Speech Intelligibility Prediction for Hearing Aids Using an Auditory Model and Acoustic Parameters

Benita Angela Titalim*, Candy Olivia Mawalim*, Shogo Okada, and Masashi Unoki
Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
Email: {s2110104, candylim, okada-s, unoki}@jaist.ac.jp

***Abstract*—Objective speech intelligibility (SI) metrics for hearing-impaired people play an important role in hearing aid development. The work on improving SI prediction also became the basis of the first Clarity Prediction Challenge (CPC1). This study investigates a physiological auditory model called EarModel and acoustic parameters for SI prediction. EarModel is utilized because it provides advantages in estimating human hearing, both normal and impaired. The hearing-impaired condition is simulated in EarModel based on audiograms; thus, the SI perceived by hearing-impaired people is more accurately predicted. Moreover, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) and WavLM, as additional acoustic parameters for estimating the difficulty levels of given utterances, are included to achieve improved prediction accuracy. The proposed method is evaluated on the CPC1 database. The results show that the proposed method improves the SI prediction effects of the baseline and hearing aid speech prediction index (HASPI). Additionally, an ablation test shows that incorporating the eGeMAPS and WavLM can significantly contribute to the prediction model by increasing the Pearson correlation coefficient by more than 15% and decreasing the root-mean-square error (RMSE) by more than 10.00 in both closed-set and open-set tracks.**

I. Introduction

Building communication and relationships with others involves hearing ability. Hearing ability also helps humans receive information from the outside world to understand the situations happening around them. Unaddressed hearing loss problems can negatively impact many aspects of life, especially for elderly individuals. Hearing loss problems can introduce communication barriers and an inability to relate to other people [1] and cause other social issues [2], [3]. The Lancet Commission has also reported that hearing loss in the mid- and late-life stages affects dementia [4], which may result in reduced quality of life.

Hearing aids are used as a solution to correct the perception of hearing impairment, specifically for those with sensorineural hearing loss. A hearing aid system collects sound from the environment, analyzes it, and adjusts based on the user's hearing loss level. Hearing loss can be explained based on the frequency selectivity in the inner ear that is simultaneously mapped to the auditory threshold with sound loudness. The auditory threshold of hearing loss is always lifted above the normal hearing (NH) threshold, depending on the degree of damage. Another explanation is related to the contribution of outer hair cells (OHCs) to the signal compression transmitted to the inner ear. The dynamic range compression in individuals with hearing loss changes the compression ratio, contributing to the elevated auditory threshold [5]. In addition, the loss of inner hair cells (IHCs) affects the perception of hearing loss by providing additional attenuation [6]. Therefore, in addition to amplifiers, some features are added to hearing aids to improve hearing ability.

The evaluation of the signal processing effects in hearing aids is based on subjective and objective assessments. Although a subjective evaluation gives the most correlated results to measure hearing loss, the process might take much time and a certain number of subjects with hearing loss. In contrast, the preparation required to conduct an objective evaluation is less expensive [7]. Moreover, objective metrics are usually preferable for assessments since they correlate highly with the quality and intelligibility measurements in subjective tests [8]. Despite these advantages, the development of objective metrics specifically for hearing aid systems is still in progress [9].

The objective intelligibility measurement that is often considered in hearing aid development is the hearing aid speech prediction index (HASPI) [10]. Despite its high accuracy, the HASPI model has evaluation limitations, an inability to handle the binaural data and invalidity for tonal languages. In the first Clarity Prediction Challenge (CPC1), the modified binaural short-time objective intelligibility (MBSTOI) model was incorporated into the MBSG hearing loss model [11] and eventually included in the baseline model. In addition, several prediction methods were proposed by CPC1 participants, such as metrics using a multibranched network (MBI-Net) [12] and the unsupervised uncertainty measurement of automatic speech recognition (ASR) [13].

The purpose of this paper is to predict speech intelligibility (SI) using the auditory periphery model and acoustic parameters. We propose a method for predicting SI using EarModel, WavLM, and the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS). The idea of utilizing the additional acoustic parameters is based on the study of predicting speech intelligibility with acoustic parameters measured from room impulse response [14]. The research found that the parameters from room acoustic highly correlate with phoneme and word recognition rate (PRR and WRR). The higher the PRR and WRR simultaneously, the more likely an increase in speech intelligibility occurs.

We hypothesize that EarModel can improve SI prediction for

*These authors contributed equally to this work

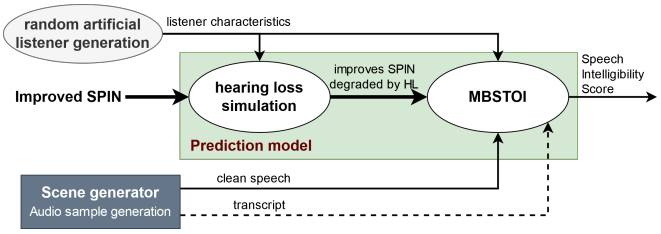

Fig. 1. Baseline model for the CPC1.

hearing aids because it estimates normal and impaired hearing. Furthermore, the additional acoustic parameters can be used to estimate utterance characteristics (e.g., sentence difficulty, the statistical parameters of speech characteristics, etc.), which are beneficial for improving the prediction accuracy of the model. To achieve our goal, we investigate the physiological auditory model and develop an SI prediction method by using a machine learning approach.

The rest of the paper is comprised of a few sections. Section 2 contains information about the CPC1, including its rules and scenarios. The details of the related work, that is, the HASPI and baseline MBSTOI, are explained in Section 3. Section 4 describes the proposed method. Section 5 describes the experimental data, evaluation method, and results. Section 6 concludes the findings.

II. Clarity Challenge

The Clarity Challenge[1] is a series of machine learning tasks that focus on improving signal processing in hearing aids. The challenge was created based on the need to solve the current problem concerning hearing aids: the difficulty of discriminating speech from noise in hearing loss cases [15], which brings dissatisfaction to users [16]. The challenge utilizes machine learning since the performance of some deep learning processes seems promising. Unfortunately, a learning algorithm should form new ways of processing audio to estimate how the sound is degraded by hearing loss. Moreover, it should consider binaural hearing since the collaboration between both ears to reduce noise works better than the use of a single ear, as it allows the brain to locate and boost speech over noise [17].

The project was divided into the Clarity Enhancement Challenge (CEC) and the CPC. In the first challenge, the enhancement challenge focused on maximizing the SI score, and the prediction challenge focused on evaluating the enhancement model based on a hearing loss simulation. The components of the baseline model for the prediction challenge are described in Fig. 1[2]. The available information for the baseline prediction model included the listener characteristics, the improved speech perception in noise (SPIN) generated from the hearing aid processing in the CEC, clean speech, and a transcript produced by a scene generator. As in the baseline figure, the prediction model consisted of a hearing loss simulation and an intelligibility model. However, the participants could reconfigure their own design for the prediction model. In the future, SI scores will be used to develop better hearing aid systems using machine learning techniques.

III. Related Work

The CPC1 challenge involved the prediction of SI when listeners perceived SPIN processed by a hearing aid. The given baseline system consisted of a hearing loss simulation and SI models for binaural hearing. Nevertheless, the configuration of the prediction model could be changed as needed. The remaining part of this section describes the existing hearing-impaired intelligibility models the HASPI and baseline MBSTOI models.

A. HASPI

The HASPI [10], [18] is an SI metric for normal and impaired hearing. It takes a monaural clean signal without processing and a monaural degraded signal considering the hearing loss condition and performs parallel processing on both signals in EarModel 2. The ability of EarModel to simulate processing with hearing loss motivates us to incorporate this model into the proposed method. The processing approach in EarModel will be explained in Section 4.

The envelope output of EarModel, in dB, is low-pass filtered and subsampled before being converted into a time-varying spectrum. As a result, the dB envelope represents the log spectrum on the frequency scale for each sample according to the subsampling rate. The resulting short-time spectra fit five basis functions with the $\frac{1}{2}$ cycle to the $2\frac{1}{2}$ cosine spanning the spectrum from 80 to 8000 Hz. Each cepstral coefficient is passed through a modulation filter bank with ten filters and generates fifty filtered sequences. For each sequence, the clean and degraded signals are compared using the normalized cross-correlation averaged across the basis functions to produce ten modulation filter outputs. Last, the intelligibility index is estimated by mapping the ten modulation filter outputs using neural networks [18].

The advantage of the HASPI mainly comes from the presence of an auditory model to consider normal and impaired hearing. In addition, reference and processed signal fidelity measurements have been evaluated over a wide range of processing conditions, including additive stationary and modulated noise, nonlinear distortion, noise suppression, dynamic range compression, frequency compression, feedback cancellation, and linear filtering. Due to these advantages, the HASPI has been used in several applications, especially for hearing aid assessment [9]. On the other hand, this metric cannot measure SI for binaural inputs, making it invalid for tonal languages, and the evaluation process is limited to the training data.

B. Baseline Modified Binaural Short-time Objective Intelligibility (MBSTOI)

The baseline prediction metric for the CPC1[2] is baseline MBSTOI as shown in Fig. 1. The baseline MBSTOI consists of

[1]https://claritychallenge.org/
[2]https://claritychallenge.org/docs/cpc1/cpc1_intro

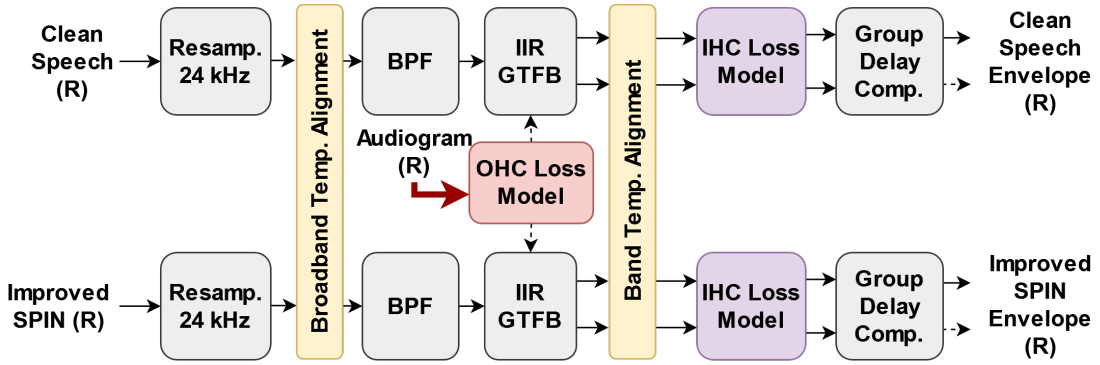

Fig. 2. EarModel processing for the right ear (R) [19].

a hearing loss module and MBSTOI as the speech intelligibility module [20]. The hearing loss module is based on the auditory model developed by Moore et al. in the Auditory Perception Group at the University of Cambridge [11]. The auditory model attenuates the binaural reference and processed signal in each frequency band based on the audiogram of each listener to simulate the auditory threshold elevation. The loudness recruitment simulates by a filterbank with two times broadening. Then, the frequency selectivity reduction performed by the hearing loss severity is defined by the average loss in dB hearing loss.

The intelligibility model is the MBSTOI, which is constructed based on the STOI metric [21]. The model utilizes the binaural processing advantages of the deterministic binaural STOI (DBSTOI) [22] for both multiple and spatially distributed interferers. The inputs of the baseline MBSTOI model are the binaural degraded $y_{l/r}(n)$and the clean reference signal $x_{l/r}(n)$. The input degraded signal in the MBSTOI is assumed to be the summation of clean input speech with an uncorrelated additive noise component $d_{l/r}(n)$.

$$y_{l/r}(n) = x_{l/r}(n) + d_{l/r}(n) \tag{1}$$

Each signal is analyzed by a short-time discrete Fourier transform (DFT) to generate coefficients. The $\frac{1}{3}$ octave equalization-cancellation (EC) operation combines and modifies the coefficient to align the interferer and distortion in both ears. Next, the combined DFT coefficient is mapped in the $\frac{1}{3}$ octave band to produce signal power envelopes. Finally, the envelopes are converted into zero-mean vectors and used to calculate the estimated correlation, as in the STOI measurement approach [21].

Based on an application, the baseline MBSTOI outperforms the other existing method, i.e., the HASPI, in predicting SI. The MBSTOI is capable of predicting intelligibility in cases with fluctuating interferers and reverberation. It corrects the delay per ear caused by hearing loss. Although the degraded signal assumption may limit its applicability in some situations, the baseline performs identical, linear, slow-changing, constant processing in the $\frac{1}{3}$ octave band for both ears, which makes its implementation relevant to a broader range of noise conditions. However, the delay induced after hearing aid processing for HI cannot be corrected by the baseline. Another problem is that the MBSTOI prediction process utilizes a correlation function and yields signal-level insensitivity, which results in a high intelligibility score when the sound level falls below the set auditory thresholds.

IV. Proposed Method

Figure 3 shows the overall block diagram of the proposed method. The method consists of EarModel to extract signal envelope, pre-trained self-supervised learning (SSL) mode and eGeMAPS extractor to obtain acoustic parameters which may introduce further speech intelligibility improvement [14]. The descriptions of the feature extraction process and the SI model are explained as follows.

A. Feature Extraction

1) EarModel: EarModel, developed by James Kates [19], is used to extract the input signal envelope. The proposed method takes clean binaural speech and improved SPIN and then splits them into left and right signals to be processed in an equal and parallel manner. The clean speech and improved SPIN of the right ear process are shown in Fig. 2. Each signal is resampled to 24 kHz to ensure equal shapes for all cochlear filters before aligning them via broadband temporal alignment. Next, processing is performed through a bandpass filter, such as the middle ear filter, and gammatone filterbank (GTFB)[3]. The decomposition into the 32 channel and the impulse response of the gammatone function in Eq. (2) is converted into a fourth-order digital filterbank whose describe in impulse-invariant transform [6] as in Eq. (3).

$$g(n) = A(nT)^{N-1} e^{-2\pi BnT} \cos(2\pi f_c nT) \tag{2}$$

$$H(z) = T^3 \times \frac{(az^{-1})(a^2z^{-2} + 4az^{-1} + 1)}{1 - 4az^{-1} + 6a^2z^{-2} - 4a^3z^{-3} + a^4z^{-4}} \tag{3}$$

with

$$a = e^{-b \times 2\pi / f_s \times B} \tag{4}$$

$$B = 1.019 \times \text{ERB} \tag{5}$$

[3]https://staffwww.dcs.shef.ac.uk/people/N.Ma/resources/gammatone/

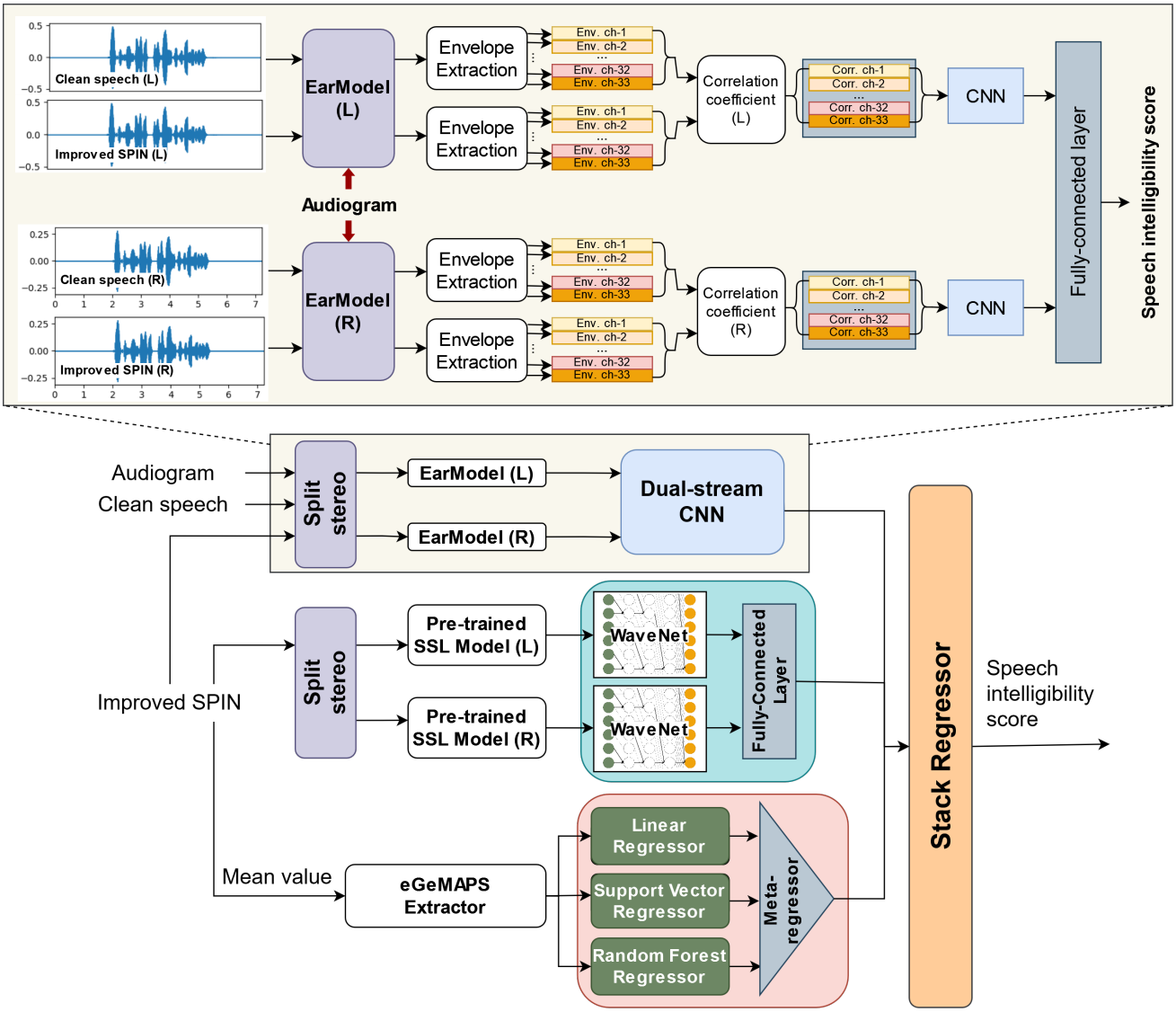

Fig. 3. Block diagram of proposed method

$$\text{ERB} = 24.7 + \frac{f_c}{9.26449} \tag{6}$$

where A is the signal amplitude; N is the filter order (which is set to 4 to model human hearing); B is the filter bandwidth based on ERB scale; f_c is the filter center frequency with a range from 80 to 8000 Hz; T is the sampling period equals to $1/f_s$; a is the filter coefficient depending on the input; and b is the bandwidth for input signal relative to ear condition. b is set to b^{NH} for clean speech and b^{HI} for improved SPIN.

The bandwidth of the GTFB for normal hearing is based on the equivalent rectangular bandwidth (ERB) scale [23]. Furthermore, the bandwidth for HI is influenced by the OHC loss model and utilizes an input audiogram to increase the cochlear filter in proportion to the dynamic range compression (DRC) reduction based on [24], [25], [26]; the compression gain for each control signal is expressed as follows.

$$G_{comp}(i) = -\text{attnOHC} - \left(1 - \frac{1}{CR}\right)\left(\theta_{\text{low}} - \hat{E}_c(i)\right) \tag{7}$$

where

$$\hat{E}_c(i) = \max\left(\theta_{\text{low}}, \left(\min\left(\hat{E}_c(i), \theta_{\text{high}}\right)\right)\right). \tag{8}$$

i, $\hat{E}_c(i)$, θ_{low}, θ_{high}, and CR are the channel number, control signal envelope in dB, a lower threshold equal to $(\text{attnOHC} + 30)$, the highest threshold equal to 100 dB, and the compression ratio, respectively [19], [27]. Then, the filter bandwidth approximation relative to normal hearing is described in Eq. (9). attnOHC indicates the loss caused by OHC damage in dB, and the maximum value is 50 dB [6], [19].

$$b^{\text{HI}} = \left(1 + \left(\frac{\text{attnOHC}}{50}\right) + 2 \times \left(\frac{\text{attnOHC}}{50}\right)^6\right) b^{\text{NH}} \tag{9}$$

The output of GTFB is 32 channel signals that are represented as temporal amplitude envelopes and temporal fine structures (BM vibration) of each channel signal. Each signal envelope is temporally aligned again with its BM vibration to reduce the loss of speech information in the envelope extraction step. In addition to OHC damage, the IHC loss that simulates the IHC firing rate adaptation [28] is added to generate additional attenuation [6] to the signal envelope based on an equivalent RC circuit model. Furthermore, the temporal envelope delay and fine structures are corrected for each channel using a group delay compensator.

2) Pre-trained Self-Supervised Learning (SSL) Model: Automatic speech recognition (ASR) is a technique for transcribing a spoken language. ASR requires clean speech to produce the ratio of the number of errors in the transcript to the total number of spoken words, which is called the word error rate (WER). However, the limitation of this technique is mainly its demand for large quantities of labeled training data. The solution for this issue is the use of an SSL framework [29].

SSL is a learning process used to build a better downstream ASR model [29], [30]. The process in pretrained SSL usually consists of two phases: pretraining on unlabeled data and fine-tuning the model on a specific dataset. WavLM [31] is used as the pretrained model in the proposed method. The WavLM model includes a masked prediction loss and a denoising process for SSL, making it different from other models, such as wav2vec2.0 [30] and HuBERT [32].

The architecture mode of WavLM contains transformer and convolutional neural network (CNN) encoders. The convolutional encoder contains seven blocks of temporal convolution with layer normalization and a Gaussian error linear unit (GELU) activation function layer. Furthermore, the transformer encoder is composed of 512 channels, resulting in output transformation and gated relative position bias in the network calculation. During training, WavLM randomly transforms the input wav file, and approximately 50% of the input signal is randomly covered and labeled according to the covered position predicted in the output [31].

3) eGeMAPS extractor: The last model is the eGeMAPS [33], which extracts the acoustic parameter set with a toolkit named openSMILE [34]. The result of the extraction process is that the frequency, energy, and related spectral parameters are specified, as shown in [33]. The reason for including this model is based on the speech perception of hearing-impaired people: distortion based on sensorineural hearing loss changes the pitch perception, frequency discrimination, and amplitude modulation detection processes. This indirectly means that the speaker's gender influences his or he speech recognition in noisy environments. Hearing-impaired people have more difficulty perceiving a female voice since it has a higher F0 average, higher spectral energy above 4 kHz, and lower energy below 4 kHz [35].

B. SI Model

After each feature extraction step, the SI score is calculated using specific machine learning techniques. First, the Pearson correlation coefficient of the signal envelope in each ear for EarModel is calculated. Then, the resulting correlation coefficients for 32 channels pass through a typical dual-stream CNN and a fully connected layer to calculate the SI score, as shown in Fig. 3. We chose CNN encoders to reduce the high-dimensional speech feature input. Consequently, the time and space complexity using the CNN could be reduced significantly from those of the traditional neural networks.

Subsequently, a WaveNet was utilized to learn the input from binaural pre-trained SSL model. WaveNet [36] stacks ten 1-D convolutional layers, doubling the dilation rate by 1,2,4,... at every layer. Before every layer, the left-padded input sequences also include the same number of zeros in the dilatation rate to maintain the same sequence length across the network.

Finally, for eGeMAPS feature extraction, the SI score is calculated by applying various regression analyses, such as linear regression, support vector regression, and random forest regression. Then, the meta-regressor combines the analysis results and compares and synthesizes them into an SI score. The overall proposed method is illustrated in Fig. 3. As the final process of the proposed model, another stack regressor synthesizes the SIs derived from the combined models to predict the final SI score.

V. Evaluation

A. Dataset

The experiment utilizes the CPC1 dataset[4]. This dataset consists of data related to scenes (the scene dataset) and metadata. The scene dataset consists of wav files of generated scenes, interference signals, clean speech (target) that is convolved with the anechoic binaural room impulse response (BRIR) for each ear, and improved SPIN obtained from hearing aid processors. Each target speech is spoken by a British English speaker. Six speakers, ten hearing aid processors from the first CEC [5], and 27 hearing-impaired listeners are involved in the dataset. The metadata provide detailed information about the scenes, listeners, and transcripts. The listener's characteristics include a pure-tone air conduction audiogram; a digit triplet test (DTT) [37]; a Glasgow hearing aid benefit profile questionnaire (GHABP) [38]; and the speech and spatial qualities of the hearing questionnaire (SSQ12) [39]. The DTT, GHABP, and SSQ12 data contain missing data; thus, further processing is required before utilized these data to develop a prediction model. Similar to the baseline system, we only utilize the audiogram for the proposed method.

The CPC1 contains two tracks: a closed set as track 1 and an open set as track 2. The test data in the closed-set track consists of all unseen scenes (the listeners and HA processors are all seen in the training data). Additionally, the test data in the open-set track consist of all unseen scenes from unseen listeners and unseen HA processors. The training and test data distributions for both tracks are different and do not overlap. For track 1, the data are split into training/development data

[4]https://claritychallenge.org/clarity_CPC1_doc/docs/cpc1_data
[5]https://claritychallenge.org/docs/cec1/cec1_intro

(4,863 scenes) and test data (2,421 scenes). Track 2 consists of 3,580 scenes for training/development data and 632 scenes for test data.

B. Evaluation Metrics

We use four metrics to evaluate our proposed method: the Pearson correlation coefficient (ρ), root-mean-square error (RMSE), coefficient of determination (R^2), and F1 score (F1). The Pearson correlation coefficient (ρ) measures the strength of a linear association between the actual correctness level of the subjective listening test and the predicted SI. The RMSE indicates the prediction errors between the predicted SI and the actual SI. R^2 is a regression score function that measures the proportion of the variance of a dependent variable (prediction score) that is explained by the independent variable (actual score) in the model. It ranges from 1 (perfect score) to a negative score (the independent variable cannot explain the variance and contributes negatively to the prediction model). The F1 score is utilized to evaluate the prediction accuracy by grouping the obtained scores into three classes (low (scores less than or equal to 30%), medium (scores between 30% and 70%), and high (scores greater than or equal to 70%)).

TABLE I
EVALUATION RESULTS OF SEVERAL INTRUSIVE SI PREDICTION MODELS: THE BASELINE (MBSTOI + MBSG MODEL), THE HASPI (LEFT AND RIGHT), AND OUR PROPOSED METHOD. FOR REFERENCE, THE RESULTS OBTAINED BY OTHER METHODS ON THE CPC1 ARE AVAILABLE HERE[6].

Method	Metrics			
	$\rho \uparrow$	RMSE $\downarrow$	$R^2 \uparrow$	F1 (%) $\uparrow$
Track 1 (close-set)				
Baseline	0.62	28.5 ± 0.58	0.39	78.7
HASPI (left)	0.60	37.7 ± 0.60	-0.08	51.9
HASPI (right)	0.60	37.7 ± 0.60	-0.07	52.1
Proposed method	**0.74**	**24.6 ± 0.50**	**0.54**	**81.2**
Track 2 (open-set)				
Baseline	0.53	36.5 ± 1.35	-0.02	68.2
HASPI (left)	0.57	37.9 ± 1.20	-0.10	52.4
HASPI (right)	0.55	38.6 ± 1.23	-0.14	53.7
Proposed method	**0.71**	**26.2 ± 1.02**	**0.48**	**77.4**

C. Results

First, we carried out a comparative analysis between our proposed method and other existing methods. Table I shows the SI prediction results of the baseline method (Subsection III-B), the HASPI (Subsection III-A), and the proposed method for the closed-set and open-set tracks. Since the HASPI is a prediction method for monaural hearing, we analyzed the results obtained from each ear (HASPI (left) for the left ear and HASPI (right) for the right ear). The results indicate that the proposed method performed better than the baseline and HASPI on both tracks. Additionally, our proposed method could achieve high ranks in terms of the ρ and RMSE metrics compared to other methods in the CPC1[6], including [13], [12], [40], [41], [42]). Figures 5 and 6 show the average prediction results per listener and system, respectively. Note that the dotted lines only help show the correlations between the methods and do not represent relationships between listeners. The results in Fig. 5 indicate that the proposed method could better predict SI scores from various HI listeners. Moreover, the results in Fig. 6 show that the predictions obtained by the proposed method had an almost perfect association with the actual SI scores from the listening test ($\rho = 0.997$). This result is clarified by the system E005 and E018 where the error between actual and proposed method is lower than the error between actual and baseline.

TABLE II
COMPARISON BETWEEN THE INTRUSIVE AND NONINTRUSIVE VERSIONS OF THE PROPOSED METHODS. THE INTRUSIVE MODEL USED EARMODEL, THE EGEMAPS, AND WAVLM AS THE INPUT FEATURES. THE NONINTRUSIVE MODEL USED THE EGEMAPS AND WAVLM AS THE INPUT FEATURES.

Model	Metrics			
	$\rho \uparrow$	RMSE $\downarrow$	$R^2 \uparrow$	F1 (%) $\uparrow$
Track 1 (close-set)				
Intrusive	0.74	24.6 ± 0.50	0.54	81.2
Non-intrusive	0.74	25.0 ± 0.51	0.53	80.4
Track 2 (open-set)				
Intrusive	0.71	26.2 ± 1.02	0.48	77.4
Non-intrusive	0.63	28.3 ± 1.12	0.39	74.4

Second, we conducted an ablation test by excluding the additional features, including the eGeMAPS and WavLM. Ablation tests are often used to analyze the contribution of each feature in a tested model. Figure 4 shows the results of the ablation test. The overall results obtained for both tracks indicate that the proposed method (without the exclusion of the proposed features) could achieve the highest correlation and lowest RMSE. A more significant difference was observed in the open-set track. In addition, these results also indicated that the eGeMAPS feature contributed more to the prediction model than WavLM. Additionally, adding both the eGeMAPS and WavLM significantly improved the prediction model (ρ increased by more than 15%, and the RMSE decreased by more than 10.00 in both tracks).

Last but not least, we considered the possibility of providing a nonintrusive (blind) prediction method in this study. The nonintrusive method is important and more applicable in realistic situations since clean speech may not always be available. From Fig. 3, the features that required clean speech were only the features extracted using EarModel. The eGeMAPS and WavLM features only required improved SPIN as the prediction model input, so this version could be regarded as a nonintrusive prediction method. Table II shows the results produced by the nonintrusive version of the proposed method. The results obtained for the closed-set track indicate that the nonintrusive method could achieve almost similar results to those of the intrusive method (using all features described in Section IV-A). However, in the open-set track, a relatively more significant prediction performance

[6]https://claritychallenge.org/clarity2022-workshop/results.html

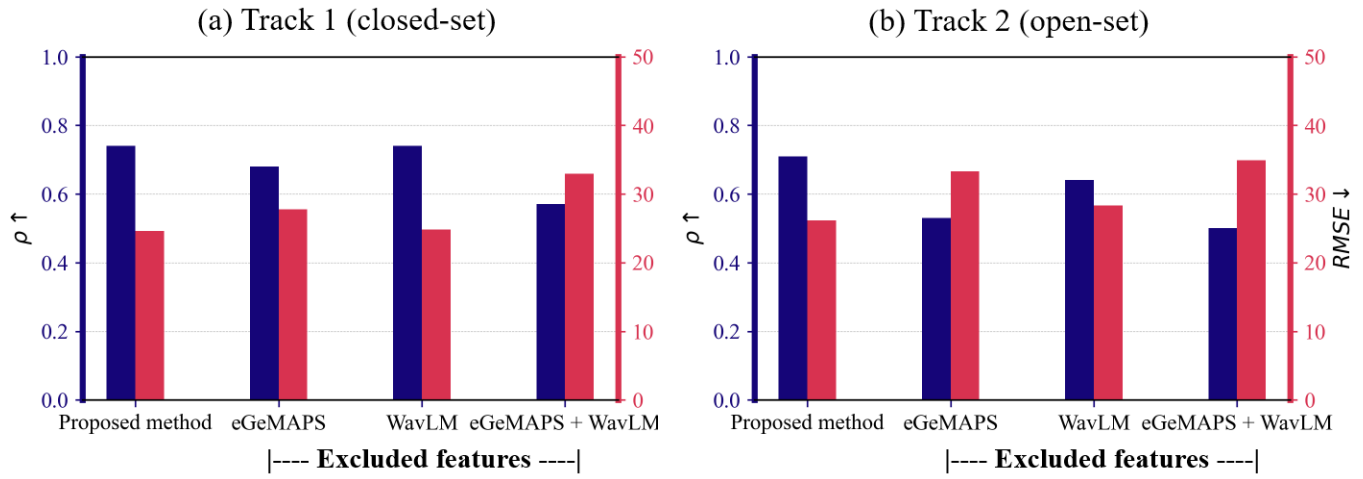

Fig. 4. Results of an ablation test conducted by excluding additional features (the eGeMAPS and WavLM).

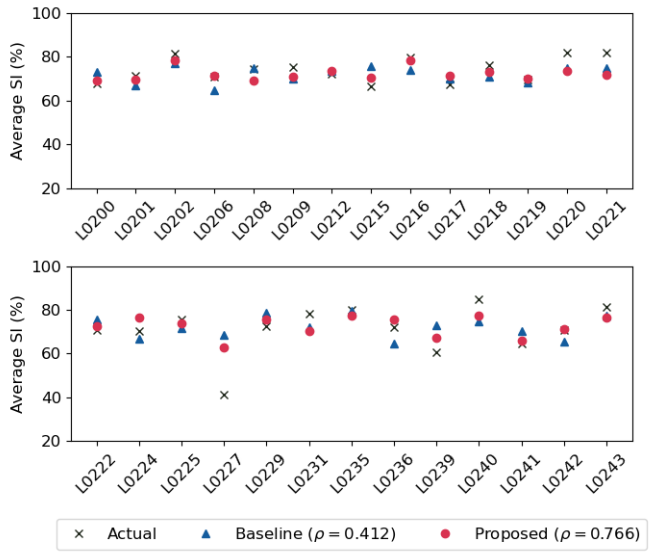

Fig. 5. Average SI prediction based on the listeners in the closed-set track.

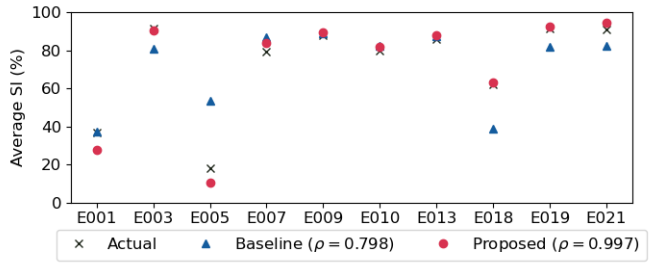

Fig. 6. Average SI prediction based on the system in the closed-set track.

reduction was obtained (ρ was 0.08 lower than the results of the intrusive method). Despite these results, the proposed nonintrusive method still outperformed the baseline method for both tracks.

VI. Conclusions

This study utilized EarModel as an auditory model for SI prediction. The inputs of EarModel were binaural clean speech, improved SPIN signals, and an audiogram for estimating the hearing loss condition. Additionally, the eGeMAPS and WavLM were included as features to estimate the acoustic parameters contributing to hearing-impaired speech perception. The evaluation utilized the CPC1 dataset and several evaluation metrics, such as the Pearson correlation coefficient, RMSE, coefficient of determination, and F1 score. A comparative analysis was performed between the proposed method, the baseline, and the HASPI. The overall results showed that the proposed method predicted SI better than the baseline and HASPI. The improvement was indicated by higher ρ, R^2, and F1 scores and lower RMSEs than those of the other methods in closed-set and open-set tracks. Furthermore, the average prediction results per listener and system appeared to be closely related to the actual SI scores.

An ablation test was also conducted to analyze the contribution of each additional feature, i.e., the eGeMAPS and WavLM. From the ablation test, it could be concluded that incorporating the eGeMAPS and WavLM could significantly improve the prediction results. Last, the possibility of modifying the proposed method into a nonintrusive version was considered, and this version could outperform the baseline method in terms of accuracy.

Acknowledgment

This work was supported by the SCOPE Program of Ministry of Internal Affairs and Communications (No. 201605002), a Grant-in-Aid for Scientific Research (B) (No. 21H03463),

and a JSPS KAKENHI grant (No. 22K21304). This work was also partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (No. 22H04860 and 22H00536) and JST AIP Trilateral AI Research, Japan (No. JPMJCR20G6).

REFERENCES

[1] A. D. Palmer, P. C. Carder, D. L. White, G. Saunders, H. Woo, D. J. Graville, and J. T. Newsom, "The impact of communication impairments on the social relationships of older adults: Pathways to psychological well-being," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 1, pp. 1–21, 2019.

[2] G. Keidser, M. Seeto, M. Rudner, S. Hygge, and J. Rönnberg, "On the relationship between functional hearing and depression," *International Journal of Audiology*, vol. 54, no. 10, pp. 653–664, 2015.

[3] Y.-k. Sung, L. Li, C. Blake, J. Betz, and F. R. Lin, "Association of hearing loss and loneliness in older adults," *Journal of aging and health*, vol. 28, no. 6, pp. 979–994, 2016.

[4] G. Livingston, A. Sommerlad, S. Costafreda, J. Huntley, D. Ames, C. Ballard, S. Banerjee, A. Burns, J. Cohen-Mansfield, N. Fox, L. Gitlin, R. Howard, H. Kales, E. Larson, K. Ritchie, K. Rockwood, E. Sampson, and N. Mukadam, "Dementia prevention, intervention, and care," *The Lancet*, vol. 390, 07 2017.

[5] C. J. Plack, V. Drga, and E. A. Lopez-Poveda, "Inferred basilar-membrane response functions for listeners with mild to moderate sensorineural hearing loss." *Journal of the Acoustical Society of America*, vol. 115 4, pp. 1684–95, 2004.

[6] B. C. J. Moore, D. A. Vickers, C. J. Plack, and A. J. Oxenham, "Inter-relationship between different psychoacoustic measures assumed to be related to the cochlear active mechanism." *Journal of the Acoustical Society of America*, vol. 106 5, pp. 2761–78, 1999.

[7] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[8] P. C. Loizou, *Speech Quality Assessment*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 623–654, accessed on 2022-07-25. [Online]. Available: https://doi.org/10.1007/978-3-642-19551-8_23

[9] J. Kates, K. Arehart, M. Anderson, R. Muralimanohar, and L. Harvey, "Using objective metrics to measure hearing aid performance," *Ear and Hearing*, vol. 39, p. 1, 03 2018.

[10] J. Kates and K. Arehart, "The Hearing-Aid Speech Perception Index (HASPI)," *Speech Communication*, vol. 65, 11 2014.

[11] Y. Nejime and B. Moore, "Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise," *Journal of the Acoustical Society of America*, vol. 102, pp. 603–15, 08 1997.

[12] R. E. Zezario, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "MBI-Net: A non-intrusive multi-branched speech intelligibility prediction model for hearing aids," *arXiv preprint arXiv:2204.03305*, 2022.

[13] Z. Tu, N. Ma, and J. Barker, "Unsupervised uncertainty measures of automatic speech recognition for non-intrusive speech intelligibility prediction," *arXiv preprint arXiv:2204.04288*, 2022.

[14] A. Alghamdi, W.-Y. Chan, and D. Fogerty, "Using acoustic parameters for intelligibility prediction of reverberant speech," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2534–2538.

[15] S. Kochkin, "Marketrak v: "why my hearing aids are in the drawer": The consumers' perspective," *The Hearing Journal*, vol. 53, pp. 34,36,39–41, 02 2000.

[16] L. V. Knudsen, M. Öberg, C. Nielsen, G. Naylor, and S. E. Kramer, "Factors influencing help seeking, hearing aid uptake, hearing aid use and satisfaction with hearing aids: A review of the literature," *Trends in Amplification*, vol. 14, pp. 127 – 154, 2010.

[17] P. Derleth, E. Georganti, M. Latzel, G. Courtois, M. Hofbauer, J. Raether, and V. Kuehnel, "Binaural signal processing in hearing aids," *Seminars in Hearing*, vol. 42, pp. 206–223, 08 2021.

[18] J. Kates and K. Arehart, "The Hearing-Aid Speech Perception (HASPI) Version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.

[19] J. Kates, "An auditory model for intelligibility and quality predictions," *The Journal of the Acoustical Society of America*, vol. 133, p. 3560, 05 2013.

[20] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. V. Muñoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Interspeech 2021*. ISCA, 2021, pp. 686–690.

[21] 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.

[22] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "A method for predicting the intelligibility of noisy and non-linearly enhanced binaural speech," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4995–4999.

[23] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns." *Journal of the Acoustical Society of America*, vol. 74 3, pp. 750–3, 1983.

[24] J. Kiessling, "Current approach to hearing aid evaluation," *Canadian Journal of Speech-Language Pathology and Audiology*, vol. 17, no. 4, pp. 39–49, 1993.

[25] C. Plack, A. Oxenham, A. Simonson, C. O'Hanlon, V. Drga, and D. Arifianto, "Estimates of compression at low and high frequencies using masking additivity in normal and impaired ears," *Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4321–4330, 2008.

[26] C. J. Plack and A. J. Oxenham, "Basilar-membrane nonlinearity estimated by pulsation threshold." *The Journal of the Acoustical Society of America*, vol. 107 1, pp. 501–7, 2000.

[27] Z. Tu, N. Ma, and J. Barker, "Dhasp: Differentiable hearing aid speech processing," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 296–300.

[28] D. M. Harris and P. Dallos, "Forward masking of auditory nerve fiber responses," *Journal of neurophysiology*, vol. 42, no. 4, pp. 1083–1107, 1979.

[29] A. Baevski and A. Mohamed, "Effectiveness of self-supervised pre-training for asr," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7694–7698.

[30] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[31] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *ArXiv*, vol. abs/2110.13900, 2022.

[32] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," 2021.

[33] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, pp. 1–1, 01 2015.

[34] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 835–838. [Online]. Available: https://doi.org/10.1145/2502081.2502224

[35] B. Larsby, M. Hällgren, L. Nilsson, and A. Mcallister, "The influence of female versus male speakers' voice on speech recognition thresholds in noise: Effects of low- and high-frequency hearing impairment," *Speech, Language and Hearing*, vol. 18, pp. 83–90, 06 2015.

[36] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *Arxiv*, 2016. [Online]. Available: https://arxiv.org/abs/1609.03499

[37] E. V. den Borre, S. Denys, A. van Wieringen, and J. Wouters, "The digit triplet test: a scoping review," *International Journal of Audiology*, vol. 60, no. 12, pp. 946–963, 2021.

[38] W. Whitmer, P. Howell, and M. Akeroyd, "Proposed norms for the Glasgow hearing-aid benefit profile (GHABP) questionnaire," *International journal of audiology*, vol. 53, 02 2014.

[39] K. Andersson, L. Andersen, J. Christensen, and T. Neher, "Assessing Real-Life Benefit From Hearing-Aid Noise Management: SSQ12 Questionnaire Versus Ecological Momentary Assessment With Acoustic Data-Logging," *American Journal of Audiology*, vol. 30, 12 2020.

[40] N. Kamo, K. Arai, A. Ogawa, S. Araki, T. Nakatani, K. Kinoshita, M. Delcroix, T. Ochiai, and T. Irino, "Conformer-based fusion of text, audio, and listener characteristics for predicting speech intelligibility of hearing aid users," 2022, accessed on 2022-07-27. [Online]. Available: https://claritychallenge.org/clarity2022-workshop/papers/CPC1_E036_report.pdf

[41] J. Roßbach, R. Huber, S. Röttges, C. F. Hauth, T. Biberger, T. Brand, B. T. Meyer, and J. Rennies, "Speech intelligibility prediction for hearing-impaired listeners with the leap model," 2022, accessed on 2022-07-27. [Online]. Available: https://claritychallenge.org/clarity2022-workshop/papers/CPC1_E022_report.pdf

[42] S. Röttges, J. Roßbach, C. F. Hauth, T. Biberger, B. T. Meyer, R. Huber, J. Rennies, and T. Brand, "Speech intelligibility prediction using the bbsim-sti model," 2022, 2022-07-27. [Online]. Available: https://claritychallenge.org/clarity2022-workshop/papers/CPC1_E019_report.pdf