IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 June 2026, accepted 17 June 2026, date of publication 23 June 2026, date of current version 29 June 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3706536

RESEARCH ARTICLE

Speech Intelligibility Prediction Inspired by Processing Auditory Information for Hearing-Impaired Listeners

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 by Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 25H01139 and Grant 25K21245.

ABSTRACT Hearing aids are a common intervention for listeners with hearing loss and used across a wide range of hearing conditions, from mild to severe impairment. However, evaluating how well different hearing-aid outputs support speech understanding remains challenging across diverse acoustic conditions and listeners. Although certain current speech intelligibility prediction methods incorporate listener with hearing loss, most still focus primarily on auditory peripheral degradation and do not adequately represent higher level auditory information. To address this problem, this study proposes a speech intelligibility prediction method inspired by processing auditory information in the human auditory system under hearing-loss conditions. At the peripheral stage of the auditory system, the method simulates hearing-loss degradations, including elevated hearing thresholds, reduced nonlinear compression, and loss of frequency selectivity. At the central stage of the auditory system, it combines neural representation learning, binaural integration, and spectro-temporal modulation analysis to capture higher-level auditory information relevant to speech understanding. These auditory features are then integrated into a hybrid speech intelligibility prediction method that combines a Vision Transformer architecture with modulation similarity cues. Experiments on the second Clarity Prediction Challenge dataset show that the proposed method achieved competitive performance, ranking second overall and first among intrusive methods. Compared with the baseline method, it improved the Pearson correlation coefficient by 16.4% and reduced the root mean squared error by 12.2%. These results indicate that jointly modeling auditory peripheral and central stages provides an effective framework for predicting speech intelligibility in hearing-impaired listeners by capturing information relevant to auditory pathway processing.

INDEX TERMS Hearing aids, hearing loss, speech intelligibility, auditory system, spectro-temporal modulation, auditory pathway processing.

I. INTRODUCTION

Hearing loss affects more than 1.3 billion people worldwide, and nearly one-third of adults over 65 experience disabling impairment. Its most immediate consequence is reduced auditory sensitivity, but its impact extends beyond audibility alone. A large body of research has shown that hearing loss is associated with accelerated cognitive decline and an increased risk of dementia, indicating long-term consequences for cognitive health in older adults [1]. Hearing loss has also been linked to a higher likelihood of depressive symptoms, with particularly pronounced effects among individuals in lower socioeconomic groups [2]. Despite these broad consequences, hearing loss has long been insufficiently recognized in clinical and public health practice, leaving many individuals untreated and at risk of reduced quality of life.

Hearing aids are one of the most common and cost-effective interventions and can help reduce loneliness and support social participation [3]. As populations age,

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

the number of individuals with moderate to severe hearing loss will continue to increase, making effective intervention increasingly important [4], [5]. Despite hearing aids substantially improving communication ability and reducing disability burden, hearing-aid uptake remains low, with nearly 80% of adults with moderate-to-severe hearing loss worldwide not receiving appropriate fitting [4]. Studies also suggest that hearing-aid use may provide broader benefits beyond speech perception: first-time users have shown improvements in executive function, attention, and verbal learning, and longitudinal evidence indicates that aided listeners tend to exhibit a slower rate of cognitive decline than untreated individuals [6]. These findings suggest that hearing aids may influence not only audibility, but also cognition and communication. Understanding these effects requires examining the core signal processing mechanisms in modern hearing aids, including enhancement and intelligibility prediction.

Modern hearing aids typically rely on two principal signal processing modules: an enhancement module designed to improve audibility and the signal-to-noise ratio (SNR), and a prediction module intended to estimate the listener's expected speech intelligibility [7], [8]. Although enhancement strategies such as dynamic-range compression, noise reduction, and frequency-lowering can improve speech audibility [9], they may also distort temporal envelopes and modulation patterns, which can negatively affect intelligibility under some conditions [10]. To evaluate these effects, various objective prediction methods have been developed to estimate speech quality and intelligibility from processed speech signals.

These objective prediction methods can be broadly categorized according to whether the prediction is mainly based on acoustic-level signal information or auditory-model-based internal representations: acoustic methods and auditory methods. Acoustic methods, such as those derived from band-importance functions or the speech intelligibility index (SII), estimate performance by quantifying the effective SNR across frequency channels [11]. Although these methods can incorporate audiometric thresholds to approximate audibility loss, they operate primarily at the acoustic level and provide limited insight into the processing differences associated with hearing impairment.

In contrast, auditory methods have been proposed to more explicitly approximate auditory information. Representative examples include the hearing aid speech quality index (HASQI), hearing aid speech perception index (HASPI), and perception-model-based quality prediction method for hearing impairments (PEMO-Q-HI). These methods construct internal auditory representations using auditory filterbanks, outer hair cell (OHC) compression, envelope-modulation analysis, and neural adaptation mechanisms, enabling quantitative assessment of signal fidelity and speech intelligibility [10]. However, these methods still do not fully reflect auditory information across multiple stages of the human auditory system under hearing-loss conditions.

Other auditory methods further simulate the auditory peripheral stage and employ decision mechanisms to predict speech-reception thresholds [12]. Recent developments have also incorporated auditory-inspired modulation analysis. For example, the gammachirp envelope distortion index (GEDI) introduces an auditory filterbank with envelope and modulation-domain distortion analysis to better capture cues [13]. Building on this line of work, the gammachirp envelope similarity index (GESI) further incorporates listener-specific characteristics, such as audiogram information and temporal processing, to improve prediction for hearing-impaired listeners [14]. When extended to include components that account for additional distortions along the auditory pathway processing, their prediction accuracy for listeners with hearing loss improves substantially, outperforming purely acoustic predictors such as short-term SII. However, most of these methods remain centered on transformations up to the auditory periphery and provide only a limited description of how speech representations are further processed at the central stage of the auditory system.

Recent developments have further moved beyond purely acoustic representations by incorporating auditory-inspired modulation analysis. For example, the gammachirp envelope distortion index (GEDI) introduces an auditory filterbank with envelope and modulation-domain distortion analysis to better capture degradation cues [13]. Building on this line of work, the gammachirp envelope similarity index (GESI) further incorporates listener-specific characteristics, such as audiogram information and temporal processing, to improve prediction for hearing-impaired listeners [14]. When extended to include components that account for additional distortions along auditory processing, their prediction accuracy for listeners with hearing loss improves substantially, outperforming purely acoustic predictors such as short-term SII. However, most of these methods remain centered on transformations up to the auditory periphery and provide only a limited description of how speech representations are further processed at the central stage of the auditory system.

Beyond the peripheral stage, auditory information is relayed through successive subcortical stages, including the cochlear nucleus, superior olivary complex, inferior colliculus, and medial geniculate body, before reaching the auditory cortex [15]. These intermediate stages are not simple transmission steps. Instead, they progressively preserve, transform, and integrate temporal, spectral, and spatial cues [16]. This processing provides a critical basis for subsequent speech analysis at the central stage of the auditory system. Across this pathway, the signal is progressively transformed from peripheral encoding into increasingly structured neural representations that support speech perception.

In the auditory cortex, spectro-temporal patterns are further analyzed and reorganized, making spectro-temporal modulation (STM) analysis particularly relevant. Previous studies have shown that combined STM patterns are represented

in the human auditory cortex, and that responses driven by speech in the superior temporal gyrus (STG) exhibit an organized distribution from posterior to anterior [17], [18]. Specifically, the posterior STG is more sensitive to temporally faster and spectrally simpler speech patterns. In contrast, the anterior STG is more sensitive to temporally slower and spectrally richer patterns. These findings suggest that speech intelligibility depends not only on the peripheral stage but also on how auditory information is progressively transformed, integrated, and reorganized across subcortical and auditory cortical stages.

Speech intelligibility depends on how auditory information is processed across multiple stages of the auditory system. Our previous work modeled hearing loss as spectral and temporal deficits and used spectro-temporal modulation (STM) analysis to capture the resulting distortions in modulation patterns [19]. However, this method was based on monaural representations at the auditory peripheral stage and did not explicitly consider how information from the left and right ears is integrated and further processed at the auditory central stage. As a result, higher-level auditory representation and binaural information integration at the auditory central stage, as well as their contribution to speech intelligibility for hearing-impaired listeners, were not fully represented.

To address this limitation, this study extends our previous hearing-loss modeling framework beyond monaural peripheral processing. For hearing-impaired listeners, degraded auditory representations and interaural differences affect how information is further represented and combined under noisy conditions. Therefore, to better characterize speech intelligibility under hearing-loss conditions, the proposed method jointly models hearing-loss-related degradation at the peripheral stage and higher-level representations at the central stage.

The main contributions of this work are summarized as follows.

1) We propose a speech intelligibility prediction method for hearing-impaired listeners for explicitly **modeling auditory information along the auditory system**, from the auditory peripheral stage to the auditory central stage, under noisy listening conditions.
2) At the peripheral stage, the proposed method incorporates **auditory degradations** associated with hearing loss, including elevated hearing thresholds, reduced nonlinear compression, and loss of frequency selectivity. At the central stage, the proposed method further accounts for **higher level auditory information** relevant to speech understanding.

The remainder of this paper is organized as follows. Section II reviews related work. Section III presents the proposed method for modeling speech processing from the auditory periphery to the central stage under hearing-loss conditions. Section IV describes the experimental settings. Sections V–VII present the experimental results and analyses, including the overall evaluation of the proposed method, ablation analyses, and interpretation of hearing-loss effects through STM representations. Finally, Section VIII concludes the paper and discusses future work.

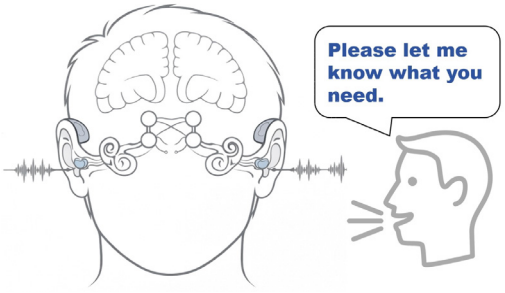

FIGURE 1. Schematic illustration of auditory system.

II. RELATED WORK

This section reviews related work from two perspectives: mechanisms of hearing loss in the auditory system and current approaches to simulating hearing loss in speech intelligibility prediction methods.

A. HEARING LOSS IN AUDITORY SYSTEM

From a functional perspective, the auditory system is commonly divided into peripheral and central stages, as shown in Fig. 1. The peripheral stage includes the outer ear, middle ear, cochlea, hair cells, and auditory nerve fibers, and converts acoustic signals into neural codes. The central stage includes brainstem nuclei, midbrain structures, and the auditory cortex, where these neural signals are further processed across time and frequency [20], [21]. In listeners with hearing loss, dysfunction often first appears in the peripheral stage. Audiograms are thus used to characterize the degree of hearing loss and evaluate hearing-aid performance. However, pure-tone audiometry primarily reflects auditory peripheral sensitivity.

In contrast, tasks such as word recognition and speech understanding in noise also depend on the auditory central stage and cognitive resources. Evidence further suggests that, although the peripheral stage plays a major role in age-related hearing loss, individual variability in speech understanding is also strongly shaped by differences in the auditory central stage and cognitive integration [21]. Therefore, hearing loss and its effects on speech understanding cannot be explained solely from the peripheral stage but must also account for the contribution of the central stage.

At the peripheral stage, age-related hearing loss is associated with multiple degenerative changes in the cochlea, including the impaired OHC function, reduced cochlear amplification, and progressively elevated high-frequency pure-tone thresholds [22]. Behavioral and physiological data indicate that older listeners typically exhibit a sloping high-frequency audiometric pattern along with markedly reduced incidence and levels of spontaneous, transient, and distortion product otoacoustic emissions, consistent with

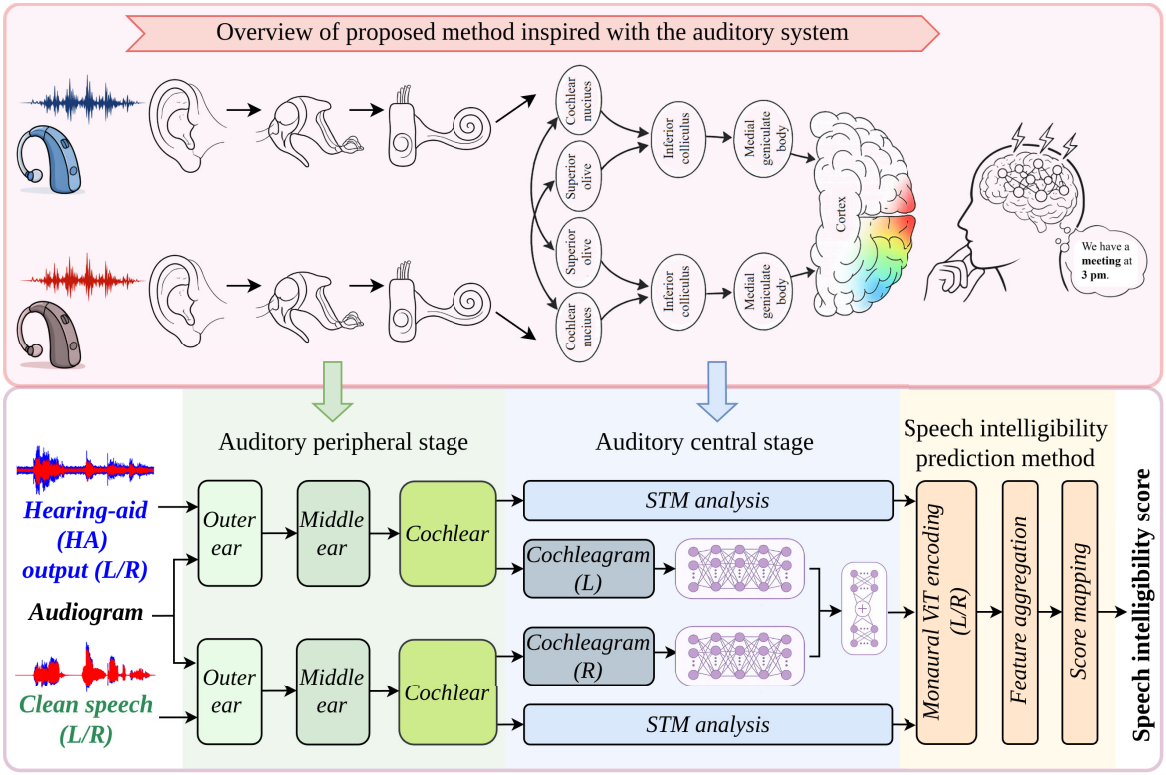

FIGURE 2. Overview of proposed speech intelligibility prediction method under hearing-loss conditions. Upper panel illustrates stages involved in processing auditory information in auditory system that motivate proposed method, and lower panel shows corresponding method design. L and R denote left and right channels; STM denotes spectro-temporal modulation; ViT denotes Vision Transformer.

OHC pathology [22]. More recent animal and human studies further demonstrate that synaptic connections between inner hair cells (IHCs) and spiral ganglion neurons can degenerate even in the absence of overt hair-cell loss or permanent threshold shifts. This hidden hearing loss leads to degraded temporal processing and poorer detection of tones or speech in noise, even when conventional audiograms are near normal [23], [24]. These findings indicate that damage in the peripheral stage not only reduces audibility but also disrupts the encoding of audible speech envelopes and the spectro-temporal structure, thus constraining later stages of auditory information [21], [22], [23].

At the central stage, age-related changes are evident at multiple levels. Psychoacoustic studies report that, even when audiometric thresholds are matched to those of younger adults, older listeners show elevated gap-detection thresholds, reduced sensitivity to interaural time differences, and age-related deficits in tasks relying on the precedence effect, pointing to temporal processing changes that are at least partly independent of peripheral hearing loss. Electrophysiological research further reveals that older adults exhibit reduced midbrain frequency-following response (FFR) amplitudes accompanied by an exaggerated cortical representation of the speech envelope, both in quiet and noise [25]. In hearing-impaired older adults, the relationship between midbrain and cortical measures is altered, suggesting that reduced input from the peripheral stage reshapes not only individual stages of processing but also the functional coupling along the auditory system. Central stage gain adjustments and cortical-stage changes are thought to reflect compensatory responses to reduced input from the peripheral stage. However, these changes can also reduce temporal precision and impair auditory scene analysis in complex acoustic environments.

For speech intelligibility, the peripheral and central stages jointly determine how well listeners communicate in real-world environments. At the peripheral stage, degradation weakens the encoding of key high-frequency speech components and temporal resolution, and the severity of these deficits is closely associated with elevated speech-recognition thresholds and reduced speech clarity. However, reduced audibility alone does not fully explain speech understanding difficulties. Even after audibility or audiometric differences are controlled, older listeners continue to experience substantial difficulties in understanding speech in noise. These difficulties are closely related to gap-detection performance, binaural integration, and the quality of cortical encoding of the speech envelope [22], [25].

Hearing loss also interferes with auditory scene analysis and selective attention, making it more difficult for listeners to form stable auditory objects and track a target talker in rapidly changing acoustic scenes. Therefore, reduced

speech intelligibility does not arise solely from the peripheral stage, but from the combined effects of cochlear degeneration, impaired temporal and spatial processing, and interactions with cognitive resources. This multi-level perspective provides an important rationale for incorporating auditory system representations into subsequent prediction methods [20], [21], [22], [23], [24], [25].

B. MODELING HEARING LOSS IN SPEECH INTELLIGIBILITY PREDICTION METHODS

For hearing-impaired listeners, a wide range of speech intelligibility prediction methods have been developed [26], [27]. These methods include early audibility-based measures, such as the SII, and environment-dependent metrics, such as the speech transmission index (STI), as well as data-driven approaches that exploit hidden representations from automatic speech recognition (ASR) systems and deep neural networks [11], [28], [29], [30]. With these methods, hearing loss is incorporated at different stages of the processing pipeline. Some methods embed audiogram information into band-importance weights or audibility calculations [31]. In contrast, others explicitly simulate the nonlinearities of OHCs and IHCs in the auditory front end [32]. Other methods use models based on the auditory periphery stage with hearing-loss parameters as feature extractors and combine them with data-driven regressors [32]. As research has shifted from generic metrics for normal hearing listeners toward tools tailored to hearing aid users, modeling how hearing loss alters speech encoding and cue extraction has become increasingly important for accurate and listener-specific speech intelligibility prediction [8].

A representative method is HASPI, which predicts speech intelligibility by comparing the internal auditory representations of clean speech and hearing-aid outputs [31]. Its core component is a model based on the auditory periphery stage that processes signals through separate normal hearing and hearing loss. This method includes middle ear filtering, an auditory filterbank, OHCs-mediated dynamic range compression, and IHCs-related adaptation. Hearing loss is incorporated by broadening auditory filters, reducing compression, and shifting auditory thresholds upward [10]. On this basis, HASPI further analyzes envelope modulation differences to predict speech intelligibility. Because hearing loss is directly represented in the internal auditory representation, HASPI provides a unified framework for evaluating both normal hearing and hearing-impaired listeners.

Another representative method is the speech-based envelope power spectrum model (sEPSM), which predicts speech intelligibility from the envelope SNR after modulation-frequency selective processing [33]. In this model, the peripheral stage is represented by a gammatone filterbank. At the same time, the temporal envelope is further analyzed using a modulation filterbank across modulation frequencies. These findings suggest that temporal-envelope information across audio- and modulation-frequency channels provides an important basis for predicting speech intelligibility. Overall, these studies suggest that incorporating auditory information from the auditory system is important not only for improving prediction accuracy but also for better characterizing the auditory information relevant to speech intelligibility.

III. PROPOSED METHOD

As shown in Fig. 2, the proposed method is inspired by processing auditory information in the auditory system under hearing loss, spanning from the peripheral stage to the central stage. The method first simulates the peripheral stage for clean speech and hearing-aid output, using hearing-loss parameters derived from the audiogram. It then extracts the central representations through neural processing, including binaural integration and STM analysis. These representations are finally combined in a speech intelligibility prediction method to estimate the speech intelligibility score. The following subsections describe each component in detail.

A. AUDITORY PERIPHERAL STAGE AND HEARING-LOSS EFFECTS ON FREQUENCY AND TEMPORAL RESOLUTION

We first incorporate the key peripheral stage, as illustrated in Fig. 3, including outer- and middle-ear filtering and cochlear analysis, because these stages determine how acoustic information is transmitted to the cochlea and define the frequency region of highest auditory sensitivity. The combined outer-ear and middle-ear transfer functions behave as a band-pass filter, emphasizing energy in the mid-frequency region where speech cues are most effectively processed and where the auditory system is particularly susceptible to impairment [34]. Within the cochlea, frequency selectivity originates from the sharp tuning of auditory filters. Psychoacoustic studies have demonstrated that sensorineural hearing loss broadens these filters, enlarges critical bandwidths, and reduces loudness summation, with the extent of broadening positively correlated with hearing-loss severity. These frequency- and time-domain degradations in the peripheral stage are well known to impair speech intelligibility, especially in noisy or complex acoustic environments, thus motivating the peripheral stage modeling components used in this study.

Psychophysical studies have further clarified how these degradations at the auditory peripheral stage depend on the degree and configuration of hearing loss. Measurements of frequency selectivity using psychophysical tuning curves, notched-noise masking, and critical bandwidth estimates show consistent patterns. Listeners with cochlear impairment have broader auditory filters and shallower filter slopes than normal hearing listeners. In some cases, the auditory filter bandwidth becomes four to five times wider. The amount of broadening also increases systematically with the severity of hearing loss [35], [36], [37], [38]. Even when stimulus level and audibility are matched, impaired listeners still show residual changes in auditory filter characteristics,

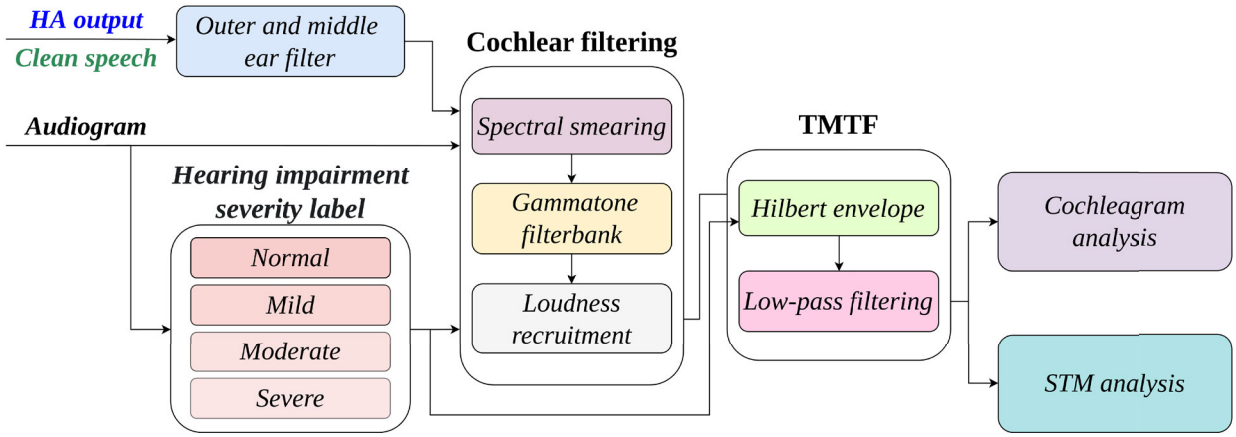

FIGURE 3. Auditory peripheral stage under hearing-loss conditions. Proposed method takes clean speech, hearing-aid output, and audiogram information as inputs, and simulates outer and middle ear filtering, cochlear processing, and temporal modulation transfer function (TMFT). Resulting cochlear representations are used as inputs for subsequent auditory central stage.

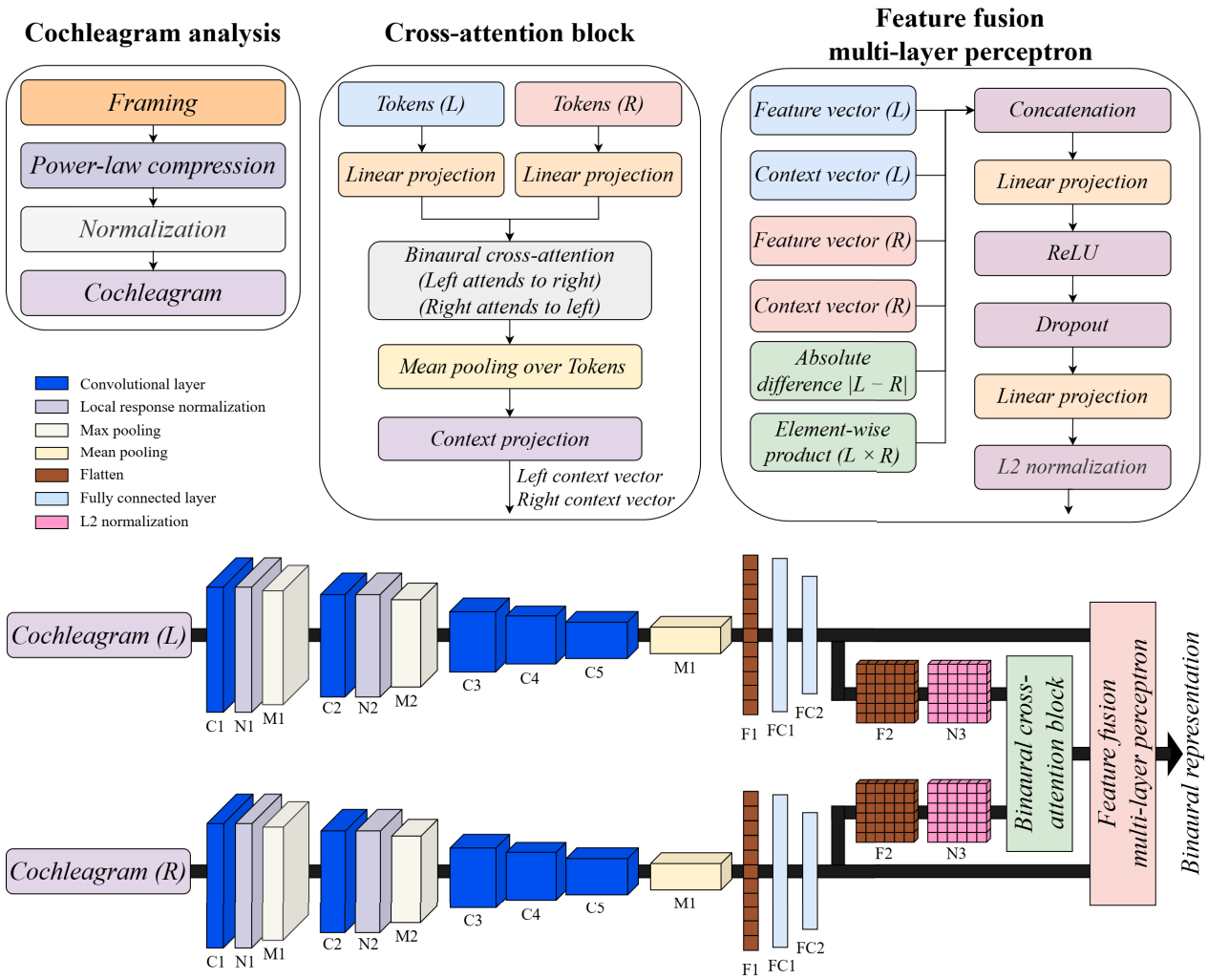

FIGURE 4. Auditory central stage with binaural integration. Cochleagram representations of left and right channels are processed through hierarchical neural network to extract central representations. Binaural cross-attention mechanism is then applied to model interactions between two channels, followed by feature fusion and prediction. ReLU denotes rectified linear unit.

confirming that cochlear pathology directly reduces frequency resolution [36], [38], [39].

In parallel, temporal processing experiments reveal similar declines in temporal resolution. Older and hearing-impaired

listeners require higher thresholds to detect envelope fluctuations and exhibit lower cut-off frequencies in temporal modulation transfer functions. These findings indicate a consistent deterioration in temporal processing as hearing loss becomes more severe. These results indicate that hearing loss produces coupled deficits in both frequency and temporal resolution in the peripheral stage. Motivated by this evidence, the following subsections describe how our method captures (i) changes in frequency resolution and (ii) changes in temporal resolution associated with different degrees of hearing loss [19].

Hearing loss and frequency resolution: In our previous study, we modeled the effect of hearing loss on frequency resolution using the Cambridge Auditory Group Moore/Stone/Baer/Glasberg (MSBG) hearing-loss model based on Nejime's research. In this model, cochlear damage is modeled by widening the Gammatone filters according to an individual's audiogram and by incorporating recruitment nonlinearities that simulate abnormal loudness growth. As hearing loss progresses from normal to severe, each auditory filter becomes broader in terms of its equivalent rectangular bandwidth (ERB). When the filters broaden, the cochlea can no longer separate nearby frequency components as effectively. Energy from adjacent spectral regions overlaps more strongly, causing the representation within each channel to become less distinct. As a result, fine details in speech, such as the separation between neighboring formants or the acoustic cues that distinguish consonants, are more easily blurred. These changes reduce a listener's ability to identify important speech components, even if the signal has been amplified to restore audibility.

Hearing loss and temporal resolution: Hearing loss also degrades temporal resolution, defined here as the ability of the auditory system to follow rapid amplitude fluctuations in the speech envelope. Building on psychoacoustic measurements of temporal modulation transfer functions (TMTFs) in normal hearing and hearing-impaired listeners, our earlier study modeled temporal resolution by applying a first-order low-pass filter to the Hilbert envelope of each cochlear channel. The filter time constant τ was selected according to hearing-loss severity, such that more severe loss corresponds to a larger τ, thus a lower 3-dB cutoff frequency. This parameterization captures the reduced sensitivity to higher modulation rates that carry information about voicing, rhythm, and phoneme boundaries. In the current study, we adopted the same set of τ values and corresponding cutoff frequencies for each hearing-loss category, ensuring that the modeled changes in temporal resolution remain consistent across all conditions.

B. AUDITORY CENTRAL STAGE: NEURAL REPRESENTATION LEARNING AND BINAURAL INTEGRATION

After the peripheral stage, the signals are further processed in the central stage. Each of these stages plays a distinct role in transforming the incoming signal. Recent studies suggest that deep convolutional neural networks (CNNs) trained on neural representations can mimic this progression. When optimized for tasks such as speech recognition, these networks exhibit behavior similar to that of listeners. They can predict activity patterns observed in both primary and non-primary auditory areas [16]. Another study combined similar networks with peripheral stage simulations to explain speech performance in noise among hearing-impaired listeners, where the network serves as an idealized central stage [40]. These findings motivate the use of a cochleagram-driven convolutional framework in the current study as a computational analog of the central stage, as illustrated in Fig. 4.

In this study, the outputs of the peripheral stage are first transformed into cochleagram representations for the left and right ears. We adopt cochleagrams because the time–frequency structure produced by cochlear filtering, compression, and envelope extraction closely resembles the firing patterns observed in auditory nerve fibers under physiological conditions. On the basis of this principle, the envelope signals from each cochlear channel are treated as time–frequency representations and segmented into short-time frames. Each frame then undergoes power-law compression, in which the envelope is raised to an exponent smaller than one. This operation simulates the compressive input–output characteristics of the basilar membrane and IHCs' transduction. The compression also reduces the dynamic range, enabling mid-level energy fluctuations to be emphasized without being overshadowed by high-intensity peaks. Previous findings have shown that methods trained directly on raw waveforms fail to develop representations consistent with cortical activity. In contrast, cochleagram-based inputs lead to substantially improved correspondence to responses across auditory cortical regions [16]. Therefore, cochleagrams provide a physiologically grounded input representation for modeling the central stage.

After generating the cochleagrams, we feed them into a hierarchical network composed of convolution, normalization, and pooling layers.[1] CNNs are used because their layer-wise processing naturally mirrors the graded transformations observed along the central stage. To clarify this correspondence, we outline the following functional roles of different network depths.

- **Shallow layers:** These layers extract local spectro-temporal patterns and resemble the response characteristics of lower-level structures such as the cochlear nucleus.
- **Intermediate layers:** With broader integration over time and frequency, these layers develop more structured representations that align with properties observed in the primary auditory cortex.
- **Deeper layers:** These layers capture complex relationships across extended temporal and spectral contexts, corresponding to processing in non-primary auditory cortical regions such as the belt and parabelt areas.

[1] https://github.com/mcdermottLab/kelletal2018

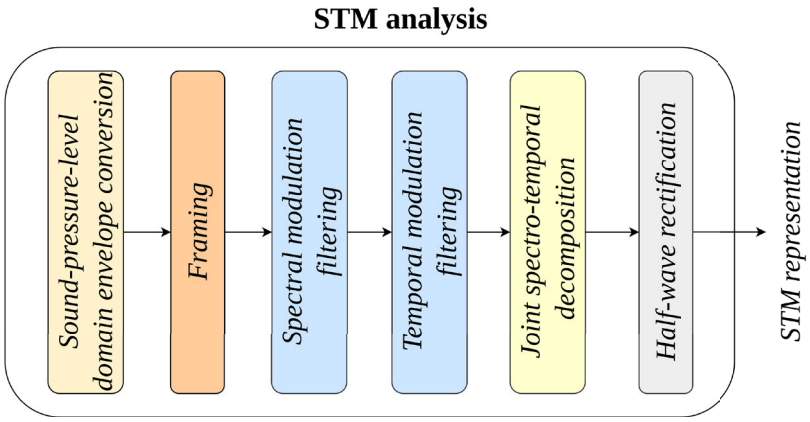

FIGURE 5. Auditory central stage with STM analysis. Cochlear outputs are processed by spectral and temporal modulation filtering and joint spectro-temporal decomposition to obtain STM representations.

Motivated by the fact that acoustic information from the two ears is integrated across multiple stages of the central stage, the proposed method extends the monaural representations with a dedicated binaural integration module. Specifically, the single-ear features extracted by the cochleagram-driven CNN are passed through two parallel convolutional branches. Each branch produces a sequence of feature vectors that summarizes the local spectro-temporal structure of the corresponding ear, providing a basis for subsequent cross-ear interactions. This mechanism enables tokens from one ear to attend to those of the other ear and to receive weighted information from them, thus generating context vectors that represent binaural relationships.

In functional terms, cross-attention enables the proposed method to update each ear's representation while referencing the other ear, analogous to how the central stage integrates interaural time differences, level differences, and correlation cues. In addition to the attention-derived context vectors, we explicitly construct two forms of binaural interaction features. The first is the element-wise absolute difference $|L - R|$, representing contrastive information similar to that carried by difference pathways. The second is the element-wise product $L \times R$, representing coherent or joint components analogous to coincidence-detection pathways. Together with the cross-attention outputs, these features provide complementary cues reflecting distinct aspects of binaural processing.

Finally, all binaural features are passed through a multi-layer perceptron (MLP) to produce a compact binaural representation. This representation serves as an abstract encoding of how the central stage may integrate signals from the two ears and is used as an important input to the speech intelligibility prediction method.

C. AUDITORY CENTRAL STAGE: SPECTRO-TEMPORAL MODULATION ANALYSIS

In the auditory cortex, particularly in Heschl's gyrus on the superior temporal plane, previous studies have shown that neurons respond selectively to specific combinations of spectral modulation frequency Ω and temporal modulation rate ω. This selectivity is typically assessed using dynamic ripple stimuli and modulation transfer function (MTF). Functional imaging results further indicate that voxels in the auditory cortex exhibit different preferred regions in the joint STM plane [18]. Most voxels are more sensitive to relatively low temporal modulation rates (around 2–4 Hz) and low spectral modulation densities. Consequently, their overall MTF is low-pass in both spectral and temporal domains and closely matches the modulation statistics of speech and natural environmental sounds. These findings support the view that, at the cortical level, sounds can be represented as energy distributed over a set of modulation "sub-bands," referred to as the STM representation. STM therefore provides a neurophysiologically grounded framework for relating speech intelligibility and hearing loss to cortical modulation sensitivity [41].

In this study, we use an STM analysis based on Gabor filters to map the envelope signals obtained after the peripheral stage into a joint modulation domain. We construct a set of two-dimensional Gabor filters in the frequency–time plane. Each filter consists of a sinusoidal carrier at a given modulation center frequency multiplied by a finite-support window, which ensures localization in both frequency and time [42], [43]. The filters are designed to have approximately constant-Q properties, so that their bandwidth increases in proportion to the center frequency of the modulation. By maintaining approximately constant overlap between neighboring filters, the filter bank provides even coverage of combinations of low spectral modulation frequencies and low temporal modulation rates. The design is thus consistent with the auditory cortex's preference for low modulation rates and with the dominant modulation components observed in speech.

On the basis of these findings, we focus on low-to-mid temporal modulation rates and relatively low spectral modulation densities, since these ranges contribute most

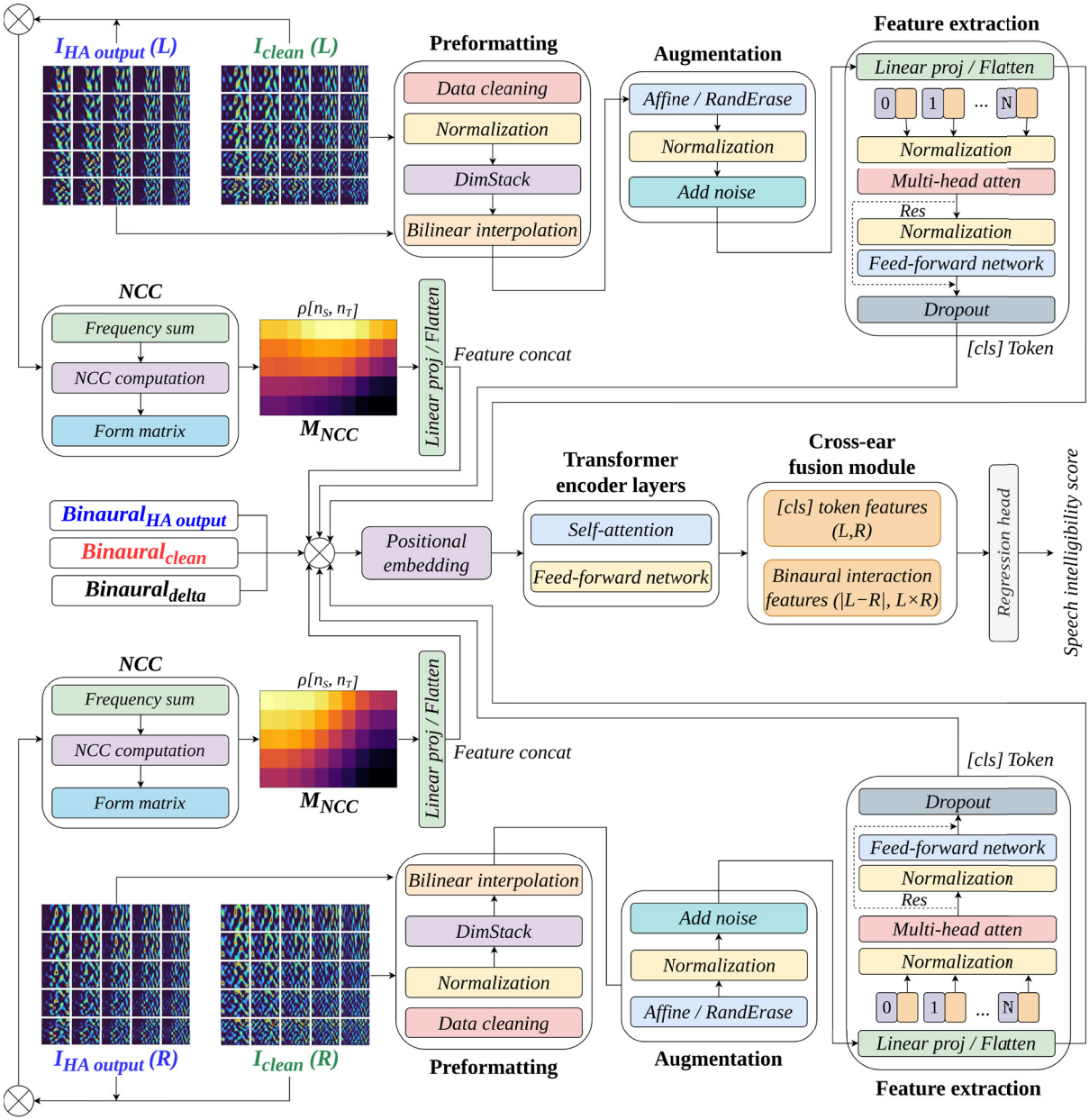

FIGURE 6. Overview of speech intelligibility prediction method. Left and right ear STM representations, NCC features, and the central stage is jointly encoded and fused through cross-ear fusion module to predict speech intelligibility score.

strongly to speech understanding. After defining the modulation center frequencies, each Gabor channel is processed by applying half-wave rectification to the filter output, then summing the resulting energy over time. Here, half-wave rectification is applied to suppress negative responses introduced by the oscillatory nature of Gabor filtering while retaining positive modulation energy. It is used as a signal processing operation for feature stabilization. This procedure produces a compact set of STM representations that retain the modulation patterns most relevant to cortical sensitivity and serve as the modulation-domain inputs for the subsequent stages.

D. SPEECH INTELLIGIBILITY PREDICTION WITH MULTI-SOURCE BINAURAL FUSION

After the auditory peripheral and central stages, the final step is to predict speech intelligibility scores, as illustrated in Fig. 6. In the input stage, we construct a pair of STM images for each ear, where each image contains two channels representing the clean and degraded signals. After bilinear interpolation, normalization, and optional data augmentation, these images are fed into two parallel Vision Transformer (ViT) base encoders [44], one for each ear.[2]

[2]https://github.com/google-research/vision_transformer

The ViT architecture consists of a patch-embedding module, a stack of transformer encoder blocks (each containing multi-head self-attention and feed-forward sublayers), and a final layer-normalization stage. The patch-embedding module divides each STM image into fixed-size patches. It projects them into a unified embedding space, enabling the subsequent transformer layers to execute token-level modeling over the time–frequency structure. The multi-head self-attention mechanism within the encoder blocks captures long-range dependencies across patches, enabling the method to represent relationships among STM sub-bands. After several layers of processing, the *[CLS]* (classification) token produced by the ViT serves as a global summary of the STM image for that ear, encoding both the overall modulation structure and cross-sub-band interactions.

To make the encoder more sensitive to modulation distortions, we augment the visual tokens with an additional normalized cross-correlation (NCC) token. The NCC is computed between corresponding clean and degraded modulation channels and measures how well the modulation structure is preserved. For a given modulation unit, the NCC can be written as

$$\rho = \frac{\langle X_{\text{clean}} - \mu_X,\ X_{\text{deg}} - \mu_Y \rangle}{\|X_{\text{clean}} - \mu_X\|\ \|X_{\text{deg}} - \mu_Y\|}, \tag{1}$$

where X_{clean} and X_{deg} respectively denote the clean and degraded STM energies and μ_X and μ_Y are their mean values. Higher ρ indicates that the corresponding modulation band is less affected by degradation, which provides a useful cue for predicting speech intelligibility.

In addition to the NCC token, we incorporate the central stage by introducing three further tokens derived from the neural representation learning module: a clean central embedding, a degraded central embedding, and their difference. These vectors are linearly projected to the ViT-embedding dimension and concatenated with the NCC token and patch tokens. The input sequence for each ear thus consists of the *[CLS]* token, one NCC token, three central tokens, and the STM patch tokens. The Transformer layers jointly process these elements so that the final *[CLS]* output for each ear reflects not only the STM structure but also modulation similarity and auditory central representations.

To model binaural integration, the left and right ear *[CLS]* vectors, denoted as $\mathbf{h}_L$ and $\mathbf{h}_R$, are passed to a cross-ear fusion module. This module does not rely solely on the two monaural representations but also forms two explicit interaction terms. The first is an element-wise absolute difference,

$$\left|\mathbf{h}_L - \mathbf{h}_R\right|, \tag{2}$$

which emphasizes contrastive information between the two ears. The second is an element-wise product,

$$\mathbf{h}_L \odot \mathbf{h}_R, \tag{3}$$

which highlights components that are jointly active and can be viewed as a simple form of coincidence detection. These four vectors (left ear, right ear, difference, and product) capture complementary aspects of binaural processing, including specific patterns, interaural disparities, and shared structure.

Finally, the four vectors are concatenated and passed through an MLP serving as the regression head. It outputs a single scalar representing the predicted speech intelligibility score for the given sentence. The proposed method thus links the peripheral and central stages within a unified framework for predicting speech intelligibility.

IV. EXPERIMENTS

A. DATASET

We use the CPC2 dataset[3] designed for speech intelligibility prediction among hearing-aid users. The released data include hearing-aid outputs; corresponding clean speech; listener information, such as bilateral pure-tone audiograms; and ground-truth correctness scores obtained from listening tests. As shown in Table 1, the training material combines the first Clarity Enhancement Challenge (CEC1) and the second Clarity Enhancement Challenge (CEC2) [45], [46], whereas the evaluation data come from CEC2.

TABLE 1. Dataset configuration of CPC2 dataset for speech intelligibility prediction.

Item	Description
Dataset	CPC2
Training source	CEC1 + CEC2
Evaluation source	CEC2
No. of utterances in Train	6297 (CEC1) + 5944 (CEC2)
No. of utterances in Test	897
Listener information	Bilateral audiograms (left/right ears)

The speech in noise mixtures were generated from spatial acoustic scenes. Each scene included a target utterance and one or more interfering sources. In CEC1, each scene included one interferer, whereas CEC2 introduced more complex conditions with two or three interferers, including competing speech, domestic noise, and music [45], [46]. The acoustic scenes were simulated in small rooms with low to moderate reverberation [47]. For each scene, the listener's position was randomly selected within the room, with constraints to keep the listener at least 1 m away from the walls. The listener height was set to either 1.2 m for a sitting position or 1.6 m for a standing position. The target talker was also randomly positioned within the room, at the same height as the listener and at least 1 m away from the listener. The interferers were placed at different spatial positions around the listener, resulting in different binaural cues at the left and right ears.

In CEC2, listener head rotation was further included. The listener was initially oriented away from the target talker by approximately 25 degrees, and the head rotation

[3]https://zenodo.org/records/17045970

TABLE 2. Implementation details of central CNN and binaural integration.

Module	Setting
Input	
Representation	Cochleagram (frame-level)
Frame rate	200 Hz
Compression	Power-law (0.3)
Monaural CNN	
Conv blocks	3 layers (9×9, 5×5, 3×3)
Channels	$1 \rightarrow 96 \rightarrow 256 \rightarrow 512$
Normalization	Local response normalization
Pooling	Max pooling (stride 2)
Output	Adaptive pooling to 6×6
Embedding	512-dim (FC layers)
Binaural Interaction	
Branches	Left/right independent CNNs
Tokens per ear	36
Cross-attention	4 heads
Explicit cues	$\lvert L - R\rvert, L \times R$
Fusion	
Fusion layer	MLP (hidden dim 512)
Output	512-dim binaural embedding

started around the onset of the target speech. The rotation lasted approximately 200 ms, after which the listener was approximately oriented toward the target talker. These spatial settings are relevant to the present study because the proposed method uses binaural information from the left and right ears. The SNR was not fixed across all mixtures. For each scene, a desired SNR was randomly selected from a predefined range determined by pilot listening tests, so that the mixtures covered a suitable range of speech intelligibility. The interferer level was then adjusted using the better-ear SNR, which was calculated from the left- and right-ear speech-weighted SNRs at the reference hearing-aid microphone channel. This procedure models the better-ear effect in binaural listening and determines the final SNR stored in the scene metadata.

The hearing-aid outputs in CPC2 were generated by systems submitted to CEC1 and CEC2 rather than by a single fixed hearing-aid setting. As representative examples, the top-ranked CEC1 system used a two-stage processing framework consisting of a denoising module and an amplification module [48]. The denoising module suppressed noise and speech interference. In contrast, the amplification module was individually tailored to the listener's hearing ability for hearing-loss compensation. The top-ranked CEC2 system used a multi-channel target-speaker extraction framework with neural enhancement and beamforming, followed by listener-specific National Acoustic Laboratories Revised (NAL-R) fitting and dynamic range compression [49]. These examples indicate that the hearing-aid outputs include noise and interferer suppression, spatial processing, listener-specific amplification, and compression.

The subjective listening tests were conducted with experienced bilateral hearing-aid users. The listeners completed the experiment in a quiet room using a universal serial bus (USB) stereo headset rather than loudspeakers. During the test, each sentence was presented once through web-based listening-test software. After each presentation, the listeners repeated what they heard from the target talker, and their spoken responses were recorded with the headset microphone. The responses were then scored offline based on the number of correctly recognized words. The playback setting specified that 0 decibels relative to full scale (dB FS) corresponded to 100 decibels sound pressure level (dB SPL), while the listeners were allowed to adjust the volume to a comfortable level.

Specifically, bilateral audiograms are particularly important for this study because they provide listener-specific hearing thresholds at different frequencies for the left and right ears [8]. This information allows simulation of not only different degrees of hearing loss but also asymmetric hearing profiles across ears. In this study, hearing loss severity is categorized using the better-ear hearing threshold as mild (≤35 dB), moderate (35–56 dB), and severe (>56 dB). These listener-specific hearing profiles are used to control the hearing-loss simulation with the proposed method.

B. EXPERIMENTAL SETTINGS

1) NEURAL REPRESENTATION LEARNING AND BINAURAL INTEGRATION

In the neural representation learning module of the central stage, the envelope outputs from the peripheral stage are first converted into frame-level cochleagram representations at a frame rate of 200 Hz. A power-law compression with an exponent of 0.3 is then applied to stabilize the representation. For each ear, the cochleagram is fed into an independent CNN to extract features. The network contains three convolutional blocks. Each block includes convolution, local response normalization, and max-pooling. These operations gradually capture and compress the time–frequency structure.

After the convolutional stages, the feature maps are processed by an adaptive average pooling layer to obtain a fixed 6×6 representation. The pooled features are then flattened and projected through fully connected layers to produce a 512-dimensional monaural embedding. To model binaural interaction, the pooled features from the left and right ears are reshaped into token sequences and passed into a cross-attention module with four attention heads. In addition to the attention outputs, two explicit interaction terms are introduced. These are the element-wise absolute difference and element-wise product between the left and right embeddings. Finally, all monaural and binaural features are fused by an MLP to produce a 512-dimensional binaural representation for the prediction method. The implementation details are summarized in Table 2.

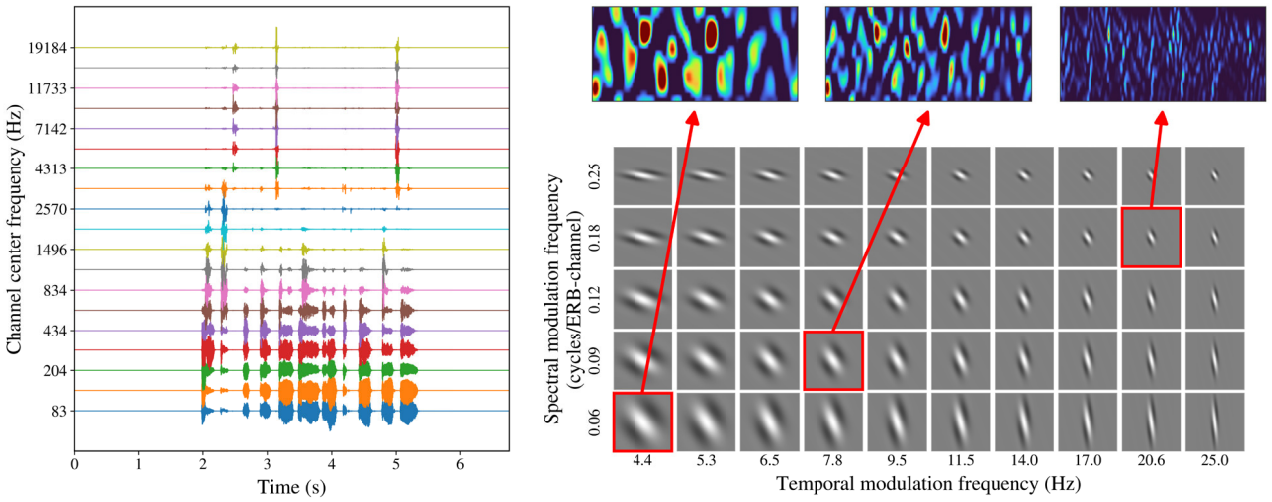

FIGURE 7. Example of STM extraction process for listener with severe hearing loss. Left: output of peripheral stage represented as channel-wise responses over time. Right: separable Gabor filter grid used for STM extraction, organized by spectral modulation frequency (vertical axis) and temporal modulation frequency (horizontal axis). Red boxes indicate example modulation channels, and corresponding responses are shown above to illustrate how different combinations of spectral modulation frequency and temporal modulation frequency emphasize different spectro-temporal patterns.

TABLE 3. Parameter settings for STM extraction.

Component	Setting
Envelope frame rate	100 frames/s
Spectral Gabor filter g_{Ω_i}	1D separable Gabor filters
Temporal Gabor filter h_{ω_j}	1D separable Gabor filters
Half-waves under envelope ν	3.5 (spectral and temporal)
Spectral modulation range c_i	0.025–0.25 cycles/channel
Temporal modulation range f_j	4.38–25 Hz
Spectral modulation channels i	5
Temporal modulation channels j	10
Spectral axis definition Ω_i	$\Omega_i = 2\pi c_i$
Temporal axis definition ω_j	$\omega_j = 2\pi f_j/100$
Post-processing	Half-wave rectification
Final STM grid size	5×10

2) SPECTRO-TEMPORAL-MODULATION ANALYSIS

In the experimental setting, we focus on how to construct and parameterize the STM representation in practice. Let the envelope representation after the preceding auditory information stages be denoted as $E(f,t) \in \mathbb{R}^{C\times T}$, where $E(f,t)$ denotes the auditory envelope at frequency band f and time frame t, C is the number of frequency bands, and T is the number of time frames. We extract STM features using the separable Gabor filters. Each modulation channel is defined by a pair of spectral and temporal modulation center frequencies, (Ω_i, ω_j). For each channel, we first apply the spectral Gabor filters g_{Ω_i} to the envelope representation, then apply the temporal Gabor filters h_{ω_j}. The resulting response is written as

$$\mathrm{STM}_{i,j} = h_{\omega_j}\left(g_{\Omega_i}(E)\right), \tag{4}$$

where E denotes the envelope representation after processing auditory information, and g_{Ω_i} and h_{ω_j} denote the one-dimensional Gabor filters for spectral and temporal modulation analysis, respectively. The STM representation thus forms a modulation-frequency grid, in which each channel corresponds to a specific STM sub-band.

Due to the oscillatory nature of Gabor filters, the resulting STM responses may contain both positive and negative values that do not directly correspond to modulation energy. To obtain a stable and interpretable representation, we apply half-wave rectification to each modulation channel, defined as

$$\mathrm{STM}^{+}_{i,j} = \max\left(\mathrm{STM}_{i,j}, 0\right). \tag{5}$$

This operation suppresses negative responses and retains only positive modulation energy, thereby reducing sign ambiguity introduced by the filtering process. It should be noted that half-wave rectification is used here solely as a signal processing step for feature stabilization.

To keep the modulation analysis consistent across scales, the one-dimensional Gabor filter can be written as

$$g_\omega(x) = w_b(x)\cos(\omega x), \tag{6}$$

where $w_b(x)$ is a Hann window with support width b. We set

$$b = \frac{2\pi}{|\omega|}\cdot\frac{\nu}{2}, \tag{7}$$

where ν is the number of half-waves under the envelope. With this setting, the filter width varies with the center frequency and approximately follows a constant-Q property. Lower modulation frequencies thus use broader analysis windows, whereas higher modulation frequencies use more localized filters. This design provides stable coverage of low spectral

TABLE 4. Comparison of proposed method with other methods on CPC2 dataset.

Data	Rank	Method	Intrusive	$\rho\uparrow$	RMSE $\downarrow$	Theory	Features
CPC2	1	E011 [50]	No	0.78	25.1	Noise-robust foundation models **encode signal–noise disentanglement** across layers, motivating hierarchical intelligibility modeling	**Multi-layer Whisper and WavLM representations** with temporal, cross-layer, binaural, and audiogram information
	2	**Proposed Method**	**Yes**	**0.78**	**25.2**	Speech intelligibility depends on processing across the **auditory system** and can be better predicted by **jointly modeling auditory peripheral stage and central stage**	Hearing-aid outputs and clean speech processed into **STM representations, auditory central stage, NCC-based modulation similarity features, and binaural interaction cues** under hearing-loss simulation
	3	E002 [51]	No	0.77	25.3	**Intermediate ASR representations** encode speech-intelligibility-relevant information, enhanced by exemplar-based human memory modeling	**Multi-layer Whisper decoder features** aggregated over time, together with exemplar similarity information
	4	E009 [52]	Yes	0.78	25.4	Speech intelligibility depends on **complementary acoustic, phonetic, linguistic, and audiometric cues**	**Hand-crafted multi-source features**: short-time objective intelligibility correlations, phone-lattice correlations, sentence probability, and audiometric thresholds
	5	E022 [53]	Yes	0.77	25.7	ASR ability can act as a proxy for speech intelligibility through **representation similarity and decoding uncertainty**	**Noise-robust ASR-decoder representations and uncertainty-related cues** from decoding outputs
	6	E023 [54]	Yes	0.76	26.4	Joint modeling of **hearing perception and acoustic-information** benefits intelligibility prediction, further strengthened by multi-task learning	**Cross-domain features** combining short-time Fourier transform, log filterbank, and Whisper representations with MSBG hearing-loss mode
	7	E016 [54]	No	0.75	26.8	Joint modeling of **hearing perception and acoustic-information** benefits intelligibility prediction in a multi-branch framework	**Cross-domain features** combining short-time Fourier transform, log filterbank, and Whisper representations with MSBG hearing-loss model
	8	E025 [53]	No	0.72	27.9	**ASR-decoding** uncertainty provides a non-intrusive proxy for speech intelligibility without requiring a reference signal	**Entropy-based uncertainty features** derived from ASR-decoder probability distributions
	9	be-HASPI [26]	Yes	0.67	28.7	Intelligibility is determined by how well **auditory envelope and temporal fine structure information** are preserved after degradation	Auditory-model-based envelope modulation and temporal-fine-structure-related features extracted via **auditory peripheral stage and modulation filterbanks**
	10	E003	No	0.64	31.1	**High-level speech representations** can be mapped to intelligibility through ensemble regression	**Global WavLM embeddings** extracted from binaural signals after left–right averaging
	11	E024	No	0.62	31.7	**Binaural cues and listener-specific characteristics** are important for intelligibility and can be modeled with temporal learning	**Time-series WavLM features** after equalization–cancellation processing, combined with listener-audiogram embedding

modulation densities and low-to-mid temporal modulation rates, which are the main focus of the STM representation in this study.

In the implementation, we discretize the spectral modulation axis into 5 channels spanning 0.025–0.25 cycles/channel and discretize the temporal modulation axis into 10 channels spanning 4.38–25 Hz. We therefore organize the final STM output as a 5×10 modulation grid, with temporal modulation frequency on the horizontal axis and spectral modulation frequency on the vertical axis. Because the input-envelope representation is uniformly defined at 100 frames/s, we convert the temporal modulation center frequencies to the discrete-time domain as $\omega_j = 2\pi f_j/100$. For the spectral axis, we define the modulation center frequencies as $\Omega_i = 2\pi c_i$. Table 3 summarizes the corresponding parameter settings. This parameterization ensures that STM representations across different samples and listeners share the same physical meaning and dimensional structure, facilitating subsequent modeling of modulation channel response patterns.

Figure 7 shows a practical example for a listener with severe hearing loss and illustrates how the STM analysis used in our method is applied to a real signal. The left panel shows the output of the peripheral stage as 19 channel-wise responses over time. The right panel shows the

modulation grid together with the separable Gabor filters used for STM extraction. The highlighted examples show that different (Ω_i, ω_j) combinations emphasize different spectro-temporal patterns. Each STM sub-band in the figure can thus be directly interpreted as an analysis unit with a specific modulation meaning, consistent with the fixed modulation-grid structure used for the central stage in this study.

C. EVALUATION METRICS

The target variable is the *correctness* score provided in the listener-response metadata. This score represents the percentage of words that the listener correctly identified for each signal. The method output is therefore compared directly with the corresponding correctness value. To evaluate prediction performance, we use the *Pearson correlation coefficient* (ρ) and *root mean square error* (RMSE). These two metrics reflect complementary aspects of speech intelligibility prediction.

The ρ measures the linear relationship between the predicted scores and ground-truth correctness scores. A higher ρ indicates that the method better captures the relative variation and ranking of intelligibility across different samples. It is defined as

$$\rho = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{8}$$

where y_i denotes the ground-truth correctness score of the i-th sample, $\hat{y}_i$ denotes the corresponding predicted score, and $\bar{y}$ and $\bar{\hat{y}}$ are their mean values.

The RMSE measures the average magnitude of the prediction error between the predicted scores and ground-truth correctness scores. A lower RMSE indicates that the predicted intelligibility scores are numerically closer to the listener responses. It is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \tag{9}$$

V. ANALYSIS I: OVERALL EVALUATION OF PROPOSED METHOD

A. OVERALL SPEECH INTELLIGIBILITY PREDICTION PERFORMANCE ON CPC2 DATASET

We compared the proposed method with the methods from the CPC2 challenge listed in Table 4. All methods were evaluated using ρ and the RMSE. A better method is expected to achieve a higher ρ and lower RMSE. The comparison was conducted to assess the overall predictive performance of the proposed method on the CPC2 dataset and position it relative to both intrusive and non-intrusive methods. The baseline (be-HASPI) is based on HASPI and adopts a better-ear strategy, computing intelligibility scores for both ears and selecting the higher one as the final output.

As shown in Table 4, the proposed method achieved a ρ of 0.78 and RMSE of 25.2, ranking second overall among the compared methods. It also ranked first among the intrusive methods. Compared with the baseline (be-HASPI), the proposed method achieved clear improvements on both evaluation metrics. The ρ increased from 0.67 to 0.78, which corresponds to a relative improvement of 16.4%. The RMSE decreased from 28.7 to 25.2, which corresponds to a relative reduction of 12.2%. This comparison shows that the proposed method can predict ground-truth correctness scores more accurately than be-HASPI.

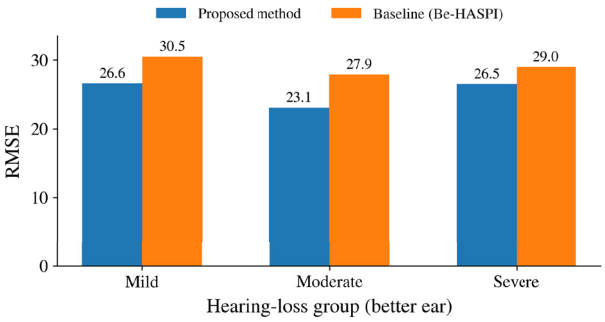

FIGURE 8. RMSE comparison between proposed method and baseline (be-HASPI) across hearing-loss groups defined by better ear.

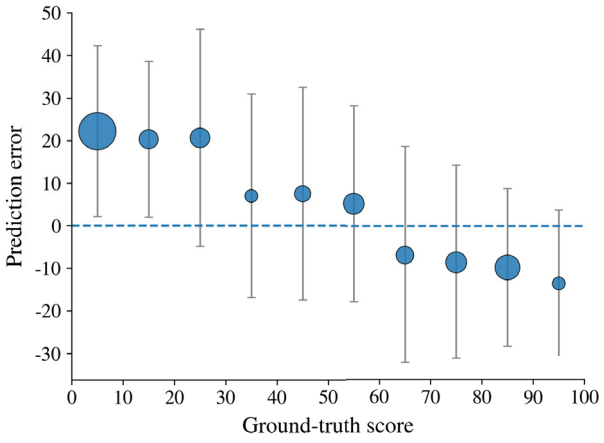

FIGURE 9. Prediction error of proposed method across ground-truth score bins. Each marker represents mean prediction error (predicted minus ground-truth score) within 10-point score bin, and error bars denote standard deviation. Marker size is proportional to number of samples in each bin.

An important reason for this improvement is that the proposed method is based on the auditory system for hearing-impaired listeners. It is inspired by auditory information processing, where speech is received, processed, and interpreted by the peripheral and central stages, rather than relying solely on high-level learned speech embeddings. The peripheral stage accounts for the degradation in hearing loss. The central stage then captures how speech information is represented after processing auditory information. The STM analysis provides an explicit description of spectro-temporal patterns that are closely related to speech perception. This is important because speech intelligibility is not determined only by

audibility at the periphery but also by how temporal and spectral modulations are encoded and integrated in the central stage. By combining the simulation of the periphery stage with the central stage, the proposed method can better capture the mechanisms that affect ground-truth correctness scores.

Table 4 also shows that many recent prediction methods follow a similar trend. A representative example is the group of ASR-based methods, including E002, E022, and E025. With these methods, it is assumed that speech intelligibility is closely related to how well a robust ASR model can represent or recognize degraded speech. Another notable trend is the repeated use of Whisper and waveform language model (WavLM) features across several competitive methods, including E002, E011, E016, E023, E003, and E024. This pattern suggests that large pretrained speech models provide feature spaces that are highly informative for intelligibility prediction. Whisper is trained in a weakly supervised manner on large-scale multilingual speech data. It produces robust representations that remain effective across different acoustic conditions. WavLM is trained in a self-supervised manner and designed to preserve not only speech content but also speaker and environment information. These properties are particularly useful for speech intelligibility prediction because the task depends not only on linguistic recoverability but also on how degradation affects the speech signal.

From a modeling perspective, the value of Whisper and WavLM lies in their hierarchical representations. Rather than relying on handcrafted signal measures, these models learn multi-level features from large, diverse corpora. Different methods then use these representations in different ways. Some methods use them as direct cues for prediction. Others further combine them with temporal modeling, cross-layer aggregation, or listener information. A notable point is that the information carried by these representations is not uniform across layers, and some methods explicitly exploit this property to improve prediction performance. Overall, the repeated use of Whisper and WavLM in recent CPC2 methods indicates that pretrained speech representations have become an important direction for prediction and can support a wide range of competitive model designs.

However, these methods and the proposed method differ in their assumptions and limitations. The top-ranked method, E011, can perform a non-intrusive assessment because it does not require clean reference speech. This is an important practical advantage when only hearing-aid outputs are available. However, its prediction is mainly based on learned speech representations from large-scale pretrained models, and the relationship between these representations and hearing-loss-related auditory processing may not be explicit. In contrast, the proposed method is intrusive because it requires clean reference speech, which limits its applicability when clean reference speech is unavailable. Nevertheless, this design allows direct comparison between clean reference speech and hearing-aid outputs in auditory-inspired representations. Therefore, the proposed method is useful not only for prediction but also for analyzing how hearing-loss-related peripheral degradation and higher-level representations contribute to speech intelligibility.

B. PERFORMANCE ACROSS HEARING-LOSS GROUPS AND SCORE RANGES

To further analyze the results, we grouped the test samples by the better-ear hearing-loss level and compared the RMSE performance of the proposed method and be-HASPI across these groups. As shown in Fig. 8, the proposed method consistently outperformed be-HASPI across all hearing-loss groups, indicating that its advantage is not limited to a specific condition but remains stable across different levels of hearing loss. A similar trend was observed in ρ, where the proposed method achieved 0.73, 0.82, and 0.74 for the Mild, Moderate, and Severe groups, compared with 0.69, 0.73, and 0.68 for the be-HASPI. These results indicate that the proposed method not only improves overall performance but also provides a more consistent characterization of correctness across different hearing-loss conditions, with the most pronounced improvement observed in the Moderate group. This suggests that incorporating the auditory system of hearing-impaired listeners improves the prediction method's adaptability to different hearing conditions, rather than simply enhancing overall average performance.

To further characterize the prediction behavior of the proposed method, we analyzed the distribution of prediction errors across different ground-truth score ranges. As shown in Fig. 9, the prediction error was positive in the lower score ranges, indicating that the methods tend to overestimate samples with low correctness. As the ground-truth score increased, the prediction error gradually shifted toward negative values, indicating that the method tends to underestimate samples with high correctness. These results suggest a mild compression of predictions toward the mid-score range. This trend is also reflected in the mid-score region, where the mean prediction error remains closer to zero, and the bias is relatively smaller. The error bars further show that the variability of the predictions changes across score ranges: it is larger for lower ground-truth correctness scores. However, it decreases at higher scores, even though a negative bias persists.

VI. ANALYSIS II: ABLATION STUDIES AND COMPONENT-WISE CONTRIBUTIONS

We further analyzed the specific role of each model variant generated with the proposed method through ablation experiments, with particular focus on the contributions of STM representations, NCC, binaural representations, and the fusion module to final speech intelligibility prediction performance. In addition to evaluating individual components and their combinations, we further compared the method's performance across left-, right-, and two-ear settings to examine the practical role of binaural information in this prediction task. Through this series of controlled experiments, we aimed to address two issues: (1) the relative contribution of each feature type to the Full Model; and (2) whether the

TABLE 5. Ablation study of model variants generated with proposed method showing contribution of STM representations, NCC, binaural representations, and fusion module.

Model Variant	Ear setting	STM	NCC	Binaural	Fusion	RMSE ↓	ρ ↑
STM + NCC	Left ear	✓	✓	×	×	26.1	0.76
STM + NCC	Right ear	✓	✓	×	×	26.5	0.75
STM + NCC	Two ears	✓	✓	×	×	25.7	0.77
STM only	Two ears	✓	×	×	×	26.3	0.75
NCC only	Two ears	×	✓	×	×	31.6	0.63
Binaural only	Two ears	×	×	✓	×	41.7	0.07
STM + Binaural	Two ears	✓	×	✓	×	26.8	0.74
NCC + Binaural	Two ears	×	✓	✓	×	28.1	0.71
STM + NCC + Binaural	Two ears	✓	✓	✓	×	25.9	0.77
Full Model (STM + NCC + Binaural + Fusion)	**Two ears**	✓	✓	✓	✓	**25.2**	**0.78**

performance change obtained by combining multiple features is due to the features or to the fusion strategy.

As shown in Table 5, the STM representations are the most stable and most critical component. When STM representations were used alone, the method already achieved strong performance, indicating that STM representations can effectively characterize distortion patterns related to speech intelligibility. This observation shows that STM representations reflect how hearing loss and speech degradation affect the auditory encoding of speech information from the modulation-domain perspective. In contrast, using NCC alone resulted in a clear performance drop, though it still outperformed using binaural representations alone. This suggests that NCC alone is insufficient to support optimal prediction but still provides complementary discriminative information. Since NCC reflects the consistency between clean speech and hearing-aid output in the STM space, it enhances the method's sensitivity to changes in speech intelligibility by comparing signals.

Across different ear settings, the model using STM representations and NCC exhibited good performance in both left- and right-ear monaural settings, indicating that either ear can provide meaningful predictive cues. However, the two-ear setting further improved performance, showing that jointly modeling information from both ears provides an additional benefit. In other words, the gain from binaural modeling does not come from the absolute dominance of one ear but from the complementarity between the two sides. This result is consistent with the actual auditory-perception process. It also supports the need to preserve the binaural structure in the current task.

Binaural representations, however, perform worse when used alone, indicating that high-level binaural representations alone are not sufficient to accomplish the prediction task. However, they can still provide useful auxiliary information when combined with other features. For example, combining NCC with binaural representations shows a clear improvement over NCC alone, suggesting that binaural representations still contain meaningful information related to binaural auditory integration.

It is also worth noting that adding binaural representations to the model already using STM representations and NCC does not lead to further improvement, and the performance instead decreases slightly. This suggests that the limitation is not due to a lack of usefulness in binaural representations but to the heterogeneity among the different feature types. Specifically, STM representations mainly describe modulation structure, NCC captures the consistency between the reference and degraded signals, and binaural representations encode higher-level binaural integration information. Because these features belong to different representational levels and statistical spaces, a simple feature combination may introduce redundancy or interfere with optimization, limiting overall performance. After introducing the fusion module, however, the Full Model achieved the best result. This indicates that its advantage comes not simply from the use of more features but from a more effective way of modeling the interactions among multiple auditory information sources. Therefore, the performance gain of Full Model is attributable not only to the complementarity among STM representations, NCC, and binaural representations but also to the effective integration enabled by the fusion module.

VII. ANALYSIS III: INTERPRETATION OF HEARING-LOSS EFFECTS THROUGH STM REPRESENTATIONS

To further explain how hearing loss affects speech intelligibility prediction, we analyzed the proposed method from the perspective of auditory system processing. Although hearing aids have enhanced speech in noise, speech intelligibility for hearing-impaired listeners still depends on subsequent encoding and analysis by the auditory system. Therefore, for the hearing-aid output, the proposed method simulates auditory information from the peripheral to the central stage, using STM representations to characterize higher-level representations.

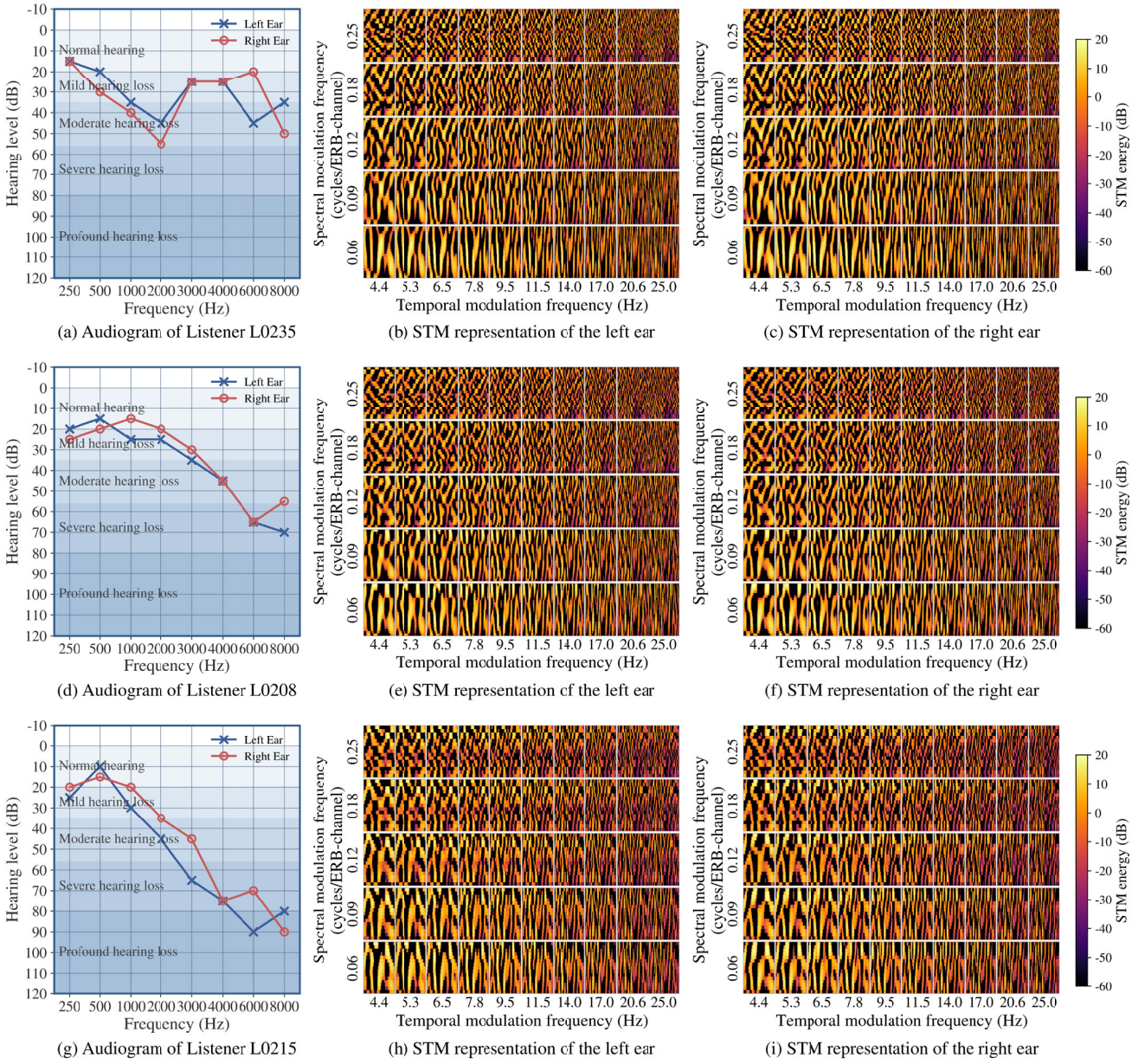

FIGURE 10. Effect of hearing loss on STM representations of clean speech. Each row corresponds to different hearing-loss condition (Mild, Moderate, and Severe), illustrated with audiogram of corresponding listener (left). Right panels show STM representations of left and right ears obtained from same clean speech signal after hearing-loss simulation.

Previous studies have shown that STM analysis has a clear physiological relationship with the auditory system. Neurons in the auditory cortex, however, exhibit selective responses to specific STM patterns. The human auditory cortex shows relatively differentiated processing of temporal and spectral variations, which supports the use of STM representations for analyzing speech information. Unlike the overall performance comparison across hearing-loss groups in Fig. 8, this section focuses on how hearing degradation affects the internal representation of the signal.

As shown in Fig. 10, the STM representations are derived from the same clean speech, namely the sentence "I decided to do it this morning" in scene S09475. Listeners L0235, L0208, and L0215 correspond to Mild, Moderate, and Severe hearing loss, respectively, as characterized by their audiograms. The STM visualizations on the right are used to observe the representational changes under different hearing conditions. In the figure, the time region containing the target words is locally enlarged, and only the STM structures corresponding to ERB-scale channels with center frequencies below 8000 Hz are displayed. The STM representations generated with the proposed method are obtained from the auditory system simulation of the same sentence under different hearing-loss conditions. Therefore,

the figure reflects the effect of hearing loss on the auditory system rather than differences in speech content.

Overall, STM representations provide an effective framework for characterizing the effects of hearing loss on speech information. Because the proposed method continues to simulate auditory system processing from the peripheral stage to higher-level modulation representation, it can capture distortion patterns directly related to speech intelligibility, rather than relying solely on global signal similarity. This also helps explain why the proposed method achieves stable performance improvements across different hearing-loss conditions, as shown in Fig. 8. With the proposed method, STM features serve as the core representation, linking auditory information to the prediction method and enabling the method to learn distortion structures associated with hearing conditions. This representation-based modeling strategy improves method robustness and generalization across different hearing-loss conditions.

VIII. CONCLUSION AND FUTURE WORK

We proposed a speech intelligibility prediction method for hearing-impaired listeners, inspired by auditory information across peripheral and central stages of the auditory system. At the peripheral stage, the method simulates hearing-loss degradations, including elevated hearing thresholds, reduced nonlinear compression, and loss of frequency selectivity. At the central stage, the method further incorporates higher-level auditory information relevant to speech understanding through neural representation learning, binaural integration, and STM analysis. By combining simulation of the auditory peripheral stage with analysis of auditory central representations, the proposed method provides a unified framework for predicting speech intelligibility under hearing-loss conditions.

Improved overall prediction performance. Experimental results on the CPC2 dataset indicate that the proposed method achieved strong overall performance, ranking second among all compared methods and first among the intrusive methods. Compared with the baseline method, be-HASPI, the proposed method increased ρ by 16.4% and reduced the RMSE by 12.2%. These results indicate that incorporating auditory information across multiple stages is effective for speech intelligibility prediction. They also suggest that integrating the auditory peripheral stage and the auditory central stage provides a useful basis for characterizing listener correctness.

Multi-stage auditory information processing under hearing-loss conditions. Beyond the overall result, the proposed method consistently outperformed be-HAPSI across all hearing-loss groups, showing that its advantage remained stable under Mild, Moderate, and Severe hearing-loss conditions. The proposed method reduced the RMSE by 12.8% in the Mild group, 17.2% in the Moderate group, and 8.6% in the Severe group. The largest improvement was observed in the Moderate group, suggesting that combining hearing-loss modeling across both peripheral and central stages improves the prediction method's adaptability across different hearing conditions. These findings indicate that the proposed method not only improves overall prediction performance but also provides a more consistent characterization of listener correctness under different degrees of hearing loss. Nevertheless, the error analysis still showed a mild compression effect toward the mid-score range, indicating that further improvement is needed for samples with very low or very high correctness scores.

Despite these improvements, this study still has several limitations. First, the current prediction method remains intrusive and relies on clean reference speech, which limits its direct applicability in practical scenarios. Second, although the proposed method models auditory information from the periphery stage to the central stage, it remains a computational approximation of the auditory system processing rather than a complete physiological method. Third, the heterogeneous nature of STM representations, normalized cross-correlation features, and binaural representations means that simple feature combination does not always lead to further improvement, suggesting that more effective integration of multi-source auditory information is still needed.

For future work, we will investigate how speech intelligibility prediction methods can better integrate complementary information from different auditory representations, especially in practical settings where clean reference signals are unavailable. We will also further refine the computational approximation of auditory information in the auditory system and evaluate the generalization of the proposed method under broader acoustic conditions and more diverse hearing-loss profiles.

REFERENCES

[1] D. S. Powell, E. S. Oh, N. S. Reed, F. R. Lin, and J. A. Deal, "Hearing loss and cognition: What we know and where we need to go," *Frontiers Aging Neurosci.*, vol. 13, Feb. 2022, Art. no. 769405.

[2] D. Tsimpida, E. Kontopantelis, D. M. Ashcroft, and M. Panagioti, "The dynamic relationship between hearing loss, quality of life, socioeconomic position and depression and the impact of hearing aids: Answers from the English longitudinal study of ageing (ELSA)," *Social Psychiatry Psychiatric Epidemiology*, vol. 57, no. 2, pp. 353–362, Feb. 2022.

[3] S. Ellis, S. Sheik Ali, and W. Ahmed, "A review of the impact of hearing interventions on social isolation and loneliness in older people with hearing loss," *Eur. Arch. Oto-Rhino-Laryngology*, vol. 278, no. 12, pp. 4653–4661, Dec. 2021.

[4] M. E. Sanders, E. Kant, A. L. Smit, and I. Stegeman, "The effect of hearing aids on cognitive function: A systematic review," *PLoS ONE*, vol. 16, no. 12, Dec. 2021, Art. no. e0261207.

[5] L. M. Haile, A. U. Orji, K. M. Reavis, P. S. Briant, K. M. Lucas, F. Alahdab, T. W. Bärnighausen, A. W. Bell, C. Cao, and X. Dai, "Hearing loss prevalence, years lived with disability, and hearing aid use in the United States from 1990 to 2019: Findings from the global burden of disease study," *Ear Hear.*, vol. 45, no. 1, pp. 257–267, 2024.

[6] N. Scarinci, M. Waite, M. Nickbakht, K. Ekberg, B. Timmer, C. Meyer, and L. Hickson, "How do adults with hearing loss, family members, and hearing care professionals respond to the stigma of hearing loss and hearing aids?" *Int. J. Audiology*, vol. 64, no. sup1, pp. S20–S27, Apr. 2025.

[7] J. Barker, M. Akeroyd, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, and Z. Podwinska, "The 1st clarity prediction challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proc. Interspeech*, 2022, pp. 3508–3512.

[8] 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.

[9] E. A. Lopez-Poveda, P. T. Johannesen, P. Pérez-González, J. L. Blanco, S. Kalluri, and B. Edwards, "Predictors of hearing-aid outcomes," *Trends Hearing*, vol. 21, Jan. 2017, Art. no. 2331216517730526.

[10] 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.

[11] 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.

[12] B. Kollmeier and J. Kiessling, "Functionality of hearing aids: State-of-the-art and future model-based solutions," *Int. J. Audiology*, vol. 57, no. sup3, pp. S3–S28, May 2018.

[13] K. Yamamoto, T. Irino, S. Araki, K. Kinoshita, and T. Nakatani, "GEDI: Gammachirp envelope distortion index for predicting intelligibility of enhanced speech," *Speech Commun.*, vol. 123, pp. 43–58, Oct. 2020.

[14] A. Yamamoto, F. Miyazaki, and T. Irino, "Predicting speech intelligibility in older adults for speech enhancement using the gammachirp envelope similarity index, Gesi," in *Proc. Speech Commun.*, Oct. 2025, Art. no. 103318.

[15] C. M. Hackney, "Anatomical features of the auditory pathway from cochlea to cortex," *Brit. Med. Bull.*, vol. 43, no. 4, pp. 780–801, 1987.

[16] A. J. E. Kell, D. L. K. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott, "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy," *Neuron*, vol. 98, no. 3, pp. 630–644.e16, May 2018.

[17] P. W. Hullett, L. S. Hamilton, N. Mesgarani, C. E. Schreiner, and E. F. Chang, "Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli," *J. Neurosci.*, vol. 36, no. 6, pp. 2014–2026, Feb. 2016.

[18] M. Schönwiesner and R. J. Zatorre, "Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 34, pp. 14611–14616, Aug. 2009.

[19] X. Zhou, C. O. Mawalim, and M. Unoki, "Modeling multi-level hearing loss for speech intelligibility prediction," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2025, pp. 1–5.

[20] S. Keshishzadeh, M. Garrett, and S. Verhulst, "Towards personalized auditory models: Predicting individual sensorineural hearing-loss profiles from recorded human auditory physiology," *Trends Hearing*, vol. 25, Jan. 2021, Art. no. 2331216520988406.

[21] A. M. H. Lesicko and D. A. Llano, "Impact of peripheral hearing loss on top-down auditory processing," *Hearing Res.*, vol. 343, pp. 4–13, Jan. 2017.

[22] J. Syka, "Plastic changes in the central auditory system after hearing loss, restoration of function, and during learning," *Physiological Rev.*, vol. 82, no. 3, pp. 601–636, Jan. 2002.

[23] J. Mazelová, J. Popelar, and J. Syka, "Auditory function in presbycusis: Peripheral vs. central changes," *Experim. Gerontology*, vol. 38, nos. 1–2, pp. 87–94, Jan. 2003.

[24] B. D. Auerbach and H. J. Gritton, "Hearing in complex environments: Auditory gain control, attention, and hearing loss," *Frontiers Neurosci.*, vol. 16, Feb. 2022, Art. no. 799787.

[25] A. Presacco, J. Z. Simon, and S. Anderson, "Speech-in-noise representation in the aging midbrain and cortex: Effects of hearing loss," *PLoS ONE*, vol. 14, no. 3, Mar. 2019, Art. no. e0213899.

[26] 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.

[27] X. Zhou, C. O. Mawalim, and M. Unoki, "Speech intelligibility prediction using binaural processing for hearing loss," *IEEE Access*, vol. 13, pp. 25817–25836, 2025.

[28] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 54, no. 2, p. 557, Aug. 1973.

[29] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Amer.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[30] C. Spille, S. D. Ewert, B. Kollmeier, and B. T. Meyer, "Predicting speech intelligibility with deep neural networks," *Comput. Speech Lang.*, vol. 48, pp. 51–66, Mar. 2018.

[31] 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.

[32] 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.

[33] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Amer.*, vol. 130, no. 3, pp. 1475–1487, 2011.

[34] J. J. Rosowski, "The effects of external- and middle-ear filtering on auditory threshold and noise-induced hearing loss," *J. Acoust. Soc. Amer.*, vol. 90, no. 1, pp. 124–135, Jul. 1991.

[35] A. Strouse, D. H. Ashmead, R. N. Ohde, and D. W. Grantham, "Temporal processing in the aging auditory system," *J. Acoust. Soc. Amer.*, vol. 104, no. 4, pp. 2385–2399, Oct. 1998.

[36] J. R. Dubno and A. B. Schaefer, "Frequency selectivity for hearing-impaired and broadband-noise-masked normal listeners," *Quart. J. Experim. Psychol. Sect. A*, vol. 43, no. 3, pp. 543–564, Aug. 1991.

[37] M. Florentine, S. Buus, B. Scharf, and E. Zwicker, "Frequency selectivity in normally-hearing and hearing-impaired observers," *J. Speech, Lang., Hearing Res.*, vol. 23, no. 3, pp. 646–669, Sep. 1980.

[38] B. C. 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.

[39] 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.

[40] S. Haro, C. J. Smalt, G. A. Ciccarelli, and T. F. Quatieri, "Deep neural network model of hearing-impaired speech-in-noise perception," *Frontiers Neurosci.*, vol. 14, Dec. 2020, Art. no. 588448.

[41] A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty, "Speech intelligibility prediction using spectro-temporal modulation analysis," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 29, pp. 210–225, 2021.

[42] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 131, no. 5, pp. 4134–4151, May 2012.

[43] M. R. Schädler and B. Kollmeier, "Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 137, no. 4, pp. 2047–2059, 2015.

[44] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–12.

[45] 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.

[46] 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. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[47] D. Schröder and M. Vorländer, "RAVEN: A real-time framework for the auralization of interactive virtual environments," in *Proc. Forum Acusticum*, vol. 27, pp. 1541–1546, 2011.

[48] Z. Tu, J. Zhang, N. Ma, and J. Barker, "A two-stage end-to-end system for speech-in-noise hearing aid processing," in *Proc. Clarity Workshop Mach. Learn. Challenges Hearing Aids (Clarity)*, Sep. 2021, pp. 3–5.

[49] S. Cornell, Z.-Q. Wang, Y. Masuyama, S. Watanabe, M. Pariente, and N. Ono, "Multi-channel target speaker extraction with refinement: The WavLab submission to the second clarity enhancement challenge," 2023, *arXiv:2302.07928*.

[50] S. Cuervo and R. Marxer, "Temporal-hierarchical features from noise-robust speech foundation models for non-intrusive intelligibility prediction," in *Proc. 4th Clarity Workshop Mach. Learn. Challenges Hearing Aids (Clarity)*, Aug. 2023, pp. 17–19.
[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. 21, 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. 4th Clarity Workshop Mach. Learn. Challenges Hearing Aids (Clarity)*, Aug. 2023, pp. 9–11.
[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. E. Zezario, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based speech intelligibility prediction model by incorporating whisper for hearing aids," in *Proc. 4th Clarity Workshop Mach. Learn. Challenges Hearing Aids (Clarity)*, Aug. 2023, pp. 4–6.

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), Indonesia, and the M.S. and Ph.D. degrees from the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), in 2022. Since April 2022, she has been a Faculty Member with the School of Information Science, JAIST, where she is currently a Senior Lecturer. Her main research interests include speech signal processing, hearing perception, voice privacy protection, and machine learning. She was selected as a Japan Society for the Promotion of Science (JSPS) Research Fellow for Young Scientists (DC1), from 2020 to 2022. She serves on the education team for 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 associated with ATR Human Information Processing Laboratories as a Visiting Researcher, from 1999 to 2000, and then a Visiting Research Associate with the Centre for the Neural Basis of Hearing (CNBH), Department of Physiology, University of Cambridge, from 2000 to 2001. He has been a Faculty Member with the School of Information Science, JAIST, since 2001, where he is currently a Professor. His research interests include auditory-motivated signal processing and the modeling of auditory systems. He was a Japan Society for the Promotion of Science (JSPS) Research Fellow, from 1998 to 2001. He is a member of the Research Institute of Signal Processing (RISP), the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Acoustical Society of America (ASA), the Acoustical Society of Japan (ASJ), and the International Speech Communication Association (ISCA). He received the Sato Prize for an Outstanding Paper from ASJ, in 1999, 2010, and 2013, as well as the Yamashita Taro "Young Researcher" Prize from the Yamashita Taro Research Foundation, in 2005.