Applied Acoustics 214 (2023) 109663

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Non-intrusive speech intelligibility prediction using an auditory periphery model with hearing loss

Candy Olivia Mawalim *, Benita Angela Titalim, Shogo Okada, Masashi Unoki

Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, 923-1292, Ishikawa, Japan

ARTICLE INFO

Keywords:
Auditory model
Speech intelligibility
Hearing loss
Non-intrusive

ABSTRACT

Speech intelligibility prediction methods are necessary for hearing aid development. However, many such prediction methods are categorized as intrusive metrics because they require reference speech as input, which is often unavailable in real-world situations. Additionally, the processing techniques in hearing aids may cause temporal or frequency shifts, which degrade the accuracy of intrusive speech intelligibility metrics. This paper proposes a non-intrusive auditory model for predicting speech intelligibility under hearing loss conditions. The proposed method requires binaural signals from hearing aids and audiograms representing the hearing conditions of hearing-impaired listeners. It also includes additional acoustic features to improve the method's robustness in noisy and reverberant environments. A two-dimensional convolutional neural network with neural decision forests is used to construct a speech intelligibility prediction model. An evaluation conducted with the first Clarity Prediction Challenge dataset shows that the proposed method performs better than the baseline system.

1. Introduction

A hearing aid (HA) is an assistive device to help hearing-impaired people. Hearing aid technology has improved rapidly, providing better hearing and high speech intelligibility in noisy, reverberant environments. In HA development, however, there are various issues related to performance analysis [1,2]. The evaluation process currently uses subjective testing, which is expensive and time-consuming. Moreover, the requirement of finding hearing-impaired listeners as test subjects may lead to other problems [2]. Consequently, there is no guarantee on how well an enhancement system will work in correcting hearing loss and improving speech intelligibility. In addition, the impact of technological sophistication on hearing aids is often underestimated. Alternatively, the development of objective tests to predict speech intelligibility has been proposed as a better way to evaluate the system performance. As the target users of hearing aids are hearing-impaired listeners, a prediction method should also facilitate signal processing based on auditory perception with hearing loss.

A well-known method to measure speech intelligibility is the hearing-aid speech perception index (HASPI) [3]. The HASPI was built with the aim of assessing the speech intelligibility perceived by both normal listeners and hearing-impaired listeners after hearing aid processing. This method mainly comprises an auditory model to extract speech features and a machine learning model to predict speech intelligibility. The auditory model, which was designed to simulate both normal and impaired hearing, provides benefits in terms of better understanding both kinds of perception in the human ear. Additionally, temporal alignment and a delay compensator are used to synchronize input signals and correct the delay introduced after hearing aid processing.

Although this auditory model design provides many benefits for speech intelligibility prediction with the HASPI, it cannot address binaural hearing, thereby limiting this method's effectiveness in noisy environments. Moreover, auditory models cannot simulate hearing loss severity outside the conditions represented by audiograms captured from individuals. HASPI evaluation has also been reported to strongly depend on the human subject data used in the training process.

Another method to measure speech intelligibility is the modified binaural short-time objective intelligibility (MBSTOI) method, which is the latest STOI method specializing in binaural processes [4]. Predicting speech intelligibility under hearing loss conditions requires performing preprocessing to generate time-domain signals after auditory processing. To predict speech intelligibility under hearing loss conditions in the first Clarity Prediction Challenge (CPC1) [5,6], the Cambridge hearing loss model was utilized as the initial stage to preprocess the input binaural signal. Then, after generating the speech signals affected by

* Corresponding author.
E-mail address: candylim@jaist.ac.jp (C.O. Mawalim).

https://doi.org/10.1016/j.apacoust.2023.109663
Received 12 April 2023; Received in revised form 14 September 2023; Accepted 20 September 2023
Available online 5 October 2023

hearing loss, speech intelligibility prediction based on a mathematical approach was performed.

Unfortunately, the baseline model in the CPC1 cannot compete well with the HASPI method in terms of speech intelligibility prediction [6]. Despite the model's capability to accept binaural input, there is no process to eliminate the delay introduced after hearing aid and hearing loss processing (a common problem with intrusive metrics). In addition, the MBSTOI's insensitivity to level changes limits its usefulness because such level changes are a critical factor in assessing speech intelligibility under various hearing loss conditions. Therefore, other measures or subjective ratings are needed to provide a more comprehensive speech intelligibility assessment in conjunction with the MBSTOI. Finally, both the HASPI and the baseline in the CPC1 are intrusive methods that require clean speech as a reference signal. However, clean speech is often unavailable in real-world applications.

Hence, this paper proposes a non-intrusive method that incorporates a hearing loss model to predict speech intelligibility. Specifically, we modified the HASPI auditory periphery model, allowing its use without reference speech. We also incorporated acoustic features that are beneficial for speech intelligibility prediction. We then built a speech intelligibility prediction model by using a parallel deep learning model to process binaural signals. We hypothesize that the proposed method can provide accurate speech intelligibility prediction by incorporating information from the hearing loss model, acoustic parameters, and binaural signals without the problem of processing the delay due to hearing aids.

The rest of this paper comprises several sections. We describe related work on speech intelligibility prediction in Section 2. Next, Section 3 covers the details of the proposed method. Then, Section 4 explains our experimental data, evaluation method, and the results of the experiments. Section 5 discusses our results, some key findings, and the limitations of the work. Finally, Section 6 summarizes our findings and concludes the paper.

2. Related work

Speech intelligibility can be described as the ability to understand target speech [7]. Because speech intelligibility measures how well an individual perceives and interprets speech, speech intelligibility assessment has become fundamental in communication systems. By measuring speech intelligibility, accurate and efficient communication can be established. An even larger impact is the improvement in individuals' quality of life.

The methods used to measure speech intelligibility in communication systems are divided into intrusive and non-intrusive methods, where the former require a clean speech signal and the latter do not. The choice between intrusive and non-intrusive methods depends on the specific goals and evaluation requirements [8].

2.1. Intrusive methods for speech intelligibility prediction

Intrusive methods are often used to evaluate human speech perception performance under controlled, ideal listening conditions, such as those in clinical settings [9]. Under ideal conditions, intrusive methods provide results that are directly related to an individual listener's subjective experience. Some intrusive metrics can also be used to determine the specific types of distortion that affect speech perception and evaluate the effectiveness of signal processing in reducing distortion.

We begin with the speech intelligibility index (SII) [10], an intrusive method for measuring speech intelligibility. The SII measures the proportion of speech information available in a given speech signal. It is based on the idea that certain frequency bands of a speech signal are more important for speech perception than others. To calculate the SII, the spectrum data of both the background noise and the speech signal are needed to calculate the signal-to-noise ratios (SNRs) in different frequency bands. These SNRs are multiplied with frequency-dependent weights based on the importance of each frequency band for speech intelligibility. The SII is the sum of the weighted SNRs across all frequency bands.

Similar to the SII, the speech transmission index (STI) [11] is used to evaluate the intelligibility of speech signals under noisy, reverberant environments. One advantage of the STI is that it provides more robust, accurate speech intelligibility prediction. The STI method is based on a standardized method for measuring the transmission of speech in a given environment. Specifically, a modulation transfer function (MTF) is used to evaluate the transmission quality of a speech signal through a communication system by incorporating the effects of background noise, reverberation, and other factors that can degrade speech intelligibility. The STI captures the modulation depth, spectral content, and temporal fluctuations of the input degraded signal in each frequency band. The standard IEC 60268-16:2020 [12] describes the setup for calculating the STI. The modulation transmission indices (MTIs) are calculated as the ratio of output modulation to input modulation for each frequency band. Then, it is reduced to a single-number STI by converting the MTIs into signal-to-noise ratios, summing them by the octave band, and applying weighting factors.

Another intrusive method is the normalized covariance metric (NCM) [13], which leverages the idea that a spectral structure contains much information on distortion and how similar a degraded signal is to the reference signal. The NCM intelligibility calculation starts with signal decomposition using a gammatone filterbank, which is followed by temporal amplitude envelope extraction via a Hilbert transform. Next, the power spectral density of the signal over different frequency bands and the cross-spectral density between the degraded and reference signals are calculated. As a result, the NCM evaluates the degree of correlation between degraded and reference speech signals that are normalized by their overall power levels.

While the NCM discards the temporal fine structure (TFS), the TFS spectrum (TFSS) index [14] is a speech intelligibility prediction method that incorporates the TFS, which has been found to be important for speech perception and intelligibility. In the TFSS measurement, the TFSs of both clean and degraded signals are derived from bandpassed signals through a decomposition process involving the Hilbert transform and fine structure analysis. The Hilbert transform generates an analytic signal with an envelope and a phase component, effectively encoding frequency information. Extracting the phase component yields the TFS, which represents fine-grained temporal variations in the signal. This TFS is used to calculate coherence indices for each band, along with articulation index weighting functions. The accumulated coherence indices produce the TFSS index, which performs effectively under nonstationary noise and reverberation conditions.

The STOI [15,16] is sometimes preferred in an evaluation over some other metrics for several reasons. One reason is that short-term processing allows frame-by-frame temporal and spectral characterization, which can be useful for analyzing distorted speech signals. Another reason is that STOI predictions have been shown to be highly correlated with speech intelligibility measured by subjective tests. There are many STOI extensions to facilitate a wider range of applications. For example, the discrete binaural STOI (DBSTOI) [17] and MBSTOI [4] enhance the initial STOI performance and accuracy and extend the model, enabling the binaural processing of stereo input signals.

Finally, to evaluate the speech perception performance of hearing aids, Kates et al. proposed the HASPI method [3]. The HASPI is a variant of the SII that incorporates the effects of hearing aid processing on speech perception. Specifically, the HASPI incorporates models for normal and impaired hearing, such as frequency shaping and compression models, to estimate the effects of auditory perception on a speech signal. The resulting estimate of a processed speech signal is then used to calculate a speech intelligibility score. This method has been widely used in hearing aid research and development because of its promising prediction accuracy. In particular, the incorporation of hearing loss

models enables the HASPI to more accurately reflect real-world hearing aid performance than some other metrics.

2.2. Non-intrusive methods for speech intelligibility prediction

In contrast, non-intrusive methods aim to estimate speech intelligibility without the involvement of reference signals in the prediction. Non-intrusive methods are often preferred in real-life situations and research settings due to their advantages over intrusive methods. While these methods may not always perfectly align with subjective experiences, they offer certain benefits, such as reduced disruption and increased feasibility. For instance, non-intrusive methods do not necessarily require the participation to utilize sensors or equipment in the communication process (for obtaining the reference signals). Thus, participants can have conversations without additional interruptions or distractions. Additionally, the environment settings for intrusive methods to get pairs of reference and noisy signals are often more complicated than the non-intrusive ones. The noisy signals for non-intrusive methods can be recorded without requiring specialized equipment, such as in everyday situations. However, according to [1], there have yet to be any non-intrusive methods that are sufficiently well established for measuring speech intelligibility.

The most well-known non-intrusive method for speech intelligibility prediction is the speech-to-reverberation modulation energy ratio (SRMR) [1]. The SRMR is calculated by dividing the energy of the modulation spectrum in the speech region by the energy of the modulation spectrum in the reverberation region. The SRMR quantifies the relationship between the modulation energy found in low-frequency channels centered between 4 and 18 Hz and the modulation energy detected in higher-frequency channels ranging from 29 to 128 Hz. The SRMR for cochlear implants (SRMR-CI) [18] and SRMR for hearing aids (SRMR-HA) [19] are variations that were specifically modified to assess the performance of cochlear implants and hearing aids, respectively.

While the SRMR focuses on measuring intelligibility by leveraging the modulation spectrum, the modulation spectrum area (ModA) [20] measures the impact of noise, particularly nonstationary noise, on speech intelligibility. The development of the ModA was based on the hypothesis that the auditory system analyzes speech signals in terms of amplitude modulation (AM) and frequency modulation (FM) patterns. Thus, the ModA calculates the modulation power spectrum of a degraded signal in each frequency band. The speech intelligibility index is then determined as the average area under the modulation spectrum curve value across all frequency bands.

In recent years, deep learning-based approaches for speech intelligibility prediction were also proposed to compensate for the limitations of signal processing approaches [21,22]. For instance, Zezario et al. [22] proposed the multiobjective speech assessment (MOSA-Net) model, which exploits acoustic information from multiple domains in model training. It was originally trained to predict speech assessment metrics, such as the STOI and perceptual evaluation of speech quality. The model was extended by integrating the Cambridge hearing loss model [23] to predict speech intelligibility for hearing aids [24].

3. Proposed method

Fig. 1 shows a block diagram of our proposed method. Because this method is a non-intrusive method, the inputs are improved speech-in-noise (SPIN) signals from an HA output and the listener's audiogram. We extract three groups of features. First, we extract the spectral envelope from an auditory model. Second, we extract the general acoustic parameters that are used as a baseline representation in many speech processing tasks, namely, the extended Geneva minimalistic acoustic parameter set (eGeMAPS) [25]. The eGeMAPS feature set is obtained by utilizing the openSMILE toolkit [26]. Finally, we also incorporate a pre-trained large-scale self-supervised learning model for speech processing, namely the wavLM model [27] that was reported to be able to learn universal representations for speech processing tasks. These three groups of features were also used in our prior work [28], and their use significantly improved speech intelligibility prediction for hearing aids. However, the auditory model proposed for intelligibility and quality predictions in [29] (further we called the EarModel) required reference signals and thus could not be considered a non-intrusive method. Moreover, each feature group was processed separately with a regression model to predict speech intelligibility scores, which were then used to predict a final speech intelligibility score with a stack regressor. In contrast, this study aims to simplify the speech intelligibility prediction model by concatenating the feature groups as inputs to a neural network model, as shown in Fig. 2.

3.1. Auditory model

The auditory model in the proposed method is a modified version of the HASPI model [29]. In the resampling process, the input degraded signal (improved SPIN signal) is adjusted to 24 kHz and passed through the middle ear. The middle ear is modeled by cascading a 2nd-order low-pass filter with a cutoff frequency of 350 Hz and a 1st-order high-pass filter with a cutoff frequency of 5000 Hz, as shown in Fig. 3.

The filtered signal is processed in two filterbanks for control and analysis. Both filterbanks decompose the input into 32-channel signals by using a gammatone filterbank. The details of the gammatone filter design will be explained later, along with the analysis filterbank. The center frequency is converted to an equivalent rectangular bandwidth (ERB) frequency, and for the control filterbank, a basal shift by 0.02 of the basilar membrane length is applied. The maximum hearing level at frequencies of 250, 500, 1000, 2000, 4000, and 6000 Hz is set to 100 dB and a maximum hearing loss reference is used for the control filterbank [29]. The input listener's audiogram is used to determine the estimated loss parameters. The output of the control filterbank is the signal envelope extracted at the maximum hearing level.

Under normal hearing conditions, an input signal with a sound pressure level (SPL) below 30 or above 100 dB undergoes linear amplification, while a signal with an SPL between 30 and 100 dB is subject to compression. The compression ratio (CR) initially varies with the frequency and increases from 1.25:1 at the lowest center frequencies to 3.5:1 at the highest, as follows:

$$\mathrm{CR} = 1.25 + 2.25 \times \frac{n-1}{N_{\mathrm{ch}}-1}, \tag{1}$$

where n is the channel number and N_{ch} is the total number of channels. The CR is adjusted to become closer to 1:1 as the outer hair cell (OHC) loss increases.

We also approximate the proportion of attenuation due to OHC and inner hair cell (IHC) damage. As suggested in [30], OHC damage causes a more significant loss of sensitivity at high frequencies, whereas IHC damage may be more important for loss of sensitivity at low frequencies. Although the percentages cannot be precisely defined, the proportion of OHC loss is higher than that of IHC loss, depending on the severity of an individual's hearing loss. This is because the OHCs are more responsible for amplifying sounds by changing shape in response to sound vibrations, causing the tectorial membrane to move and thereby stimulating the IHCs. Hence, we use the OHC attenuation attnOHC to calculate the filterbank bandwidth BW relative to normal hearing by the following equation:

$$\mathrm{BW} = \mathrm{BW} + 2 \times \mathrm{attnOHC} + \left(\frac{\mathrm{attnOHC}}{50}\right)^6, \tag{2}$$

In addition, the signal envelopes generated by the control filterbank are converted to a sound pressure level in decibels and used to compute the bandwidth increment in response to a high-level signal. If the signal is below 50 dB, there is no bandwidth adjustment (i.e., below the minimum bandwidth or the bandwidth calculated for the input audiogram). If the signal is above 100 dB, the bandwidth at the maximum OHC loss is used. For signals between 50 and 100 dB, the bandwidth is

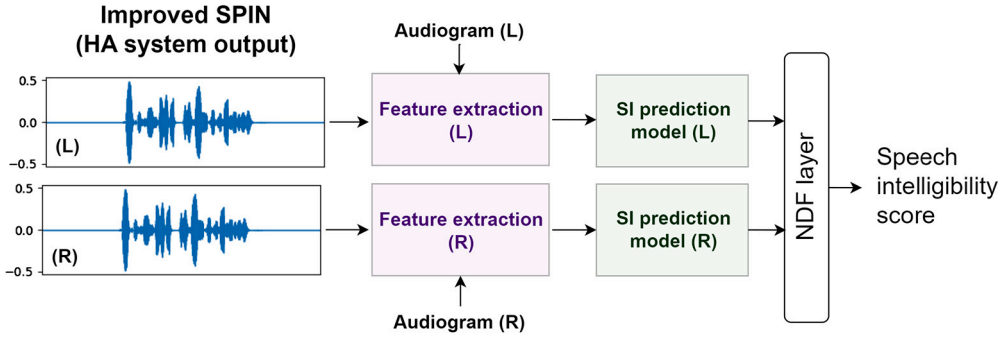


Fig. 1. Block diagram of the proposed method. Here, L and R denote the left and right ears, respectively, and NDF is a neural decision forest layer for predicting the speech intelligibility score.

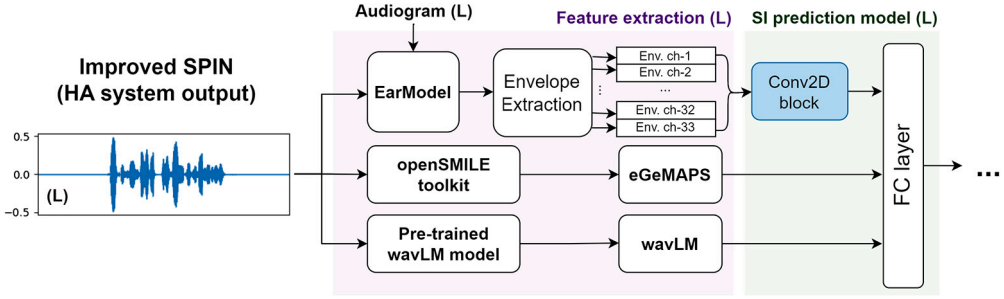


Fig. 2. Speech intelligibility prediction model for a one-channel input, specifically, the left ear. This figure is a more detailed version of the parts in Fig. 1. "Env. Ch-x" denotes the spectral envelope of channel x. The Conv2D block is a two-dimensional convolutional neural network (CNN) block, and the FC layer is a final, fully connected neural network layer.

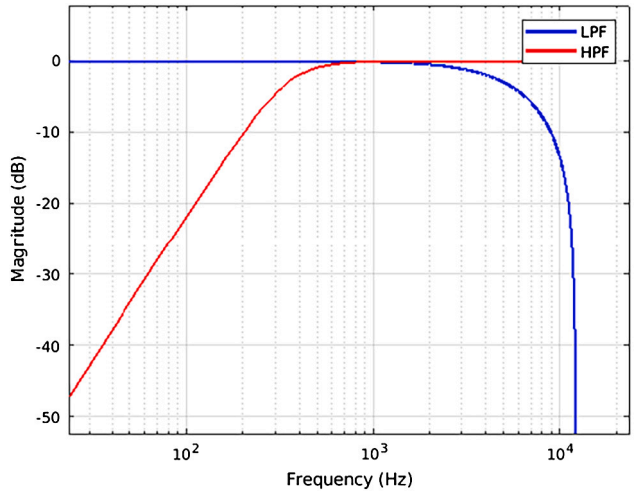


Fig. 3. Middle ear filter. LPF stands for low-pass filter, and HPF stands for high-pass filter.

linearly interpolated between the minimum and maximum bandwidths according to the controlled signal level.

The analysis filterbank implements the calculated bandwidth to simulate hearing loss due to OHC damage. This filterbank comprises a 4th-order gammatone filter ($N = 4$) based on the implementation in [31]. The impulse response of the gammatone filter is given by

$$h(t) = At^{N-1}e^{-2\pi bt}\cos(2\pi f t), \tag{3}$$

where A is the amplitude, b is the bandwidth of the filter, and f is the filter's center frequency. In this implementation, an impulse-invariant transformation of the gammatone filter is applied. The input signal is first demodulated down to the baseband by using a complex exponential and then passed through a series of four one-pole low-pass filters. The output of the analysis filterbank is the signal's temporal amplitude envelope (TAE) and temporal fine structure (TFS), which will be subject to compression. The TAE represents the magnitude of the signal and is calculated by taking the square root of the sum of the squares of the real and imaginary parts of the signal, with scaling by a gain factor.

Next, we explain the cochlear compression that is implemented in each auditory filter band. The auditory compressor works as a nonlinear system that applies a gain to the signal envelope generated by the control filterbank to compress the dynamic range of the signal. Cochlear compression includes the conversion of a control envelope to a dB sound pressure level and the computation of compression gain in dB. A control envelope is converted into a dB sound pressure level, the signal levels above the upper threshold and at the lower threshold are clipped, and then the compression gain in dB is computed. The upper threshold for clipping is set at 100 dB, and the lower threshold is determined by the low knee value, which represents the adjustment due to OHC attenuation. The gain is then converted to a linear form and filtered by a low-pass filter. Finally, it is applied to the input signals to obtain the compressed TAE and TFS.

To incorporate the effect of IHC damage, the compressed TAE is converted back to decibels, and the previously obtained IHC attenuation is subtracted from the results. The resulting gain is also applied to the compressed TFS. In addition, rapid and short-term IHC adaptation is provided because the sensitivity of IHCs to sound input changes over time in response to stimuli. These adaptations are based on an equivalent resistor–capacitor (RC) circuit model with two time constants that represent rapid (2 ms) and short-term (60 ms) adaptations and are mapped to 1st-order backward differences. This process uses the TAE in decibels and generates an adapted envelope and a function for gain versus time that mimics the adaptation of basilar membrane motion. Additional Gaussian noise is also preserved in the TFS to ensure the correct application of IHC adaptation and simulate the auditory threshold.

Finally, delay compensation is implemented in the gammatone filterbank for each filter at its center frequency. The group delay com-

pensation first computes the group delay of each gammatone filter with the ERB equation in [32]. Then, the firing rate output of the IHC model is adjusted so that all outputs have the same group delay. The resulting TAE in each frequency band is attenuated with linear gain below the lower threshold and above the upper threshold; between the thresholds, it is compressive, with a compression ratio of CR:1.

3.2. Speech intelligibility prediction model

The proposed speech intelligibility prediction model comprises two main parts. The first part contains a two-dimensional CNN (Conv2D) block and a fully connected layer. The second part contains a neural decision forest (NDF) layer.

The Conv2D block receives inputs comprising 2D spectral envelopes with time on one axis and the channel on the other, as extracted from the EarModel. The Conv2D block can be used to compress higher-dimensional features and automatically learn and extract discriminative features for speech intelligibility prediction. After passing through the Conv2D block, the discriminative features are concatenated to form eGeMAPS and wavLM features. eGeMAPS [25] includes 88 distinct low-level descriptors that are related to spectral, prosodic, voice quality, and temporal features. The eGeMAPS feature set has been shown to be effective in capturing emotions and affective states. Thus, it is commonly used as a baseline acoustic feature set in many speech processing studies.

Recent technology in automatic speech recognition (ASR) has shown remarkable improvements through self-supervised learning. WavLM is a technology that was reported to be effective in solving downstream speech tasks [27]. Because it combines a speech prediction learning process and denoising, wavLM is not only effective for ASR tasks but can also potentially improve the performance of non-ASR tasks. We incorporate the wavLM feature as a representation of speech content because the difficulty of speech content affects speech intelligibility prediction. Finally, all the features are concatenated and passed to the fully connected layer.

As mentioned above, our speech intelligibility prediction model uses an NDF layer as the last layer. The NDF [33] was chosen because it can provide a better interpretable model than traditional neural networks. It can also capture both linear and nonlinear dependencies in data. Thus, it has been reported to improve the prediction accuracy with a shorter training time.

4. Experiment

This section describes our experiment, including the dataset, the experimental and evaluation settings, and the results.

4.1. Dataset

We used the first Clarity Prediction Challenge (CPC1) dataset,[1] which was recorded under the following scenario. Each scene was simulated as a small, box-shaped room with moderate reverberation. A hearing-impaired person listened to a unique sentence with 7-10 words spoken by the target speaker. Interference noise was present in the form of another speaker or a continuous noise source. The positions of each sound source and the listener, the room dimensions, and the wall materials were generated by a scene generator, as described in the challenge.

The CPC1 dataset contains two subsets for training and evaluation. The recording process incorporated six British English speakers, 10 hearing aid processors, and 27 hearing-impaired listeners. Each scene comprises a distinct speech sample from the target speaker and interference noise. Additionally, the listener characteristics are also available, e.g., pure-tone air conduction audiograms. The challenge included two tracks: (1) a *closed-set* track (all listeners and HA processors were seen, but the recording scenes were unseen) and (2) an *open-set* track (all scenes, listeners, and HA processors were unseen).

4.2. Experimental setting

Since our proposed method is a non-intrusive method, the inputs are only the improved SPIN signals without the reference signals and the listener's audiogram. The input signals are stereo signals. We processed the input signals with dual-stream neural networks and finally combined them with a fully connected layer with a rectified linear unit (ReLU) activation function to predict the speech intelligibility score with a range from 0 to 100 (as shown in Fig. 1).

As mentioned in Section 3, we extracted three feature groups: (1) the spectral envelope, (2) eGeMAPS, and (3) wavLM. To extract the spectral envelope, we used the EarModel described in Section 3.1. In this experiment, we set the number of channels (N_{ch}) to 32. Subsequently, the frame length was set to 20 ms with a 50% overlap between consecutive frames. The output spectral envelope extracted from each improved SPIN signal was an $M \times (N_{\mathrm{ch}}+1)$ matrix, with M as the number of overlapping frames. For training, we unified the rows of the input matrices for the extracted spectral envelopes, with M' set to 2400.

Next, to extract the eGeMAPS and wavLM features, we downsampled the input signals to 16 kHz. This was because the pretrained model was trained with 16-kHz sampled speech audio. The eGeMAPS was extracted with the openSMILE toolkit in Python.[2] Specifically, we used the eGeMAPSv02 functionals extractor, which outputs an 88-dimensional feature vector for each input signal. Finally, we extracted the wavLM features by using a pretrained wavLM large model from Microsoft[3] [27]. Specifically, we used the output of the model's temporally averaged embedding, which resulted in a 1024-dimensional feature vector for each input signal.

The training process was conducted in a supervised manner with the correctness of the subjective listening test as the target label. First, we conducted a 5-fold cross-validation with the training dataset to determine the details of the neural network architecture for the speech intelligibility prediction model. We set the number of epochs and batch size to 100 and 16, respectively. Additionally, we used the adaptive moment estimation (Adam) optimizer with its default parameters and the mean squared error as the loss function in the Keras framework. Early stopping regularization was applied to avoid overfitting. Second, we evaluated the best model in the training phase by using the CPC1 evaluation data as test data. As mentioned above, the dataset includes two tracks. The *closed-set* track was used when we assumed that the listener and HA system were known. This track contains 4,863 scenes of training data and 2,421 scenes of test data. When both the listener and HA system were unknown, we used the *open-set* track, which contains 3,580 scenes of training data and 632 scenes of test data.

4.3. Evaluation

We compared our proposed method with the existing speech intelligibility prediction methods introduced in CPC1, including the baseline MBSTOI with a hearing loss model, the HASPI, and the first winner of CPC1, namely, the multi-branched speech intelligibility prediction model (MBI-Net) [24]. Recall that the HASPI is a well-known metric that is used for hearing aid development and is trained with the IEEE sentence dataset. To map the HASPI output to the listening test results in the CPC1 dataset, we used a logistic mapping model with a sigmoid function. Mathematically, the mapping function $f(x)$ can be expressed as follows:

[1] https://claritychallenge.org/clarity_CPC1_doc/docs/cpc1_data.

[2] https://audeering.github.io/opensmile-python/.

[3] https://huggingface.co/microsoft/wavlm-large.

Table 1
Overall results from the speech intelligibility prediction models: the baseline model in CPC1 [6], the HASPI [3], MBI-Net [24], and our proposed method. We also showed the results of the ablation test by excluding (excl.) each feature used in the proposed method. 'n/a' indicates that the score is not available.

Method	Binaural	Non-intrusive	Track 1 (closed-set)			Track 2 (open-set)		
			$\rho\uparrow$	RMSE (%) $\downarrow$	$R^2\uparrow$	$\rho\uparrow$	RMSE (%) $\downarrow$	$R^2\uparrow$
Baseline	Yes	No	0.62	28.52 ± 0.58	0.39	0.53	36.52 ± 1.35	-0.02
HASPI (left)	No	No	0.62	28.67 ± 0.58	0.38	0.54	30.72 ± 1.22	0.28
HASPI (right)	No	No	0.62	28.68 ± 0.58	0.39	0.51	31.48 ± 1.25	0.24
MBI-Net [24]	Yes	Yes	0.74	24.70 ± 0.50	n/a	0.59	30.72 ± 1.22	n/a
Proposed method	**Yes**	**Yes**	**0.75**	**24.34 ± 0.49**	**0.55**	**0.60**	**28.89 ± 1.15**	**0.36**
Proposed method (excl. eGeMAPS)	Yes	Yes	0.72	25.68 ± 0.53	0.51	0.41	32.32 ± 1.28	0.26
Proposed method (excl. wavLM)	Yes	Yes	0.74	24.60 ± 0.50	0.54	0.55	30.29 ± 1.25	0.28
Proposed method (excl. EarModel)	Yes	Yes	0.71	26.11 ± 0.56	0.45	0.40	33.64 ± 1.29	0.21

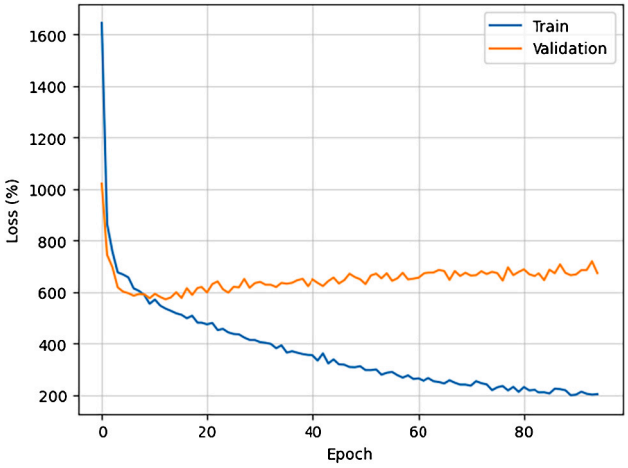


Fig. 4. Loss function during the training phase of the proposed method for the *closed-set track*. The blue and orange lines indicate the losses for the training and validation sets, respectively.

$$f(x) = \frac{1}{(1 + e^{-k(x-x_0)})}, \tag{4}$$

where k is the growth rate of the curve, x is the HASPI output value, and x_0 is the x value at the logistic function's midpoint.

To compare the methods, we used three common evaluation metrics for regression tasks: (1) the Pearson correlation coefficient (ρ), (2) the root-mean-square error (RMSE), and (3) the coefficient of determination (R^2). First, ρ measures the linear correlation between the actual listening test results and the predicted speech intelligibility. Second, the RMSE measures the difference or error between the predicted and actual speech intelligibility. The standard deviation error is also calculated to estimate the variability across multiple samples of the prediction. Third, R^2 measures how much of the data's variation can be explained by the model. We also conducted an ablation study to analyze the performance of each part of our proposed method.

4.4. Results

In this section, we describe the experimental results to evaluate the overall performance of our proposed method and the performance of every component included in its architecture.

4.4.1. Overall performance

Fig. 4 shows the loss function plotted against the number of epochs during the training phase of the proposed method. It is observed that the loss function values for both the training and validation sets decreased sharply during the first few epochs, indicating that the model quickly learned to fit the training data. The loss function continued to decrease for the training set, whereas after approximately 10 epochs, the loss function for the validation set began to plateau, indicating that the model was no longer improving with the validation data. Hence, we set the number of epochs to 20 for building our final speech intelligibility prediction model.

Table 2
Ablation test results for evaluating the auditory model.

Method	Track 1		
	$\rho\uparrow$	RMSE (%) $\downarrow$	$R^2\uparrow$
Proposed Method	**0.75**	**24.34 ± 0.49**	**0.55**
NH model (a)	0.71	26.18 ± 0.53	0.48
EarModel only (b)	0.68	28.77 ± 0.56	0.37
Better ear (c)	0.74	24.69 ± 0.50	0.54

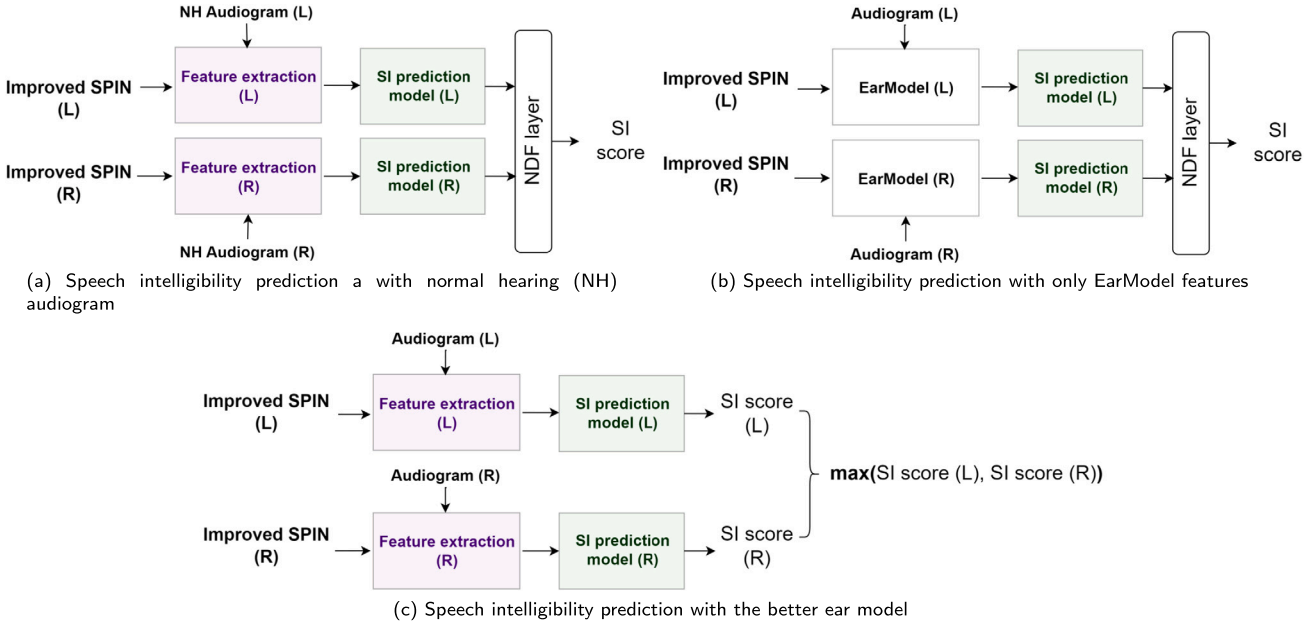
Next, Table 1 shows a summary of the comparative evaluation results from our proposed method, the baseline method in CPC1 (Cambridge hearing loss model + MBSTOI), HASPI, and MBI-Net. Because the HASPI algorithm receives a monaural signal input, we separately analyzed the results obtained from each ear, denoted as "HASPI (left)" and "HASPI (right)." Both the baseline method and the HASPI are considered intrusive metrics because they require clean speech for alignment in the speech intelligibility prediction model. MBI-Net [24] is also a binaural and non-intrusive method because it utilizes the output signals of the hearing loss model for both the left and right channels. In contrast to our prior works [28,34], here, we developed a non-intrusive metric that does not require clean speech to predict speech intelligibility. In most cases, intrusive metrics provide more precise speech intelligibility prediction. Nevertheless, processing in HA systems may cause temporal or frequency shifts, making it difficult for existing intrusive speech intelligibility metrics to obtain accurate results.

Overall, the experimental results showed that the proposed method performed better than the other compared methods in terms of the ρ, RMSE, and R^2 metrics. For instance, our method obtained an R^2 value that was approximately 0.16 higher than the R^2 values of the baseline and HASPI, which means that 16% more of the variation in the dependent variable was explained by the independent variables. In summary, this result indicates that the proposed method yielded a better fit than that of the comparison methods. The bottom part of Table 1 shows the performance of the proposed method when one set of features was excluded. In all scenarios, excluding one set of features causes a reduction in ρ and R^2. Moreover, it increases the RMSE, which indicates that each feature contributed positively to the proposed method. Hence, to further support our conclusions, we performed an ablation study to analyze the performance of each component in our proposed method.

4.4.2. Auditory model performance

As shown in Fig. 1, our proposed method is constructed using an auditory model that incorporates hearing loss phenomena. To evaluate the significance of the proposed EarModel, we performed an ablation study with the three variations shown in Figs. 5(a–c).

First, we investigated the hearing loss model in the auditory periphery model by inputting a normal hearing audiogram into the EarModel



(a) Speech intelligibility prediction a with normal hearing (NH) audiogram

(b) Speech intelligibility prediction with only EarModel features

(c) Speech intelligibility prediction with the better ear model

Fig. 5. Ablation test for evaluating the auditory model performance.

(referred to later as the "NH model"). The resulting model is illustrated in Fig. 5a. We defined the characteristic of normal hearing by specifying 0 dB as the hearing loss level at all frequencies. The results are listed in Table 2, and they indicate that by ignoring the hearing loss level, speech intelligibility prediction by the NH model became less accurate (increased RMSE and decreased ρ).

Next, we investigated the performance by using only the non-intrusive binaural EarModel without the additional acoustic features, as illustrated in Fig. 5b. A comparison of these results with those of the existing methods, as listed in Table 1, indicates that our EarModel could provide comparable results, with a slightly higher ρ but a slightly lower RMSE.

Finally, Fig. 5c shows the speech intelligibility prediction model obtained by extracting the maximum value of the speech intelligibility score of left and right signals. We further named this model as the 'better ear' model. The results in Table 1 indicate that the better ear model is slightly less effective than the proposed method, which infers the speech intelligibility score with the NDF layer and could provide better speech intelligibility prediction.

4.4.3. Significance of additional acoustic features

In the last part of this study, we investigated the impact of incorporating additional acoustic features, i.e., the eGeMAPS and wavLM features. Our prior study [28] showed that the inclusion of both eGeMAPS and wavLM features significantly improved the prediction model. Specifically, we observed an approximate increase of more than 15% in ρ and an approximate decrease of over 10.00% in the RMSE for both the closed- and open-set tracks. The experimental results here agreed well with the previously reported results. We verified the significance of the additional acoustic features by performing an analysis of variance (ANOVA) test between the proposed method's results and the "EarModel only" results. The ANOVA results showed a statistically significant difference in the mean scores between the two groups ($F(2, 2421) = 107.83, p < 0.05$). This F value indicates that the variation among the sample means was higher than the variation within the samples.

In this study, we enhanced the analysis by investigating which specific acoustic features contributed more to the prediction. Specifically, we used NDFs to identify the most important or informative acoustic features for predicting speech intelligibility. Fig. 6 shows the importance of each eGeMAPS feature in the prediction model. The NDF determines the most important features or variables that contribute to a model's accuracy. A feature's importance is measured by how much it decreases the overall entropy or impurity of the decision trees when it is used to split nodes.

As seen in Fig. 6, the ten most important features were "mfcc3V_sma3nz_stddevNorm," "mfcc2_sma3_stddev-Norm," "hammarbergIndexV_sma3nz_amean," "alpha-RatioUV_sma3nz_amean," "loudness-PeaksPerSec," "loudness_sma3_amean," "alphaRatioV_sma3nz_amean," "mfcc3_sma3_stddevNorm," "MeanUnvoicedSegment-Length," and "StddevUnvoicedSegmentLength." In summary, the most important acoustic features in the eGeMAPS were those based on mel-frequency cepstral coefficients (MFCCs), the Hammarberg index [35], the alpha ratio, and the loudness. MFCCs are commonly used features in speech processing that capture the spectral characteristics of speech. Meanwhile, the Hammarberg index and alpha ratio are related to spectral balance parameters. Specifically, the Hammarberg index measures the signal intensity difference between the maximum intensities that are present in a lower frequency range of $0-2000$ Hz and in a higher frequency range of $2000-5000$ Hz. The alpha ratio measures the ratio of the total energy present in a lower frequency range of $0-1000$ Hz to that in a higher frequency range of $1000-5000$ Hz. Finally, loudness is a factor that is highly related to speech intelligibility and estimates the perceived signal intensity from an auditory spectrum.

4.4.4. Evaluation grouped by listener, HA system, and interferer

Fig. 7 shows the average results for speech intelligibility prediction based on a hearing-impaired listener. Generally, the proposed method showed significant improvements in speech intelligibility prediction for all listeners. The ρ value of the proposed method was approximately 0.4 higher than that of the baseline system. This indicates that the proposed method can be used as a more accurate, reliable prediction model for hearing aid development. Despite the improvements, our method still had limitations in terms of predicting the speech intelligibility perceived by a particular listener, such as L0227. Our preliminary analysis of the data suggested that the other listeners exhibited normal hearing to mild hearing loss when hearing the low-frequency pure tones, as shown in their audiograms. In contrast, listener L0227 exhibited moderate to severe hearing loss for pure tones from lower to higher frequencies, which might have caused difficulties in perceiving any target speech.

Next, Fig. 8 shows the distribution of the speech intelligibility prediction result outcomes grouped by HA system. Ten HA systems were in-

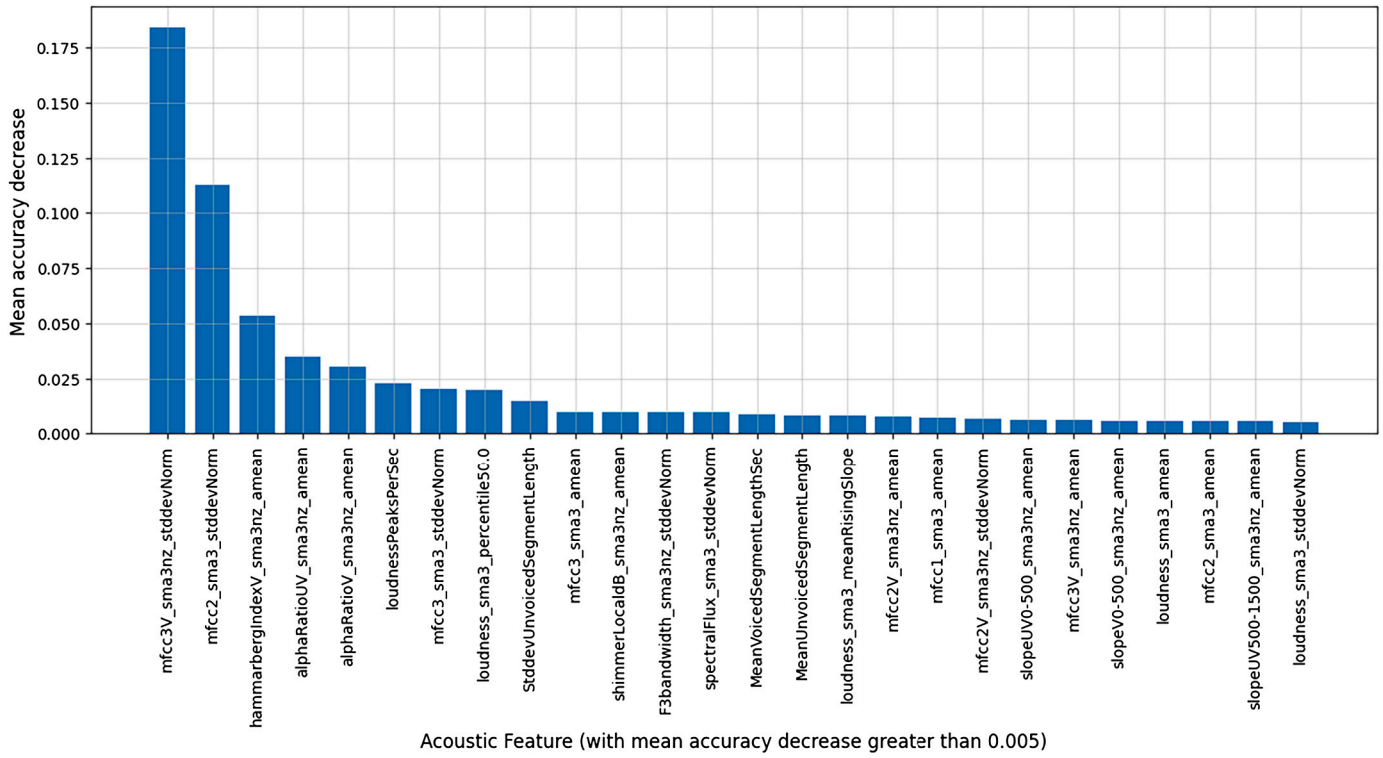


Fig. 6. Importance of each eGeMAPS feature. The y-axis indicates the mean accuracy decrease when the corresponding feature was excluded. The more important the feature is, the more the accuracy decreases.

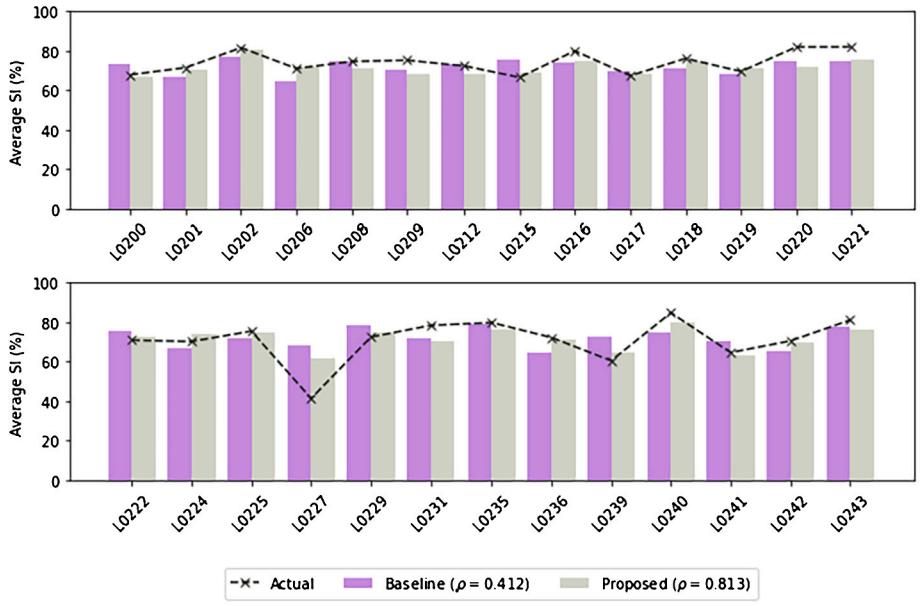


Fig. 7. Average speech intelligibility predictions as grouped by listener (*closed-set* track).

cluded in the data. These systems were labeled E001, E003, E005, E007, E009, E010, E13, E018, E019, and E021. When comparing the distributions of the actual results and the model predictions, the proposed method was generally better than the baseline method. For instance, the distribution centers, spreads, and overall ranges of the proposed method were more similar to the actual SI than those obtained using the baseline method. Although some of the RMSEs of the predictions by the proposed method were higher than those of the baseline method, the results might be due to a narrow range around the most common labels of the predicted values.

We also analyzed how the interferer type in a scene affected speech intelligibility perception. The CPC1 dataset has seven types of known interferers, namely, a vacuum cleaner, microwave, kettle, fan, dishwasher, hairdryer, and washing machine. We excluded approximately 300 test samples with unknown interferers. Fig. 9 shows the distribution of the speech intelligibility prediction results in terms of the interferer type. The results showed that our proposed method could provide more highly correlated prediction than the baseline methods for all types of known interferers. In particular, scenes with a vacuum cleaner or a hairdryer as the noise interferer caused more difficulties in predict-

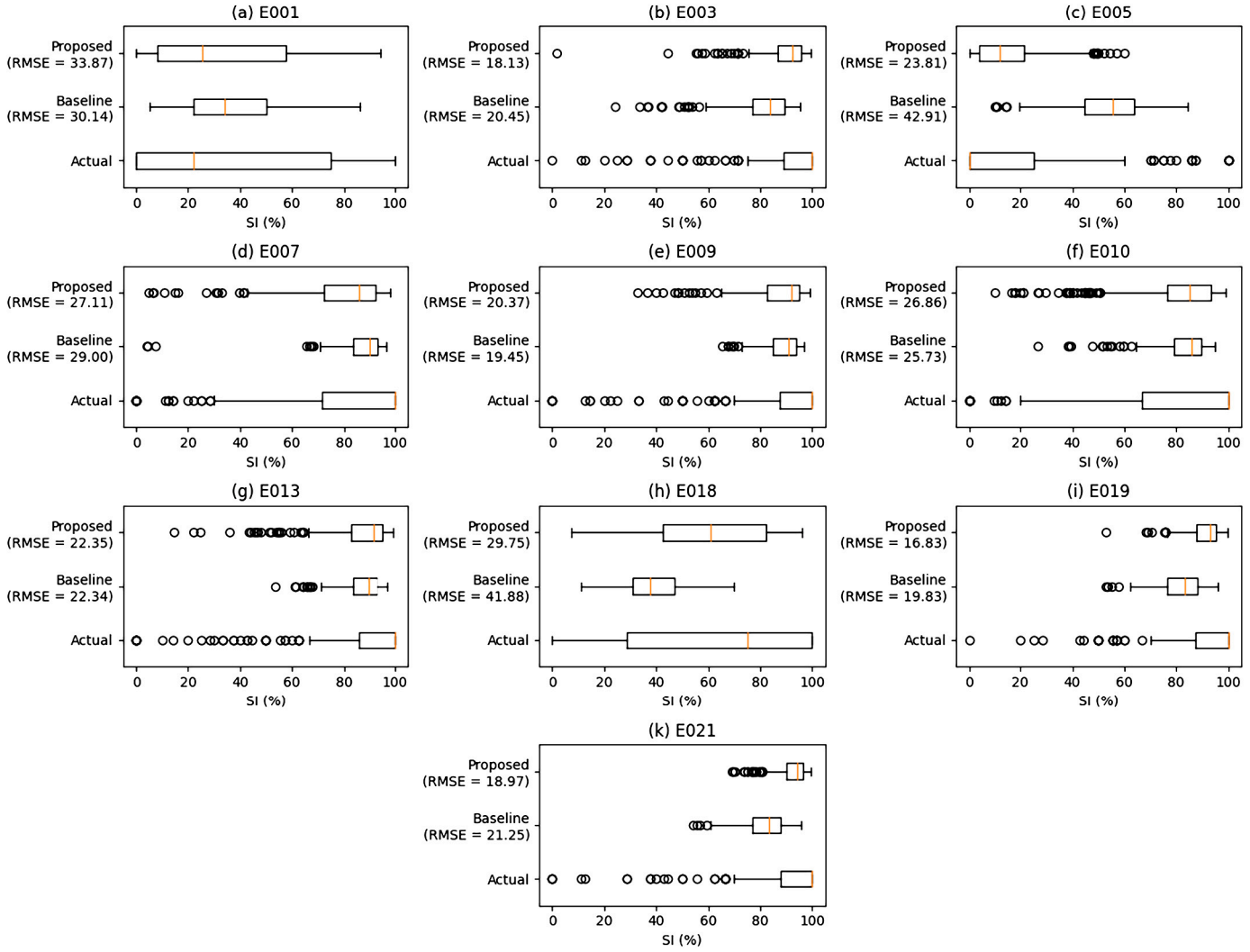


Fig. 8. Speech intelligibility prediction results as grouped by HA system (*closed-set* track).

ing speech intelligibility in hearing-impaired listeners. Meanwhile, the scenes with a fan caused the fewest difficulties (the lowest RMSEs).

5. Discussion

We discuss the key findings obtained in this study. We also discuss the progress of speech intelligibility prediction models for hearing aids and which issues we addressed. Finally, the limitations and future directions are discussed.

From the experimental results, as shown in Subsection 4.4, we can discuss three main points.

- Design of auditory model for non-intrusive speech intelligibility prediction
  This study demonstrates the use of an auditory periphery model for predicting speech intelligibility in hearing aids without a reference signal. Audiograms, which represent listeners' hearing loss, were considered to model the OHC and IHC damage in the ears. The evaluation results of our proposed method that utilized the features extracted from the auditory model ('EarModel only' (b) in Table 2) showed a comparative performance with that of an earlier model that required a reference signal (HASPI in Table 1). Although we received binaural signals, the model does not consider binaural cues in the human auditory system. A further investigation of these cues will be addressed in future work.
- Significance of components and features in the proposed method
  To justify the significance of each component and each feature set in the proposed method, we also carried out ablation tests and the one-way ANOVA test for statistical hypothesis testing. Each component and feature set positively improved the performance of the proposed method. The improvement in correlation between the predicted scores and actual speech intelligibility scores from the listening test validates the model's ability to capture the essential aspects of speech intelligibility.
  Subsequently, we hypothesize that the combination of the spectral features from the hearing loss model and the additional acoustic features contributed to this improvement. Acoustic features related to speech recognition, such as wavLM features, have also been reported to be beneficial in speech intelligibility prediction by other proposed methods in CPC1 [24,36].
- Robustness of the prediction under various listeners' characteristics, hearing aid systems, and interferers
  This study shows that the proposed method that incorporates the features from the auditory model and acoustic parameters into the speech intelligibility prediction model has generally more robust performance than that of the baseline MBSTOI method in various settings, including listeners' characteristics, hearing aid systems, and interferers (as shown in the experimental results in Subsubsection 4.4.4).

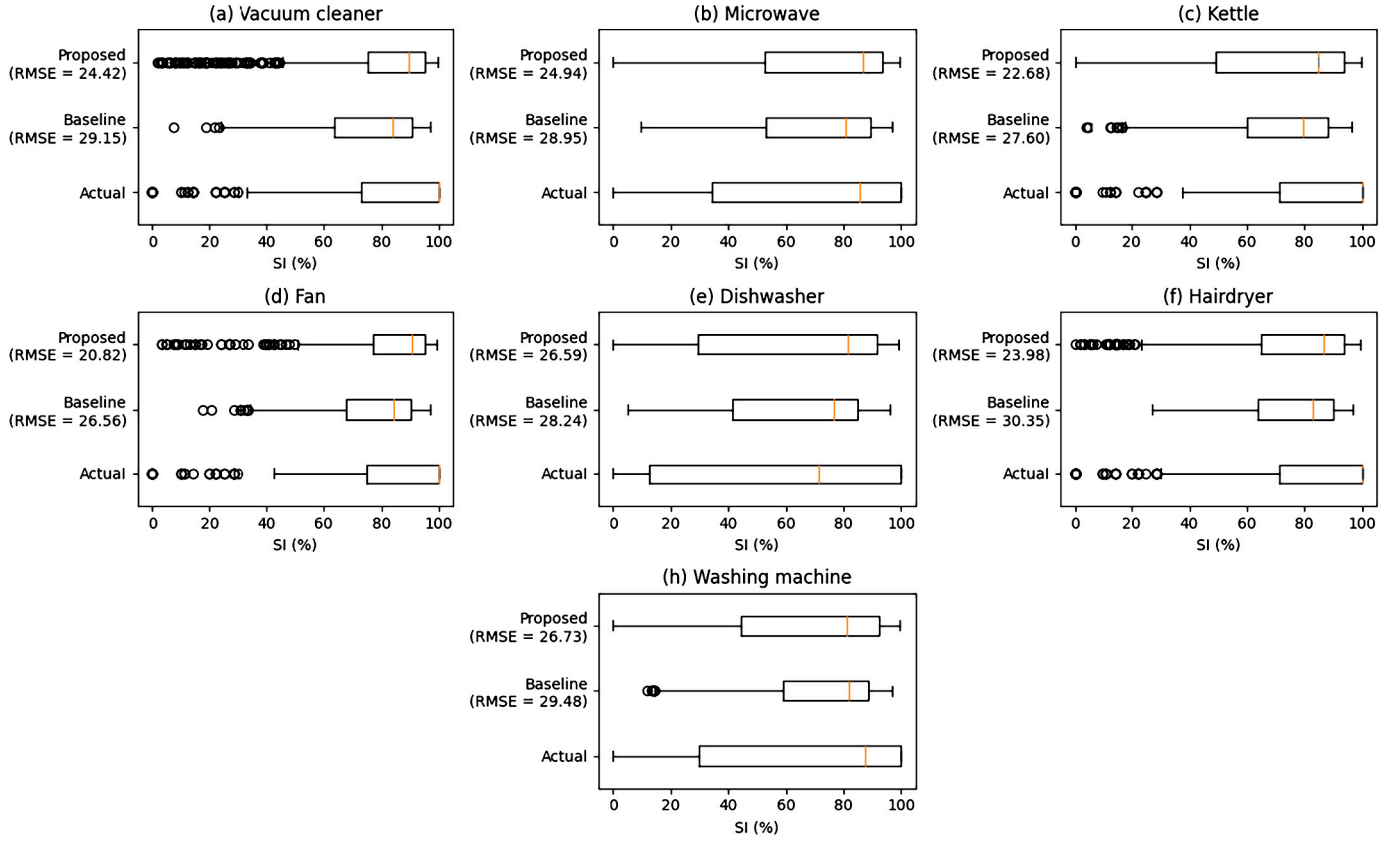


Fig. 9. Speech intelligibility prediction results as grouped by interferer type (*closed-set* track).

The progress of speech intelligibility prediction models for hearing aids has significantly advanced [1,4,17,37]. Earlier studies focused on the use of specific signal-related features, such as spectral features and modulation features [1,9,18,37]. Some recent studies, however, reported that machine learning models could outperform traditional approaches [21,22,24]. Unfortunately, the proposed machine learning models were generally sophisticated and difficult to explain (black box models). Moreover, these models depend heavily on the training data. Our proposed method attempted to balance the benefits of machine learning models while minimizing the disadvantages by integrating an auditory model and the common features used in machine learning-based models. A neural decision forest was also utilized as a machine learning model to improve the explainability of a feature's effectiveness.

Although our proposed method showed promising results for predicting outcomes in the CPC1 dataset, we should note several limitations. First, the CPC1 dataset was recorded under a specific scenario in a cuboid room with low to moderate simulated reverberation. Second, although the CPC1 dataset is the most comprehensive dataset involving hearing-impaired listeners to date, the proposed method might not be generalizable to other populations or to datasets recorded under different scenarios. Third, we only used hearing loss conditions based on pure-tone audiograms to develop the auditory model. Hearing loss conditions are complex, and the audiograms might not have been sufficient to represent them. Further research in hearing loss modeling and experiments with a wider variety of datasets will be crucial to extend the proposed method's applicability.

6. Summary and future work

In this study, we developed a non-intrusive method that incorporates an auditory periphery model to predict speech intelligibility under hearing loss conditions. The proposed method only requires binaural improved speech-perception-in-noise (SPIN) signals and an audiogram representing a given listener's hearing loss. The features extracted for the speech intelligibility prediction model are the spectral envelopes and additional acoustic features (eGeMAPS and wavLM). The model was constructed using a two-dimensional CNN module combined with an NDF layer. To evaluate the proposed method, the CPC1 dataset was used with three evaluation metrics, the Pearson correlation coefficient, RMSE, and coefficient of determination. A comparative analysis of multiple methods was performed to further evaluate the proposed method. The experimental results showed that our method outperformed the other methods for both the closed- and open-set tracks of the CPC1 dataset.

We also performed an ablation study to analyze the auditory model performance. These results demonstrated that each component in the proposed auditory model could positively contribute to improving speech intelligibility prediction. The additional acoustic features also significantly improved the prediction results. For future directions, we will investigate a better binaural processing model for more realistic scenes, and we aim to incorporate the proposed method in hearing aid development.

CRediT authorship contribution statement

Candy Olivia Mawalim: Conceptualization, Methodology, Software, Investigation, Validation, Visualization, Writing – original draft, Review and editing, Funding acquisition. **Benita Angela Titalim:** Methodology, Software, Writing – original draft, Review and editing. **Shogo Okada:** Methodology, Validation, Review and editing, Funding acquisition. **Masashi Unoki:** Conceptualization, Validation, Funding acquisition, Review and editing.

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.

Data availability

The data is available online.

Acknowledgements

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), the Japan Society for the Promotion of Science (JSPS) KAKENHI grant (No. 22K21304, No. 22H04860, and No. 22H00536), and JST AIP Trilateral AI Research, Japan (No. JPMJCR20G6).

References

[1] Falk TH, Zheng C, Chan W-Y. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Trans Audio Speech Lang Process 2010;18(7):1766–74.
[2] Mendel L. Objective and subjective hearing aid assessment outcomes. Am J Audiol 2007;16:118–29.
[3] Kates J, Arehart K. The hearing-aid speech perception index (HASPI) version 2. Speech Commun 2021;131:35–46.
[4] Andersen AH, de Haan JM, Tan Z-H, Jensen J. Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions. Speech Commun 2018;102:1–13.
[5] Graetzer S, et al. Clarity-2021 challenges: machine learning challenges for advancing hearing aid processing. In: Proc. of interspeech, ISCA; 2021. p. 686–90.
[6] Barker J, Akeroyd M, Cox TJ, Culling JF, Firth J, Graetzer S, et al. The 1st clarity prediction challenge: a machine learning challenge for hearing aid intelligibility prediction. In: Proc. of interspeech, ISCA; 2022. p. 3508–12.
[7] Munro MJ, Derwing TM. Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. Lang Speech 1995;38(3):289–306.
[8] Edraki A, Chan W-Y, Jensen J, Fogerty D. Speech intelligibility prediction using spectro-temporal modulation analysis. IEEE/ACM Trans Audio Speech Lang Process 2021;29:210–25.
[9] Janbakhshi P, Kodrasi I, Bourlard H. Spectral subspace analysis for automatic assessment of pathological speech intelligibility. In: Proc. of interspeech, ISCA; 2019. p. 3038–42.
[10] ANSI S3.5. Methods for calculation of the speech intelligibility index. American National Standard Institute; 1997.
[11] Houtgast T, Steeneken HJM. The modulation transfer function in room acoustics as a predictor of speech intelligibility. J Acoust Soc Am 1973;54:557.
[12] IEC 60268–16:2020, Sound system equipment – part 16: objective rating of speech intelligibility by speech transmission index; 2020.
[13] Goldsworthy RL, Greenberg JE. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. J Acoust Soc Am 2004;116(6):3679–89.
[14] Chen F, Wong L, Hu Y. A Hilbert-fine-structure-derived physical metric for predicting the intelligibility of noise-distorted and noise-suppressed speech. Speech Commun 2013;55:1011–20.
[15] Jensen J, Taal CH. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. IEEE/ACM Trans Audio Speech Lang Process 2016;24(11):2009–22.
[16] Taal CH, Hendriks RC, Heusdens R, Jensen J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans Audio Speech Lang Process 2011;19(7):2125–36.
[17] Andersen AH, de Haan JM, Tan Z-H, Jensen J. Predicting the intelligibility of noisy and nonlinearly processed binaural speech. IEEE/ACM Trans Audio Speech Lang Process 2016;24(11):1908–20.
[18] Santos JF, Cosentino S, Hazrati O, Loizou PC, Falk TH. Objective speech intelligibility measurement for cochlear implant users in complex listening environments. Speech Commun 2013;55(7):815–24. https://doi.org/10.1016/j.specom.2013.04.001.
[19] Suelzle D, Parsa V, Falk TH. On a reference-free speech quality estimator for hearing aids. J Acoust Soc Am 2013;133(5):EL412–8.
[20] Chen F, Hazrati O, Loizou P. Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure. Biomed Signal Process Control 2013;8:311–4.
[21] Cooper E, Huang W, Toda T, Yamagishi J. Generalization ability of MOS prediction networks. In: IEEE international conference on acoustics, speech and signal processing, Speech and signal processing, ICASSP 2022, virtual and Singapore, 23–27 May 2022. IEEE; 2022. p. 8442–6.
[22] Zezario RE, Fu S, Chen F, Fuh C, Wang H, Tsao Y. Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features. IEEE/ACM Trans Audio Speech Lang Process 2023;31:54–70. https://doi.org/10.1109/TASLP.2022.3205757.
[23] Nejime Y, Moore B. Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise. J Acoust Soc Am 1997;102:603–15.
[24] Zezario RE, Chen F, Fuh C-S, Wang H-M, Tsao Y. MBI-Net: a non-intrusive multi-branched speech intelligibility prediction model for hearing aids; 2022. p. 3944–8.
[25] Eyben F, Scherer K, Schuller B, Sundberg J, Andre E, Busso C, et al. The Geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE Trans Affect Comput 2016;7:190–202. https://doi.org/10.1109/TAFFC.2015.2457417.
[26] Eyben F, Wöllmer M, Schuller Opensmile B. The Munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on multimedia. New York, NY, USA: Association for Computing Machinery; 2010. p. 1459–62.
[27] Chen S, Wang C, Chen Z, Wu Y, Liu S, Chen Z, et al. WavLM: large-scale self-supervised pre-training for full stack speech processing. arXiv:2110.13900, 2021.
[28] Titalim BA, Mawalim CO, Okada S, Unoki M. Speech intelligibility prediction for hearing aids using an auditory model and acoustic parameters. In: 2022 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC); 2022. p. 1076–84.
[29] Kates J. An auditory model for intelligibility and quality predictions. J Acoust Soc Am 2013;133:3560.
[30] Moore BCJ, Vickers DA, Plack CJ, Oxenham AJ. Inter-relationship between different psychoacoustic measures assumed to be related to the cochlear active mechanism. J Acoust Soc Am 1999;106(5):2761–78.
[31] Cooke M. Modelling auditory processing and organisation. PhD Dissertation. University of Sheffield, Computer Science Department; 1991.
[32] Moore BCJ, Glasberg BR. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J Acoust Soc Am 1983;74(3):750–3.
[33] Kontschieder P, Fiterau M, Criminisi A, Bulò SR. Deep neural decision forests. In: 2015 IEEE international conference on computer vision (ICCV). IEEE Computer Society; 2015. p. 1467–75.
[34] Mawalim CO, Titalim BA, Unoki M, Okada S. OBISHI: objective binaural intelligibility score for the hearing impaired. In: Proc. of SST; 2022. p. 111–5.
[35] Hammarberg B, Fritzell B, Gauffin J, Sundberg J, Wedin L. Perceptual and acoustic correlates of abnormal voice qualities. Acta Oto-Laryngol 1980;90(5–6):441–f51.
[36] Tu Z, Ma N, Barker J. Unsupervised uncertainty measures of automatic speech recognition for non-intrusive speech intelligibility prediction. In: Proc. of interspeech, ISCA; 2022.
[37] Feng Y, Chen F. Nonintrusive objective measurement of speech intelligibility: a review of methodology. Biomed Signal Process Control 2022;71:103204.