IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 November 2024, accepted 30 January 2025, date of publication 4 February 2025, date of current version 11 February 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3538708

RESEARCH ARTICLE

Speech Intelligibility Prediction Using Binaural Processing for Hearing Loss

XIAJIE ZHOU, CANDY OLIVIA MAWALIM, (Member, IEEE), AND MASASHI UNOKI, (Member, IEEE)

Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Ishikawa 923-1292, Japan

Corresponding author: Xiajie Zhou (xiajie@jaist.ac.jp)

This work was supported in part by the Strategic Information and Communications Research and Development Promotion Programme (SCOPE) of Ministry of Internal Affairs and Communications in Japan under Grant 201605002; in part by the Grant-in-Aid for Scientific Research (B) under Grant 21H03463; and in part by Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 20KK0233, Grant 21H03463, and Grant 22K21304.

ABSTRACT As the global issue of hearing loss becomes increasingly severe, developing effective speech intelligibility prediction methods is crucial for improving the performance of hearing aids. However, current methods struggle in noisy environments and overlook individual differences in hearing loss between ears, which impacts prediction accuracy. Therefore, this study proposes a non-intrusive speech intelligibility prediction method that incorporates the binaural processing for hearing loss. The proposed method simulates the multi-stage binaural processing of the outer, middle, and inner ear and integrates binaural cues through an equalization-cancellation model to mitigate masking effects in noisy environments. Key features extracted from speech signals serve as inputs for a hybrid speech intelligibility model combining long short-term memory (LSTM) and light gradient boosting machine (LightGBM) models. The proposed method captures the critical features of speech signals, especially in challenging environments and for different types of hearing loss. Experimental results show that, compared to the baseline system of the second Clarity Prediction Challenge (CPC2) dataset, the proposed method achieves an 8.3% reduction in root mean squared error (RMSE). Notably, the proposed method reduces RMSE by 12.8% when predicting inconsistent hearing loss compared to listeners with consistent hearing levels, confirming the potential of combining hearing loss modeling with binaural processing.

INDEX TERMS Hearing loss, speech intelligibility prediction, binaural processing, equalization-cancellation model.

I. INTRODUCTION

Hearing loss is a growing public health challenge globally. According to the World Hearing Report, by 2050, nearly 2.5 billion people will experience hearing loss, with at least 700 million requiring rehabilitation services [1], [2]. This condition, especially among older adults, leads to communication difficulties, isolation, and frustration, significantly impacting quality of life [3], [4]. Economically, hearing loss increases medical costs related to mental health and cognitive impairment, and leads to an earlier exit from the labor market, reduced income, and greater dependence on social services. Studies show that individuals with hearing loss face a 52% higher risk of social isolation, a 47% greater likelihood of depression, and an unemployment rate twice that of people with normal hearing [5]. Hence, hearing health is a critical issue in the medical field and a key factor influencing social well-being, demanding urgent global efforts.

As a wearable and easy-to-use device, hearing aids can effectively improve speech intelligibility for individuals with hearing loss [6], [7]. Indeed, a study comparing hearing aids with different technical designs found that both advanced and basic models could significantly improve speech understanding in daily life [8], suggesting that even more affordable hearing aids can substantially benefit users. However, a noticeable gap in global hearing aid services still

The associate editor coordinating the review of this manuscript and approving it for publication was Akansha Singh.

exists, particularly in low-income countries, where only 17% of those in need of hearing aids use them [1]. The gap is not only due to the high cost of advanced hearing aids but also to the lack of knowledge about hearing tests and the public's insufficient awareness of the benefits of hearing aids, particularly their role in improving speech intelligibility.

Traditionally, the assessment of speech intelligibility relies on subjective hearing tests, which are both time-consuming and resource-intensive. With the development of machine learning technologies, researchers have begun exploring intrusive and non-intrusive methods for objectively predicting speech intelligibility. These methods learn how the auditory system functions and capture key speech features to predict speech intelligibility. For example, Andersen et al. proposed a non-intrusive speech intelligibility prediction model based on a convolutional neural network (CNN), which predicted speech intelligibility by analyzing speech signals [9], [10].

However, subjective hearing tests assess speech intelligibility by relying on the brain's ability to balance spatial auditory cues, integrating both binaural and spectral cues [11], [12], [13]. Binaural cues, including interaural time differences (ITD) and interaural level differences (ILD), play a crucial role in horizontal localization (azimuth). In contrast, spectral cues are generated by the frequency-filtering effects of the outer ear (pinna) and provide essential information for vertical localization (elevation). This combined spatial information allows for more precise and reliable localization, especially in real-world scenarios.

Both binaural and spectral cues enhance speech intelligibility. Research shows that binaural hearing aids improve speech recognition in noisy environments, particularly when sounds and noise come from different directions [14]. However, for hearing aids, horizontal localization tends to be more accurate [15]. This difference arises because the human ear and brain have robust mechanisms for processing time and phase differences for horizontal localization. In contrast, vertical localization lacks similar physiological signals and processing pathways, reducing accuracy. Consequently, this study emphasizes the importance of binaural cues, specifically ITD and ILD, as they enhance sound localization and noise suppression and thereby provide a natural auditory experience [16]. Specifically, the roles of binaural cues include the following aspects:

- Head Shadow Effect: When noise arrives from different directions, the head causes diffraction of sound waves, generating ITD and ILD. These differences allow listeners to accurately localize sound sources and enhance speech understanding in noisy environments.
- Effect of Noise Azimuth: The direction of noise sources affects speech intelligibility. When noise comes from the side (e.g., at a 90-degree azimuth), binaural hearing can provide a greater signal-to-noise ratio (SNR) gain than monaural hearing. This gain stems from the ILD effect on the ear with better high-frequency hearing, with increases ranging from 0 to over 7 dB in individuals with normal hearing.
- Combined Effect of ITD and ILD: When noise signals contain ITD and ILD, speech intelligibility gains range from 2 to 2.5 dB across all listener groups. Studies have shown that binaural cues can enhance speech intelligibility in individuals with hearing loss [16].

To further explore how hearing loss affects speech perception, auditory spectrograms can be utilized to visualize the impact of different types and degrees of hearing impairment on the frequency content and energy of speech signals. In particular, gammatone filterbank (GTFB) spectrograms, which model the human auditory filterbank, provide detailed insights into how hearing loss affects specific frequency channels [17]. By simulating hearing loss using audiograms and analyzing the resulting GTFB spectrograms, we can better comprehend the challenges individuals face and develop more effective hearing aid algorithms.

Considering an objective speech intelligibility prediction model for hearing aids, this study introduces a model that simulates hearing loss in both ears using audiograms. Binaural processing is incorporated through the equalization-cancellation (EC) model to eliminate masking effects and improve speech intelligibility [18]. Advanced feature extraction methods are employed, with long short-term memory (LSTM) and light gradient boosting machine (LightGBM) models combined in a hybrid approach for final predictions.

The proposed method is trained on the second Clarity Prediction Challenge (CPC2) dataset, which includes processed signals from hearing aids and listener information such as age, gender, and audiograms [19]. Each listener's audiogram is stored as a frequency-level pair for both ears, and all listeners have experience with binaural hearing aids. The evaluation set consists of listeners who do not overlap with the training set, verifying the model's generalization ability to unseen listeners. The final prediction is evaluated using root mean squared error (RMSE), with lower values indicating a better prediction performance.

In summary, the main contributions of this work are as follows:

- We propose a speech intelligibility prediction method that addresses the challenge of combining binaural processing with hearing loss simulations. This approach captures the ability to process speech information across different levels of hearing loss in both ears.
- The proposed method demonstrates strong generalization capabilities for predicting speech intelligibility in unknown and noisy environments.

Section II of this paper reviews related work. In Section III, we introduce the proposed method with binaural processing for hearing loss, and in Section IV, we describe the dataset, experimental setup, and evaluation metrics. Section V provides an overall evaluation of the proposed method's effectiveness in predicting speech intelligibility. In Section VI, we analyze the impact of different types and severities of hearing loss, demonstrating the advantages of the proposed method. Section VII examines the impact of datasets in unknown noisy environments with varying

TABLE 1. Evolution of speech intelligibility prediction models.

Method	Type	Description	Strengths	Limitations
Articulation index [20]	Intrusive	**SNR-based** speech intelligibility prediction, originally for telephone lines.	Works well in stationary noise.	Poor in complex noise, sensitive to distortions.
Speech intelligibility index (SII) [21]	Intrusive	**Extended articulation index model** with frequency importance functions (FIFs) and masking effects for noisy environments.	Better than articulation index in varied conditions.	Struggles with non-linearity, assumes idealized conditions.
Hearing aid speech perception index (HASPI) [22]	Intrusive	Auditory model analyzing **envelope and fine structure** for speech intelligibility prediction.	Excellent in noisy/nonlinear conditions (e.g., frequency compression).	Requires clean reference signal, high computational cost.
Modified binaural short-time objective intelligibility (MBSTOI) [23]	Intrusive	Integrates left and right signals using the equalization-cancellation stage to **optimize short-time objective intelligibility (STOI)**.	Improves binaural speech intelligibility prediction, effective in simple environments.	Requires clean reference signal, high computational cost.
Non-intrusive methods [24]	Non-intrusive	Machine learning-based methods predicting speech intelligibility **without clean reference signal**.	Flexible, practical in noisy environments.	Generally less accurate in controlled conditions.

interference sources. Finally, we conclude in Section VIII with future directions.

II. RELATED WORK

A. EVOLUTION OF SPEECH INTELLIGIBILITY PREDICTION MODELS

Although traditional intrusive methods, such as the articulation index and speech intelligibility index (SII), provide a solid theoretical and practical foundation, they face limitations in handling complex noise and nonlinearities. These methods, detailed in Table 1, are less accurate in controlled environments but offer greater flexibility and practicality for real-world applications. As shown in the table, the adaptability of non-intrusive methods has become important in speech intelligibility prediction.

Models such as the articulation index and SII have been mentioned as methods for speech intelligibility prediction in hearing aids [21], [25], [26]. The articulation index was initially proposed to predict the effects of telephone line variations on speech comprehension. However, it has since been used to evaluate the performance of speech communication systems [20]. In predicting speech intelligibility, the articulation index is a weighted score representing an effective proportion of the speech signal, which is available to the listener under the given speech channel and noise conditions. The articulation index calculates the SNR contribution of speech signals in different frequency bands by analyzing acoustic measurements. However, it has limited applicability in dealing with complex non-stationary noise, such as intermittent noise. In addition, it is sensitive to frequency and amplitude distortions in speech signals.

To address the limitations of the articulation index, the SII was proposed as an extended version, providing a more flexible and general framework for calculating the availability of speech information [21]. SII inherits the basic theory of the articulation index, positing that speech intelligibility is directly related to the proportion of audible speech information. However, SII introduces more variables and correction factors in its calculation process, such as frequency importance functions, upward spread of masking effects, and the influence of high sound pressure levels.

Although the articulation index and SII provide theoretical foundations and practical tools for speech intelligibility prediction, they have limitations in practical applications [27]. Studies have shown that while they can predict speech recognition scores through empirical transfer functions, they do not directly measure speech intelligibility; instead, they reflect the audibility and availability of speech signals under specific conditions. This means that the articulation index and SII models rely on idealized acoustic conditions and cannot adequately cope with noisy environments or nonlinear distortions. In particular, when applied to the prediction evaluation of hearing aids, their accuracy may be affected by different noise [24].

To address the limitations of traditional models, researchers have developed a series of prediction tools in recent years, aiming to overcome the shortcomings of these conventional methods. These tools can be divided into two categories: intrusive and non-intrusive methods. Intrusive methods require a clean reference signal, while non-intrusive methods do not. Intrusive methods have the advantage of being able to accurately compare the difference between the clean reference signal and the test signal, especially when dealing with complex noise and nonlinear distortions.

For example, intrusive methods such as the hearing aid speech perception index (HASPI) utilize processing based on the auditory periphery model to better simulate the effects of hearing impairments [22], [28], [29]. These methods perform well under high SNR conditions and can adapt to the complex signal variations introduced during hearing aid processing. In particular, HASPI has been shown to

provide excellent predictions of speech intelligibility under challenging conditions such as frequency compression and noise suppression. The better ear HASPI (be-HASPI) can simulate the auditory changes in both normal hearing and hearing loss [30]. This model compares the temporal amplitude envelope and temporal fine structures of degraded signals with unprocessed clean reference signals to accurately simulate the effects of hearing loss on speech signals [28].

The short-time objective intelligibility (STOI) measure is utilized to predict speech intelligibility by assessing the correlation between short-time segments of clean and degraded signals. The modified binaural STOI (MBSTOI) measure uses a modified equalization cancellation stage to optimize speech intelligibility predictions [9], [23]. In MBSTOI, left and right ear signals are processed through an EC stage, which aligns and cancels interfering noise by maximizing the correlation between signals. This approach calculates intelligibility by adjusting parameters for each time frame and frequency band, achieving an optimal correlation between the clean and improved speech signals. **However, despite the high precision of intrusive methods, they also have some notable limitations.** First, these methods require a clean reference signal, which is often difficult to obtain in practical operations. Second, in real-world hearing aid applications, the process of acquiring and synchronizing clean reference signals may be limited by environmental factors and device performance.

To address the limitations of intrusive methods, researchers have explored non-intrusive methods. These methods use advanced signal processing techniques and machine learning algorithms, enabling the models to capture critical information directly from the output signal without requiring a clean reference signal [24]. This mitigates the need for synchronization between reference and target signals, making non-intrusive models particularly advantageous in variable acoustic environments and complex noise conditions.

For example, researchers have developed a non-intrusive speech intelligibility prediction method that utilizes hierarchical and temporal features in noise-robust models to address the limitations of traditional models. This method utilizes a pre-trained model such as the whisper and waveform language model (WavLM) to extract speech signal features and predict intelligibility [31]. It also uses a transformer model to process the obtained temporal features and incorporates audiogram information from listeners to improve the prediction. These processing steps enable the model to separate speech signals in noisy environments.

B. INTEGRATION OF NON-INTRUSIVE SPEECH INTELLIGIBILITY PREDICTION METHODS WITH THE AUDITORY SYSTEM

As shown in Fig. 1, our motivation for this study is twofold: first, to clarify the communication barriers caused by hearing loss; second, to create a binaural signal processing model designed for hearing aids. The figure illustrates the challenges

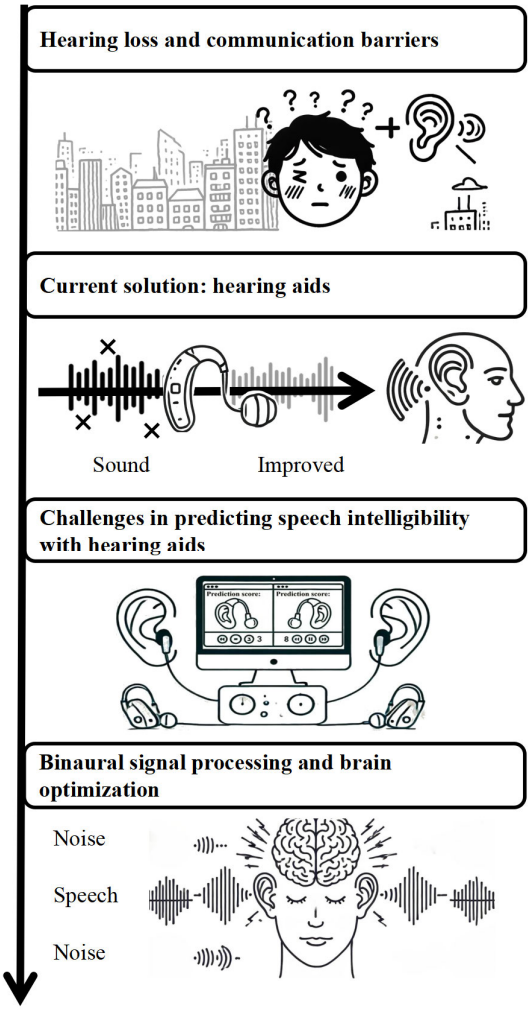


FIGURE 1. A holistic approach to enhancing speech intelligibility for hearing aids: from hearing aids to binaural signal processing.

associated with current hearing aid technology and highlights points of inspiration for combining prediction methods with binaural signal processing. In this section, we discuss combining the auditory system capabilities of the human ear with advanced machine learning techniques to predict speech intelligibility even in noisy environments.

The global lack of hearing health knowledge has left many people without the help they need. At the same time, research by Cox et al. has highlighted that while both new and experienced users believe hearing aids greatly improve communication in daily life, there are still issues related to overall user satisfaction [8]. Therefore, in addition to technological advancements, it is equally important to ensure that individuals with hearing loss trust and actively use hearing aids. Despite existing users believing that hearing aids greatly improve communication, many people remain skeptical about their effectiveness, leading to lower usage rates than expected. This is why it is important to develop objective prediction methods that help individuals with hearing loss predict speech intelligibility in unknown environments.

In 1991, researchers conducted experiments comparing the subjective and objective speech clarity measurements of older adults with hearing loss [32], [33]. Although there was a clear correlation between the two, subjective scores tended to be lower than objective scores. This suggests that older individuals may underestimate their speech comprehension abilities, thereby leading them to question the effectiveness of hearing aids. Additionally, over-reliance on subjective feedback during fitting and adjustment may result in inaccurate settings that fail to fully utilize the device's capabilities. Therefore, combining subjective and objective measurement methods during the fitting process is critical to ensure optimal hearing compensation for older adults with hearing loss [34].

Non-intrusive speech intelligibility prediction methods are practical in assessing speech intelligibility without needing a clean reference signal [35]. **However, these non-intrusive methods still face many challenges when dealing with noisy environments and different types of hearing loss.** For example, they ignore the specific needs of auditory systems, so individual differences in hearing loss patients are not adequately considered [36], [37]. The human ear can analyze and understand sound in various environments, and its mechanism is very complex and delicate [38], [39]. Non-intrusive methods can only analyze the external characteristics of signals, and it is difficult to deeply simulate the process of the auditory system in processing speech signals. Due to the lack of simulation of the internal mechanisms, the prediction ability of these methods in high-noise environments or complex hearing loss situations is limited.

Combining binaural signal processing with non-intrusive methods considers the link between objective and subjective prediction methods. Incorporating hierarchical processing of the ear and neural signal analysis into these objective prediction models is intended to mimic the processing of signals by the auditory system of an individual with hearing loss. This treatment is also useful for different types of hearing loss, where we can adapt the model to an individual's unique hearing characteristics. By considering prediction models that incorporate the auditory system, we aim to make the model work better in noisy environments and better understand how individuals with hearing loss experience sound in noisy environments.

III. PROPOSED METHOD

We propose a non-intrusive speech intelligibility prediction method to enhance prediction accuracy for hearing aids, as shown in Fig. 2. In this section, we present a comprehensive overview of the proposed method.

A. BINAURAL HEARING LOSS SIMULATION USING THE MSBG MODEL

This study utilizes the Cambridge Auditory Group Moore/Stone/Baer/Glasberg (MSBG) hearing loss model based on Nejime's research, which plays a crucial role in adjusting speech signals to reflect the anticipated hearing loss effects [40]. The model processes stereo speech signals and requires audiograms representing the hearing thresholds of both ears. Audiograms provide essential frequency-specific information on hearing sensitivity, allowing the model to replicate the listener's auditory experience under various hearing loss conditions. By adjusting the signal processing parameters, the model simulates the effects of different degrees of hearing loss (e.g., mild, moderate, severe) on the perceived sound quality. This model simulates how individuals with hearing loss experience audio, especially in speech intelligibility, where the degradation of sound clarity often hinders effective communication. The model is suited to simulating hearing aids and other assistive devices for hearing loss.

1) OUTER AND MIDDLE EAR PROCESSING

The initial stage of the model involves simulating the effects of the outer and middle ear. The outer ear, also known as the pinna, plays a crucial role in spatial hearing by emphasizing higher frequencies, which helps in sound localization. The middle ear, on the other hand, performs the essential function of impedance matching between the air-filled ear canal and the fluid-filled cochlea, ensuring efficient sound energy transfer.

The model applies filters that replicate the acoustic transfer from the free field to the eardrum and through the middle ear. These filters are designed to reflect the natural physical properties of the human auditory system. The processing is independent of audiogram data and simulates the response of the outer and middle ear to sound. The outer ear emphasizes higher frequencies, which are essential for spatial hearing [41]. The middle ear, on the other hand, acts as an impedance match, ensuring that sound waves are adequately transmitted from the air-filled ear canal to the cochlea. The signal is thus attenuated or amplified appropriately before reaching the cochlea, which ensures that the higher frequencies (important for consonant recognition) are preserved or enhanced [41]. This processing stage ensures that the signal entering the cochlea reflects the natural hearing experience as accurately as possible before the hearing loss is simulated.

2) AUDIOGRAM SELECTION AND ITS IMPACT ON HEARING LOSS SIMULATION

The MSBG model utilizes audiograms to define hearing thresholds at various frequencies for each ear. These audiograms are crucial for tailoring the hearing loss simulation to individuals, as they provide frequency-specific information about hearing sensitivity. Audiograms plot the hearing threshold levels (typically measured in hearing level) against the frequency spectrum, ranging from low to high frequencies. In this study, audiograms were selected based on the severity of hearing loss (e.g., mild, moderate, severe) and were applied to both ears independently. Each

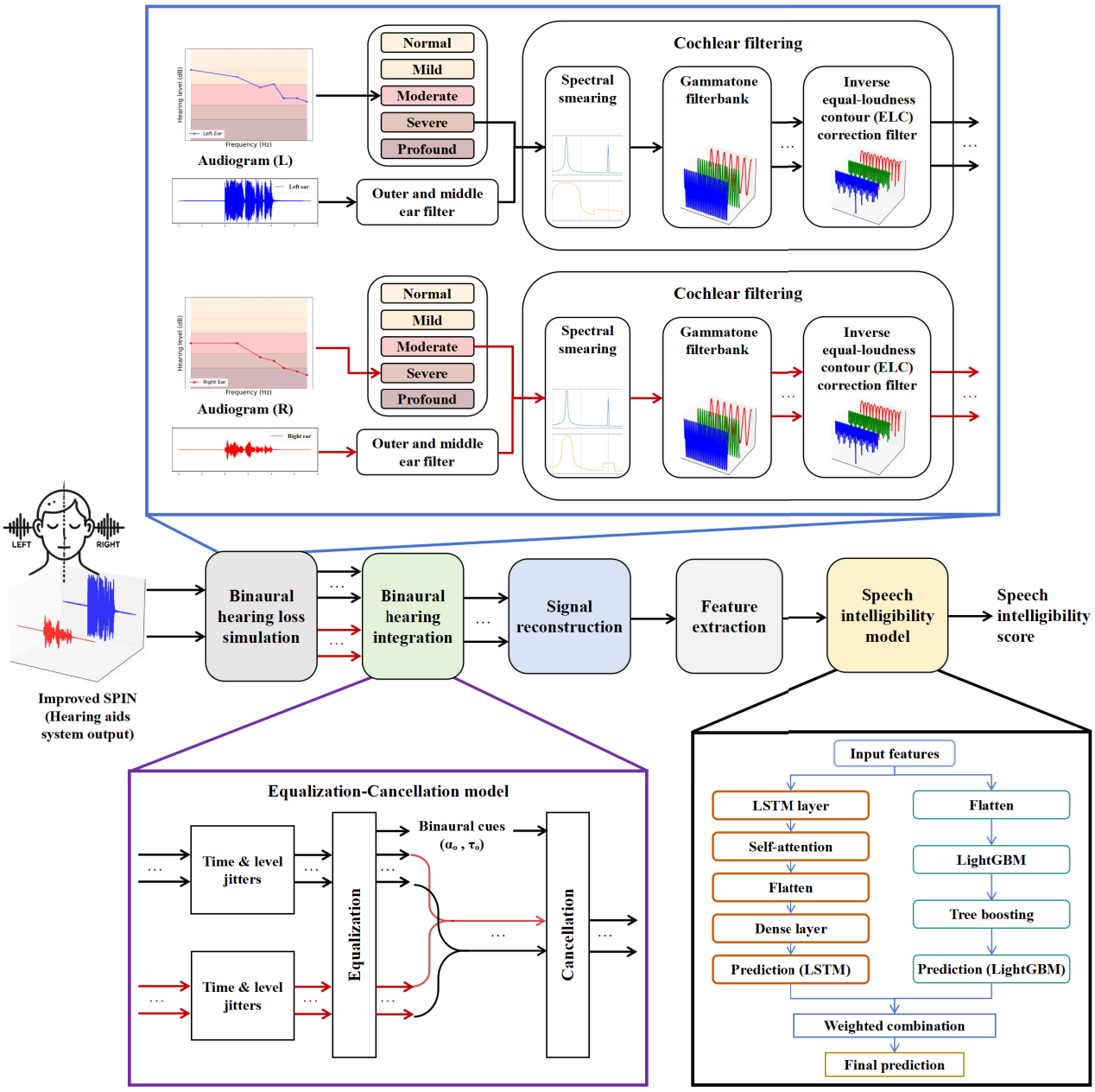


FIGURE 2. Block diagram of proposed method. (L) and (R) respectively denote left and right ear signals. SPIN represents speech in noise.

audiogram serves as a frequency-dependent attenuation map that determines how the audio signal is processed through the model.

3) COCHLEAR FILTERING AND HEARING LOSS SIMULATION

The cochlea plays a critical role in hearing by converting sound waves into electrical signals that the brain can interpret. In a healthy ear, the cochlea functions with high frequency selectivity, allowing individuals to distinguish between closely spaced frequencies. However, damage to the cochlea, such as through aging, noise exposure, or other factors, has several detrimental effects on hearing. These effects include increased hearing threshold, loudness recruitment, and reduced frequency selectivity, all of which are simulated in the MSBG hearing loss model.

a: INCREASED HEARING THRESHOLD

When the cochlea is damaged or auditory nerves are compromised, the hearing threshold increases, meaning that softer sounds become inaudible unless amplified. This is a common effect of sensorineural hearing loss, where the ability to perceive soft sounds is significantly diminished. In the MSBG model, this effect is simulated by applying frequency-dependent attenuation to the audio signal based on

the audiogram. The audiogram provides hearing thresholds at different frequencies, and the model attenuates signals according to the severity of hearing loss in each frequency band. For example, mild hearing loss primarily affects high frequencies, whereas severe hearing loss may impact a broad range of frequencies, making speech sounds imperceptible, especially high-frequency consonants.

b: LOUDNESS RECRUITMENT

Another important aspect of cochlear damage is loudness enhancement, in which a person with hearing loss may feel that sounds become abnormally loud once they exceed a certain threshold. In a healthy cochlea, outer hair cells help modulate the intensity of incoming sounds, but the dynamic range is reduced with hearing loss. The MSBG model simulates this effect by applying expansive nonlinearity to the audio signal. In this way, once a sound surpasses the hearing threshold (as defined by the audiogram), the perceived loudness grows rapidly, reflecting the experience of individuals with loudness recruitment. This phenomenon can cause discomfort as certain sounds are perceived as uncomfortably loud, even at moderate volumes.

c: REDUCED FREQUENCY SELECTIVITY

Frequency selectivity refers to the cochlea's ability to distinguish between sounds that are close in frequency. Hearing loss, particularly in the high frequencies, reduces the cochlea's frequency selectivity. This is especially problematic for speech understanding, as it becomes difficult to differentiate between speech sounds like "s" and "sh." In the MSBG model, this effect is simulated by broadening the Equivalent Rectangular Bandwidth (ERB) of the cochlear filters [42]. For individuals with hearing loss, the ERB is typically widened, reducing the cochlea's ability to resolve different frequencies finely. The model reflects this broadening by adjusting the filterbank parameters, making it more challenging to separate sounds that are close in frequency, thereby simulating the decreased ability to discriminate speech in noisy environments.

Together, these processes-*increased hearing threshold*, *loudness recruitment*, and *reduced frequency selectivity*-ensure that the MSBG model provides a realistic simulation of hearing loss. The model's ability to adjust the speech signal based on the listener's audiogram makes it a powerful tool for understanding how different degrees of hearing loss affect speech perception and the overall auditory experience.

B. BINAURAL HEARING AND THE EC MODEL FOR SPEECH INTELLIGIBILITY

1) THE EC MODEL AND BINAURAL CUES PROCESSING

The EC model, developed by Wan [18], utilizes binaural cues to enhance the target SNR. It processes left and right ear signals separately across frequency bands and dynamically adjusts binaural cues through modulation of ITD and ILD [18], [43]. Binaural hearing plays a critical role in noisy environments by allowing listeners to localize sound sources and improve speech intelligibility, especially in the presence of background noise. The EC model mimics the brain's processing of these binaural cues, aiming to replicate how humans naturally enhance target sounds in noisy environments.

TABLE 2. Symbols and their corresponding meanings in the EC model.

Symbol	Description
$U(i,j)$	Time-frequency unit in the i-th frequency channel and j-th time frame
$\tau_o(i,j)$	Optimal interaural time difference (ITD) for unit $U(i,j)$
$\alpha_o(i,j)$	Optimal interaural level difference (ILD) for unit $U(i,j)$
$\rho_{i,j}(\tau)$	Normalized cross-correlation function for left and right ear signals
$E_{\mathrm{NL}_{i,j}}$	Masker energy for the left ear in unit $U(i,j)$
$E_{\mathrm{NR}_{i,j}}$	Masker energy for the right ear in unit $U(i,j)$
$L_i(t)$	Jittered waveform for the left ear in the i-th frequency channel
$R_i(t)$	Jittered waveform for the right ear in the i-th frequency channel
$W_j(t)$	Rectangular window function for the j-th time frame
ω_i	Angular frequency corresponding to the i-th frequency channel

In each time-frequency cell, the EC model optimizes noise cancellation by selecting the ideal ITD ($\tau_o(i,j)$) and ILD ($\alpha_o(i,j)$) values. These parameters, along with others used in the model, are summarized in Table 2. These parameters are calculated as follows:

$$\tau_o(i,j) = \underset{\tau}{\mathrm{argmax}}\{\rho_{i,j}(\tau)\},\ |\tau| < \frac{\pi}{\omega_i}, \tag{1}$$

$$\alpha_o(i,j) = \sqrt{\frac{E_{\mathrm{NL}_{i,j}}}{E_{\mathrm{NR}_{i,j}}}}, \tag{2}$$

where $\rho_{i,j}(\tau)$ is the normalized cross-correlation function that represents the synchronization of jittered signals between the left and right ears. $E_{\mathrm{NL}_{i,j}}$ and $E_{\mathrm{NR}_{i,j}}$ represent the masker energies for the left and right ears, respectively. The purpose of this process is to align the left and right ear signals in such a way that maximizes the cancellation of unwanted noise while preserving the target speech signal.

Once the optimal values for $\tau_o(i,j)$ and $\alpha_o(i,j)$ are obtained, the EC model applies these parameters in a noise cancellation process. The resulting output signal for each time-frequency cell $Y_{i,j}(t)$ is computed as

$$Y_{i,j}(t) = W_j(t)\left\{\frac{1}{\sqrt{\alpha_o(i,j)}}\left(L_i\left(t+\frac{\tau_o(i,j)}{2}\right) - \sqrt{\alpha_o(i,j)}R_i\left(t-\frac{\tau_o(i,j)}{2}\right)\right)\right\}, \tag{3}$$

where $L_i(t)$ and $R_i(t)$ represent the jittered waveforms from the left and right ears, respectively, and $W_j(t)$ is the rectangular window function. This dynamic adjustment process improves speech intelligibility in noisy environments with competing maskers by optimizing the binaural cues in each time-frequency unit.

In noisy environments, binaural hearing offers significant advantages, especially in helping the listener to discriminate and focus on target sounds. By comparing the time difference (ITD) and volume difference (ILD) between the arrival of a sound in the left and right ears, the brain can localize the source of the sound and improve speech intelligibility. This mechanism, known as the ''cocktail party effect,'' helps listeners focus on specific target sounds (e.g., the speaker's voice) even when other distracting sounds are in the background. The proposed method simulates the brain's enhancement of important auditory information during this process by dynamically adjusting the ITD and ILD cues. This processing makes the target speech clearer while attenuating the effects of interfering sounds.

2) AUDITORY MASKING AND THE BRAIN'S ROLE

Auditory masking occurs when one sound makes it difficult to hear another. Through selective attention, the brain prioritizes the target sound from a specific direction and ignores other distracting sounds. However, when the masker is similar to the target sound (e.g., multiple speakers), it is more difficult for the brain to distinguish the target from the masker, and this is the situation in which auditory masking is most likely to affect speech comprehension. Competing maskers (e.g., other conversations or noises) can interfere with the listener's ability to understand the target speech. The proposed method reduces the effects of these maskers by incorporating binaural processing. Specifically, the binaural cues in the EC model improve speech intelligibility in noisy environments by spatially separating the target signal from the masked signal.

3) PREDICTING SPEECH INTELLIGIBILITY

The EC model aims to predict listeners' ability to understand speech in environments where competing sounds are present. The model is particularly effective when there is a clear spatial separation between the target and masking sounds, as the brain can utilize spatial cues to improve speech intelligibility. However, when the masking sound is close to the target sound or has similar spectral characteristics (e.g., a masking sound similar to the speech), the masking effect is stronger, making it more difficult to understand the target speech. This dynamic tuning of binaural cues is critical because it directly reflects the brain's strategy in solving complex auditory scenarios by selectively enhancing target sounds and suppressing masking sounds to improve speech intelligibility. By simulating this process, the EC model helps to understand the effects of masking and spatial separation on listeners' speech perception abilities.

C. SPEECH INTELLIGIBILITY MODEL

To effectively predict speech intelligibility, this study uses the pre-trained WavLM to extract speech features. WavLM is a Transformer-based self-supervised learning model designed for speech processing tasks, capable of learning rich speech representations directly from raw waveforms [44], [45], [46]. In our experiments, we first load the single-channel speech signals processed through the MSBG hearing loss and EC model. The audio signal is segmented at fixed time intervals, and the WavLM extracts 1024-dimensional deep features for each audio segment. The model then averages these features across the time dimension to produce a compact and representative speech feature vector.

The WavLM extracts features that capture low-level acoustic features such as spectral properties, pitch, and resonance peaks, which reflect articulatory clarity and speech fluency. The model also pays attention to the context and can learn semantic and syntactic information in speech to help assess the coherence and completeness of speech content. In addition, WavLM demonstrates noise immunity when processing speech data in different environments and can extract key information related to the goal of the proposed method (speech intelligibility) in noisy environments. Compared to traditional speech feature extraction methods (e.g., Mel frequency cepstral coefficient) [47], WavLM provides a more comprehensive and efficient end-to-end feature representation [44].

By utilizing the WavLM, we avoid the complexity of manual feature engineering while being able to capture multi-level information from speech signals. As a result, WavLM plays a key role in speech intelligibility prediction tasks, supporting more accurate intelligibility scores. In the proposed method, we used a combination of two complementary models: LSTM and LightGBM. LSTM networks are particularly good at capturing sequential dependencies in time-series data, which is crucial for speech intelligibility since it involves understanding how speech features change over time. **Using LSTM, we can better interpret variations in the speech signal, such as pauses, clarity of articulation, and changes in pitch or tone, which are key factors in assessing intelligibility.**

As for LightGBM, a gradient boosting model based on decision trees, it is excellent at handling large feature sets and learning complex non-linear relationships [48]. In our model, the features extracted by WavLM capture different aspects of the speech signal, and **LightGBM efficiently learns the relationship between these features and the speech intelligibility scores. Its ability to handle high-dimensional data without overfitting makes it an ideal choice for this task.**

To improve the robustness of our predictions, we combined the strengths of both models by weighting their outputs. The weights were assigned based on each model's performance on the validation set, using the inverse of their RMSE values as a measure. This weighted approach ensures that we take full advantage of LSTM's sequential learning capabilities

TABLE 3. Data distribution for the CPC2 dataset. The dataset is split into three test sets (Set 1, Set 2, and Set 3) for cross-validation, each paired with a unique training subset, allowing evaluation on unseen listeners and hearing aid systems. The combined set includes all signals across the three test sets.

Data	No. of utterances in Train		No. of utterances in Test
	CEC1 [49]	CEC2 [50]	
Set 1	5820	2779	305
Set 2	5124	2772	294
Set 3	5239	2796	298
Combined	6297	5944	897

and LightGBM's feature learning power, resulting in a more accurate and generalized speech intelligibility prediction model.

IV. EXPERIMENTS

A. DATASET

In predicting speech intelligibility for hearing aids, it is crucial to use datasets that reflect real scenarios and capture speech signal characteristics under multiple interference sources and environmental conditions. For this study, we utilized the dataset from the CPC2, a competition aimed at improving methods for predicting speech intelligibility for hearing aids [19]. As shown in Table 3, the dataset is divided into two parts: the first Clarity Enhancement Challenge (CEC1) and the second Clarity Enhancement Challenge (CEC2), both consisting of speech enhanced by different systems in noisy environments [49], [50]. The training set includes three cross-training subsets (Set1, Set2, Set3), each containing different improved speech in noise (SPIN) used as references during model training. In this study, we focused on the complete combined dataset, which includes 6297 signals from CEC1 and 5944 signals from CEC2, while the combined test set contains 897 complete signals.

The dataset used in this hearing aid enhancement task includes multiple auditory scenarios, providing rich training and testing speech data. The target speech consists of sentences with with seven to ten words, overlapped with one to three interfering noise sources. Each scenario is generated using a geometric room acoustic model, with all sound sources randomly distributed within a typical living room. The test scenarios include multiple interference sources, such as music, speech, or household appliance noises. To further simulate hearing experiences, the dataset incorporates head rotations and head-related transfer function (HRTF) effects, capturing the impact of multiple interference sources on the speech signal and offering a diverse and challenging test environment. This dataset was selected because speech intelligibility prediction requires validation across multiple noisy environments and types of hearing loss to ensure the model's generalization capability. The dataset also accounts for different types of hearing loss, with participants having prior experience wearing binaural hearing aids.

TABLE 4. Parameter settings in proposed method.

Parameter name (unit)	Parameter settings
MSBG [42]	
Center frequencies (Hz)	72.6, 97.9, 129.4, 168.2, 214.9, 270.5, 335.7, 411.1, 497.7, 596.1, 707.2, 831.7, 970.6, 1124.9, 1299.1, 1495.7, 1717.6, 1968.1, 2250.8, 2569.9, 2930.0, 3336.5, 3795.2, 4313.1, 4897.5, 5557.1, 6301.7, 7142.0, 8090.5, 9161.0, 10369.2, 11733.0, 13272.2, 15009.5, 16970.3, 19183.5
Hearing loss type	**ERB, Channels**
Mild	(1.5, 36)
Moderate	(2.0, 28)
Severe	(3.0, 19)
Profound	(3.0, 19)
EC model	
Sampling frequency (Hz)	44100
Window length (seconds)	0.02
Window function type	Rectangular
Overlap ratio	50%
Frame step (seconds)	0.01
Center frequency (Hz)	1000
Padding windows	10
Feature extraction	
No. of WavLM features	1024 per second
LSTM model	
Window size	1024
LSTM units	128
Batch size	6
Epochs	50
Loss function	Mean squared error (MSE)
Optimizer	Adam
LightGBM model	
Objective	Regression
Boosting type	Gradient boosted decision trees
No. of leaves	40
Learning rate	0.05
Feature fraction	0.9
No. of boosting rounds	800
Evaluation metric	RMSE
Loss function	MSE

Each participant's hearing profile was determined through audiograms measured at 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 3000 Hz, 4000 Hz, 6000 Hz, and 8000 Hz in both ears. Inclusion criteria for participants were hearing loss in two or more consecutive frequency bands not exceeding 80 dB and using binaural hearing aids. Listeners with Ménière's disease, hyperacusis, or severe tinnitus were excluded. The proposed method categorizes hearing loss into mild ($\leq$35 dB), moderate (35–56 dB), and severe ($>$56 dB) based on the hearing loss data. Most participants' audiograms

TABLE 5. Performance comparison of proposed method with other methods on CPC2 dataset.

Data	Rank	Method	Intrusive	$\rho \uparrow$	RMSE $\downarrow$	Features	Models
CPC2	1	E011 [31]	No	0.78	25.1	Whisper, WavLM	Temporal and layer-wise transformers with cross-attention for binaural processing
	2	E002 [51]	No	0.77	25.3	Whisper	Bidirectional long short-term memory (Bi-LSTM) with attention pooling and memory-informed exemplar module
	3	E009 [52]	Yes	0.78	25.4	STOI, phone lattice, audiometric data	Non-linear regression
	4	E022 [53]	Yes	0.77	25.7	Pretrained noise-robust automatic speech recognition (ASR) hidden layers	Similarity-based prediction with logistic mapping
	5	**Proposed method**	**No**	**0.75**	**26.3**	**WavLM**	**Hybrid model with LSTM and LightGBM, combining MSBG hearing loss and EC model outputs**
	6	E023 [54]	Yes	0.76	26.4	Whisper	An extended system of enhanced multi-branched intelligibility network (MBI-Net+), MBI-Net++ with dual branches for frame-level and HASPI predictions
	7	E016 [54]	No	0.75	26.8	Whisper	MBI-Net+ with convolutional neural network-bidirectional long short-term memory with attention (CNN-BLSTM-ATT) for frame-level intelligibility
	8	E025 [53]	No	0.72	27.9	ASR-derived uncertainty (negative entropy)	Non-intrusive ASR uncertainty with logistic mapping
	9	be-HASPI [30]	Yes	0.67	28.7	Envelope fidelity, auditory coherence	Auditory model comparisons
	10	E003	No	0.64	31.1	WavLM	Stack regressor ensemble of linear, support vector machine (SVM), and random forest regressors
	11	E024	No	0.62	31.7	WavLM	LSTM with one-hot listener embedding and EC processing

showed a sloping pattern with less low-frequency loss, which is typical of age-related hearing loss. The improved SPIN from the hearing aids was validated through subjective speech intelligibility tests conducted in the experimental settings, providing us with correct scores.

B. EXPERIMENTAL SETTINGS

The proposed method utilized the following parameter settings (listed in Table 4. The center frequencies and bandwidths of the MSBG hearing loss model[1] covered the typical frequency range affected by hearing loss. By adjusting the width and spacing of auditory filters, these parameters simulated varying levels of hearing loss, including normal, mild, moderate, and severe [55]. In the EC model, for each channel,[2] continuous time-domain signals are discretized, with windowing, overlap, padding, and frame step settings applied to smooth signal variations over time, thereby enhancing the speech intelligibility. The feature extraction uses intermediate layer outputs from WavLM, processing each second of the signal [46].

The LSTM network then predicted speech intelligibility, using a window size to control the feature set. An input layer was added, followed by a reshaping layer, an LSTM layer, and a SeqSelfAttention layer to learn global dependencies. Flattened and dense layers generated the final prediction. On the other hand, the LightGBM model uses a decision tree structure to learn the complex nonlinear relationships between these features and speech intelligibility scores by processing the set of multi-level features extracted by WavLM. After both models have been trained, weights are assigned to each by calculating the RMSE on the validation set. Specifically, the weights are computed based on the inverse of the RMSE of each model, and the final weighted predictions are obtained by combining the temporal learning capability of LSTM with the complex feature learning capability of LightGBM.

[1] https://github.com/claritychallenge/clarity/tree/main/clarity/evaluator/msbg

[2] https://github.com/achabotl/pambox

C. EVALUATION METRICS

In this work, we utilized to assess the performance of the proposed model: the *Pearson correlation coefficient* (ρ) and the *RMSE*. These two metrics contribute to a comprehensive assessment of the intelligibility prediction task by measuring linear correlation and prediction accuracy, respectively.

1) ρ

The ρ measures the linear relationship between the predicted and correct scores. A value close to 1 suggests a strong positive correlation, signifying that the model is able to capture the linear trend present in the correct scores. The formula for calculating ρ is

$$\rho = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{4}$$

where y_i is the correct score, $\hat{y}_i$ is the predicted score, and $\bar{y}$ and $\bar{\hat{y}}$ are the means of the predicted and correct scores respectively.

2) RMSE

The RMSE measures the average magnitude of the error between the predicted and correct scores, providing insight into the overall prediction accuracy. Lower RMSE values indicate more accurate predictions. The RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{5}$$

where N is the total number of samples, $\hat{y}_i$ is the predicted score, and y_i is the correct score. The RMSE quantifies the deviation between predicted scores and correct scores to accurately predict hearing aids. Given that subjective intelligibility scores can vary from one listener to another, the RMSE measures the extent to which the proposed model replicates these subjective assessments.

V. ANALYSIS I: OVERALL EVALUATION OF PROPOSED METHOD IN PREDICTING SPEECH INTELLIGIBILITY

We compare the performance of our proposed method against several existing methods for predicting speech intelligibility on the second Clarity Prediction Challenge (CPC2) dataset [19]. The comparison is based on two key metrics: ρ and RMSE. A higher ρ indicates a stronger correlation between the predicted and correct scores, while a lower RMSE suggests more accurate predictions.

The CPC2 dataset comprises improved SPIN from various hearing aid algorithms and correct scores from listening tests with hearing-impaired participants. The objective is to develop models that can accurately predict these intelligibility scores based solely on the improved SPIN without access to the clean reference speech signals (non-intrusive method).

Table 5 summarizes the performance of several speech intelligibility prediction methods employed in the CPC2 dataset, each utilizing different strategies for modeling speech in noisy environments. The methods vary in their reliance on clean reference speech signals (intrusive vs. non-intrusive), the complexity of their models (e.g., machine learning vs. traditional signal processing), and whether they account for binaural hearing. E011, which uses deep CNNs to capture speech features that link to intelligibility [31], has achieved good performances without needing the clean speech signals.

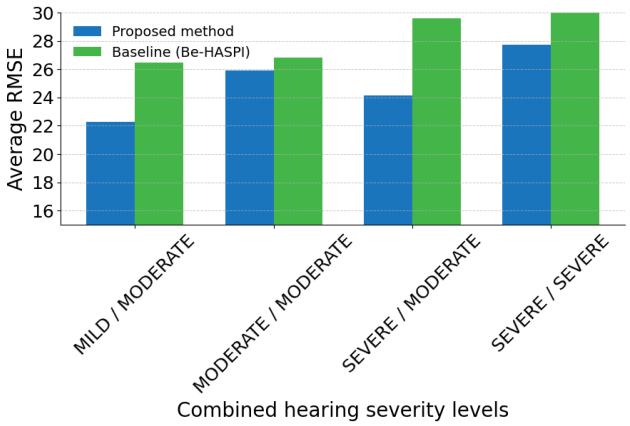


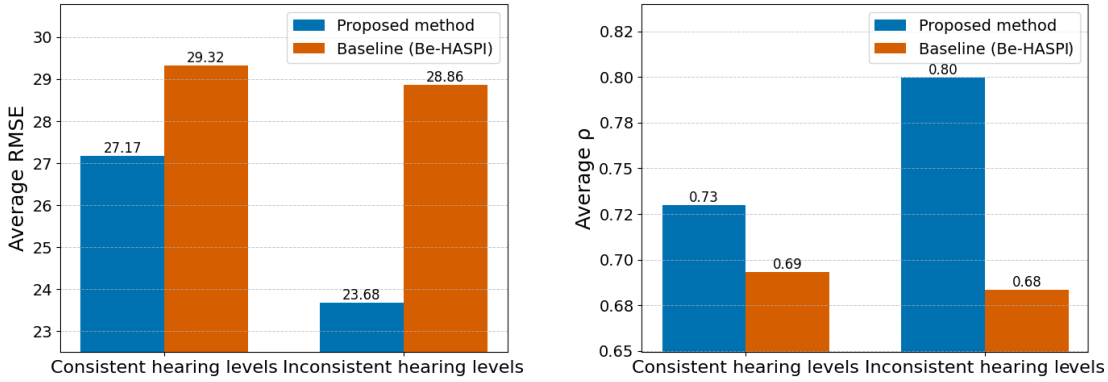
FIGURE 3. Average RMSE for different combined hearing severity levels using the proposed and baseline methods. Lower RMSE values indicate better prediction accuracy across various combinations of hearing loss severity.

Non-intrusive methods such as E002 and E023 rely on machine learning models and engineered features, including spectral and temporal dynamics, to predict speech intelligibility without needing a clean reference signal [51], [54]. In contrast, methods like E009 and E022, which are intrusive [52], [54], require access to the clean signal and use techniques like STOI or auditory models to mimic human hearing. While the be-HASPI method also accurately models auditory perception, it may be less practical in real scenarios where clean signals are unavailable.

As shown in Table 5, the proposed method achieves a ρ of 0.75 and an RMSE of 26.3. Unlike intrusive methods, which require a clean reference signal, our non-intrusive approach is more suitable for real-life applications. **When comparing our proposed method to the baseline method, be-HASPI, a significant improvement is evident.** The be-HASPI method achieves a ρ value of 0.67 and the RMSE of 28.7. In contrast, our non-intrusive method outperforms be-HASPI by a substantial margin in both correlation and error metrics. **Specifically, our method shows an improvement of 11.9% in the ρ and a reduction of 8.3% in the RMSE.**

A. PERFORMANCE ANALYSIS OF PROPOSED METHOD

In this experiment, 15 listeners with different types of binaural hearing loss were tested in the combined test set. The binaural hearing loss in these listeners ranged from mild to severe, and the significant difference in binaural hearing loss affected the accuracy of the speech intelligibility prediction. A comparative analysis of RMSE using different combinations of hearing loss when predicting



(a) Average RMSE for consistent vs. inconsistent hearing levels

(b) Average ρ for consistent vs. inconsistent hearing levels

FIGURE 4. Performance of proposed and baseline methods across consistent and inconsistent hearing levels. In (a), lower RMSE indicates better accuracy in the prediction of speech intelligibility scores, while in (b), higher ρ reflects stronger agreement between predicted and correct scores.

speech intelligibility is crucial. Different hearing loss types (e.g., combinations of mild and moderate hearing loss) affect speech processing differently. By analyzing RMSE for these combinations, the prediction accuracy of the proposed method can be evaluated for different hearing loss types.

1) IMPACT OF DIFFERENT HEARING LOSS COMBINATIONS

As shown in Fig. 3, although the CPC2 dataset contains a variety of scenarios covering SNRs ranging from −12 dB to 6 dB, the lower SNR implies that the speech signal is masked by the noise to a greater extent, making speech intelligibility prediction more difficult. In addition, the type of hearing loss of an individual further increases the complexity of prediction, especially in noisy environments.

The proposed method exhibits different levels of RMSE on different combinations of hearing loss types. Fig. 3 shows the average RMSE for different combinations of hearing severity levels, comparing the proposed method with the baseline (Be-HAPSI). As we can see, the proposed method shows a significant improvement over the Be-HAPSI for cases with differing hearing loss types between the ears. For example, in the MILD/MODERATE combination, the proposed method achieves relatively low RMSE values, with an RMSE reduction of approximately 15.9% compared to the Be-HAPSI, indicating that the proposed method has high prediction accuracy with this type of hearing loss. In contrast, as the hearing loss increases (e.g., SEVERE/SEVERE), the RMSE value increases significantly, indicating that the model has a higher prediction error when dealing with complex hearing loss combinations.

Moreover, in both MILD/MODERATE and SEVERE/MODERATE combinations, the proposed method's RMSE values remain relatively low, reducing prediction errors associated with hearing loss. This indicates that integrating binaural information plays a crucial role, particularly in asymmetric hearing loss scenarios, where different degrees of hearing loss between the ears require precise binaural processing. As a result, the proposed method leverages binaural cues to enhance the speech intelligibility prediction performance.

2) CONSISTENT VS. INCONSISTENT HEARING LEVELS

Based on the results of the proposed method and Be-HASPI, we compared the speech intelligibility prediction performance of listeners at consistent and inconsistent hearing levels (see Fig. 4). In Fig. 4(a), listeners at inconsistent listening levels show lower RMSE values (23.68) compared to listeners at consistent listening levels (27.17) using the proposed method, which indicates that listeners at inconsistent listening levels are more accurate in speech intelligibility prediction. The proposed method is thus better able to handle different levels of hearing loss.

Meanwhile, using the proposed method, the ρ values in Fig. 4(b) further validate this conclusion. The ρ of the listeners with inconsistent hearing levels is as high as 0.80, while that of the listeners with consistent hearing levels is only 0.73. This indicates a better match between the predicted scores of the model and the correct scores at inconsistent hearing levels. This result suggests that the proposed method can accurately reflect the speech comprehension ability of listeners when dealing with complex hearing loss combinations. In particular, using binaural information plays an important role in improving the accuracy of speech intelligibility prediction when there are significant differences in binaural hearing loss.

Compared to the Be-HASPI, the proposed method significantly improves the accuracy of speech intelligibility prediction under complex hearing loss conditions by integrating binaural cues. The prediction performance of listeners with inconsistent hearing levels was significantly

improved compared to those with consistent hearing levels, as evidenced by a reduction in RMSE values of about 12.8% and an improvement in ρ of about 9.6%. This suggests that the integrated processing of binaural information is particularly effective in the case of asymmetric hearing loss, enhancing the adaptability and robustness of the model.

B. DISCUSSION

Our proposed method shows a strong performance in predicting speech intelligibility without requiring access to clean speech signals. We capture linear and non-linear relationships in the data by using models like LightGBM, which handles complex datasets with high-dimensional features, and LSTM, which captures temporal dependencies in speech signals.

In analyzing the impact of different hearing loss combinations, the proposed method shows varying levels of prediction accuracy. As discussed earlier, our method exhibits lower RMSE for listeners with inconsistent hearing levels, indicating that the model is better at predicting speech intelligibility. However, as the severity of hearing loss increases, particularly in SEVERE/SEVERE combinations, the RMSE increases, indicating greater prediction challenges.

Moreover, while this performance is competitive among non-intrusive methods, it is slightly below the top-ranking methods such as E011 and E009, which achieve lower RMSE scores of 25.1 and 25.4, respectively. This indicates that there is still room for improvement in the prediction accuracy of the proposed method, particularly in comparison to the top-ranking methods.

VI. ANALYSIS II: IMPACT OF HEARING LOSS ON AUDIO SIGNAL CHARACTERISTICS

A. ANALYSIS OF HEARING LOSS GROUPS IN CPC2

Our proposed method is valid for predicting speech intelligibility at different levels of hearing loss severity. To further explore how different levels of hearing loss affect the received speech signal and speech intelligibility, this section analyzes the effects of hearing loss in depth and simulates various hearing conditions. First, we present an overview of the hearing loss groups in the CPC2, highlighting the distribution of hearing loss severity across the hearing loss groups. The audiogram data are then examined in detail to understand how hearing thresholds change in different frequency bands and the impact of these changes on speech perception. Then, using the proposed method, we simulate different types of hearing loss and compare their effects on the same speech signals. Finally, the proposed method uses the EC model to take into account the union of binaural information and improve speech intelligibility for listeners with binaural hearing loss.

The dataset includes 31 listeners with varying degrees of hearing loss severity, as detailed in Table 6. The listener with the highest number of test instances (listener L0218) appeared 633 times, while most listeners appeared around 200–400 times. These counts reflect the degree of performance of the different listeners in the dataset.

TABLE 6. Listener counts and hearing loss severity for 31 listeners (left and right ear).

Listener ID	Count	Left ear severity	Right ear severity
L0200	626	SEVERE	SEVERE
L0201	641	SEVERE	SEVERE
L0202	220	MODERATE	MODERATE
L0206	231	SEVERE	SEVERE
L0208	229	MODERATE	MODERATE
L0209	635	SEVERE	SEVERE
L0212	627	SEVERE	SEVERE
L0215	226	SEVERE	SEVERE
L0216	237	SEVERE	SEVERE
L0217	226	MODERATE	SEVERE
L0218	633	SEVERE	MODERATE
L0219	629	SEVERE	MODERATE
L0220	234	MODERATE	MODERATE
L0221	645	MODERATE	MODERATE
L0222	628	SEVERE	SEVERE
L0224	228	SEVERE	SEVERE
L0225	238	SEVERE	MODERATE
L0227	228	SEVERE	MODERATE
L0229	238	MODERATE	MODERATE
L0231	232	SEVERE	MODERATE
L0235	227	MILD	MILD
L0236	232	SEVERE	SEVERE
L0239	238	SEVERE	SEVERE
L0240	630	MODERATE	MODERATE
L0241	243	SEVERE	SEVERE
L0242	643	SEVERE	SEVERE
L0243	639	SEVERE	MODERATE
L0249	398	MODERATE	MODERATE
L0250	400	MILD	MODERATE
L0252	399	SEVERE	SEVERE
L0254	363	SEVERE	SEVERE

The listeners had varying degrees of binaural hearing impairment. Specifically, 68% had severe hearing loss in the left ear, and 58% had severe hearing loss in the right ear. This means that 21 out of 31 listeners had severe hearing loss in at least one ear. The rest of the listeners were categorized as having moderate or mild hearing loss, with 29% having moderate hearing loss in the left ear and 35% in the right ear. Only a small percentage of listeners had mild hearing loss in both ears. For the older age group, this distribution of hearing loss is close to the real scenarios, i.e., severe hearing loss is more common.

B. ANALYSIS OF AUDIOGRAMS

In the CPC2 study, listener information is obtained through audiometric testing. Audiograms are used to measure an individual's response to sounds and speech across different frequencies. These tests include tones and speech sounds that assess sensitivity at specific frequencies (often represented by beeps), which help us understand what a person with hearing

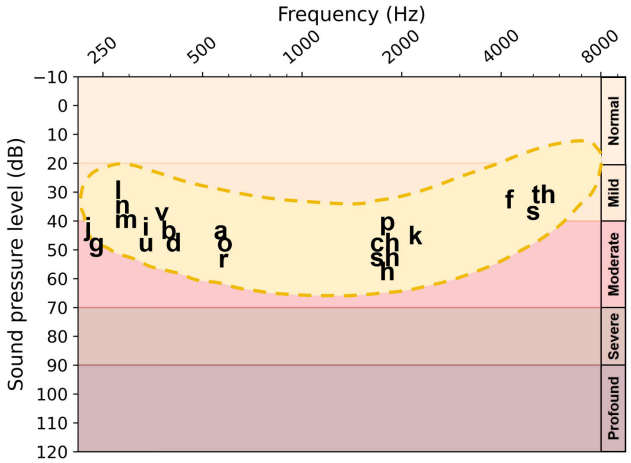


FIGURE 5. Adapted from the original chart by the American academy of audiology at audiology.org, this speech banana chart illustrates how various levels of hearing loss affect the perception of speech sounds across different frequency bands. The yellow "banana" shape represents the typical frequency and sound pressure range of speech phonemes audible to individuals with normal hearing, including common vowels and consonants. As hearing loss severity increases (from mild to profound), hearing thresholds in different frequency bands rise, making it difficult for individuals to perceive certain phonemes. For example, individuals with moderate hearing loss may struggle to hear high-frequency consonants even within the conversational sound pressure range (40–70 dB), leading to misunderstandings during conversations.

loss can or cannot hear. Figure 5 illustrates how hearing loss affects perception across various frequencies.

The yellow "banana" shape on the audiogram represents the typical frequency and loudness range (20–70 dB) of human speech sounds, including vowels and consonants. This "speech banana" highlights the critical frequencies necessary for everyday speech comprehension, indicating which sounds are likely audible or challenging for individuals with hearing loss, depending on their specific hearing threshold curve. This graph helps us visualize which sounds may be missed by someone with hearing loss, depending on their hearing threshold across different frequencies. It also emphasizes that speech sounds vary in pitch and loudness, contributing to an understanding of how hearing loss affects communication.

For individuals with hearing loss, their audiogram curve shifts downward at different frequencies, meaning that louder sound levels are required to detect the same sounds. As hearing loss progresses-from mild to profound-the threshold curve begins to cover more of the "banana" area, indicating that fewer speech sounds remain audible. This reduction in accessible speech sounds impacts overall speech comprehension. The "speech banana" effectively demonstrates the impact of hearing loss by overlaying each consonant and vowel in terms of pitch and relative loudness. The range of human speech typically forms a banana-like shape across the upper third of the audiogram, where the essential frequencies for speech clarity are found.

The following sections provide further analysis of how different frequency ranges contribute to speech comprehension:

1) LOW FREQUENCIES (500 Hz)

This range includes low-pitched sounds like vowels and environmental noises. People with mild hearing loss generally retain sensitivity to these low-frequency sounds, which contribute significantly to speech volume. However, as hearing loss worsens, these sounds may become difficult to hear.

2) MID FREQUENCIES (500 Hz TO 2000 Hz)

Mid frequencies are crucial for understanding speech, as they contain many vowel and consonant sounds. Individuals with moderate hearing loss often struggle with these sounds, even when volume is increased. Loss in this range makes it challenging to understand speech clearly, especially in conversations.

3) HIGH FREQUENCIES (2000 Hz TO 8000 Hz)

High-frequency phonemes like "s," "f," and "th" are essential for speech clarity and word distinction. These are often the first frequencies affected by hearing loss, particularly age-related or noise-induced loss. The figure shows that the ability to perceive these high frequencies declines rapidly as hearing loss progresses. In cases of severe hearing loss, much higher volumes are needed to hear these sounds, if they can be heard at all. This explains common difficulties distinguishing words like "sat," "fat," and "that."

Age-related hearing loss, or presbycusis, often results in more significant loss at high frequencies than low frequencies. Audiogram data reflect this trend, showing a steep decline in high-frequency sensitivity for older adults. This loss impacts speech clarity because many critical speech sounds, particularly consonants, fall in this range. Reduced perception of these sounds leads to common misunderstandings in communication, particularly in noisy environments or when the listener is far from the speaker. Even mild hearing loss can affect one's ability to understand speech, as shown in the audiogram's "speech banana" area. This effect is especially noticeable when background noise or the listener is distant from the speaker. Noise and distance amplify the difficulty of picking up subtle details in speech, leaving individuals wondering why they can "hear" but not fully "understand."

C. PROPOSED METHOD: SIMULATING DIFFERENT TYPES OF HEARING LOSS

In this section, we simulate the same speech signal with different types of hearing loss to better understand how these impairments alter auditory perception. The input signal is the sentence "The following comment is from one member of this board," as shown in Fig. 6, with the clean reference signal serving as a baseline.

Simulations are conducted for normal, mild, moderate, and severe hearing loss by adjusting the auditory filter characteristics, such as the ERB and channel count, to reflect

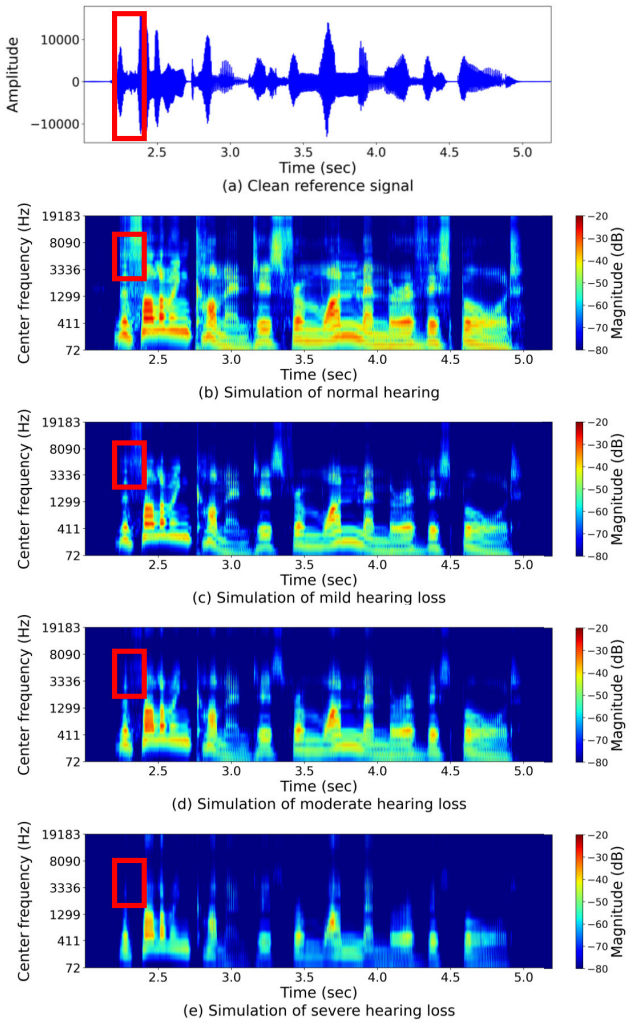


FIGURE 6. Spectrogram comparison of simulated hearing loss at various severity levels. (a) shows the time-domain waveform of the clean reference signal, while (b) to (e) display GTFB spectrograms under different simulated hearing loss conditions. The y-axis represents the center frequencies of the auditory filters (Hz), covering frequency components from low to high. For different levels of hearing loss, the auditory filter bandwidths are broadened to simulate the reduction in frequency resolution due to hearing impairment (e.g., mild hearing loss by 1.5 ERB). The red boxes highlight the frequency regions where high-frequency phonemes such as "th," "e," and "f" are located, demonstrating the challenges faced by individuals with hearing loss in distinguishing these high-frequency sounds. As the severity of simulated hearing loss increases from (b) to (e), the reduction in frequency selectivity leads to noticeable attenuation and spectral blurring, especially in cases of moderate and severe hearing loss.

decreased frequency resolution. As the severity of hearing loss increases, the filter bandwidth broadens and the number of channels reduces, simulating the effect of inner ear damage that diminishes frequency selectivity. For example, normal hearing corresponds to 1 ERB with more channels, mild hearing loss to approximately 1.5 ERB, moderate to 2.0 ERB, and severe to 3.0 ERB with progressively fewer channels. This broadening and reduction in channels leads to the ''blurring'' of frequency components in the spectrogram, making it harder to distinguish specific speech sounds. By comparing spectrograms across these simulated conditions, we can observe how different levels of hearing impairment alter the frequency content and intensity of the signal, thereby illustrating the impact on speech intelligibility.

The sentence ''The following comment is from one member of this board'' contains a variety of consonant and vowel phonemes encompassing a broad range of speech sounds, making it an excellent example for assessing the impact of hearing loss. Consonants such as ''f'' and ''s'' are particularly susceptible to the effects of hearing loss, while vowels like ''i'' and ''a'' are less affected. This combination of sounds allows us to observe how hearing loss affects the perception of different speech components. Fig. 6(a) illustrates the clean signal, with speech sounds represented across the frequency channels, demonstrating optimal speech intelligibility.

In the simulated scenarios, different types of hearing loss exhibit distinct impacts on speech intelligibility, as seen in the GTFB spectrograms (Fig. 6). For normal hearing (Fig. 6(b)), minimal distortion is observed, with most spectral components preserved and speech clarity largely unaffected. As the severity increases to mild hearing loss (Fig. 6(c)), there is noticeable attenuation in certain frequency channels, making some speech sounds harder to distinguish despite the overall intelligibility of speech. In the case of moderate hearing loss (Fig. 6(d)), the energy levels across these channels are substantially weakened, reducing speech clarity and requiring listeners to rely more on contextual cues to understand sentences. Finally, in severe hearing loss (Fig. 6(e)), the energy levels across many channels are significantly reduced, leading to a blurred spectrogram and making speech intelligibility nearly impossible to guarantee.

As hearing loss increases, the effects extend beyond the attenuation of specific speech sounds due to the following factors:

1) INCREASED HEARING THRESHOLDS

Due to damage to the cochlea or auditory nerve, signals must reach a higher loudness to be perceived. This is reflected in the simulation as a decrease in overall signal energy in the spectrogram. The attenuation is evident in Figs. 6(c) and 6(d), where the energy levels across the frequency channels are reduced.

2) LOUDNESS RECRUITMENT

Hearing loss can lead to loudness recruitment, where softer sounds are difficult to perceive, but sudden loud sounds appear overly loud. This phenomenon is challenging to fully visualize on a spectrogram but can be reflected by sudden increases in energy in certain channels. Although this is less apparent in mild hearing loss, some sounds may appear abnormally loud to the listener in the moderate and severe losses shown in Figs. 6(d) and 6(e).

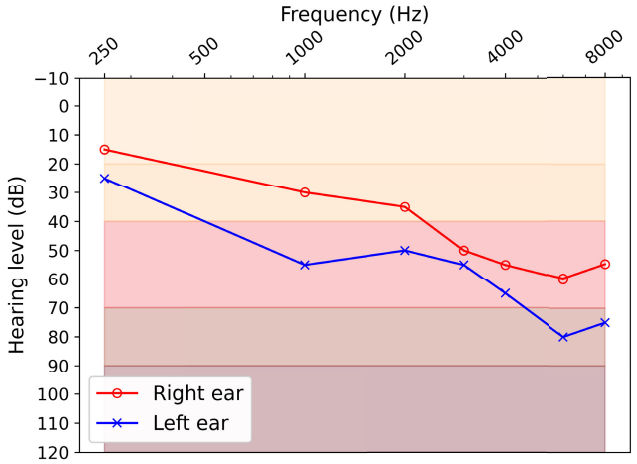


FIGURE 7. Audiogram showing hearing thresholds for listener L0219 in each ear, with right (red) and left (blue) ears plotted across frequency bands. The shaded areas represent different levels of hearing loss severity: normal (beige), mild (light orange), moderate (pink), severe (brown), and profound (dark brown). This visual aids in understanding the degree of hearing loss at various frequencies for each ear.

3) REDUCED FREQUENCY SELECTIVITY

Hearing loss diminishes the listener's ability to discriminate between different spectral components, especially in noisy environments where distinguishing adjacent sounds becomes difficult. In the context of GTFB spectrograms, the y-axis represents distinct frequency channels, each corresponding to the center frequency of a specific gammatone filter that models the human auditory system's frequency selectivity. This simulation demonstrates that with increasing severity of hearing loss, there is a significant reduction in energy levels across these frequency channels, particularly affecting higher frequencies. In Fig. 6(e) (severe hearing loss), the spectrogram shows a marked decrease in color intensity across multiple channels, indicating a loss of energy in those frequency bands. This reduction in energy makes it challenging to clearly distinguish speech sounds, thereby directly impacting speech intelligibility.

D. PROPOSED METHOD: COMBINING BINAURAL INFORMATION

In the analyses in Figs. 3 and 4, we noticed that the prediction error (RMSE) varied under different combinations of hearing loss. The proposed method better predicted the results, especially in the case of asymmetric right and left ear hearing loss. Taking this finding, we further validate the effectiveness of the proposed method in dealing with different combinations of hearing loss by simulating the binaural hearing loss of listener L0219 and observing the integration effect of the EC model.

1) AUDIOGRAM ANALYSIS OF LISTENER L0219

From Fig. 7, we can see that listener L0219 has severe hearing loss in the left ear, while the loss in the right ear is relatively mild. In particular, the hearing in the left ear shows a moderate loss from the low frequency (250 Hz) and becomes severe in the middle and high frequency bands (above 2000 Hz). At the same time, the loss in the high-frequency region has a greater impact on the discrimination of consonants, especially for consonants such as "s" and "th", which may not be fully distinguished by the listener. In addition, the reduced frequency selectivity also makes it difficult to distinguish sounds of similar frequencies, which reduces speech intelligibility.

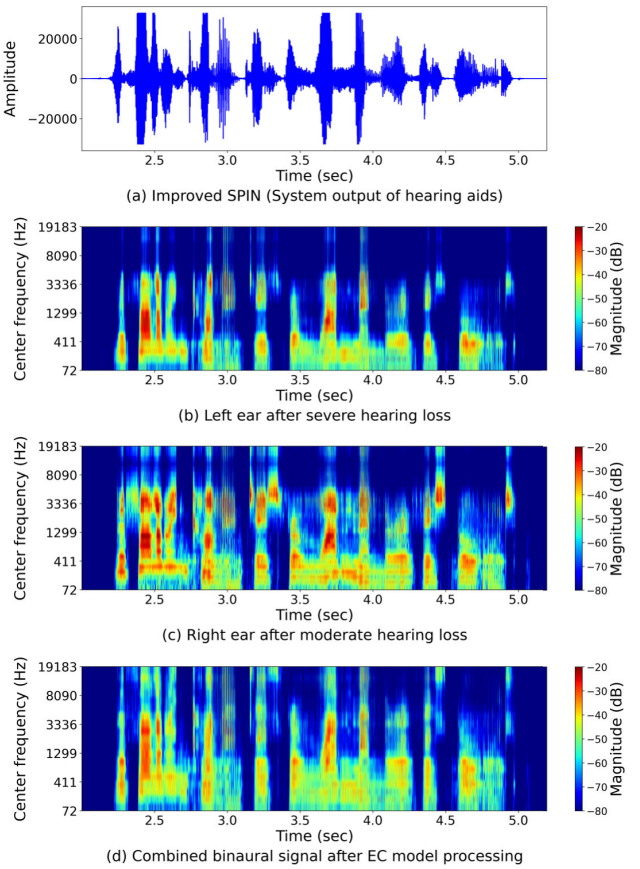


FIGURE 8. Spectrogram comparison after hearing loss simulation and EC model processing for listener L0219. (a) shows the speech signal processed by the hearing aid enhancement module, containing some residual noise. (b), (c), and (d) simulate listener L0219 under different hearing loss conditions. Specifically, (b) and (c) display GTFB spectrograms for the left and right ears with severe and moderate hearing loss, respectively, where the y-axis represents the center frequencies of the auditory filters affected by hearing loss. (d) combines signals from (b) and (c) using the EC model, which leverages binaural cues to improve intelligibility. The EC model's advantage in utilizing binaural information is that it preserves the high-frequency details essential for speech intelligibility.

2) SIGNAL SIMULATION AND ANALYSIS OF EC MODEL

After completing the audiogram analysis for listener L0219 (discussed above), we here simulate the tests conducted for this listener using the proposed method. Before simulating the hearing loss, the input signal was processed through an enhancement module, resulting in the improved SPIN shown in Fig. 8(a). Although the enhancement module reduces some

TABLE 7. Overview of noise types and details of CEC1 and CEC2 datasets.

Dataset	Total instances	Noise types	Details
CEC1	6297	Noise	Includes household appliance noise like dishwashers, fans, hairdryers, kettles, microwaves, vacuums, and washing machines. Each noise type is uniquely labeled by its category and ID.
CEC2	5944	Noise	Similar appliance noise types as in CEC1, with the same labeling convention.
		Speech	Includes speech interference labeled by accent and speaker ID (e.g., "mif_02484" for a Midlands-accented female speaker).
		Music	Includes music tracks from the MTG Jamendo database [56], labeled by unique track IDs.

TABLE 8. Comparison of the proposed model's performance across the CEC1 and CEC2 datasets, highlighting how varying interference complexities influence prediction accuracy in speech intelligibility tasks. CEC1 contains simpler, appliance-generated noise sources, while CEC2 incorporates a broader range of real-world interference, including music, speech, and noise, making it a more challenging dataset. The combination of CEC1+CEC2 models highlights the importance of training in different settings.

Dataset	$\rho \uparrow$	RMSE $\downarrow$
CEC1+CEC2	0.70	28.46
CEC1	0.32	42.74
CEC2	0.70	28.18

noise, the signal still contains various types of residual noise. The MSBG hearing loss model was then applied to simulate both the left and right ears separately, resulting in the signals shown in Figs. 8(b) and 8(c). Subsequently, the proposed method employed the EC model to process and integrate the signals from both ears, as shown in Fig. 8(d).

As illustrated in Fig. 8, by simulating the left and right ear signals and integrating them using the EC model, we can observe the following:

Figure 8(b) shows the left ear signal after the simulation of severe hearing loss. In this GTFB spectrogram, significant attenuation is observed across multiple frequency channels. The decreased energy levels indicate that auditory cues are weakened or lost, especially in channels corresponding to higher frequencies.

Figure 8(c) depicts the simulated right ear signal, reflecting moderate hearing loss. Compared to the left ear in Fig. 8(b), the right ear retains more auditory information. However, the spectrogram exhibits attenuation across several frequency channels, making it challenging to discern certain speech components. The energy levels in these frequency channels are reduced, affecting the perception of speech sounds.

Figure 8(d) displays the result of applying the EC model to integrate the binaural information from both ears. Unlike the monaural simulations in Figs. 8(b) and 8(c), which show notable attenuation and loss of spectral detail in certain channels due to hearing loss, the EC model combines residual auditory information from each ear, effectively compensating for inconsistent hearing loss. The EC model enhances specific channels where auditory cues were weakened or lost in individual ear simulations by optimizing interaural time delays and amplitude differences. As a result, some attenuated channels regain partial energy, as observed in the figure, demonstrating the model's ability to restore essential speech information. This selective restoration does not uniformly recover all frequency bands. Instead, it enhances regions where residual binaural cues allow improved perception.

The effectiveness of this binaural integration was further validated through subjective testing. The results showed that the listener achieved a perfect intelligibility score of 100 in this scenario, even though L0219 has severe hearing loss in both ears. Despite the significant hearing impairments, she was able to accurately hear and comprehend the speech content. In the prediction generated by the proposed model, a score of 85.59 was achieved. This demonstrates that the EC model effectively compensates for the impact of hearing loss, especially when the hearing loss is inconsistent between the two ears, enabling the extraction and processing of critical speech information.

VII. ANALYSIS III: INVESTIGATING THE IMPACT OF DIFFERENT INTERFERENCE SOURCES

This section evaluates the impact of different interference sources in the training scenarios on the speech intelligibility prediction results of hearing aids, providing insights for selecting more suitable training data in future work. Specifically, we compare the CEC1 and CEC2 datasets from the Clarity Challenge, which both consist of speech signals with added noise but with different interference conditions representing different auditory challenges. By evaluating the proposed method on these two datasets, we can observe its ability to handle different levels of complexity.

In analyzing the CEC1 and CEC2 datasets, we focus on the effects of single and multiple interference sources on the speech intelligibility model. As shown in Table 7, the CEC1 dataset has single interference sources, mainly noise generated by various household appliances, providing a relatively controlled environment to evaluate baseline performance. The CEC2 dataset is more complex and contains multiple interference sources, such as noise, speech, and music, modeling a complex environment that is more

closely aligned with a real scenario. Such scenarios with multiple interference sources are typically more challenging.

By comparing the performance of the CEC1 and CEC2 datasets in the speech intelligibility prediction task, we find that the model on the CEC2 dataset performs better in terms of both relevance and error of prediction. The experiments used a combined test set of 897 signals provided by the CPC2 challenge to simulate unknown predictions for real scenarios. We explored the impact of the dataset using a more straightforward speech intelligibility model that includes an input layer, LSTM layer, and attention layer designed to capture temporal features and significant parts of the speech signal and use them to predict intelligibility scores.

According to Table 8, the results of the model trained solely on the CEC1 dataset and the model trained solely on the CEC2 dataset show that the CEC2 dataset reduced the RMSE by approximately 34.07%. The performance of the model with the same parameters on different datasets indicates that training with CEC1 alone did not significantly improve prediction performance on the unknown test set. In contrast, the model trained on a combination of CEC1 and CEC2 better handled mixed interference conditions. This demonstrates that diverse training in noisy environments helps to improve the model's ability to handle multiple interference sources in reality.

To determine whether the inclusion of the CEC1 and CEC2 datasets versus the CEC2 dataset alone yields statistically significant performance improvements, we conducted paired t-tests and non-parametric Mann-Whitney U tests on the error metrics across these datasets. Results from both tests ($p > 0.05$) indicate that the differences in prediction error between the two training scenarios are not statistically significant. This finding suggests that although training with the combined CEC1 and CEC2 datasets reduces RMSE, the reduction is not significant enough to bring about a robust performance gain. Thus, while the CEC1 dataset may add variety in interference types, its simpler noise structure may limit its contribution to predictive accuracy in real-world, complex noise environments, where the CEC2 dataset alone is effective. Future efforts might focus on further diversifying complex interference conditions to enhance model robustness in multi-noise scenarios.

VIII. CONCLUSION AND FUTURE WORK

In this study, we proposed a non-intrusive speech intelligibility prediction method that integrates binaural processing for hearing loss. Hearing loss is modeled by simulating the multi-stage process of how the outer ear, middle ear, inner ear, and binaural cues process information. This is combined with LSTM and LightGBM models. Our method demonstrates a strong ability to capture key features of speech signals, especially in noisy environments and for different types of hearing loss.

- **Speech intelligibility prediction using binaural processing and hearing loss simulation:**

We confirmed that hearing loss affects key features of the speech signal, particularly the attenuation of high-frequency components, shown in spectrograms. Simulating hearing loss through the MSBG model includes raised hearing thresholds, loudness reconstruction, and reduced frequency selectivity. We found that these effects result in the attenuation or blurring of high-frequency consonants in speech, making it difficult for listeners to distinguish between similar sounds. This loss of high-frequency information directly affects the accuracy of speech intelligibility predictions. With the EC model for binaural processing, we mitigate these effects, allowing the model to focus on the better-hearing ear and reduce masking effects, thereby improving speech intelligibility in noisy environments.

- **Improved prediction accuracy and robustness:**

Compared to the baseline method (be-HASPI), our model demonstrated an 8.3% reduction in RMSE, indicating a meaningful improvement in prediction accuracy. Our method demonstrated a substantial improvement for individuals with asymmetric hearing loss, with RMSE reduced by 12.8% compared to symmetric hearing loss cases. This improvement is largely due to the integration of binaural processing, which allows the model to focus attention on the ear with better hearing, compensating for the weaker ear and enhancing overall speech intelligibility. This capability makes our approach more effective in real scenarios where hearing loss often differs between ears, providing a more personalized prediction for users with uneven auditory profiles.

However, despite these strengths, the proposed method still has areas for improvement. When predicting speech intelligibility for previously unseen listeners in noisy environments, our model's performance, though better than be-HASPI, lags by approximately 4.6% in accuracy compared to the leading the state-of-the-art method (E011) [31].

Future work will focus on refining our model to bridge this gap by improving the robustness of feature extraction in noisy environments and extending the model's adaptability to handle even more complex auditory scenarios. In addition, we will incorporate more diverse datasets, especially listener datasets with higher subjective speech intelligibility. Such datasets help to narrow the performance gap with intrusive models and make non-intrusive models more reliable and universally applicable.

REFERENCES

[1] *World Report on Hearing*, World Health Org., Geneva, Switzerland, 2021.
[2] D. G. Loughrey, M. E. Kelly, G. A. Kelley, S. Brennan, and B. A. Lawlor, "Association of age-related hearing loss with cognitive function, cognitive impairment, and dementia: A systematic review and meta-analysis," *JAMA Otolaryngol.-Head Neck Surg.*, vol. 144, no. 2, pp. 115–126, 2018.
[3] A. Ciorba, C. Bianchini, S. Pelucchi, and A. Pastore, "The impact of hearing loss on the quality of life of elderly adults," *Clin. Interventions Aging*, vol. 2012, pp. 159–163, Jun. 2012.
[4] A. Shukla, M. Harper, E. Pedersen, A. Goman, J. J. Suen, C. Price, J. Applebaum, M. Hoyer, F. R. Lin, and N. S. Reed, "Hearing loss, loneliness, and social isolation: A systematic review," *Otolaryngol.-Head Neck Surg.*, vol. 162, no. 5, pp. 622–633, May 2020.

[5] F. R. Lin, R. Thorpe, S. Gordon-Salant, and L. Ferrucci, "Hearing loss prevalence and risk factors among older adults in the United States," *J. Gerontol. A, Biol. Sci. Med. Sci.*, vol. 66A, no. 5, pp. 582–590, May 2011.

[6] M. A. Ferguson, P. T. Kitterick, L. Y. Chong, M. Edmondson-Jones, F. Barker, and D. J. Hoare, "Hearing aids for mild to moderate hearing loss in adults," *Cochrane Database Systematic Rev.*, vol. 2017, no. 9, Sep. 2017.

[7] X. Wu, Y. Ren, Q. Wang, B. Li, H. Wu, Z. Huang, and X. Wang, "Factors associated with the efficiency of hearing aids for patients with age-related hearing loss," *Clin. Interventions Aging*, vol. 14, pp. 485–492, Feb. 2019.

[8] R. M. Cox, J. A. Johnson, and J. Xu, "Impact of advanced hearing aid technology on speech understanding for older listeners with mild to moderate, adult-onset, sensorineural hearing loss," *Gerontology*, vol. 60, no. 6, pp. 557–568, 2014.

[9] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Predicting the intelligibility of noisy and nonlinearly processed binaural speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 1908–1920, Nov. 2016.

[10] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Nonintrusive speech intelligibility prediction using convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1925–1939, Oct. 2018.

[11] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1648–1661, Mar. 1992.

[12] B. Razavi, W. E. O'Neill, and G. D. Paige, "Both interaural and spectral cues impact sound localization in azimuth," in *Proc. 2nd Int. IEEE EMBS Conf. Neural Eng.*, Jun. 2005, pp. 587–590.

[13] T. Rodemann, G. Ince, F. Joublin, and C. Goerick, "Using binaural and spectral cues for azimuth and elevation localization," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2008, pp. 2185–2190.

[14] D. M. Markle and W. Aber, "A clinical evaluation of monaural and binaural hearing aids," *AMA Arch. Otolaryngol.*, vol. 67, no. 5, pp. 606–608, 1958.

[15] P. Derleth, E. Georganti, M. Latzel, G. Courtois, M. Hofbauer, J. Raether, and V. Kuehnel, "Binaural signal processing in hearing aids," *Seminars Hearing*, vol. 42, no. 3, pp. 206–223, Aug. 2021.

[16] A. W. Bronkhorst and R. Plomp, "Binaural speech intelligibility in noise for hearing-impaired listeners," *J. Acoust. Soc. Amer.*, vol. 86, no. 4, pp. 1374–1383, Oct. 1989.

[17] T. Irino and R. D. Patterson, "A dynamic compressive gammachirp auditory filterbank," *IEEE Trans. Audio, Speech Language Process.*, vol. 14, no. 6, pp. 2222–2232, Nov. 2006.

[18] R. Wan, N. I. Durlach, and H. S. Colburn, "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," *J. Acoust. Soc. Amer.*, vol. 128, no. 6, pp. 3678–3690, Dec. 2010.

[19] J. Barker, M. A. Akeroyd, W. Bailey, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, and G. Naylor, "The 2nd clarity prediction challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 11551–11555.

[20] K. D. Kryter, "Methods for the calculation and use of the articulation index," *J. Acoust. Soc. Amer.*, vol. 34, no. 11, pp. 1689–1697, Nov. 1962.

[21] B. W. Y. Hornsby, "The speech intelligibility index: What is it and what's it good for?" *Hearing J.*, vol. 57, no. 10, pp. 10–17, Oct. 2004.

[22] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Commun.*, vol. 65, pp. 75–93, Nov. 2014.

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

[24] 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 Process. Mag.*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[25] C. A. da Silva Andrade, M. R. F. de Souza, and M. C. M. Iorio, "Speech recognition and speech intelligibility index in intra-aural hearing aids users: A comparative study," *Audiol.-Commun. Res.*, vol. 25, p. 2362, Jun. 2021.

[26] U. Ariöz and B. Günel, "Evaluation of hearing loss simulation using a speech intelligibility index," *TURKISH J. Electr. Eng. Comput. Sci.*, vol. 24, pp. 4193–4207, Jan. 2016.

[27] N. Ellaham, "Binaural speech intelligibility prediction and nonlinear hearing devices," Ph.D. dissertation, School Elect. Eng. Comput. Sci. (EECS), Univ. Ottawa, Ottawa, ON, Canada, 2014.

[28] J. Kates and K. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Amer.*, vol. 115, no. 5, p. 2604, May 2004.

[29] J. M. Kates and K. H. Arehart, "An overview of the HASPI and HASQI metrics for predicting speech intelligibility and speech quality for normal hearing, hearing loss, and hearing aids," *Hearing Res.*, vol. 426, Dec. 2022, Art. no. 108608.

[30] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI) version 2," *Speech Commun.*, vol. 131, pp. 35–46, Jul. 2021.

[31] S. Cuervo and R. Marxer, "Temporal-hierarchical features from noise-robust speech foundation models for non-intrusive intelligibility prediction," in *Proc. ISCA*, 2023.

[32] R. M. Cox, G. C. Alexander, and I. M. Rivera, "Comparison of objective and subjective measures of speech intelligibility in elderly hearing-impaired listeners," *J. Speech, Lang., Hearing Res.*, vol. 34, no. 4, pp. 904–915, Aug. 1991.

[33] B. C. J. Moore, "Effects of hearing loss and age on the binaural processing of temporal envelope and temporal fine structure information," *Hearing Res.*, vol. 402, Mar. 2021, Art. no. 107991.

[34] A. Schmidt-Nielsen, "Intelligibility and acceptability testing for speech technology," Nav. Res. Lab., Washington, DC, USA, Tech. Rep. ADA252015, May 1992.

[35] J. Roßbach, S. Röttges, C. F. Hauth, T. Brand, and B. T. Meyer, "Non-intrusive binaural prediction of speech intelligibility based on phoneme classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 396–400.

[36] I. Holube and B. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *J. Acoust. Soc. Amer.*, vol. 100, no. 3, pp. 1703–1716, Sep. 1996.

[37] J. G. W. Bernstein, V. Summers, E. Grassi, and K. W. Grant, "Auditory models of suprathreshold distortion and speech intelligibility in persons with impaired hearing," *J. Amer. Acad. Audiol.*, vol. 24, no. 4, pp. 307–328, Apr. 2013.

[38] J. Zaar and L. H. Carney, "Predicting speech intelligibility in hearing-impaired listeners using a physiologically inspired auditory model," *Hearing Res.*, vol. 426, Dec. 2022, Art. no. 108553.

[39] M. S. A. Zilany and I. C. Bruce, "Predictions of speech intelligibility with a model of the normal and impaired auditory-periphery," in *Proc. 3rd Int. IEEE/EMBS Conf. Neural Eng.*, May 2007, pp. 481–485.

[40] Y. Nejime and B. C. J. Moore, "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. Amer.*, vol. 102, no. 1, pp. 603–615, Jul. 1997.

[41] B. C. J. Moore and B. R. Glasberg, "A revised model of loudness perception applied to cochlear hearing loss," *Hearing Res.*, vol. 188, nos. 1–2, pp. 70–88, Feb. 2004.

[42] B. C. J. Moore and B. R. Glasberg, "A model of loudness perception applied to cochlear hearing loss," *Auditory Neurosci.*, vol. 3, no. 3, pp. 289–311, 1997.

[43] R. Wan, N. I. Durlach, and H. S. Colburn, "Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers," *J. Acoust. Soc. Amer.*, vol. 136, no. 2, pp. 768–776, Aug. 2014.

[44] 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, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

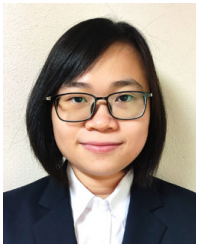
[45] R. Mogridge, G. Close, R. Sutherland, T. Hain, J. Barker, S. Goetze, and A. Ragni, "Non-intrusive speech intelligibility prediction for hearing-impaired users using intermediate ASR features and human memory models," 2024, *arXiv:2401.13611*.

[46] C. O. Mawalim, B. A. Titalim, S. Okada, and M. Unoki, "Non-intrusive speech intelligibility prediction using an auditory periphery model with hearing loss," *Appl. Acoust.*, vol. 214, Nov. 2023, Art. no. 109663.

[47] K. L. Kadi, S. A. Selouani, B. Boudraa, and M. Boudraa, "Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge," *Biocybernetics Biomed. Eng.*, vol. 36, no. 1, pp. 233–247, 2016.

[48] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Dec. 2017, p. 52.
[49] 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 *Proc. Interspeech*, Aug. 2021, pp. 686–690.
[50] M. A. Akeroyd, W. Bailey, J. Barker, T. J. Cox, J. F. Culling, S. Graetzer, G. Naylor, Z. Podwinska, and Z. Tu, "The 2nd clarity enhancement challenge for hearing aid speech intelligibility enhancement: Overview and outcomes," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[51] R. Mogridge, G. Close, R. Sutherland, S. Goetze, and A. Ragni, "Pre-trained intermediate asr features and human memory simulation for non-intrusive speech intelligibility prediction in the clarity prediction challenge 2," *Evaluation*, vol. 1, no. 6, pp. 6–21, 2023.
[52] M. Huckvale and G. Hilkhuysen, "Combining acoustic, phonetic, linguistic and audiometric data in an intrusive intelligibility metric for hearing-impaired listeners," in *Proc. ISCA*, 2023.
[53] Z. Tu, N. Ma, and J. Barker, "Intelligibility prediction with a pretrained noise-robust automatic speech recognition model," 2023, *arXiv:2310.19817*.
[54] R. Zezario, "Deep learning-based speech intelligibility prediction model by incorporating whisper for hearing aids," in *Proc. ISCA*, 2023.
[55] B. C. J. Moore and B. R. Glasberg, "Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech," *J. Acoust. Soc. Amer.*, vol. 94, no. 4, pp. 2050–2062, Oct. 1993.
[56] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The MTG-jamendo dataset for automatic music tagging," in *Proc. ICML*, Jan. 2019.

XIAJIE ZHOU received the M.S. degree from Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, Japan, in 2024, where she is currently pursuing the Ph.D. degree with the School of Information Science. Her research interests include auditory models, speech intelligibility prediction, hearing aid signal processing, and the application of machine learning techniques to auditory perception and hearing loss rehabilitation.



CANDY OLIVIA MAWALIM (Member, IEEE) received the B.S. degree in computer science from Institut Teknologi Bandung (ITB), Bandung, Indonesia, in 2017, and the M.S. and Ph.D. degrees from the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), in 2019 and 2022, respectively. She was selected as a Research Fellow of Young Scientists DC1 (JSPS), from 2020 to 2022. Since April 2022, she has been working as an Assistant Professor with the School of Information Science, Research Center for Biological Function and Sensory Information, JAIST. Her research interests include speech signal processing, hearing perception, voice privacy preservation, and machine learning. She is also on the Education Team of the ISCA Special Interest Group of Security and Privacy in Speech Communication (SIG-SPSC) Committee.

MASASHI UNOKI (Member, IEEE) received the M.S. and Ph.D. degrees in information science from Japan Advanced Institute of Science and Technology (JAIST), in 1996 and 1999, respectively. He was a Research Fellow at Japan Society for the Promotion of Science (JSPS), from 1998 to 2001. He was a Visiting Researcher at ATR Human Information Processing Laboratories, from 1999 to 2000, and a Visiting Research Associate at the Centre for the Neural Basis of Hearing (CNBH), Department of Physiology, University of Cambridge, from 2000 to 2001. He has been working as a Faculty Member with the School of Information Science, JAIST, since 2001, where he is currently working as a Professor. His research interests include auditory-motivated signal processing and modeling of auditory systems. He is a member of the Research Institute of Signal Processing (RISP); the Institute of Electronics, Information, and Communication Engineers (IEICE) of Japan; and the Acoustical Society of America (ASA). He is also a member of the Acoustical Society of Japan (ASJ) and the International Speech Communication Association (ISCA). He received the Sato Prize for an Outstanding Paper from the ASJ, in 1999, 2010, and 2013; and the Yamashita Taro "Young Researcher" Prize from the Yamashita Taro Research Foundation, in 2005.