paper are as follows:

1) Introduction of an extended embedding strategy based on the audiograms of hearing-impaired patients, enabling the network to analyze the hearing loss information of different individuals and improve the accuracy of speech quality evaluation for hearing aids.
2) Application of a multi-head self-attention mechanism to enable the network to adaptively select relevant information from frame-level features and better utilize global context information.

II. Method

A. Overall Architecture of Speech Quality Evaluation Network

The proposed network comprises two main modules: the frame level feature extraction module and the score prediction module, as depicted in Fig. 1. The input features of the network consist of two components. Firstly, speech signals are transformed by the Short-Time Fourier transform (STFT). Then, spectrogram is obtained from the magnitude of the STFT, which is a distribution of energy in the time-frequency plane. Secondly, the audiogram of hearing-impaired patients is extended and embedded along the frequency axis. These two features are concatenated and fed into the feature extraction module. The feature extraction module, based on CNNs, extracts deep representations from the input features. The extracted deep information is then passed to the score prediction module for determining the quality evaluation score. The score prediction module consists of a BiLSTM, a multi-head self-attention layer, and fully connected layers that serve as mapping units. The module generates a quality evaluation score between 0 and 1 through a sigmoid activation function.

B. Extended Embedding of Audiograms

The patient's hearing condition is typically represented by an audiogram, which depicts the hearing loss (or hearing threshold) of a patient at specific testing frequencies. Audiograms are obtained through hearing tests conducted by doctors or audiologists. If the audiogram is only dimensionally transformed and overlaid with the spectral feature, the embedding method can only utilize the distribution differences between audiograms. However, the corresponding relationship between the audiogram and frequency bands cannot be learned by the network. To address this, the audiogram is first extended along the frequency band and then superimposed with the spectral features to achieve alignment on the frequency axis. Because the audiogram of hearing-impaired patients in this study includes six frequency points (250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz, and 8000 Hz), the entire frequency axis is divided into six frequency bands. The correspondence between the audiogram and frequency bands is shown in Table I. Within a frequency band, the patient's hearing threshold is set to be the same. Here, the frame length is 512 samples and the frame hopping is 256 samples. So, a 512-point FFT is used.

TABLE I
Correspondence Between Audiograms and Frequency Bands

Audiogram gain/dB	Corresponding frequency point	frequency range/Hz
HL_1	0∼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∼256	4000∼8000

Note: Here, H_i represents the i-th frequency band.

C. Feature Extraction Module

The feature extraction module takes the combined extended embedding of the audiogram and the spectrogram of the input speech as input. It utilizes two-dimensional convolutional layers and pooling layers to extract deep representations from the input features. This module consists of five two-dimensional convolutional networks, each comprising a batch normalization layer, a two-dimensional convolutional layer, and a Leaky ReLU activation function. The Leaky ReLU activation function is chosen over the ReLU activation function because it assigns smaller linear components to negative inputs (with a negative coefficient of 0.1 in this paper), allowing for gradient adjustment of negative values. Power average pooling layers are inserted intermittently between multiple convolutional modules. These pooling layers calculate the p-th root of the p-th power sum of all data within a moving window:

$$f(x) = \left(\sum_{x \in X} x^p \right)^{\frac{1}{p}} \tag{1}$$

Here, the coefficient p is set to 4.

D. Score Prediction Module

The structure of the score prediction module is illustrated in the dotted box in the lower right corner of Fig. 1. It consists

of a BiLSTM, a multi-head self-attention mechanism layer, and a fully connected layer for mapping. The input to this module is the deep frame-level feature, which is obtained by combining the channel dimension and feature dimension of the output from the feature extraction module. The output is the predicted quality score for distorted speech.

LSTMs dynamically control the flow of information by incorporating gate mechanisms, thereby addressing the long-term dependency problem of traditional Recurrent Neural Networks (RNNs). They are widely employed for modeling temporal information. Considering the strong contextual associations in speech signals, BiLSTM is utilized to model frame-level features. BiLSTM consists of a forward LSTM and a backward LSTM, enabling the learning of bidirectional dependencies in time series data and leveraging context information. Taking the forward LSTM as an example, the calculation process is as follows:

$$\begin{cases} \mathbf{i}_t = \sigma(\mathbf{W}_\mathrm{i} \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_\mathrm{i}) \\ \mathbf{f}_t = \sigma(\mathbf{W}_\mathrm{f} \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_\mathrm{f}) \\ \mathbf{c}_t = \mathbf{f}_t \otimes \mathbf{c}_{t-1} + \mathbf{i}_t \otimes \tanh(\mathbf{W}_\mathrm{c} \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_\mathrm{c}) \\ \mathbf{o}_t = \sigma(\mathbf{W}_\mathrm{o} \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_\mathrm{o}) \\ \mathbf{h}_t = \mathbf{o}_t \otimes \tanh(\mathbf{c}_t) \end{cases} \quad (2)$$

Here, t is the index number of the frame, σ is the sigmoid function, $\mathbf{i}$, $\mathbf{f}$ and $\mathbf{o}$ are the input gate, forgetting gate, and output gate respectively, $\mathbf{c}$ and $\mathbf{h}$ are the cell state and hidden output, $\mathbf{W}$ and $\mathbf{b}$ are the weight matrix and bias vector, and $\otimes$ represents element by element multiplication.

Although BiLSTM effectively captures bidirectional semantic dependencies, redundant information may exist between multiple speech frames. Moreover, it is desirable to consider the importance of each frame's information in the input speech for the target task. For instance, a silence segment should receive less attention (i.e., lower weight), while frames containing more effective information in the speech should receive higher attention. To address this, self-attention mechanisms are employed to enable the model to learn weighted combinations of different time frames. Multi-head self-attention mechanisms are utilized instead of traditional single-head self-attention to obtain information from multiple representation subspaces and enhance the model's performance, as depicted in Fig. 2. In this case, the input keys of Q, K, and V are set to all time step outputs of the BiLSTM.

After being weighted by multi-head self-attention layers, the model uses an adaptive maximum pooling layer to compress the time frame dimension, and then is input into the fully connected layer. The fully connected layer as a mapper has 256 input nodes and one output node. The output of the full connection layer is passed through the sigmoid activation function to get the final quality score.

III. Experimental Setup

A. Model of the Simulation System

The simulation system, as depicted in Fig. 3, is constructed to generate simulated speech signals and corresponding HASQI. The auditory model employed in the system is the simulated auditory model [19]. During the training process, the clean

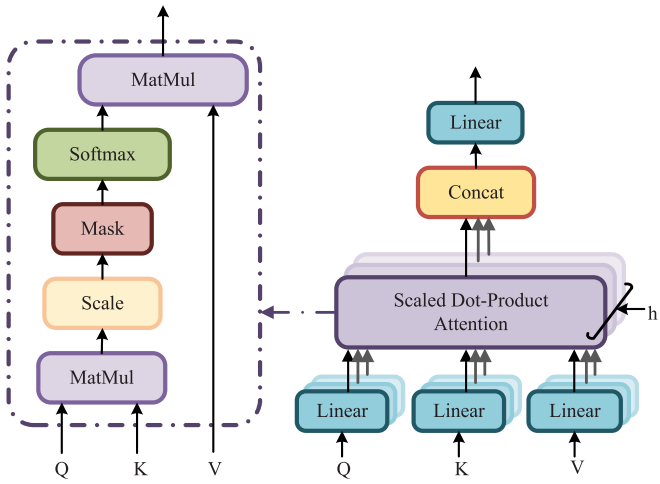

Fig. 2. Block diagram of multi-head self-attention layer.

speech is initially processed by adding noise and performing speech enhancement to obtain the corresponding distorted speech. Then, under hearing loss conditions, wide dynamic range compression (WDRC) [44], [45], [46] is applied to perform loudness compensation separately for the clean and distorted speech. Subsequently, the HASQI of the compensated speech samples is calculated [19]. Signal features are then extracted from the distorted speech samples and input into the network, with the HASQI as the learning target to train the network. During the testing process, the distorted speech to be evaluated is passed through WDRC, and the signal features are extracted and input into the trained network to obtain the estimated HASQI.

The WDRC algorithm is a commonly used algorithm for hearing compensation in digital hearing aids. Its basic principle involves decomposing the speech signal into frequency bands and compensating the speech signal in each frequency band based on the patient's audiogram and signal intensity through the hearing aid fitting formulas.

The specific steps are illustrated in Fig. 4. Firstly, the speech to be compensated is segmented into frames and windowed, and then transformed using the Fast Fourier Transform (FFT). Secondly, the frequency domain signal is divided into 16 bands, and the sound pressure level within each band is calculated. Based on the audiogram, the gain for each band is calculated according to the FIG6 formula [47] and applied to the corresponding band. Thirdly, the compensated signal is obtained by performing the Inverse Fast Fourier Transform (IFFT) and frame overlap. The gain based on the FIG6 formula is calculated using the input-output curve of the sound pressure level. This curve is divided into three regions with 40 dB SPL and 60 dB SPL as two compression inflection points, and different compensation rules are applied in each region according to the FIG6 formula. The signal processing is performed with a frame length of 512, 50% overlap, and a Hanning window.

B. Database

The dataset used in this study is derived from a Chinese speech dataset containing 13,388 read-out speeches from 25 speakers in

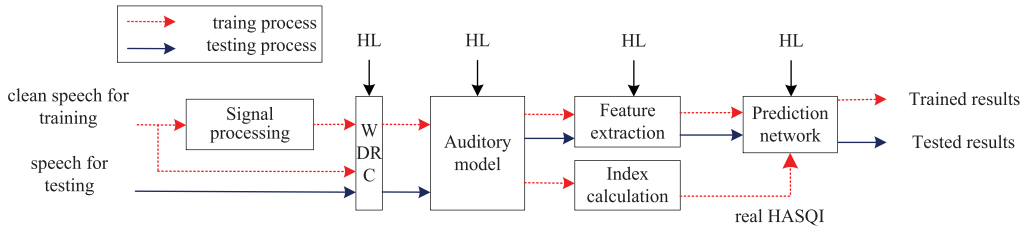

Fig. 3. Flow chart of the simulation system.

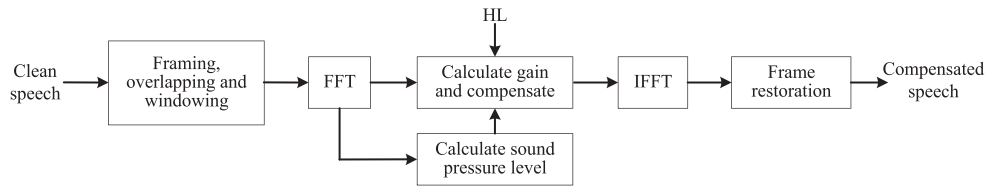

Fig. 4. Flow chart of loudness compensation.

the Interspeech 2021 Deep Noise Suppression Challenge [48]. Among these speeches, 1000 speeches spoken by two speakers are selected as the test set, while the remaining 12,388 speeches spoken by other speakers are used to construct the training set.

In the training set, one noise sample is randomly selected from the NoiseX-92 noise set [49] that contains 15 noise samples with a uniform distribution for each clean speech. A signal-to-noise ratio (SNR) between −5 and 15 dB is added to each clean speech to obtain noisy speech. Then, each noisy speech is processed by one of three methods: traditional Wiener filtering [50], Wiener filtering based on a prior SNR [50], and multi-band spectral subtraction [51] to obtain enhanced speech. Using different speech enhancement algorithms to process the speech enriches the types of distortion, thereby enhancing the network's robustness to distortion. The constructed training set consists of 12,388 * 4 samples, including a group of unprocessed noisy speech and three groups of enhanced speech processed by different speech enhancement algorithms. In the test set, clean speeches are also overlapped with noise and processed by one randomly selected method from the three speech enhancement methods. A total of 1,000 test samples are obtained.

All signals were resampled to 16 kHz. The experimental samples were evaluated for HASQI under hearing loss conditions. In this experiment, a total of 114 audiograms were obtained from 57 patients with binaural hearing loss. Among these patients, 100 audiograms from 50 individuals were used to construct the training set, while 14 audiograms from the remaining 7 patients were randomly selected to construct the test set. Each test sample was combined with a randomly selected audiogram for speech compensation and quality assessment. The average age of the hearing-impaired patients in this experiment was 68.4 years old, consisting of 27 males and 30 females. According to the World Health Organization's classification of hearing loss levels in 2021 [1], the patients exhibited varying degrees of hearing loss. Specifically, 2 patients had normal hearing in their left ears, 1 patient had mild hearing loss in both ears, 3 patients had mild hearing loss in their left ears, 1 patient had mild hearing loss in their right ear, 2 patients had moderate hearing loss in their left ears, 3 patients had moderate hearing loss in their right ears, 13 patients had moderate or severe hearing loss in both ears, 6 patients had moderate or severe hearing loss in their left ears, 8 patients had moderate or severe hearing loss in their right ears, 8 patients had severe hearing loss in both ears, 5 patients had severe hearing loss in their left ears, 7 patients had severe hearing loss in their right ears, 8 patients had extremely severe hearing loss in both ears, 9 patients had extremely severe hearing loss in their left ears, and 8 patients had extremely severe hearing loss in their right ears.

C. Network Parameters and Optimization

The speech quality evaluation network consists of two modules: frame-level feature extraction and score prediction. The specific parameters are shown in Table II.

The loss function of the speech quality evaluation network is calculated based on the mean squared error (MSE) between the predicted score and the actual score. The Adam optimizer is employed to train the network, with a batch size of 16. The initial learning rate is set to 0.0006, and it decays by a rate of 0.6 every 25 epochs. A total of 300 epochs are trained.

D. Evaluation Indicators

The performance of the model is evaluated using two metrics that measure the correlation between the predicted HASQI and the actual HASQI: the Pearson correlation coefficient (PCC) and the Spearman rank order correlation coefficient (SROCC). Additionally, the root mean square error (RMSE) is used to assess the prediction error of the HASQI.

PCC is defined as follows:

$$\rho = \frac{\sum_{i=1}^{N}\left(MOS_{\mathrm{o}}(i)-\overline{MOS_{\mathrm{o}}}\right)\left(MOS_{\mathrm{s}}(i)-\overline{MOS_{\mathrm{s}}}\right)}{\sqrt{\sum_{i=1}^{N}\left(MOS_{\mathrm{o}}(i)-\overline{MOS_{\mathrm{o}}}\right)^{2}\sum_{i=1}^{N}\left(MOS_{\mathrm{s}}(i)-\overline{MOS_{\mathrm{s}}}\right)^{2}}} \tag{3}$$

Among them, N is the number of samples, MOS_{o} is the objective score, MOS_{s} is the subjective score, $\overline{MOS_{\mathrm{o}}}$ and $\overline{MOS_{\mathrm{s}}}$ are the average of the objective score and the subjective score, respectively. The PCC ranges between −1 and 1 and is utilized to

TABLE II
PARAMETER SETTINGS FOR THE PROPOSED NETWORK

Module name	Layer name	Input size	Hyperparameter	Output size
Frame level feature extraction module	Conv_1	$2 \times T \times 257$	3×3, (1, 1), 32	$32 \times T \times 257$
	LPPool2d	$32 \times T \times 257$	$1 \times 4{,}4$	$32 \times T \times 64$
	Conv_2	$32 \times T \times 64$	3×3, (1, 1), 128	$128 \times T \times 64$
	Conv_3	$128 \times T \times 64$	3×3, (1, 1), 128	$128 \times T \times 64$
	LPPool2d	$128 \times T \times 64$	$1 \times 4{,}4$	$128 \times T \times 16$
	Conv_4	$128 \times T \times 16$	3×3, (1, 1), 128	$128 \times T \times 16$
	Conv_5	$128 \times T \times 16$	3×3, (1, 1), 16	$128 \times T \times 16$
	LPPool2d	$128 \times T \times 16$	$1 \times 4{,}4$	$128 \times T \times 4$
	Reshape	$128 \times T \times 4$	/	$T \times 512$
Score prediction module	BiLSTM	$T \times 512$	128	$T \times 256$
	Multi-head Attention	$T \times 512$	256, h=4	$T \times 256$
	MaxPool1d	$T \times 256$	1	1×256
	Dense	1×256	1	1

Note: In the following description, T represents the number of frames. The hyperparameters of the convolutional layers are presented in the format of (kernelSize, strides, outputChannels), and the input/output size is provided as channels × timeSteps × featureDims. For multi-head self-attention layers, the first parameter denotes the number of hidden layer nodes, while the last parameter represents the number of taps.

describe the linear correlation between two variables. A positive PCC indicates a positive correlation between the variables, while a negative PCC suggests a negative correlation. The closer the PCC is to 1, the stronger the correlation between the predicted scores and the actual scores.

SROCC represents the correlation coefficient of the strength of the monotonic relationship between two variables. If a variable is a strict monotonic function of another variable, the coefficient is 1 or −1, indicating complete correlation. The SROCC between calculated variables is equivalent to the Pearson correlation coefficient between calculated variable data ranks.

$$\text{SROCC} = 1 - \frac{6\sum_{i=1}^{N}(v_i - p_i)^2}{N(N^2-1)} \tag{4}$$

In the formula, N is the number of sample pairs in the predicted and real values, v_i and p_i are the sorting positions of the real and predicted values, respectively.

RMSE quantifies the prediction accuracy of the algorithm and is defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[MOS_{\text{o}}(i) - MOS_{\text{s}}(i)\right]^2} \tag{5}$$

E. Explanation of the Comparison Algorithms

To assess the effectiveness of the proposed network, SVR [29], Quality-Net [31], and HASA-Net [43] were chosen as the comparative algorithms. In order to evaluate the impact of the extended embedding strategy on network prediction accuracy, a comparison algorithm called SQINetNE-HLlevel was used, which employs fully connected layers for dimension transformation of the audiograms without extended embedding. Additionally, SQINetNA-HLlevel, a network that does not utilize an attention module, was also included as a comparison algorithm. Based on the structure depicted in Fig. 1, the multi-head self-attention module was removed in the score prediction module, and the hidden output of BiLSTM was directly fed into a dense layer to obtain the predicted score.

TABLE III
PERFORMANCE INDICATORS OF SQINETNE-HLLEVEL AND SQINET-HLLEVEL

Net	PCC	SROCC	RMSE
SQINetNE-HLlevel	0.973	0.970	0.061
SQINet-HLlevel	0.985	0.984	0.045

SVR [29] is a well-known machine learning algorithm that has been applied to quality scoring based on FBE features. Quality-Net [31] is an end-to-end non-invasive speech quality evaluation model composed of BiLSTM and fully connected layers. Considering that HASQI ranges from 0 to 1, the unbounded output of Quality-Net was constrained during the experiments. HASA-Net [43] is a network that combines BiLSTM and attention mechanisms to jointly predict HASQI and HASPI. The network was trained using a joint prediction approach. The parameter settings of the aforementioned algorithms used in this experiment are consistent with those in the original papers.

IV. RESULTS AND ANALYSIS

A. Extended Embedding Analysis of Audiograms

To analyze the impact of the introduction of the audiograms extended embedding strategy on predictive performance, the performance of SQINetNE-HLlevel, which directly utilizes fully connected layers for audiogram dimension transformation, was compared with that of SQINet-HLlevel. The results are shown in Table III.

The evaluation results indicate that the introduction of the extended embedding strategy along the frequency axis improves the prediction performance of the network. Compared with SQINetNE-HLlevel, the PCC and SROCC of SQINet-HLlevel are improved by 0.012 and 0.014, respectively, while the RMSE is reduced by 0.016. This demonstrates that the proposed strategy outperforms SQINetNE-HLlevel in all indicators. This improvement is attributed to the fact that hearing-impaired patients have different listening gain requirements in different frequency

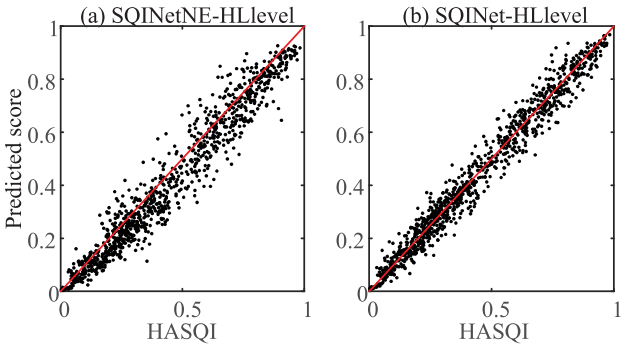

Fig. 5. Scatter plots of SQINetNE-HLlevel and SQINet-HLlevel.

TABLE IV
PERFORMANCE INDICATORS OF SQINET-HLLEVEL AND SQINETNA-HLLEVEL

Net	PCC	SROCC	RMSE
SQINetNA-HLlevel	0.943	0.940	0.088
SQINet-HLlevel	0.985	0.984	0.045

ranges. Directly using fully connected layers for audiogram dimension transformation fails to effectively capture the correspondence between audiograms and frequency ranges. In this study, specific values are assigned to each frequency range based on audiograms and then combined with spectrogram features before being fed into the network. This enables the network to achieve better prediction results that cater to the needs of hearing-impaired patients, leveraging the correspondence between frequency ranges.

Fig. 5 displays the scatter plots of the predicted results of the two networks on the test set. Overall, the scatter distributions of both networks align with the trend of the diagonal line $y = x$, indicating a high correlation between the predicted scores and the real scores. The average distance between the predicted values of SQINet-HLlevel and the diagonal line is 0.025, slightly smaller than the 0.039 for SQINetNE-HLlevel. This reflects that the addition of the extended embedding strategy can slightly improve the prediction performance of the network, which is consistent with the analysis results of performance indicators.

B. Impact of Attention

To verify the role of the multi-head self-attention mechanism in speech quality prediction networks, the performance of SQINetNA-HLlevel, which does not utilize an attention mechanism, was compared with that of SQINet-HLlevel. The results are presented in Table IV.

Compared with SQINetNA-HLlevel, SQINet-HLlevel exhibits increments of approximately 0.042 and 0.044 in PCC and SROCC, respectively, while reducing the prediction error by approximately 0.043. This indicates that the use of the attention mechanism effectively improves the correlation between the predicted scores and the real scores while reducing the prediction error.

Fig. 6 illustrates the scatter plots of the predicted scores of SQINetNA-HLlevel and SQINet-HLlevel on the test set.

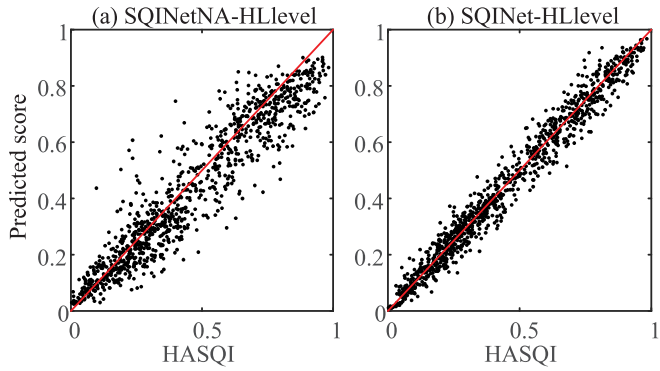

Fig. 6. Scatter plots of SQINetNA-HLlevel and SQINet-HLlevel.

TABLE V
PERFORMANCE INDICATORS WITH THE DIFFERENT NUMBER OF ATTENTION HEADS

Net	PCC	SROCC	RMSE
SQINet-HLlevel(2)	0.984	0.983	0.046
SQINet-HLlevel(4)	0.985	0.984	0.045
SQINet-HLlevel(8)	0.980	0.978	0.052

Although both networks closely align with the trend of the diagonal line $y = x$, the scatter distribution of SQINetNA-HLlevel is noticeably more dispersed. In contrast, the scatter distribution of SQINet-HLlevel is evenly distributed on both sides of the diagonal line and is closer to the diagonal line. Overall, the average distance between the scatter points of SQINet-HLlevel and the diagonal line is 0.025, considerably lower than the 0.049 for SQINetNA-HLlevel, representing a 49% reduction. This indicates that the use of attention modules effectively reduces the prediction error of the network by assigning greater weights to speech frames and relatively smaller weights to silent or noisy frames. Consequently, the final prediction of speech quality is more heavily influenced by speech frames, which aligns with the understanding that speech segments have a greater impact on overall speech quality. Therefore, the attention mechanism equips SQINet-HLlevel with the ability to distinguish between speech frames and non-speech frames, enabling the network to evaluate speech quality based on features extracted from speech frame segments. This is the main reason why its performance is considerably better than that of SQINetNA-HLlevel.

C. Impact of the Number of Attention Heads

To analyze the impact of the number of attention heads on the predictive performance of the proposed network, the effects of different numbers of attention heads on the prediction performance are investigated. As shown in Table V, the prediction results for attention head numbers of 2, 4, and 8 are presented.

From the evaluation results, it is observed that increasing the number of attention head groups from 2 to 4 leads to a slight increase of 0.001 in both the PCC and SROCC metrics, while the RMSE decreases by 0.001. This indicates a slight improvement in all metrics, but when the number of groups continues to increase to 8, the metrics start to decline. This

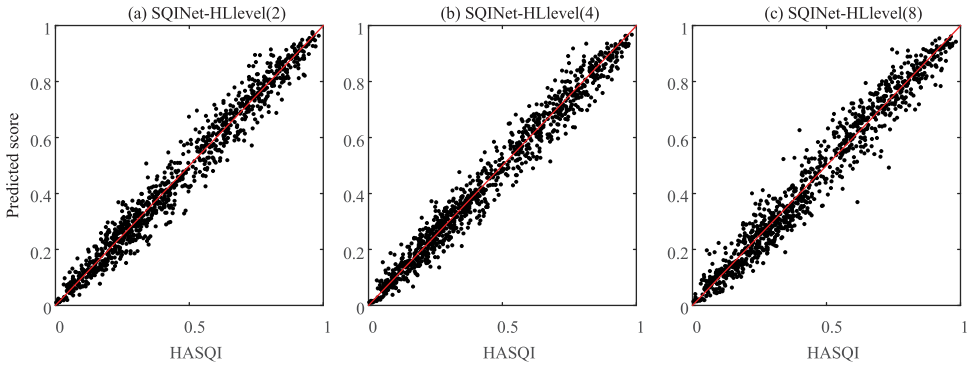

Fig. 7. Scatter plots of SQINet-HLlevel wtih the different attention head numbers.

TABLE VI
PERFORMANCE INDICATORS OF FOUR ALGORITHMS

Net	PCC	SROCC	RMSE
SVR	0.474	0.457	0.232
Quality-Net	0.823	0.820	0.165
HASA-Net	0.966	0.965	0.068
SQINet-HLlevel	0.985	0.984	0.045

suggests that controlling the number of attention heads in an appropriate range helps the model learn relevant information in different representation subspaces, while too many or too few attention heads can have a negative impact.

Fig. 7 shows the scatter plots of the prediction results for different numbers of attention heads on the test set. Overall, the scatter distributions of the three group numbers are closer to the diagonal line $y = x$. When the number of attention heads is 4, the average distance between the prediction scatter and the diagonal line is 0.025, which is smaller than the distances of 0.028 for a group number of 8 and 0.0251 for a group number of 2. This reflects that setting an appropriate number of attention heads can improve the predictive performance of the network, consistent with the analysis of the performance metrics.

D. Prediction Accuracy Analysis

The evaluation results of the proposed network and the comparative algorithms on the test set are presented in Table VI.

From Table VI, it can be observed that the PCC and SROCC of SVR exhibit a nearly double gap compared with other algorithms, and the RMSE also differs by approximately 0.1 in comparison. This indicates that SVR has limited prediction accuracy when dealing with large-scale data without effective regression. It further demonstrates that deep learning algorithms, as compared with traditional machine learning algorithms, are more effective in processing input features, learning scoring mapping rules, and achieving more accurate speech quality predictions. Quality-Net performs worse than SQINet-HLlevel in terms of both prediction score correlation and error. Although both correlation indicators demonstrate high correlation evaluation, i.e., PCC $>$ 0.8, the proposed network still exhibits a correlation advantage of 0.162 over Quality-Net. This is attributed to the lack of convolutional layers in Quality-Net for further feature extraction, making it challenging to effectively extract deep information from input features. In comparison to Quality-Net, HASA-Net demonstrates advantages in all indicators due to its introduction of a multitask learning strategy, which jointly predicts speech comprehensibility and quality scores, thereby enhancing the model's ability to extract key information through the attention mechanism. SQINet-HLlevel improves PCC and SROCC by 0.019 compared with HASA-Net, while reducing RMSE by 0.023. This indicates that the extended embedding strategy for audiograms and the introduction of the multi-head self-attention mechanism can effectively enhance the prediction ability of quality scores.

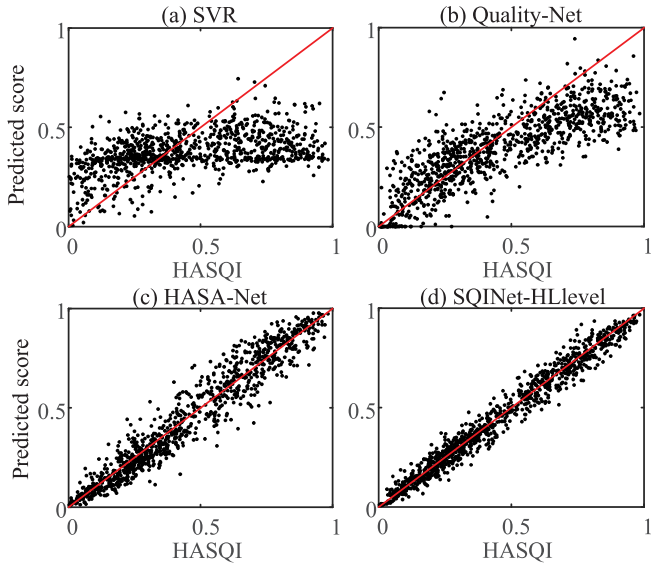

Fig. 8. Scatter plots of four algorithms.

TABLE VII
PARAMETER SIZE AND TIME COMPLEXITY FOR THREE NETWORKS

	Quality-Net	HASA-Net	SQINet-HLlevel
Parameters/million	0.12	0.45	1.401
FLOPs/million	0.24	0.63	17

Fig. 8 presents a scatter plot of the predicted results of the aforementioned networks on the test set. To facilitate the observation of deviations between the predicted values and the real

TABLE VIII
NETWORK PARAMETER SETTINGS OF DIFFERENT NETWORK CHANNELS (128/64/32)

Module name	Layer name	Input size	Hyperparameter	Output size
Frame level feature extraction module	Conv_1	$2 \times T \times 257$	3×3, (1, 1), 32	$32 \times T \times 257$
	LPPool2d	$32 \times T \times 257$	$1 \times 4{,}4$	$32 \times T \times 64$
	Conv_2	$32 \times T \times 64$	3×3, (1, 1), (128/64/32)	$(128/64/32) \times T \times 64$
	Conv_3	$(128/64/32) \times T \times 64$	3×3, (1, 1), (128/64/32)	$(128/64/32) \times T \times 64$
	LPPool2d	$(128/64/32) \times T \times 64$	$1 \times 4{,}4$	$(128/64/32) \times T \times 16$
	Conv_4	$(128/64/32) \times T \times 16$	3×3, (1, 1), (128/64/32)	$(128/64/32) \times T \times 16$
	Conv_5	$(128/64/32) \times T \times 16$	3×3, (1, 1), 16	$(128/64/32) \times T \times 16$
	LPPool2d	$(128/64/32) \times T \times 16$	$1 \times 4{,}4$	$(128/64/32) \times T \times 4$
	Reshape	$(128/64/32) \times T \times 4$	/	$T \times (512/256/128)$
Score prediction module	BiLSTM	$T \times (512/256/128)$	128	$T \times 256$
	Multi-head Attention	$T \times 512$	256, h=4	$T \times 256$
	MaxPool1d	$T \times 256$	1	1×256
	Dense	1×256	1	1

TABLE IX
COMPUTATION COMPLEXITY OF SQINET-HLLEVEL WITH DIFFERENT NETWORK CHANNELS (128/64/32)

	Frame level feature extraction module	Score prediction module	Total
Parameters/million	0.481/0.130/0.0328	0.921/0.659/0.528	1.401/0.789/0.5608
FLOPs/million	16.723/4.901/1.272	0.659/0.398/0.266	17/5.2991.538

HASQI, the diagonal line $y = x$ is also plotted in the figure. The scatter plot reveals that the predicted scores of SVR have a dense distribution around 0.3, indicating that SVR tends to provide a prediction value of approximately 0.3 regardless of the real HASQI. In comparison, Quality-Net's scatter distribution closely follows the trend of the diagonal line and exhibits fewer outlier points. HASA-Net's scatter plot is evenly distributed on both sides of the diagonal line, indicating a higher correlation with real HASQI. SQINet-HLlevel's scatter plot is more consistent with the diagonal line, suggesting more accurate predictions for various speech quality scores.

E. Discussion on Algorithm Complexity

To assess the performance of the algorithm more comprehensively, we compare the parameter size and time complexity of three networks. We employ Python's profile tool to automatically analyze the performance of these networks. The specific indicators are shown in Table VII.

From the table, it is evident that SQINet-HLlevel has a larger parameter size and time complexity compared to Quality-Net and HASA-Net. To further investigate the reasons behind this, we separately calculate the computational complexity of the frame level feature extraction module and score prediction module. Additionally, we modify the number of some network channels (128, 64 and 32) and accordingly adjust the number of nodes in BiLSTM. The specific modifications are outlined in Table VIII.

The computation complexity is summarized in Table IX. As the number of channels decreases, the parameter size and time complexity decrease rapidly. Notably, compared to the 128-channel case, the parameter size at 64 channels and 32 channels decreases by 1.78 and 2.5 times, respectively, while the time complexity decreases by 3.2 and 11 times, respectively. It is observed that the score prediction module mainly impacts the parameter size, while the frame level feature extraction module affects computational complexity. This is because, although convolutional networks require fewer parameters, they tend to be time-consuming. In addition, when the number of channels is 32, the parameter size and time complexity of the SQINet-HLlevel is still higher than that of HASA-Net. However, the computational complexities of the two algorithms are essentially on the same order of magnitude.

TABLE X
PERFORMANCE INDICATORS WITH DIFFERENT NETWORK CHANNELS

Net	PCC	SROCC	RMSE
SQINet-HLlevel-32	0.983	0.981	0.048
SQINet-HLlevel-64	0.985	0.983	0.046
SQINet-HLlevel-128	0.985	0.984	0.045

However, as shown in Table X, compared to the 128- channel case, the PLCC, SROCC, and RMSE values at 32 channels decrease by 0.002, 0.003, and 0.003, respectively. Based on the above analysis, it is evident that algorithm complexity is not the primary factor influencing the performance of the proposed algorithm. Instead, the appropriate model can be selected based on a balance between computational complexity and performance.

V. CONCLUSION

This paper proposed a non-invasive speech quality evaluation network specifically designed for hearing aids. The network takes the spectrogram of distorted speech and the audiograms of the hearing-impaired patient as input, and predicts the HASQI as the output. The proposed network consists of several key components. Firstly, frame-level deep features are extracted from the input features using a CNN. The CNN is responsible for capturing local patterns and extracting high-level representations from the input speech signal. Next, the deep features are fed into a BiLSTM network, which models the temporal dependencies in the speech signal. The BiLSTM network is able

to capture the long-term dependencies and context information, enabling the network to understand the sequential nature of speech. To further enhance the network's ability to distinguish the importance of different speech frames, a multi-head self-attention layer is employed. This layer integrates the context information of the entire sequence and assigns different weights to different frames based on their relevance to the overall speech quality. This attention mechanism allows the network to focus on the most informative frames and disregard irrelevant or noisy frames. Finally, the weighted features are linearly mapped to a quality score using a fully connected layer. This mapping process converts the learned representations into a quantitative measure of speech quality, which can be used to evaluate the performance of hearing aids. Experimental results demonstrated that the proposed network achieves higher accuracy in speech quality evaluation for hearing-impaired individuals and exhibits good robustness to noise.

However, it is important to acknowledge the limitations of the proposed network. To address these limitations, future research directions can be pursued.

Firstly, conducting hearing tests to obtain speech and subjective rating labels processed by real hearing aids would provide more realistic and reliable data for training and testing the model. This would enable the modification of the model structure based on the results, leading to improved performance.

Secondly, further enriching the model structure and exploring additional types of auxiliary objectives can help the network adapt to the speech quality evaluation of hearing aids in complex environments. This would involve incorporating additional components or objectives that capture specific aspects of speech quality, such as voice activity or quality level, etc.

Thirdly, in low resource scenarios, the proposed algorithm is still relatively complex. Therefore, reducing network complexity while maintaining accuracy is also a valuable research direction.

By addressing these research directions, the proposed network can be further optimized and enhanced to better serve the speech quality evaluation needs of hearing aids in real-world scenarios.

Declaration of competing interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C. S. K. K. C. A., "World report on hearing," World Health Org., Geneva, Switzerland, 2021.

[2] E. M. Picou, "MarkeTrak 10 (MT10) survey results demonstrate high satisfaction with and benefits from hearing aids," in *Seminars in Hearing*, New York, NY, USA: Thieme Med. Publishers, 2020, pp. 021–036.

[3] J. M. Kates and K. H. Arehart, "The hearing-aid audio quality index (HAAQI)," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 2, pp. 354–365, Feb. 2016.

[4] J. Richard, V. Zimpfer, and S. Roth, "Comparison of objective and subjective methods for evaluating speech quality and intelligibility recorded through bone conduction and in-ear microphones," *Appl. Acoust.*, vol. 211, 2023, Art. no. 109576. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X23003742

[5] 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.

[6] M. S. Kavalekalam, J. K. Nielsen, J. B. Boldt, and M. G. Christensen, "Model-based speech enhancement for intelligibility improvement in binaural hearing aids," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 1, pp. 99–113, Jan. 2019.

[7] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, "Personalized speech enhancement: New models and comprehensive evaluation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 356–360.

[8] D. Sharma, L. Meredith, J. Lainez, D. Barreda, and P. A. Naylor, "A non-intrusive PESQ measure," in *Proc. IEEE Glob. Conf. Signal Inf. Process.*, 2014, pp. 975–978.

[9] Q. Li, W. Lin, Y. Fang, and D. Thalmann, "Bag-of-words representation for non-intrusive speech quality assessment," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, 2015, pp. 616–619.

[10] F. Rahdari, R. Mousavi, and M. Eftekhari, "An ensemble learning model for single-ended speech quality assessment using multiple-level signal decomposition method," in *Proc. 4th Int. Conf. Comput. Knowl. Eng.*, 2014, pp. 189–193.

[11] M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel, and L.-T. Chia, "Nonintrusive quality assessment of noise suppressed speech with mel-filtered energies and support vector regression," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 4, pp. 1217–1232, May 2012.

[12] M. Hakami and W. B. Kleijn, "Machine learning based non-intrusive quality estimation with an augmented feature set," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 5105–5109.

[13] M. H. Soni and H. A. Patil, "Effectiveness of ideal ratio mask for non-intrusive quality assessment of noise suppressed speech," in *Proc. 25th Eur. Signal Process. Conf.*, 2017, pp. 573–577.

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

[15] 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.

[16] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[17] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "MBNET: MOS prediction for synthesized speech with mean-bias network," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 391–395.

[18] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (HASQI) version 2," *J. Audio Eng. Soc.*, vol. 62, no. 3, pp. 99–117, 2014.

[19] J. Kates, "An auditory model for intelligibility and quality predictions," *Proc. Meetings Acoust.*, vol. 19, no. 1, 2013, Art no. 050184.

[20] J. M. Kates, K. H. Arehart, M. C. Anderson, R. K. Muralimanohar, and L. O. Harvey Jr, "Using objective metrics to measure hearing-aid performance," *Ear Hear.*, vol. 39, no. 6, pp. 1165–1175, 2018.

[21] J. M. Kates and K. H. Arehart, "An overview of the HASPI and HASQI metrics for predicting speech intelligibility and speech quality for normal hearing, hearing loss, and hearing aids," *Hear. Res.*, vol. 426, 2022, Art. no. 108608. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378595522001769

[22] R. Huber, V. Parsa, and S. Scollie, "Predicting the perceived sound quality of frequency-compressed speech," *PLoS One*, vol. 9, no. 11, 2014, Art. no. e110260.

[23] R. Huber and B. Kollmeier, "PEMO-Q–A new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

[24] D. Suelzle, V. Parsa, and T. H. Falk, "On a reference-free speech quality estimator for hearing aids," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. EL412–EL418, 2013.

[25] Y. Feng and F. Chen, "Nonintrusive objective measurement of speech intelligibility: A review of methodology," *Biomed. Signal Process. Control*, vol. 71, 2022, Art. no. 103204. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421008016

[26] T. H. Falk et al., "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[27] H. Salehi and V. Parsa, "On nonintrusive speech quality estimation for hearing aids," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2015, pp. 1–5.

[28] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, P. A. Naylor, M. Brookes, and M. Huckvale, "Data driven method for non-intrusive speech intelligibility estimation," in *Proc. IEEE 18th Eur. SignalProcess. Conf.*, 2010, pp. 1899–1903.
[29] H. Salehi, D. Suelzle, P. Folkeard, and V. Parsa, "Learning-based reference-free speech quality measures for hearing aid applications," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 12, pp. 2277–2288, Dec. 2018.
[30] J. Ooster and B. T. Meyer, "Improving deep models of speech quality prediction through voice activity detection and entropy-based measures," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 636–640.
[31] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *Proc. Annu. Conf. Int. Speech Commun. Assoc., INTERSPEECH*, Sep. 2018, pp. 1873–1877.
[32] X. Dong and D. S. Williamson, "An attention enhanced multi-task model for objective speech assessment in real-world environments," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 911–915.
[33] R. E. Zezario, S.-W. Fu, C.-S. Fuh, Y. Tsao, and H.-M. Wang, "STOI-net: A deep learning based non-intrusive speech intelligibility assessment model," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 482–486.
[34] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, 2000.
[35] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 7125–7129.
[36] C.-C. Lo et al., "MOSNet: Deep learning based objective assessment for voice conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc., INTERSPEECH*, Sep. 2019, pp. 1541–1545.
[37] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "MBNet: MOS prediction for synthesized speech with mean-bias network," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 391–395.
[38] B. Cauchi, K. Siedenburg, J. F. Santos, T. H. Falk, S. Doclo, and S. Goetze, "Non-intrusive speech quality prediction using modulation energies and LSTM-network," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 7, pp. 1151–1163, Jul. 2019.
[39] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 886–890.
[40] R. K. Jaiswal and R. K. Dubey, "Multiple time-instances features based approach for reference-free speech quality measurement," *Comput. Speech Lang.*, vol. 79, 2023, Art. no. 101478.
[41] Y. Choi, Y. Jung, and H. Kim, "Neural MOS prediction for synthesized speech using multi-task learning with spoofing detection and spoofing type classification," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 462–469.
[42] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 54–70, 2023.
[43] H.-T. Chiang et al., "HASA-net: A non-intrusive hearing-aid speech assessment network," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2021, pp. 907–913.
[44] F. Drakopoulos and S. Verhulst, "A neural-network framework for the design of individualised hearing-loss compensation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 2395–2409, 2023.
[45] T. van de Laar and B. de Vries, "A probabilistic modeling approach to hearing loss compensation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 11, pp. 2200–2213, Nov. 2016.
[46] T. Ma, Y. Wei, and X. Lou, "Reconfigurable nonuniform filter bank for hearing aid systems," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 758–771, 2022.
[47] M. C. Killion, "The 3 types of sensorineural hearing loss: Loudness and intelligibility considerations," *Hear. J.*, vol. 46, no. 11, pp. 31–36, 1993.
[48] C. K. Reddy et al., "Interspeech 2021 deep noise suppression challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc., INTERSPEECH*, 2021, vol. 2, pp. 801–805.
[49] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.
[50] L. Zhang and W.-G. Gong, "Improved wiener filtering speech enhancement algorithm," *Jisuanji Gongcheng yu Yingyong (Comput. Eng. Appl.)*, vol. 46, no. 26, pp. 129–131, 2010.
[51] X. Fang and M. J. Nilsson, "Noise reduction apparatus and method," 2004, U.S. Patent 6 757 395.

Ruiyu Liang (Member, IEEE) received the Ph.D. degree from Southeast University, Nanjing, China, in 2012. He is currently a Professor with the Nanjing Institute of Technology, Nanjing. His research interests include speech signal processing and signal processing for hearing aids.

Yue Xie (Member, IEEE) received the Ph.D. degree from Southeast University, Nanjing, China, in 2021. He is currently a Lecturer with the Nanjing Institute of Technology, Nanjing. His research interests include speech signal processing and speech emotion recognition.

Jiaming Cheng is currently working toward the Ph.D. degree with Southeast University, Nanjing, China. His research interests include speech enhancement and machine learning.

Cong Pang is currently working toward the Ph.D. degree with Southeast University, Nanjing, China. His research interests include multichannel speech enhancement and machine learning.

Björn Schuller (Fellow, IEEE) received the Diploma in electrical engineering and information technology, the doctoral degree in electrical engineering and information technology (automatic speech and emotion recognition), and the Habilitation degree and Adjunct Teaching Professorship in electrical engineering and information technology (signal processing and machine intelligence) from the Technical University of Munich, Munich, Germany, in 1999, 2006, and 2012, respectively. He is currently a Full Professor of artificial intelligence, and the Head of GLAM – the Group on Language, Audio & Music, Imperial College London, London, UK. He is a Full Professor and the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany. He is also a co-founding CEO and the current Chief Scientific Officer of audEERING, Gilching, Germany, and an Associate of the Swiss Center for Affective Sciences with the University of Geneva, Geneva, Switzerland. He has authored or coauthored five books and more than 700 publications in peer reviewed books, journals, and conference proceedings which led to more than 20000 citations (h-index = 68). Dr. Schuller is the President Emeritus of the Association for the Advancement of Affective Computing (AAAC), an elected Member of the IEEE Speech and Language Processing Technical Committee, and a Senior Member of ACM.