

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/189605/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Yang, Yang, Liang, Ruiyu, Ni, Ye, Zou, Cairong, Sun, Song, Yan, Jing, Zhou, Wei, Ding, Weiping and Hao, Xiaoshuai 2027. BiSQA Fusion: Multi-level fusion for personalized binaural speech quality assessment in hearing aids. Information Fusion: An International Journal on Multi-Sensor, Multi-Source Information Fusion 139 (Part A) , 104765. 10.1016/j.inffus.2026.104765

Publishers page: <https://doi.org/10.1016/j.inffus.2026.104765>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



BiSQA Fusion: Multi-level fusion for personalized binaural speech quality assessment in hearing aids

Yang Yang^a, Ruiyu Liang^{a,b,*}, Ye Ni^a, Cairong Zou^a, Song Sun^c, Jing Yan^d, Wei Zhou^e, Weiping Ding^{f,g}, Xiaoshuai Hao^{h,*}

^a School of Information Science and Engineering, Southeast University, Nanjing, 210096, China

^b School of Communication Engineering, Nanjing Institute of Technology, Nanjing, 211167, China

^c School of Computer and Information Science, Chongqing Normal University, Chongqing, 401331, China

^d Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China

^e Cardiff University, Cardiff, United Kingdom

^f School of Artificial Intelligence and Computer Science, Nantong University, Nantong, 226019, China

^g Faculty of Data Science, City University of Macau, Macau, 999078, China

^h Xiaomi EV, Beijing, 100085, China

ARTICLE INFO

Keywords:

Hearing aids

Cross-modal fusion modulation

Acoustic and audiogram fusion

Binaural speech quality assessment

Self-supervised learning

ABSTRACT

Accurate speech quality assessment is essential for optimizing hearing aid (HA) algorithms, enabling personalized fitting, and improving user satisfaction. Existing HA assessment models, however, are largely limited to monaural processing, lack in-depth cross-modal fusion between acoustic features and audiogram priors, and exhibit insufficient robustness across diverse HA configurations. To address these challenges, we propose BiSQA Fusion, a multi-level fusion network that jointly integrates multi-layer self-supervised speech representations, individual audiogram priors, interaural interaction cues, and HA system identifiers for personalized non-intrusive binaural HA speech quality assessment. At its core, a hierarchical two-stage acoustic-audiogram fusion module is designed to adaptively inject personalized audiogram priors into dynamic acoustic features by coupling gated feature-wise linear modulation with a deep CNN-BiLSTM backbone. To overcome the inherent limitations of monaural assessment, a lightweight binaural fusion module is proposed to explicitly decouple interaural coherence-, asymmetry-, and correlation-like cues via parameter-free algebraic operations. Furthermore, systematic prediction biases arising from diverse HA processing configurations are effectively mitigated by a configuration-aware System ID embedding. An auxiliary quality-level module is formulated to impose coarse-grained classification constraints on continuous score prediction via multi-task learning. Extensive experiments on the CPC1 and DNS challenge datasets demonstrate that BiSQA Fusion significantly outperforms state-of-the-art non-intrusive baselines in both prediction accuracy and cross-scenario generalization.

1. Introduction

Hearing loss has emerged as a major global public health challenge. According to the World Health Organization (WHO), more than 700 million people are projected to require hearing rehabilitation services by 2050 [1]. Untreated hearing loss is closely associated with cognitive decline, social isolation, and mental health disorders [2,3]. Among available hearing assistive technologies, hearing aids (HAs) remain the most widely adopted intervention due to their high accessibility, fitting flexibility, and cost-effectiveness [4,5]. HAs compensate for

hearing loss through personalized signal processing algorithms, where the output speech quality directly dictates user satisfaction and long-term wearing compliance for hearing-impaired (HI) patients. Consequently, accurate and efficient assessment of HA speech quality is not merely central to evaluating fitting success, but also a fundamental technological prerequisite for the iterative optimization of HA algorithms.

Conventional HA quality assessment relies heavily on large-scale subjective listening tests conducted by audiologists, with representative metrics including the mean opinion score (MOS), degradation MOS, and diagnostic acceptability measure (DAM) [6]. Such procedures are,

* Corresponding authors.

E-mail addresses: yangy@seu.edu.cn (Y. Yang), 101101220@seu.edu.cn (R. Liang), niye@seu.edu.cn (Y. Ni), cairong@seu.edu.cn (C. Zou), sunsong@cqu.edu.cn (S. Sun), jennifer.yan@connect.polyu.hk (J. Yan), zhou26@cardiff.ac.uk (W. Zhou), dwp9988@163.com (W. Ding), haoxiaoshuai@xiaomi.com (X. Hao).

<https://doi.org/10.1016/j.inffus.2026.104765>

Received 17 May 2026; Received in revised form 10 August 2026; Accepted 5 September 2026

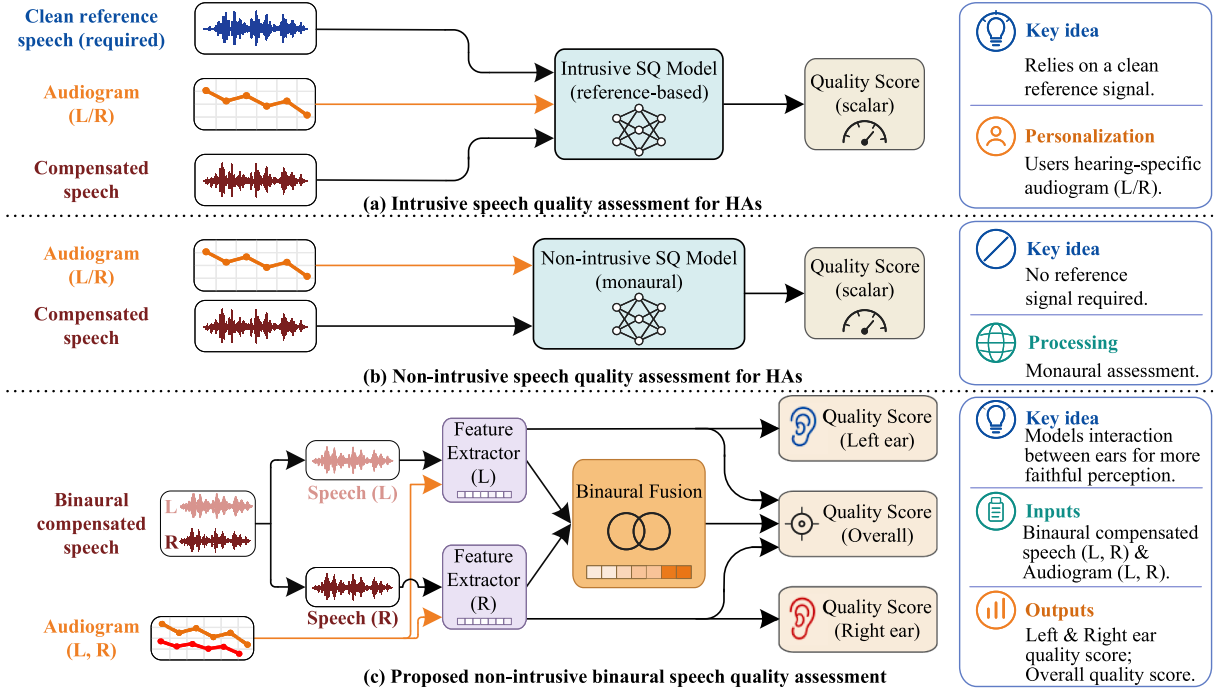


Fig. 1. Examples of HA speech quality assessment. (a) Intrusive: Assessment metrics rely on a clean reference signal and evaluate each ear independently; (b) Non-intrusive: Eliminates the reliance on a reference signal but still evaluates each ear in isolation; (c) Parallel processing of binaural channels: Not only predicts the independent quality of each ear but also outputs a comprehensive overall quality score.

however, prohibitively costly, time-consuming, and individually variable, motivating the development of objective alternatives that can be broadly categorized into intrusive and non-intrusive methods [7]. Intrusive metrics, as illustrated in Fig. 1(a), assess quality by quantifying the discrepancies between the compensated signal and the clean reference [8–11], yet such a reference is virtually unattainable in real-world wearing scenarios. Furthermore, since the passage of the 2017 U.S. Over-the-Counter Hearing Aid Act, consumers with mild-to-moderate hearing loss can purchase HAs directly without an audiologist’s prescription [12]. This trend further reinforces the urgent need for non-intrusive quality assessment methods that operate solely on HA output signals and individual audiogram priors, as shown in Fig. 1(b).

Recent non-intrusive HA speech quality assessment has shifted from extending normal-hearing (NH) metrics to deep-learning-based end-to-end regression models, in which pre-trained self-supervised learning (SSL) models, such as WavLM [13], have shown strong promise in this task by providing rich speech representations and complementary multi-layer hidden features [14]. From an information fusion perspective, however, current non-intrusive HA assessment pipelines still suffer from three structural limitations. First, audiogram and acoustic features are typically combined through shallow operations such as concatenation, replication, or simple addition [4,12,15,16], while existing SSL-based methods predominantly rely on the static final-layer output. This limits their capacity to model how band-specific hearing loss dynamically distorts speech perception across hierarchical acoustic abstractions. To this end, we adopt the learnable, layer-wise aggregation strategy that integrates multi-layer SSL representations, subsequently refined by an SSL-Encoder to preserve perceptually relevant cues while minimizing redundancy. In parallel, an HL-Encoder projects a sparse audiogram vector into high-dimensional, personalized auditory-prior embeddings. These encoded representations are then fused by the Acoustic-Audiogram Fusion Module (AAFM), which combines gated, feature-wise linear modulation with deep CNN-BiLSTM integration, thereby injecting personalized audiogram priors into the SSL acoustic features. Second, every existing non-intrusive HA model is restricted to monaural processing [4,12,15–19]: preliminary binaural quality work in the

general audio domain targets NH listeners only [20–23], and binaural intelligibility studies in the HA domain are confined to the better-ear effect [24–26], leaving comprehensive binaural quality assessment tailored to HA scenarios unresolved, as illustrated in Fig. 1(c), a gap we close with a lightweight binaural fusion module (BFM) that explicitly decouples three perceptually motivated interaural components via parameter-free algebraic operations. Third, the diversity of prescription formulas (e.g., NAL-R, DSL v5) or speech enhancement pipelines induces substantial spectral distribution shifts, causing systematic prediction biases when training data span multiple HA configurations [16], a bias we mitigate by tokenizing each HA system configuration as an explicit System ID embedding to model configuration-dependent shifts. These coupled limitations collectively expose the absence of a systematic, multi-source heterogeneous information fusion design in current HA quality assessment models.

Motivated by the aforementioned challenges, we propose BiSQAFusion, a multi-level information fusion network that jointly integrates multi-layer SSL representations, individual audiogram priors, interaural interaction cues, and HA System IDs for personalized non-intrusive binaural HA speech quality assessment. The network processes binaural channels in parallel through weight-shared encoders, fuses cross-modal information at both unilateral and binaural levels via the proposed AAFM and BFM modules, and outputs both unilateral and overall quality scores. To further regularize the fused representation within the score prediction module (SPM), an auxiliary quality-level module (QLM) is introduced to impose coarse-grained classification boundaries through multi-task learning (MTL) [27]. The main contributions of this paper are summarized as follows:

- We propose BiSQAFusion, a systematic multi-level fusion network for non-intrusive binaural HA speech quality assessment, significantly extending the applicability and generalization of objective assessment models across diverse hearing-loss profiles and HA processing configurations.
- We design a hierarchical AAFM that couples gated FiLM with CNN-BiLSTM to bridge the semantic and structural gap between

dynamic SSL representations and static audiograms. Complementarily, a lightweight BFM is developed to decouple perceptually motivated interaural components, namely coherence-, asymmetry-, and correlation-like cues, providing an interpretable inductive bias for comprehensive binaural assessment.

- We introduce a configuration-aware System ID embedding strategy that effectively mitigates the systematic prediction biases induced by configuration diversity and significantly enhances cross-configuration prediction robustness.
- We formulate a coarse-to-fine joint optimization framework in which the QLM, as an auxiliary task, imposes classification boundary constraints on the primary SPM under MTL, achieving synergistic optimization and superior continuous-score prediction.

The remainder of this paper is organized as follows: [Section 2](#) reviews related work. [Section 3](#) elaborates on the overall architecture and functional modules of the proposed network. [Section 4](#) describes the experimental setup. [Section 5](#) presents in-depth analyses and ablation studies. [Section 6](#) concludes the paper and discusses future directions.

2. Related work

2.1. Objective hearing aid speech quality metrics

Objective HA quality metrics fall into intrusive and non-intrusive categories according to their dependence on a clean reference [7]. Kates and Arehart proposed HASQI [8], which embeds individual audiograms into a peripheral auditory model to simulate quality perception, and its robustness has been extensively validated by Kressner et al. [28,29]. HASQI Version 2 [9] (hereafter HASQI) further refines the cochlear model for complex degradations across HI and NH listeners, computing linear (long-term spectral distortion) and nonlinear (time-frequency envelope modulation and fine-structure changes) components whose product yields a score in [0, 1]. HAAQI [10] targets music quality, while PEMO-Q-HI [11] extends PEMO-Q to HI listeners. Nevertheless, the requirement of clean references and strict time-frequency alignment fundamentally precludes their deployment in realistic wearing scenarios.

These limitations have motivated non-intrusive HA speech quality assessment, which has evolved through two phases. The first phase injected hearing loss characteristics into NH-oriented metrics, such as SRMR-HA [30], LCQA-HA [31], and gammatone filterbank energies (FBEs) coupled with regression mapping [32], yet these pipelines suffer from handcrafted feature dependence and high computational cost [16]. The second phase shifted toward end-to-end deep learning. Although early models such as Quality-Net [33] and MOS-Net [34] were originally designed for NH-oriented applications, they provided insight into HA assessment. Specifically tailored for HA scenarios, Chiang et al. proposed HASA-Net [15], which integrates a BiLSTM with multi-head attention (MHA) within an MTL framework to simultaneously predict HASQI and the HA Speech Perception Index (HASPI) [35]. Subsequent work [17] incorporated SSL representations to further enhance feature extraction. Other notable advancements include a DNN-LSTM attention ensemble model based on Log-Mel Filter Bank (LMFB) energy and audiograms [18], an MTL framework utilizing Gammatone FBEs [4], a joint magnitude spectrum-audiogram architecture integrating CNN-BiLSTM with MHA [12], a bias mitigation strategy via speech label assistance [16], and HAAQI-Net, which focuses on music quality assessment [19]. Despite these advancements, these methods remain confined to monaural processing, fuse audiograms and acoustic features through shallow operations, and fail to sufficiently account for configuration-induced prediction biases.

2.2. Binaural speech quality assessment

While HA quality assessment has been confined to monaural approaches, binaural evaluation has been preliminarily explored in the

general audio domain. Schäfer et al. [20] extended the perceptual evaluation of audio quality with a binaural hearing model. Reimes [21] employed a Siamese CNN guided by psychoacoustic priors for binaural MOS prediction. Fleßner et al. [22] proposed BAM-Q based on interaural level and time differences. Their subsequent work [23] demonstrated that overall perceived quality is dominated by the poorer of the signal quality or binaural spatial quality. These methods, however, target NH listeners and do not incorporate individualized audiograms. Within the HA domain, binaural objective evaluations have focused heavily on speech intelligibility. Andersen et al. [24] extended Short-Time Objective Intelligibility (STOI) to modified binaural STOI (MBSTOI), modeling the better-ear effect via window-wise reference correlation. The Clarity Prediction Challenge (CPC) series [25,26] further catalyzed this direction, with systems predominantly adopting the maximum unilateral score as the overall binaural intelligibility score. Intelligibility, however, does not capture quality attributes such as signal distortion, leaving comprehensive binaural HA quality assessment with personalized audiogram priors a largely unexplored problem.

2.3. Cross-modal fusion for personalized speech quality assessment

Effective integration of audiograms and acoustic representations underpins personalized HA speech quality assessment. Existing studies predominantly treat audiograms as static features and fuse them with acoustic representations using shallow operations such as concatenation, replication, or simple addition at the input layer [4,12,15,16], which fails to model how band-specific hearing loss dynamically distorts speech across hierarchical acoustic abstractions. Beyond the HA domain, feature-wise linear modulation (FiLM) [36] has demonstrated strong cross-modal conditioning capabilities in speech synthesis [37] and style transfer [38], generating channel-wise scaling and shifting parameters from auxiliary inputs. Yet, applying FiLM in isolation embeds audiogram information only implicitly and is prone to dilution during deep propagation. Moreover, no prior work has systematically coupled FiLM-based modulation with SSL representations for deep acoustic-audiogram integration.

In summary, the three lines of research reviewed above expose three limitations in current non-intrusive HA speech quality assessment. First, shallow acoustic-audiogram fusion on static SSL output fails to exploit hierarchical SSL complementarity for personalized perception. Second, monaural-only modeling neglects interaural cues essential to HI listeners. Third, systematic biases induced by heterogeneous HA configurations remain effectively unaddressed. These three limitations collectively motivate a unified multi-level fusion network that hierarchically integrates SSL representations, audiogram priors, interaural interaction cues, and HA system identifiers within an end-to-end pipeline, which we instantiate as the proposed BiSQA-Fusion in [Section 3](#).

3. Proposed methods

3.1. Overview

The proposed non-intrusive binaural HA speech quality assessment network, BiSQA-Fusion, aims to construct an end-to-end mapping function $F_\theta : (S, HL) \mapsto (\widehat{QS}_{bin}, \widehat{QS}_L, \widehat{QS}_R)$, which takes binaural degraded speech $S \in \mathbb{R}^{2 \times Len}$ and binaural audiogram $HL \in \mathbb{R}^{2 \times N}$ as inputs to predict comprehensive binaural $\widehat{QS}_{bin}$ and unilateral quality scores $\{\widehat{QS}_L, \widehat{QS}_R\}$. Here, Len denotes the sampling length and N represents the number of tested frequency points. As illustrated in [Fig. 2](#), BiSQA-Fusion comprises five functional modules: (1) An input feature encoding module consisting of an SSL-Encoder and a hearing loss encoder (HL-Encoder). The former reduces the dimensionality of SSL-weighted representations to focus on key features sensitive to quality assessment, while the latter transforms sparse, low-dimensional hearing threshold curves into a high-dimensional personalized auditory prior. (2) The

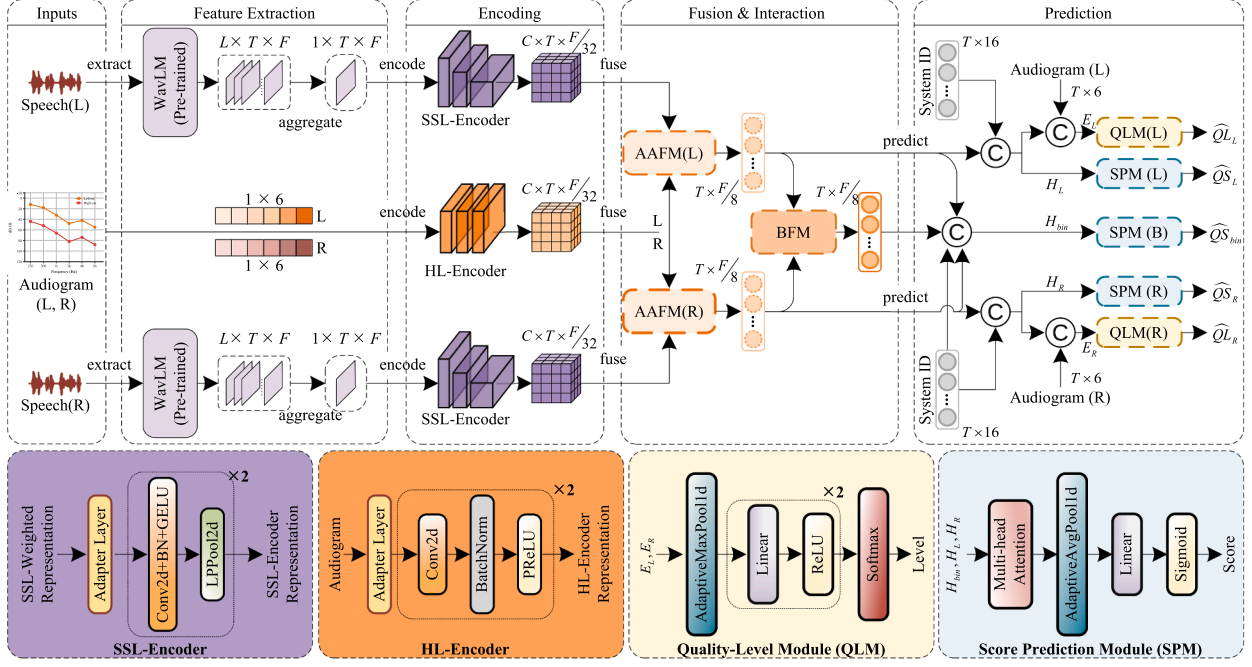


Fig. 2. Schematic diagram of binaural speech quality assessment network.

AAFM bridges the cross-modal fusion gap between post-coded acoustic features and a priori audiogram information through a two-stage fusion process: a shallow gated FiLM fusion stage and a deep CNN-BiLSTM fusion stage, thereby achieving deep interaction between acoustic features and audiogram. (3) The BFM is designed to capture interaction cues for predicting overall binaural speech quality. (4) The SPM maps high-dimensional features to scalar quality scores. (5) The QLM serves as an auxiliary supervised task, optimized in conjunction with the SPM through an MTL mechanism.

3.2. Input features and encoded representations

This paper employs the pre-trained WavLM Large [13] model as the acoustic feature extractor. Here, the representations for the left and right ear signals are extracted independently. To adaptively mine phonetic representations highly correlated with quality assessment from multi-layer SSL features, a learnable, layer-wise, weighted-aggregation strategy is adopted. Taking the unilateral signal as an example, the HA degraded speech is transformed into multi-layer feature representations $S \in \mathbb{R}^{L \times T \times F}$, where $L = 25$, T , and F denote the number of layers, temporal frames, and feature dimensions, respectively, with $F = 1024$. The specific aggregation process is formulated as:

$$w_l = \frac{\exp(\lambda_l)}{\sum_{k=0}^{L-1} \exp(\lambda_k)}, \quad (1)$$

$$Z = \sum_{l=0}^{L-1} w_l S_l,$$

where $\lambda \in \mathbb{R}^L$ represents the learnable parameter vector, denoted as $\lambda = [\lambda_0, \lambda_1, \dots, \lambda_{L-1}]$. w_l denotes the normalized attention weight for the l -th layer after softmax activation. $S_l \in \mathbb{R}^{T \times F}$ denotes the l -th layer output. The aggregated representation $Z \in \mathbb{R}^{T \times F}$ is fed into the SSL-Encoder, comprising a fully connected (FC) adapter layer, two cascaded 2D convolutional blocks, and two L_p pooling layers, formulated as:

$$SSL_0 = \mathcal{R}_{ssl}(Z \cdot W_a + b_a), \quad (2)$$

$$SSL_i = LPPool_i(\delta_i(B_i(Convd2(SSL_{i-1}; \theta_i); p))), \quad i \in \{1, 2\}$$

where $\mathcal{R}_{ssl}$ represents the reshape and dimension reduction operation applied to the SSL features. W_a and b_a refer to the weights and bias of the FC layer, respectively. $LPPool_i(\cdot; p)$, δ_i , B_i , and $Convd2(\cdot; \theta_i)$ denote L_p pooling with $p = 2$, GELU activation, batch normalization (BN), and convolution operations of the i -th layer, respectively.

As prior information reflecting the individual degree of hearing loss in HI patients, the audiogram consists of specific frequency hearing thresholds obtained through standard pure-tone audiometry. The input audiogram $HL \in \mathbb{R}^{1 \times 6}$ utilized in this study covers six key octave-band frequency points ranging from 250 Hz to 8 kHz. To bridge the modal gap between low-dimensional discrete hearing thresholds and high-dimensional SSL representations, the HL-Encoder applies an FC adapter followed by two convolutional blocks:

$$A_0 = \mathcal{R}_{hl}(HL \cdot W_a^{(hl)} + b_a^{(hl)}), \quad (3)$$

$$A_j = \phi_j(B_j(Convd2(A_{j-1}; \theta_j^{(hl)}))), \quad j \in \{1, 2\}$$

where $\mathcal{R}_{hl}$ denotes the reshape and dimension expansion operations applied to HL by the FC adapter layer. $W_a^{(hl)}$ and $b_a^{(hl)}$ are the parameters of the FC layer. ϕ_j , B_j , and $Convd2(\cdot; \theta_j^{(hl)})$ denote the PReLU activation, BN, and convolution operations of the j -th layer, respectively.

3.3. Acoustic-audiogram fusion module

To achieve deep cross-modal integration of the encoded SSL representations and the audiograms, a hierarchical two-stage AAFM is designed. As detailed in Fig. 3, a gated FiLM stage is followed by CNN-BiLSTM deep fusion. Before modulation, the encoded audiogram representation is repeated along the time axis to match the temporal dimension of the SSL features, yielding $\chi^{(A)} \in \mathbb{R}^{C \times T \times \tilde{F}}$.

Gated FiLM fusion stage, the audiogram representation $\chi^{(A)}$ dynamically modulates the SSL features $\chi^{(S)} \in \mathbb{R}^{C \times T \times \tilde{F}}$ via channel-wise affine transformation with adaptive gating:

$$\{\gamma, \beta\} = \{\chi^{(A)} \cdot W_\gamma + b_\gamma, \chi^{(A)} \cdot W_\beta + b_\beta\}, \quad (4)$$

$$g = \sigma(\chi^{(A)} \cdot W_g + b_g),$$

$$\chi^{(F)} = g \odot (\gamma \odot \chi^{(S)} + \beta) + (1 - g) \odot \chi^{(S)},$$

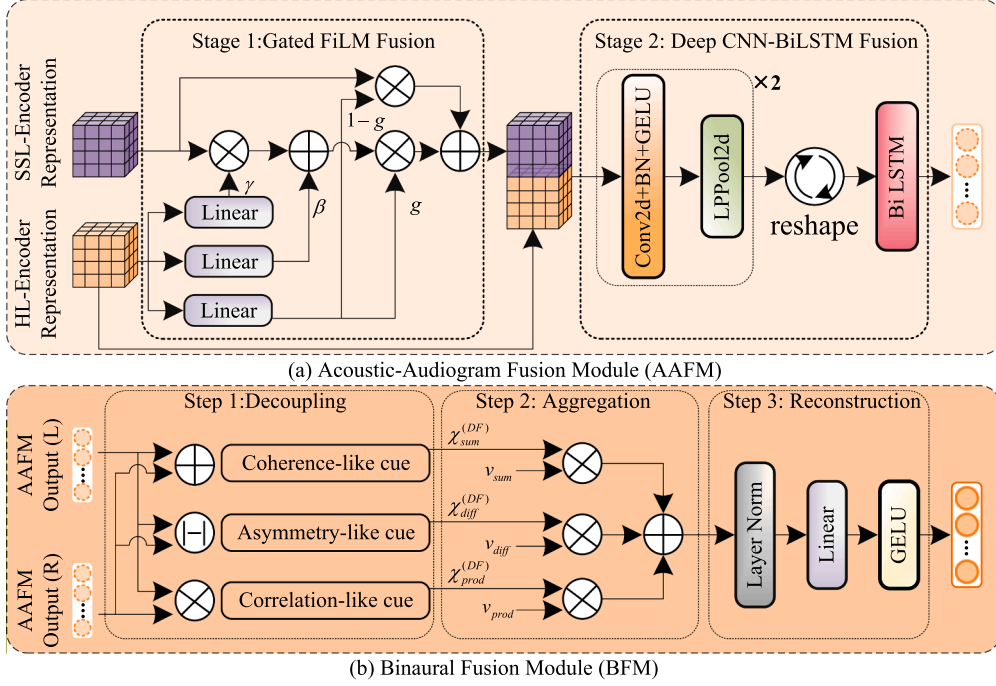


Fig. 3. (a) The network architecture of AAFM; (b) The network architecture of BFM.

where $\gamma, \beta \in \mathbb{R}^{C \times T \times \bar{F}}$ are the scaling and shifting parameters linearly mapped from the audiogram features, respectively. $\sigma(\cdot)$ denotes the sigmoid activation function, which squashes the gate pre-activation into the $(0, 1)$ interval so that g acts as a continuous mixing coefficient. $\odot$ denotes the Hadamard product. The gating mechanism enables adaptive balance between preserving original acoustic representations ($g \rightarrow 0$) and incorporating personalized hearing loss information ($g \rightarrow 1$), thereby avoiding information distortion caused by excessive modulation.

CNN-BiLSTM deep fusion stage. To further capture the high-dimensional correlations between the modulated acoustic features and personalized hearing loss profiles, the modulated feature $\chi^{(F)}$ and the encoded audiogram priors $\chi^{(A)}$ are concatenated along the channel dimension and fed into a deep fusion architecture formed by cascaded CNN blocks and a BiLSTM network. The forward process is expressed as:

$$\begin{aligned} \kappa^{(0)} &= [\chi^{(F)}; \chi^{(A)}]_{\bar{F}}, \\ \kappa^{(k)} &= LPPool_k(\delta_k(B_k(Conv2D(\kappa^{(k-1)}; \theta_k))); p), \quad k \in \{1, 2\} \\ \tilde{\kappa} &= [\tilde{\kappa}_t]_{t=1}^T = \mathcal{R}_s(\kappa^{(2)}), \\ \chi_t^{(DF)} &= [\overleftarrow{LSTM}(\tilde{\kappa}_t, \tilde{h}_{t-1}); \overrightarrow{LSTM}(\tilde{\kappa}_t, \tilde{h}_{t+1})], \end{aligned} \quad (5)$$

where $[\cdot; \cdot]_{\bar{F}}$ denotes concatenation along the feature dimension $\bar{F}$, doubling the feature axis of the convolved maps while preserving the channel and temporal dimensions. $\kappa^{(0)}$ is the initial concatenated input. $LPPool_k(\cdot; p)$, δ_k , B_k , and $Conv2D(\cdot; \theta_k)$ denote L_p pooling, GELU activation, BN, and convolution operations of the k -th layer, respectively. $\mathcal{R}_s$ acts as a sequence reconstruction operation, flattening the convolved 2D feature maps into a frame-level sequence $\tilde{\kappa}$. $\overleftarrow{LSTM}$ and $\overrightarrow{LSTM}$ denote the forward and backward LSTM units along the time axis. The output $\chi^{(DF)} \in \mathbb{R}^{T \times D}$ integrates personalized audiogram priors with deep spatiotemporal acoustic abstractions.

3.4. Binaural fusion module

Interaural interaction cues are essential for overall binaural speech quality assessment, as they not only encode bilateral auditory

confirmation of the same source but also reflect listener-specific binaural processing characteristics in HI populations. According to classical binaural psychoacoustics [39–41], the perception of binaural speech quality emerges from three distinct cues: the shared spectral energy captured by both ears (coherence), the interaural intensity disparity induced by asymmetric hearing loss (asymmetry), and the nonlinear interaural coupling that encodes fine-grained phase and envelope correlations (correlation). However, existing binaural fusion strategies either average the two channels or naively concatenate them along the feature axis, both of which entangle these three cues into an opaque representation and limit the discriminability of the fused feature. To explicitly disentangle the corresponding components in the learned feature space, we design the BFM with a three-step decouple-aggregate-reconstruct workflow, as illustrated in Fig. 3, which serves as a perceptually motivated inductive bias.

In Step 1 (Decoupling), three algebraic operations are applied to the per-channel deep representations $\chi_L^{(DF)}$ and $\chi_R^{(DF)}$. The element-wise sum $\chi_{sum}^{(DF)} = \chi_L^{(DF)} + \chi_R^{(DF)}$ captures the bilaterally consistent coherent component and serves as a noise-robust reference component. The absolute difference $\chi_{diff}^{(DF)} = |\chi_L^{(DF)} - \chi_R^{(DF)}|$ explicitly encodes interaural asymmetry, which is particularly informative under asymmetric hearing loss or unbalanced binaural HA processing. The Hadamard product $\chi_{prod}^{(DF)} = \chi_L^{(DF)} \odot \chi_R^{(DF)}$ encodes the nonlinear coupling between the fine-grained envelopes of the two ears. Crucially, all three operations are parameter-free and perceptually motivated, assigning well-defined psychoacoustic semantics to the decoupled components rather than leaving them as learned yet opaque abstractions. Note that these operations are analogous to physical measurements. They act on learned deep features and serve as inductive biases.

Decoupling alone, however, is insufficient for effective fusion, since the perceptual salience of each cue varies across listeners, hearing loss profiles, and acoustic scenarios. An adaptive reweighting mechanism is required. To this end, in Step 2 (Gated Aggregation), we introduce a learnable channel-level gating mechanism. Each gate vector $v_d = \sigma(w_d)$ ($d \in \{sum, diff, prod\}$) is independently parameterized and optimized end-to-end, allowing the network to dynamically suppress unreliable cues (e.g., $\chi_{diff}^{(DF)}$ under bilateral NH cases where interaural asymmetry

is negligible) while amplifying perceptually dominant ones. The aggregated representation is then obtained as:

$$F_{bin} = \sum_{d \in \{sum, diff, prod\}} v_d \otimes \chi_d^{(DF)}, \quad (6)$$

where $\otimes$ denotes broadcast multiplication along the feature dimension.

In Step 3 (Reconstruction), finally, the aggregated binaural unified representation F_{bin} is fed into an enhanced network comprising layer normalization (LN) and FC layers to achieve nonlinear mapping and information reorganization of features within a deep semantic space. The detailed process is formulated as follows:

$$\chi_{bin}^{(DF)} = Dropout\left(\delta(LN(F_{bin}) \cdot W_e + b_e)\right), \quad (7)$$

where δ , W_e , and b_e denote the GELU activation, weight matrix, and bias vector of the FC projection layer, respectively. $\chi_{bin}^{(DF)}$ represents the final fused binaural feature.

3.5. System ID embedding and score prediction module

To mitigate prediction biases arising from HA processing configurations, each configuration is tokenized as an integer System ID, projected into a 16-dimensional embedding χ_{ID} , broadcast along the time axis, and concatenated with the feature streams entering the SPM and QLM. This explicit tokenization enables the network to disentangle configuration-specific spectral patterns from the underlying perceptual quality, thereby improving robustness when training data span heterogeneous HA processing pipelines. During inference, a configuration absent from training retains its initialized embedding and thus acts as an unknown token, while the shared acoustic pathway, which is disentangled from configuration bias during training, still benefits from the embeddings of the seen configurations.

The SPM maps high-dimensional features to the overall binaural quality score, along with the independent left and right ear quality scores. As depicted in Fig. 2, the SPM comprises an MHA layer, an Adaptive Average Pooling (AAP) layer, and an FC layer. For the overall binaural quality prediction, the input is constructed by concatenating the binaural representations with the unilateral features and the System ID along the feature dimension, $H_{bin} = [\chi_L^{(DF)}; \chi_R^{(DF)}; \chi_{bin}^{(DF)}; \chi_{ID}]$. Similarly, the inputs for the independent left and right ear predictions are formulated as: $H_L = [\chi_L^{(DF)}; \chi_{ID}]$ and $H_R = [\chi_R^{(DF)}; \chi_{ID}]$, respectively.

Taking the binaural quality assessment as a representative example, the MHA mechanism effectively identifies critical temporal frames that dominate perceptual scoring while capturing contextual information across diverse representational subspaces. The input representation H_{bin} is linearly projected into m distinct query, key, and value matrices. The mathematical expression for the i -th attention head h_i is formulated as follows:

$$Q_i = H_{bin} W_i^Q, \quad K_i = H_{bin} W_i^K, \quad V_i = H_{bin} W_i^V, \\ h_i = Attention(Q_i, K_i, V_i) = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i, \quad (8)$$

$$H_{att} = Concat(h_1, h_2, \dots, h_m) W^O,$$

where Q_i , K_i , and V_i are the query, key, and value vectors of the i -th head, respectively. W_i^Q , W_i^K , and W_i^V are the corresponding learnable projection matrices. W^O is the projection matrix that linearly maps the concatenated outputs of all heads to the final output sequence H_{att} .

Subsequently, AAP is performed over the time sequence, aggregating all temporal feature representations across different time spans to reflect the overall speech quality representation at an utterance level:

$$G_A = AAP(H_{att}). \quad (9)$$

Finally, vector G_A is fed into a fully connected layer to perform dimension reduction into the spatial domain. To ensure the predicted scores strictly fall within the $[0, 1]$ interval, a sigmoid activation function is

applied at the output end:

$$\widehat{QS}_{bin} = \sigma(G_A \cdot W_f + b_f), \quad (10)$$

where W_f and b_f are the weight and bias of the FC layer, and $\widehat{QS}_{bin}$ represents the predicted overall binaural quality score. Similarly, $\widehat{QS}_L$ and $\widehat{QS}_R$ represent the predicted independent scores for the left and right ears.

3.6. Quality-level module

To further enhance the generalization capability of the shared features, the QLM is introduced as an auxiliary classification task. Under an MTL framework, the QLM imposes structural boundary constraints on the latent feature space, compelling the network to acquire more discriminative representations. This synergistic optimization consequently elevates the prediction accuracy of the primary continuous score regression. Referencing the ITU-T P.800 standard for subjective MOS grading, we map the continuous HASQI numerical intervals into a 5-level speech quality grading scale, as detailed in Table 1.

Table 1
Speech quality-level classification intervals and labels.

Quality-Level	Quality-Level Label	MOS Value	HASQI Value
Excellent	4	[4.5, 5]	[0.875, 1]
Good	3	[3.5, 4.5]	[0.625, 0.875]
Fair	2	[2.5, 3.5]	[0.375, 0.625]
Poor	1	[1.5, 2.5]	[0.125, 0.375]
Bad	0	[1, 1.5]	[0, 0.125]

Taking the left ear branch as an example, the input to the QLM concatenates the deeply fused acoustic-audiogram feature $\chi_L^{(DF)}$, the System ID encoding χ_{ID} , and the raw audiogram HL_L along the feature dimension, where $HL_L \in \mathbb{R}^{1 \times 6}$ is first repeated along the time axis to $\mathbb{R}^{T \times 6}$. As depicted in Fig. 2, the QLM employs an Adaptive Max Pooling (AMP) layer to capture the most salient responses along the time axis, followed by a two-layer FC network. The feature transformation process is defined as:

$$E_L = [\chi_L^{(DF)}; \chi_{ID}; HL_L], \\ E_L^{(0)} = AMP(E_L), \\ E_L^{(n)} = ReLU\left(B(E_L^{(n-1)} \cdot W_n + b_n)\right), \quad n \in \{1, 2\} \quad (11)$$

where $[\cdot; \cdot; \cdot]$ denotes concatenation along the feature dimension. $E_L^{(0)}$ is the global feature after AMP along the time axis, focusing on the most discriminative responses. W_n and b_n represent the weights and bias parameters of the n -th FC layer, respectively, while B denote the corresponding BN layer. The feature $E_L^{(2)}$ is normalized via the Softmax function to generate a probability distribution $P_L = [p_1, p_2, \dots, p_O]_L$, where O is the number of quality level classes. The category index associated with the maximum probability, determined by the $\arg \max(\cdot)$ function, serves as the predicted label $\widehat{QL}_L$:

$$p_o^L = Softmax(E_L^{(2)})_o, \\ \widehat{QL}_L = \arg \max_{o \in \{1, 2, \dots, O\}} (p_o^L), \quad (12)$$

where $\widehat{QL}_L$ represents the predicted discrete quality label for the left ear. An identical procedure is symmetrically applied to predict the right ear quality level $\widehat{QL}_R$.

3.7. Loss function

To achieve comprehensive supervision of the overall binaural quality score, unilateral independent quality scores, and quality levels, the model's training objective is designed as a multi-task composite loss function. This function is composed of weighted prediction errors for the

binaural overall quality score, unilateral ear quality scores, and quality-level labels, with the mathematical expression as follows:

$$\mathcal{L}_{Total} = \alpha \times \mathcal{L}_{MAE}(QS_{bin}, \widehat{QS}_{bin}) + \mu \times \sum_{c \in \{L, R\}} \mathcal{L}_{MAE}(QS_c, \widehat{QS}_c) + \epsilon \times \sum_{c \in \{L, R\}} \mathcal{L}_{CE}(QL_c, P_c), \quad (13)$$

where $\mathcal{L}_{MAE}(QS_{bin}, \widehat{QS}_{bin})$ denotes the Mean Absolute Error (MAE) between the predicted binaural comprehensive score $\widehat{QS}_{bin}$ and the corresponding ground-truth QS_{bin} , which constrains the model's fit to global auditory perception. For the left channel branch, $\mathcal{L}_{MAE}(QS_L, \widehat{QS}_L)$ denotes the MAE between the predicted and ground-truth scores, while $\mathcal{L}_{CE}(QL_L, P_L)$ denotes the Cross-Entropy (CE) loss between the ground-truth quality-level label QL_L and the predicted probability distribution P_L . The loss terms for the right channel branch are defined identically. α , μ , and ϵ serve as balancing weights that modulate the gradient contributions of each loss term. In this study, $\alpha = \mu = 1$ for the score prediction losses and $\epsilon = 0.02$ for the quality-level CE loss. The specific process is as shown in Algorithm 1.

4. Experiment settings

4.1. Data preparation

4.1.1. Datasets

Two publicly available benchmark datasets are used to assess the generalization and robustness of the proposed BiSQA Fusion across heterogeneous hearing loss profiles and HA system configurations.

The first dataset is drawn from the CPC1 Challenge, which provides binaural speech processed by HA algorithms and therefore faithfully reproduces the acoustic characteristics of HA binaural outputs [25]. CPC1 defines two evaluation tracks, *Seen* and *Unseen*, corresponding to whether the listener (binaural audiogram) and the HA configuration (System ID) are observed during training. The *Seen* track contains 4879 training and 2421 test samples spanning 10 HA configurations and 27 listeners. The *Unseen* track comprises 3580 training samples (9 configurations, 22 listeners) and 632 test samples (10 configurations, 27 listeners). Note that the two tracks are not strictly disjoint in factor coverage: the *Unseen* test split shares 9 of its 10 HA configurations and 22 of its 27 listeners with the *Unseen* training split, and only configuration (E018) and the remaining 5 listeners are strictly unobserved during training. The *Unseen* track thus evaluates generalization to novel configurations and listeners under partial factor overlap, rather than to an entirely unseen condition. The 10 HA configurations are indexed from 0 to 9 and used as System IDs during training and inference. To meet the input specification of WavLM-Large, all signals are downsampled from 44.1 kHz to 16 kHz, no further preprocessing is applied.

Because CPC1 contains exclusively bilateral HL samples, a second simulated corpus is built upon the DNS Challenge [42] to additionally cover bilateral NH and unilateral HL scenarios. The source corpus provides over 500h of clean monaural speech from 2150 speakers at 16 kHz. Segments with a voice activity ratio above 60% and energy within $[-35, -15]$ dB are first retained. 16,000 segments are randomly selected and either cropped or zero-padded to 6s using silence-aware segmentation to preserve complete semantic units. Noise samples from NoiseX-92 [43] are mixed at signal-to-noise ratios from -5 to 15 dB in 5 dB steps, and the monaural mixtures are duplicated to form binaural noisy inputs. We note that this construction contains no spatial cues (no ITD/ILD/HRTF). Interaural differences arise solely from the per-ear, audiogram-driven gain compensation. Simulated HA degradation and compensation are subsequently applied following the procedure described in [16]. Specifically, DeepFilterNet3 [44] performs front-end denoising, after which listener-specific gain compensation is realized by combining four prescription formulas with multi-channel wide dynamic range compression. The prescriptions include two non-linear (DSL v5, FIG6) and two linear (NAL-R, POGO) formulas, indexed 0–3 as System

Algorithm 1: Training procedure of the proposed BiSQA Fusion.

Input: Binaural HA compensated speech $S = \{S_L, S_R\}$; binaural audiogram $HL = \{HL_L, HL_R\}$; HA system identifier ID ; ground-truth scores $QS = \{QS_{bin}, QS_L, QS_R\}$ and quality levels $QL = \{QL_L, QL_R\}$, loss weights α, μ, ϵ , epoch number E , learning rate η .

Output: Trained model parameters Θ^* , predicted quality scores $\widehat{QS} = \{\widehat{QS}_{bin}, \widehat{QS}_L, \widehat{QS}_R\}$ and levels $\widehat{QL} = \{\widehat{QL}_L, \widehat{QL}_R\}$.

- 1 Initialize all trainable parameters Θ of BiSQA Fusion;
- 2 **for** $epoch = 1, 2, \dots, E$ **do**
- 3 **for each** mini-batch $\{S, HL, ID, QS, QL\}$ in training set **do**
- 4 // 1. Multi-layer SSL feature extraction.
- 5 **for** $c \in \{L, R\}$ **do**
- 6 $Z_c \leftarrow \text{WeightedLayerAgg}(\text{WavLM}(S_c));$
- 7 **end**
- 8 // 2. Modality-specific encoding.
- 9 **for** $c \in \{L, R\}$ **do**
- 10 $\chi_c^{(S)} \leftarrow \text{SSL-Encoder}(Z_c);$
- 11 $\chi_c^{(A)} \leftarrow \text{HL-Encoder}(HL_c);$
- 12 **end**
- 13 // 3. Hierarchical acoustic-audiogram fusion via AAFM.
- 14 **for** $c \in \{L, R\}$ **do**
- 15 /* Stage 1: gated FiLM fusion. */
- 16 $\chi_c^{(F)} \leftarrow \text{GatedFiLM}(\chi_c^{(S)}, \chi_c^{(A)});$
- 17 /* Stage 2: deep CNN-BiLSTM fusion. */
- 18 $\chi_c^{(DF)} \leftarrow \text{CNN-BiLSTM}([\chi_c^{(F)}; \chi_c^{(A)}]);$
- 19 **end**
- 20 // 4. Binaural fusion via BFM.
- 21 /* Decouple three perceptually motivated interaural components. */
- 22 $\chi_{bin}^{(DF)} \leftarrow \text{BFM}(\chi_L^{(DF)}, \chi_R^{(DF)});$
- 23 // 5. Configuration-aware System ID embedding.
- 24 $\chi_{ID} \leftarrow \text{Broadcast}(\text{Embedding}(ID));$
- 25 // 6. Score and quality-level prediction.
- 26 $\{\widehat{QS}_{bin}, \widehat{QS}_L, \widehat{QS}_R\} \leftarrow \text{SPM}(\chi_L^{(DF)}, \chi_R^{(DF)}, \chi_{bin}^{(DF)}, \chi_{ID});$
- 27 $\{\widehat{QL}_L, \widehat{QL}_R\} \leftarrow \text{QLM}(\chi_L^{(DF)}, \chi_R^{(DF)}, HL, \chi_{ID});$
- 28 // 7. Multi-task joint optimization.
- 29 Compute total loss $\mathcal{L}_{Total}$ via Eq. (13);
- 30 $\Theta \leftarrow \Theta - \eta \cdot \nabla_{\Theta} \mathcal{L}_{Total};$
- 31 **Result:** Updated parameters Θ for the current mini-batch;
- 32 **end**
- 33 Decay learning rate η via cosine annealing;
- 34 **end**

IDs, with the NH condition labeled 4. The simulated corpus is partitioned into training, validation, and test subsets at a ratio of 80%:13%:7%.

4.1.2. Audiograms

The 27 binaural audiograms released with CPC1 are used as prior information. To align the audiogram representation with the granularity of the target metric (HASQI), pure-tone thresholds at six octave frequencies from 250 Hz to 8 kHz are extracted. Following the WHO 2021 criteria [1], the per-ear pure-tone average (PTA) over 500 Hz, 1 kHz, 2 kHz, and 4 kHz is used to grade HL into five levels: NH (PTA < 20 dB HL), Mild ($20 \leq \text{PTA} < 35$), Moderate ($35 \leq \text{PTA} < 50$), Moderately Severe ($50 \leq \text{PTA} < 65$), and Severe ($\text{PTA} \geq 65$ dB HL). Three binaural scenarios are further defined [1]: (i) Bilateral NH, both ears PTA < 20 dB; (ii) Bilateral HL, both ears PTA ≥ 20 dB; (iii) Unilateral HL, better-ear PTA < 20 dB and worse-ear PTA ≥ 35 dB HL.

Since all CPC1 listeners fall within the Bilateral HL category, 57 binaural audiograms from the National Health and Nutrition Examination Survey (NHANES) database [45] are incorporated into the DNS Challenge-based simulation to broaden the coverage of asymmetric HL and NH conditions. A topological expansion strategy is then applied: 57 unilateral HL audiograms are synthesized by forcing one ear of each original pair to 0 dB HL, and a single ideal bilateral NH (0 dB HL) audiogram is appended, yielding 115 binaural audiograms in total. To rigorously assess generalization to unseen listeners, the audiograms are partitioned in a non-overlapping manner: 51 original pairs, together with their derived variants and the ideal bilateral-NH audiogram (103 in total), are used for training and validation; the remaining 6 pairs and their variants (12 in total) are held out as unseen. The test partition spans all 115, jointly assessing in-distribution accuracy and unseen-profile generalization.

4.2. Experimental procedure and ground-truth generation

During training, the HA compensated binaural signals are decoupled into two independent channels and fed into WavLM Large to extract per-layer hidden representations, which are then adaptively integrated via a learnable layer-wise weighting scheme and projected by the SSL-Encoder. In parallel, the HL-Encoder distills abstract representations from the audiogram and aligns their dimensionality with that of the SSL-encoded features. The two encoded streams are subsequently passed to the AAFM for cross-modal fusion, and the resulting dual-branch outputs are forwarded to the BFM for interaural cue extraction. The SPM concurrently produces left, right, and binaural quality scores. In parallel, the QLM regularizes shared representations via coarse-grained five-class supervision. In inference, the trained network operates solely on the HA-processed binaural speech, individual audiograms, and the System ID to produce all three scores in a single forward pass, without requiring any clean reference. Clean references are used exclusively to compute the HASQI-based supervision signals during training and evaluation.

Given the absence of publicly available large-scale datasets for subjective quality assessment of monaural/binaural HA speech, this study adopts HASQI [9], an objective metric extensively validated in international challenges and auditory research [28,29,46], as the ground-truth for model optimization. We emphasize that HASQI serves as a validated perceptual proxy for subjective quality, not a subjective measurement per se, and defer validation against subjective binaural ratings to future work upon public data availability. The left and right ear HASQI values are computed independently from the clean reference and the corresponding HA outputs, and serve as channel-level ground-truth scores. The binaural overall ground-truth is constructed via an asymmetric perceptual fusion that mirrors clinical findings on dominant-ear integration:

$$QS_{bin} = \zeta \times \max(QS_L, QS_R) + (1 - \zeta) \times \min(QS_L, QS_R). \quad (14)$$

where ζ represents the contribution weight of the dominant ear (the side with the higher HASQI score). In this study, ζ is fixed a priori to 0.7 following psychoacoustic evidence: the binaural loudness model of Moore et al. [39] assigns the louder ear an effective weight of approximately 0.7 in overall loudness perception, and Bronkhorst et al. [40,41] further report that listeners with hearing impairment rely on the higher-SNR ear at a comparable level ($\sim 70\%$), as compromised neural processing weakens the fine interaural comparisons required for binaural noise cancellation.

To examine the sensitivity of BiSQAFusion to this a priori choice, Table 2 reports a stability check in which the network is retrained and evaluated on the CPC1 *Seen* test set under perturbed values $\zeta \in \{0, 0.3, 0.5, 0.7, 1\}$. Three regimes emerge. When $\zeta \in \{0, 1\}$, the supervision degenerates into fitting a single ear, which biases the model toward local features and suppresses the learning of interaural cues, yielding the lowest accuracy. When $\zeta \in \{0.3, 0.5\}$, correlation metrics improve but remain suboptimal given the non-linear nature of binaural perception.

When $\zeta = 0.7$, both the absolute error on the binaural composite score (RMSE = 0.039, MAE = 0.030) and the per-channel prediction accuracy reach their global optima, indicating that dominant-ear-weighted supervision simultaneously refines global and local representations. This optimum is consistent with established findings in binaural psychoacoustics. Accordingly, fixed a priori from the psychoacoustic evidence above, $\zeta = 0.7$ is used for all subsequent experiments.

4.3. Comparative baselines

To assess the effectiveness and robustness of BiSQAFusion in binaural scenarios, a comparative baseline set is constructed that comprises two classical non-intrusive metrics for NH listeners and four state-of-the-art (SOTA) models tailored to HA scenarios. The NH baselines include Quality-Net [33], which learns a non-linear mapping from acoustic features to PESQ, and MOS-Net [34], which directly regresses subjective MOS. The HA-oriented baselines comprise SQINet-HLlevel [12] for HASQI prediction, HASA-Net [15] and its enhanced variant HASA-Net⁺ [17] for joint HASQI/HASPI prediction, and MTSE-LA [16] likewise targeting HASQI/HASPI. All of them have reported competitive accuracy on degraded speech within their original benchmarks. Notably, both HASA-Net⁺ and BiSQAFusion adopt WavLM-Large weighted representations and audiograms as inputs. Since all selected baselines are designed for monaural inputs, their direct application to binaural assessment would introduce uncontrolled architectural bias. To enable a fair comparison, each baseline is therefore extended into a weight-sharing Siamese architecture that processes the two channels in parallel. All baselines were re-trained from scratch on our binaural training data. Aside from this minor modification, the network configurations, loss functions, and hyperparameters strictly follow the settings reported in their original literature.

4.4. Evaluation metrics and implementation details

This study employs four performance evaluation metrics to validate the effectiveness of BiSQAFusion. To measure correlation, the Pearson correlation coefficient (PCC) and Spearman’s rank correlation coefficient (SRCC) are utilized to assess the linear and monotonic correlations between the network’s predicted outputs and the ground-truth values, respectively. To measure error, the root mean square error (RMSE) and the mean absolute error (MAE) are used to quantify discrepancies between predictions and ground truths. Significance protocol. Unless otherwise stated, all models are independently trained with five different random seeds. For each run, the checkpoint with the best validation performance is evaluated on the test set, and all reported metrics are averaged over the five runs. For significance testing, the five per-run metric values of BiSQAFusion and each competing model are compared using a two-tailed paired Student’s *t*-test. Differences are deemed significant at $p < .05$.

BiSQAFusion predicts independent quality scores and level labels for each channel, as well as the overall binaural composite score. The network consists of five key modules: SSL/HL encoders, AAFM, BFM, SPM, and QLM. Detailed parameter settings for each module are provided in Table 3. The entire network is trained end-to-end using the Adam optimizer to minimize the total loss $\mathcal{L}_{Total}$ via backpropagation. The initial learning rate is set to 1×10^{-4} , with a cosine annealing schedule employed for dynamic adjustment throughout training. This approach mitigates the risk of local optima and accelerates convergence. Experiments use a batch size of 16 and 200 total training epochs.

5. Experimental results and analysis

5.1. Feature-level ablation and hierarchical SSL fusion validation

To systematically validate the rationality of input-feature selection and the necessity of hierarchical SSL aggregation, comprehensive

Table 2

Prediction performance of BiSQA-Fusion under varying weights for the dominant ear.

ζ	Binaural				Left Channel				Right Channel			
	PCC	SRCC	RMSE	MAE	PCC	SRCC	RMSE	MAE	PCC	SRCC	RMSE	MAE
0	0.928	0.939	0.044	0.033	0.929	0.936	0.048	0.035	0.927	0.933	0.051	0.037
0.3	0.932	0.942	0.043	0.032	0.923	0.930	0.045	0.033	0.926	0.932	0.047	0.035
0.5	0.934	0.938	0.042	0.032	0.926	0.934	0.044	0.033	0.930	0.936	0.046	0.034
0.7	0.942	0.944	0.039	0.030	0.935	0.941	0.041	0.030	0.933	0.939	0.042	0.032
1	0.925	0.925	0.048	0.035	0.925	0.933	0.047	0.035	0.925	0.930	0.049	0.036

Table 3

Parameter settings of the core modules in the proposed framework.

Module	Module Layer	Input size	Hyperparameter	Output size
SSL-Encoder	Adapter Layer	$1 \times T \times 1024$	256	$1 \times T \times 256$
	Conv2d_SSL_1	$1 \times T \times 256$	$3 \times 3, (1, 1), 16$	$16 \times T \times 256$
	LPPool2d_SSL_1	$16 \times T \times 256$	$2, 1 \times 2$	$16 \times T \times 128$
	Conv2d_SSL_2	$16 \times T \times 128$	$3 \times 3, (1, 1), 32$	$32 \times T \times 128$
	LPPool2d_SSL_2	$32 \times T \times 128$	$2, 1 \times 4$	$32 \times T \times 32$
HL-Encoder	Adapter Layer	$1 \times 1 \times 6$	256	$1 \times 1 \times 256$
	Conv2d_HL_1	$1 \times 1 \times 256$	$1 \times 3, (1, 2), 16$	$16 \times 1 \times 128$
	Conv2d_HL_2	$16 \times 1 \times 128$	$1 \times 3, (1, 4), 32$	$32 \times 1 \times 32$
AAFMM (L/R)	Repeat	$32 \times 1 \times 32$	—	$32 \times T \times 32$
	Gate FiLM	$32 \times T \times 32; 32 \times T \times 32$	—	$32 \times T \times 32$
	Cat	$32 \times T \times 32; 32 \times T \times 32$	—	$32 \times T \times 64$
	Conv2d_1	$32 \times T \times 64$	$3 \times 3, (1, 1), 64$	$64 \times T \times 64$
	LPPool2d_1	$64 \times T \times 64$	$2, 1 \times 4$	$64 \times T \times 16$
	Conv2d_2	$64 \times T \times 16$	$3 \times 3, (1, 1), 32$	$32 \times T \times 16$
	LPPool2d_2	$32 \times T \times 16$	$2, 1 \times 4$	$32 \times T \times 4$
	BiLSTM	$T \times 128$	64	$T \times 128$
BFM	LayerNorm	$T \times 128$	128	$T \times 128$
	Linear	$T \times 128$	128	$T \times 128$
	GELU	$T \times 128$	—	$T \times 128$
System ID	Embedding	1×1	(5/10, 16)	1×16
	Expand	1×16	—	$T \times 16$
SPM (B)	Cat	$T \times 128(B); T \times 128(L, R); T \times 16(ID)$	—	$T \times 400$
	MHA	$T \times 400$	$400, h = 8$	$T \times 400$
	AAP	$T \times 400$	1	1×400
	Dense	1×400	1	1×1
SPM (L/R)	Cat	$T \times 128(L/R); T \times 16(ID)$	—	$T \times 144$
	MHA	$T \times 144$	$144, h = 8$	$T \times 144$
	AAP	$T \times 144$	1	1×144
	Dense	1×144	1	1×1
QLM (L/R)	Cat	$T \times 128(L/R); T \times 6(HL); T \times 16(ID)$	—	$T \times 150$
	AMP	$T \times 150$	1	1×150
	Linear_1	1×150	75	1×75
	Linear_2	1×75	5	1×5
	Softmax	1×5	—	1×5

comparative experiments are conducted across six acoustic feature configurations, with audiograms held as a fixed conditional input. Three classical features: Magnitude [12,15,16], LMFB [18], and Gammatone FBE [4], are first compared with three large-scale SSL representations: Wav2Vec 2.0 Large¹ [47], HuBERT Large² [48], and WavLM Large³ [13]. For all SSL variants, only the 1024-dimensional final-layer outputs are extracted and denoted as "(Last)" variants, in contrast to the proposed adaptive multi-layer aggregation strategy adopted by BiSQA-Fusion. Quantitative results are reported in Table 4. All results in the table were validated via a two-tailed Student's *t*-test ($p < .05$).

Three insights emerge from the comparison. First, among the classical acoustic features, LMFB and Gammatone FBE exhibit pronounced cross-scenario fragility: despite relatively moderate predictive performance on the CPC1 *Seen* test set, their *Unseen* PCC collapses to 0.757 and 0.567, respectively. The uncompressed Magnitude spectrum, despite lacking auditory-band weighting, retains comparatively richer low-level acoustic textures, yielding more stable cross-scenario performance (*Un-*

seen PCC = 0.824). This dichotomy reveals that handcrafted compressed features inadvertently discard distortion cues critical for HA quality assessment when distribution shifts occur. Second, the superiority of SSL features is conditional rather than universal. Wav2Vec 2.0 Large (Last) and HuBERT Large (Last) underperform Magnitude under *Unseen* conditions (PCC: 0.803 and 0.709 vs. 0.824), because their ASR-oriented pre-training objectives drive top-layer representations toward purely linguistic semantics, stripping away the low-level acoustic textures and distortion patterns that quality assessment fundamentally relies upon. In contrast, WavLM Large (Last) achieves PCC = 0.915/0.862 owing to its richer noise and speaker modeling pre-training objectives, which explicitly preserve quality-relevant acoustic cues in non-semantic dimensions. Third, the adopted hierarchical aggregation strategy delivers further substantial gains over WavLM Large (Last), boosting PCC/SRCC by 0.027/0.026 (*Seen*) and PCC/SRCC by 0.044/0.040 (*Unseen*) while reducing RMSE/MAE by 0.008/0.005 (*Seen*) and 0.007/0.005 (*Unseen*). The mechanism behind this gain is that lower WavLM layers carry fine-grained acoustic distortion information, whereas upper layers encode high-level phonetic abstraction. The learnable weights w_i in Eq. (1) adaptively balance these complementary semantic levels for the quality-assessment objective, fully exploiting the hierarchical complementarity reported in [14]. To visually demonstrate the gains from this

¹ <https://huggingface.co/facebook/wav2vec2-large-960h>² <https://huggingface.co/facebook/hubert-large-ls960-ft>³ <https://huggingface.co/microsoft/wavlm-large>

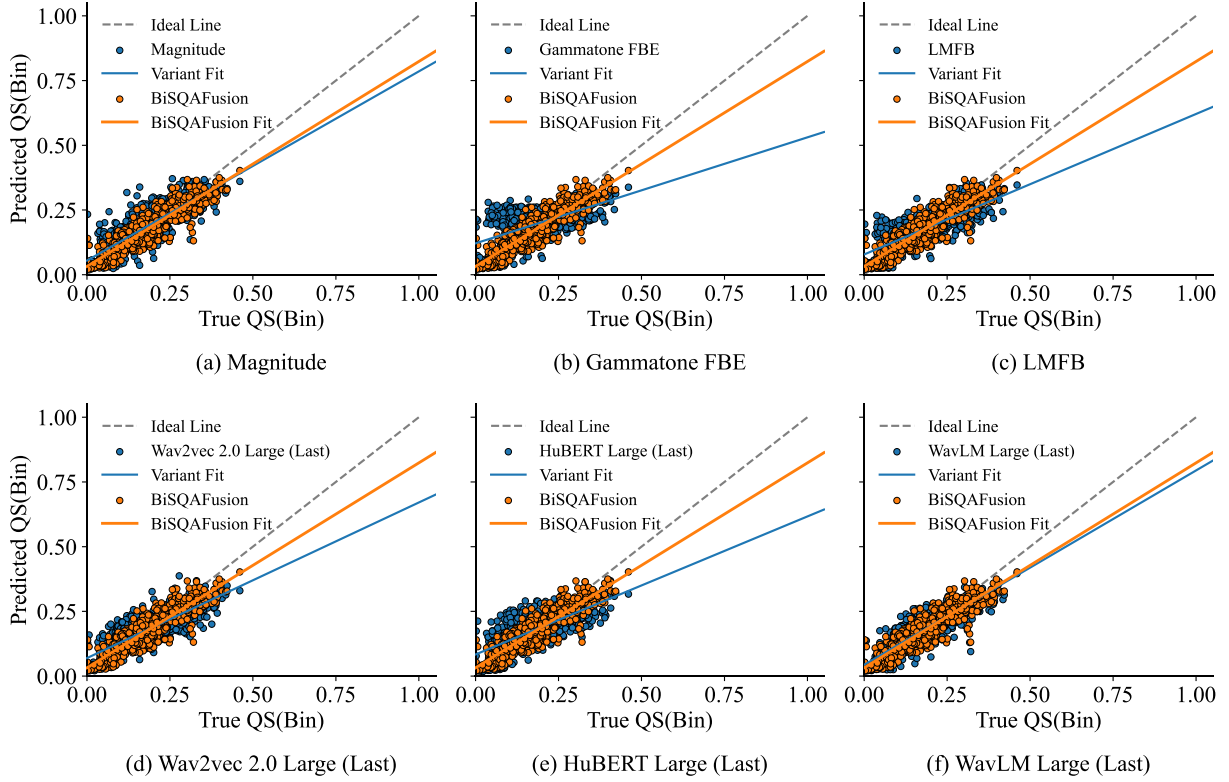


Fig. 4. Scatter plots of predicted versus ground-truth binaural composite quality scores on the CPC1 Unseen test set: each input-feature variant (blue) versus BiSQA-Fusion with adaptive multi-layer SSL aggregation (yellow). The dashed line denotes the ideal diagonal.

Table 4
Prediction performance of BiSQA-Fusion using different input features.

Module	Seen				Unseen			
	PCC↑	SRCC↑	RMSE↓	MAE↓	PCC↑	SRCC↑	RMSE↓	MAE↓
Magnitude	0.883	0.883	0.053	0.040	0.824	0.806	0.054	0.042
Gammatone FBE	0.889	0.889	0.053	0.039	0.567	0.429	0.080	0.063
LMFB	0.894	0.896	0.051	0.038	0.757	0.737	0.061	0.048
Wav2vec 2.0 Large (Last)	0.884	0.886	0.055	0.042	0.803	0.786	0.056	0.044
HuBERT Large (Last)	0.894	0.901	0.052	0.039	0.709	0.679	0.066	0.052
WavLM Large (Last)	0.915	0.918	0.047	0.035	0.862	0.861	0.048	0.037
BiSQA-Fusion	0.942	0.944	0.039	0.030	0.906	0.901	0.041	0.032

Note: ↑ (↓) indicates that higher (lower) values represent better performance.

hierarchical fusion strategy, Fig. 4 shows scatter plots of the predicted and ground-truth binaural composite quality scores for BiSQA-Fusion and its input-feature variants on the CPC1 *Unseen* test set. The gray dashed line represents the ideal diagonal ($slope = 1$) where predictions perfectly match ground truths. The blue and yellow solid lines represent the fitting lines for the input-feature variants and the proposed BiSQA-Fusion, respectively. It is observable that the prediction scatter points of BiSQA-Fusion cluster most tightly along the ideal diagonal with the smallest discrete variance. This further validates the rationality of using WavLM as an acoustic feature extractor and the significant effectiveness of the adaptive hierarchical fusion strategy in improving the accuracy of speech quality prediction.

5.2. Effectiveness of binaural fusion architecture: monaural Vs. binaural networks

To rigorously validate the superiority of the proposed BiSQA-Fusion over independent monaural training schemes in terms of binaural comprehensive quality estimation and independent unilateral quality prediction, this section systematically compares the predictive performance

of the monaural and BiSQA-Fusion training pipelines. Specifically, the experimental setup comprises six configurations: Mon-L, Mon-R, and Mon-Bin, corresponding to independent training on the left channel, right channel, and their subsequent binaural post-fusion composite, respectively. For these baselines, the training network directly reuses the unilateral branch structure of BiSQA-Fusion. Conversely, BiSQA-Fusion-L, BiSQA-Fusion-R, and BiSQA-Fusion-Bin correspond to the left-ear, right-ear, and binaural comprehensive quality scores output under the proposed joint training framework.

As shown in Table 5, in the CPC1 test sets, the quality scores of the binaural composite predicted by BiSQA-Fusion significantly outperform those obtained by combining scores from independently trained monaural models. On the *Seen* test set, BiSQA-Fusion-Bin achieves PCC and SRCC of 0.942 and 0.944, respectively, with RMSE and MAE as low as 0.039 and 0.030, demonstrating excellent predictive precision. To visually illustrate these discrepancies, Fig. 5 provides a bar chart comparison of relevant metrics. On the *Unseen* test set, the performance gap between BiSQA-Fusion-Bin and Mon-Bin widens further, with BiSQA-Fusion-Bin improving the PCC and SRCC by 0.024 and 0.023 over Mon-Bin, while reducing the RMSE and MAE by 0.006 and 0.003, respectively. This

Table 5
Prediction performance of BiSQA Fusion and monaural Network.

Module	Params(M)	FLOPs(G)	Seen				Unseen			
			PCC↑	SRCC↑	RMSE↓	MAE↓	PCC↑	SRCC↑	RMSE↓	MAE↓
Mon-L	0.51	1.13	0.930	0.938	0.045	0.033	0.883	0.888	0.047	0.036
Mon-R	0.51	1.13	0.927	0.934	0.049	0.036	0.817	0.810	0.055	0.043
Mon-Bin	1.02	2.26	0.936	0.940	0.045	0.032	0.882	0.878	0.047	0.035
BiSQA Fusion-L	1.17	2.67	0.935	0.941	0.041	0.030	0.896	0.893	0.043	0.033
BiSQA Fusion-R			0.933	0.939	0.042	0.032	0.892	0.891	0.044	0.035
BiSQA Fusion-Bin			0.942	0.944	0.039	0.030	0.906	0.901	0.041	0.032

indicates that the proposed binaural architecture exhibits stronger generalization capabilities when confronted with unseen conditions. Notably, despite the significant performance gains achieved by BiSQA Fusion-Bin, its parameter size and computational complexity remain on the same order of magnitude as Mon-Bin, introducing no excessive computational overhead.

Furthermore, under the joint optimization framework of BiSQA Fusion, the unilateral quality prediction branches (BiSQA Fusion-L and BiSQA Fusion-R) achieve significant improvements over their independently trained counterparts (Mon-L and Mon-R). This implies that binaural joint modeling not only elevates comprehensive quality prediction but also effectively enhances the discriminability of unilateral quality representations. In particular, on the *Unseen* test set, while Mon-R shows noticeable degradation, the performance deterioration of BiSQA Fusion-R is significantly mitigated. This phenomenon illuminates the core information fusion mechanism of the network: through interaural information interaction, BiSQA Fusion fully excavates the complementarity and consistency between left- and right-channel features. This deep integration not only improves the accuracy of binaural comprehensive quality estimation but also provides stable and effective information for unilateral quality assessment.

Finally, to evaluate the cross-scenario generalization stability of BiSQA Fusion, Fig. 6 illustrates the performance gap between the *Seen* and *Unseen* test sets for each model. The visualizations clearly demonstrate that, for both unilateral and binaural comprehensive predictions, the performance gap of BiSQA Fusion is markedly smaller than that of independent monaural training schemes, especially for the PCC and SRCC metrics. This further validates the efficacy of the proposed binaural information fusion architecture in bolstering the model’s generalization robustness.

5.3. Ablation and comparative analysis of network architecture

To verify the contribution of each functional module, systematic ablation experiments are conducted under incremental module stacking (Table 6). Given that the SPM is responsible for the final mapping of predicted scores, it was established as the foundational module, yielding a baseline model, BiSQA Fusion-Base, in which input features are concatenated and dimensionally reduced via FC layers. Starting from BiSQA Fusion-Base, the SSL/HL-Encoders, AAFM, BFM, QLM, and System ID are introduced sequentially. The incremental results reveal a consistent improvement at each integration step, with the largest single gain contributed by the encoders and each subsequent module providing a further statistically significant increase.

5.3.1. Impact analysis of encoders

Replacing the FC-based concatenation in BiSQA Fusion-Base with the SSL/HL-Encoders yields the largest single-step gain across the entire ablation chain. On the *Seen* test set, the correlation metrics PCC and SRCC improved by 0.041 and 0.037, respectively, while the error metrics RMSE and MAE decreased by 0.007 and 0.004, respectively. Furthermore, on the *Unseen* test set, the encoder-integrated network exhibited substantially greater robustness than BiSQA Fusion-Base,

strongly corroborating the effectiveness of the encoder design. This performance leap is attributed to the encoders’ superior capacity for learning task-relevant representations. First, the SSL-Encoder, through cascaded Conv2D and L_p pooling, suppresses high-dimensional acoustic redundancy by exploiting locality and translation-invariance inductive biases, thereby concentrating modeling capacity on the perceptually salient acoustic regions. Second, the HL-Encoder, conversely, reverses the dimensional asymmetry: it expands the sparse 6-D audiogram into a 32-D structured embedding via Conv2D with progressive strides, alleviating the manifold-level imbalance that previously overwhelmed the audiogram with SSL features in shallow concatenation. Notably, with the encoders integrated, the performance gap between the *Seen* and *Unseen* narrows considerably, suggesting that the encoders learn more stable, transferable representations rather than merely fitting the training distribution.

5.3.2. Impact analysis of acoustic-audiogram fusion module

Building on the encoder-integrated network, the contribution of the AAFM is further isolated. As reported in Table 6, introducing the AAFM consistently improves the predictive accuracy of BiSQA Fusion: PCC and SRCC rise by 0.003 and 0.006, while RMSE and MAE drop by 0.003 and 0.002, respectively, providing preliminary evidence of its cross-modal modeling capability on heterogeneous inputs.

To dissect the hierarchical fusion strategy underlying the AAFM, four representative variants are compared within a unified framework: feature concatenation (Concat), feature addition (Add), Gated FiLM Fusion only (GFF), and CNN-BiLSTM Fusion only (CBF). Dimensional inconsistencies are resolved by fully-connected mapping layers, and all other components and training configurations are kept identical to ensure a controlled comparison. The corresponding results are summarised in Table 7, from which three observations emerge.

First, static shallow fusion suffers from limited generalization. Concat and Add yield slightly inferior accuracy on the *Seen* test set, and Concat further deteriorates sharply on the *Unseen* set (PCC = 0.742), indicating that direct aggregation cannot adequately disentangle the non-linear coupling between speech representations and hearing-loss priors. Second, single-mechanism fusions consistently outperform shallow ones but exhibit distinct weaknesses. GFF strengthens the conditioning of acoustic representations on audiogram priors through gated modulation, achieving robust cross-scenario performance (PCC = 0.883, SRCC = 0.882 on the *Unseen* set). CBF, in turn, attains the highest known-scenario correlation (PCC = 0.931) via deep convolutional and temporal modeling, yet remains susceptible to scenario shifts (PCC = 0.768 on *Unseen*), revealing pronounced overfitting to acoustic artifacts. Third, the proposed AAFM (GFF + CBF) achieves the best overall performance through its dual-stage hierarchical design: GFF first performs prior-guided modulation and noise suppression on the acoustic stream, followed by CBF, which refines the modulated features through spatiotemporal contextualization. The two stages are complementary: prior-driven gating mitigates the artifact overfitting observed in CBF-only, while deep contextual modeling lifts the representational ceiling beyond GFF-only. As a result, AAFM improves PCC and SRCC on the *Unseen* set to 0.906 and 0.901, and lowers RMSE and MAE to 0.041 and

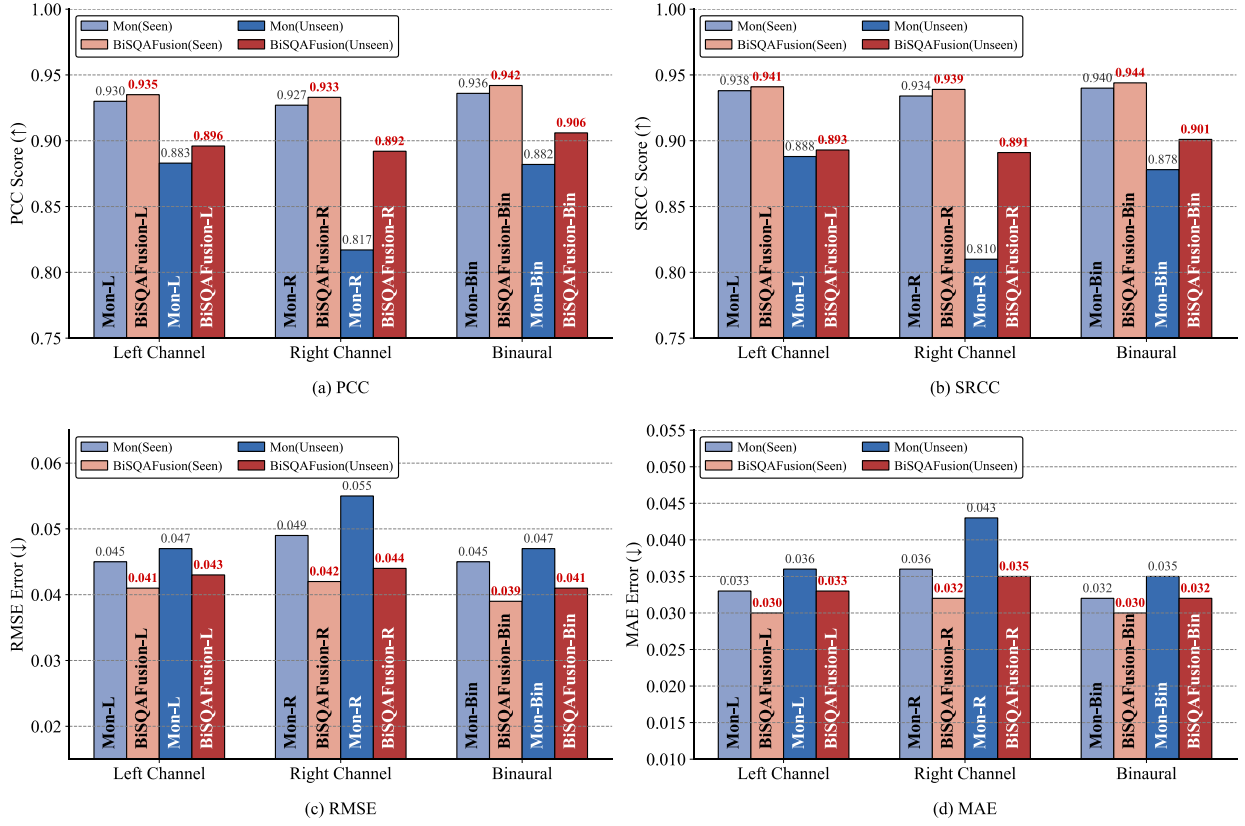


Fig. 5. Predictive performance comparison between BiSQA Fusion and monaural baselines. The comparisons are conducted across Left channel (L), Right channel (R), and Binaural (Bin) inputs under both seen and unseen test scenarios. The bold red values indicate the superior performance within each comparison group.

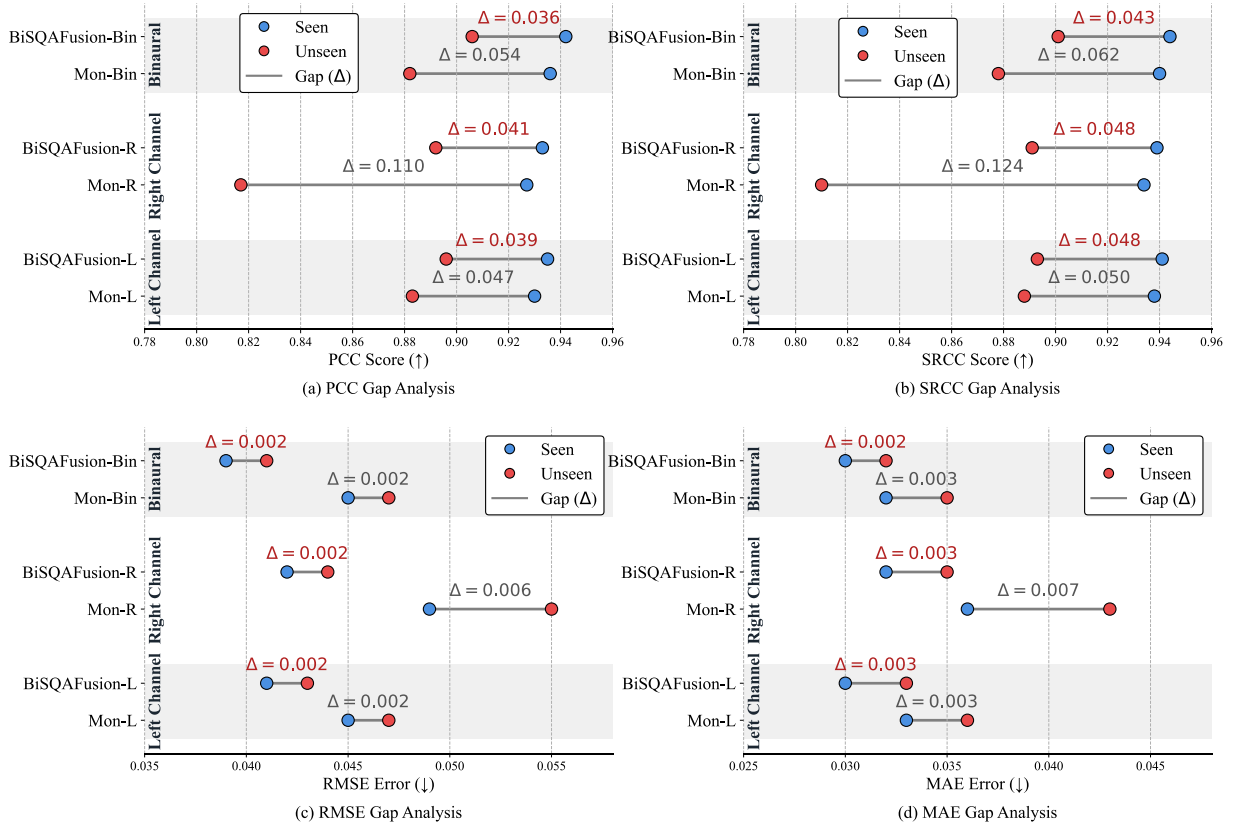


Fig. 6. Comparative analysis of performance gaps between monaural baselines and BiSQA Fusion.

Table 6

Predictive performance of BiSQAFusion with different module combinations.

Module	Params(M)	FLOPs(G)	Seen				Unseen			
			PCC↑	SRCC↑	RMSE↓	MAE↓	PCC↑	SRCC↑	RMSE↓	MAE↓
BiSQAFusion-Base	0.39	0.40	0.883	0.886	0.055	0.040	0.805	0.803	0.057	0.044
+ SSL-Encoder & HL-Encoder	0.73	1.21	0.924	0.923	0.048	0.036	0.855	0.856	0.049	0.039
+ SSL-Encoder & HL-Encoder & AAFM	0.74	2.43	0.927	0.929	0.045	0.034	0.877	0.875	0.046	0.037
+ SSL-Encoder & HL-Encoder & AAFM & BFM	1.09	2.62	0.933	0.937	0.043	0.033	0.887	0.884	0.044	0.035
+ SSL-Encoder & HL-Encoder & AAFM & BFM & QLM	1.10	2.62	0.935	0.938	0.042	0.032	0.890	0.888	0.043	0.034
+ SSL-Encoder & HL-Encoder & AAFM & BFM & QLM & ID	1.17	2.67	0.942	0.944	0.039	0.030	0.906	0.901	0.041	0.032

Table 7

Prediction performance of BiSQAFusion under different acoustic-audiogram fusion strategies.

Fusion	Seen				Unseen			
	PCC↑	SRCC↑	RMSE↓	MAE↓	PCC↑	SRCC↑	RMSE↓	MAE↓
Concat	0.923	0.925	0.045	0.034	0.742	0.726	0.064	0.051
Add	0.914	0.920	0.047	0.035	0.840	0.848	0.054	0.043
GFF	0.927	0.929	0.043	0.032	0.883	0.882	0.045	0.036
CBF	0.931	0.935	0.044	0.033	0.768	0.766	0.062	0.049
AAFM	0.942	0.944	0.039	0.030	0.906	0.901	0.041	0.032

Table 8

Prediction performance of BiSQAFusion for symmetric vs. asymmetric hearing loss scenarios.

Network	All				Symmetric HL (85.2%)				Asymmetric HL (14.8%)			
	PCC↑	SRCC↑	RMSE↓	MAE↓	PCC↑	SRCC↑	RMSE↓	MAE↓	PCC↑	SRCC↑	RMSE↓	MAE↓
BiSQAFusion-NBFM	0.928	0.931	0.044	0.032	0.928	0.928	0.045	0.033	0.918	0.919	0.037	0.028
BiSQAFusion	0.942	0.944	0.039	0.030	0.942	0.943	0.040	0.030	0.930	0.924	0.034	0.026

0.032, simultaneously sustaining high accuracy in seen scenarios and robust generalization to the novel configurations and listeners within the *Unseen* track.

5.3.3. Impact analysis of binaural fusion module

To validate the effectiveness of the BFM, an incremental evaluation was conducted by integrating the BFM into the network that already incorporated the Encoders and AAFM. As shown in Table 6, the BFM significantly improves predictive performance. On the *Seen* test set, PCC and SRCC increase by 0.006 and 0.008, while RMSE and MAE decrease by 0.002 and 0.001, respectively. Notably, this performance gain is consistently maintained on the *Unseen* test set, demonstrating that the BFM not only improves fitting on seen configurations and listeners but also retains its benefit on the novel ones within the *Unseen* track. This advantage stems primarily from the BFM’s explicit modeling of interaural interactive relationships. By extracting shared, differential, and non-linear interaural components between both ears, the BFM provides the model with a perceptually grounded inductive bias. This lowers the complexity of learning binaural relationships and enables the effective modeling of perceived quality fluctuations induced by interaural cues.

Furthermore, to evaluate the BFM’s robustness across different interaural hearing loss patterns, the *Seen* test set was partitioned based on the difference in PTA between the two ears (ΔPTA) [49]: Asymmetric HL ($\Delta\text{PTA} \geq 15$ dB HL) and Symmetric HL ($\Delta\text{PTA} < 15$ dB HL). As shown in Table 8, the Symmetric HL and Asymmetric HL subsets account for 85.2% and 14.8% of the test set, respectively. BiSQAFusion-NBFM refers to a variant of BiSQAFusion that removes only the BFM, in which the binaural fusion operation is replaced by directly concatenating the outputs of the two AAFMs. Without the BFM, BiSQAFusion-NBFM shows lower correlation metrics on the Asymmetric HL subset, suggesting limited predictive consistency under interaural imbalance conditions. Upon integrating the BFM, the PCC and SRCC on the Asymmetric HL subset improve to 0.930 and 0.924, respectively, with RMSE and MAE reduced to 0.034 and 0.026. This demonstrates the strong suppression of

error accumulation caused by interaural asymmetry. Similarly, the Symmetric HL subset also benefits from BFM integration, confirming that the performance gains are not restricted to any particular hearing loss type. Instead, the BFM enhances feature representation robustness and regression prediction stability across both symmetric and asymmetric scenarios.

5.3.4. Impact analysis of Quality-level module and System ID

To quantitatively assess the contributions of the QLM and System ID to network prediction performance, both components were incrementally introduced upon the network integrating Encoders, AAFM, and BFM. As illustrated in Table 6, the introduction of the QLM yields significant performance gains. Specifically, on the *Seen* test set, PCC and SRCC increase by 0.002 and 0.001, while both RMSE and MAE decrease by 0.001. This improvement is primarily attributed to the MTL strategy, which serves as an auxiliary task. The QLM imposes additional structural constraints on feature learning, fostering synergistic task-level optimization and driving the model to acquire more generalizable representations. Furthermore, the System ID also significantly impacts model performance. As shown in Table 6, its integration yields PCC and SRCC gains of 0.007 and 0.006 on the *Seen* test set, with reductions in RMSE and MAE of 0.003 and 0.002, respectively. On the *Unseen* test set, PCC and SRCC further improve by 0.016 and 0.013, respectively, while RMSE and MAE each decrease by 0.002. This performance improvement is attributed to the System ID embedding, which enables the model to learn detailed speech features across different HA configurations, thereby significantly enhancing its feature representation capabilities.

5.4. Performance analysis under different hearing loss configurations

To assess the robustness of BiSQAFusion under heterogeneous auditory conditions, predictive performance is examined along two complementary dimensions: three binaural scenarios (Bilateral HL, Unilateral HL, Bilateral NH) and five severity grades (Normal, Mild, Moderate,

Table 9
Predictive performance across different hearing loss patterns.

Hearing Loss Patterns	DNS-Challenge				CPC1-Challenge			
	PCC	SRCC	RMSE	MAE	PCC	SRCC	RMSE	MAE
Bilateral HL	0.988	0.986	0.033	0.022	0.942	0.944	0.039	0.030
Unilateral HL	0.983	0.953	0.011	0.007	—	—	—	—
Bilateral NH	0.952	0.844	0.012	0.009	—	—	—	—

Table 10
Predictive performance across different degrees of hearing loss.

Hearing Loss Levels	Left Ear				Right Ear			
	PCC	SRCC	RMSE	MAE	PCC	SRCC	RMSE	MAE
DNS-Challenge								
Normal	0.977	0.859	0.013	0.004	0.918	0.812	0.016	0.006
Mild	0.958	0.937	0.034	0.026	0.942	0.921	0.050	0.040
Moderate	0.920	0.936	0.067	0.051	0.832	0.832	0.058	0.044
Moderately severe	0.939	0.939	0.046	0.037	0.958	0.955	0.042	0.032
Severe	0.976	0.927	0.018	0.009	0.973	0.942	0.026	0.013
CPC1-Challenge								
Normal	—	—	—	—	—	—	—	—
Mild	0.933	0.920	0.044	0.032	0.937	0.938	0.044	0.033
Moderate	0.938	0.945	0.041	0.031	0.931	0.935	0.043	0.033
Moderately severe	0.895	0.890	0.038	0.029	0.926	0.930	0.040	0.030
Severe	—	—	—	—	0.876	0.883	0.021	0.016

Table 11
Predictive performance of BiSQA Fusion and related frontier models.

Input Features	Params(M)	FLOPs(G)	Seen				Unseen			
			PCC↑	SRCC↑	RMSE↓	MAE↓	PCC↑	SRCC↑	RMSE↓	MAE↓
Quality-Net [33]	0.30	0.01	0.723	0.714	0.113	0.094	0.545	0.364	0.093	0.077
MOS-Net [34]	1.18	7.44	0.712	0.706	0.115	0.096	0.535	0.362	0.094	0.077
HASA-Net [15]	0.45	0.45	0.815	0.812	0.068	0.050	0.683	0.656	0.069	0.054
HASA-Net ⁺ [17]	0.71	0.53	0.925	0.929	0.045	0.034	0.866	0.866	0.047	0.037
SQINet-HLlevel [12]	1.40	18.11	0.810	0.808	0.066	0.049	0.742	0.758	0.065	0.050
MTSE-LA [16]	1.61	9.11	0.833	0.828	0.062	0.046	0.778	0.763	0.061	0.047
BiSQA Fusion	1.17	2.67	0.942	0.944	0.039	0.030	0.906	0.901	0.041	0.032

Note: Bold values indicate the best performance.

Moderately Severe, Severe). Since all CPC1 listeners belong to the Bilateral HL category, the DNS Challenge-based binaural simulation dataset introduced in Section 4 is used to extend the evaluation to Unilateral HL and Bilateral NH conditions.

Table 9 reports the per-scenario performance on the DNS simulated test set and the CPC1 *Seen* test set. BiSQA Fusion generalizes consistently across NH and HL scenarios, which we attribute to its multi-level fusion mechanism, which jointly exploits acoustic features and audiogram priors to extract cues sensitive to individualized auditory pathology. One notable exception arises in the Bilateral NH subset of the DNS data: prediction error remains low, yet SRCC drops markedly. Bilateral NH samples bypass personalized gain compensation, so their quality scores cluster tightly within the high range, and the resulting low score variance limits SRCC’s capacity to resolve fine-grained rank differences.

Table 10 further breaks the performance down by severity. BiSQA Fusion remains stable across all grades, with the only deviation observed at the two extremes: in the Normal and Severe subgroups, low absolute error coexists with mildly reduced SRCC. The cause is consistent with the Bilateral NH case: scores in the Normal subgroup concentrate near the upper bound, whereas those in the Severe subgroup, even after compensation, concentrate near the lower bound. Such local low-variance ranges compress the ranking signal available to SRCC. To mitigate this in future work, an ordinal rank loss will be incorporated to enforce monotonicity between predicted and reference scores, thereby improving discriminability in low-variance regimes.

5.5. Comparison and analysis of BiSQA Fusion and related frontier models

Table 11 reports a comparison between BiSQA Fusion and several SOTA non-intrusive speech quality assessment models on the CPC1 *Seen* and *Unseen* test sets. All reported differences pass a two-tailed Student’s *t*-test ($p < .05$). The two NH-oriented baselines, Quality-Net and MOS-Net, underperform on both correlation and error metrics, primarily because their reliance on global or adaptive average pooling for score regression dilutes the perceptually salient frames, and the lack of internal normalization (BN/LN) further restricts convergence on high-variance pathological speech. The four HA-oriented baselines (HASA-Net, HASA-Net⁺, SQINet-HLlevel, and MTSE-LA) substantially close this gap through domain-specific designs: HASA-Net, HASA-Net⁺, and MTSE-LA exploit MTL over HASQI and HASPI to regularize deep acoustic representations through task-level physiological correlations; SQINet-HLlevel integrates deep convolutions with MHA to capture cross-frequency and cross-temporal dependencies; and MTSE-LA leverages pseudo-label-assisted training to mitigate biases arising from HA configuration diversity.

Notably, even when compared to the strong SOTA baselines, BiSQA Fusion consistently delivers superior accuracy on both CPC1 tracks, with the gain most pronounced on the more challenging *Unseen* set. Relative to MTSE-LA, PCC and SRCC improve by 0.128 and 0.138, and RMSE and MAE drop by 0.020 and 0.015, respectively. Even against HASA-Net⁺, which shares the same WavLM-Large-weighted representations and au-

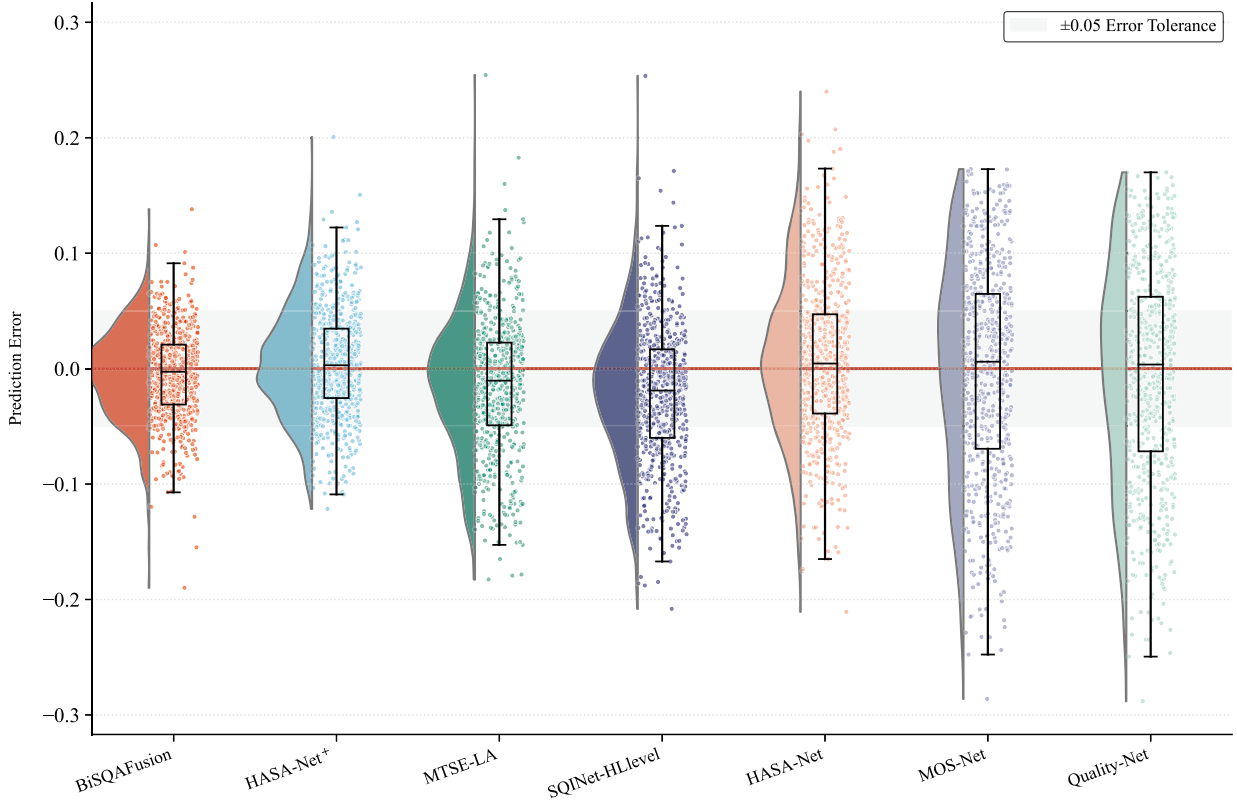


Fig. 7. Comprehensive statistical profiling of prediction errors across BiSQAFusion and frontier models.

diagrams as inputs, BiSQAFusion improves PCC and SRCC by 0.040 and 0.035, respectively, and lowers RMSE and MAE by 0.006 and 0.005, respectively. This advantage is attributable to the multi-level fusion design in this study: the SSL branch aggregates multi-layer hidden features to yield acoustic representations robust to background noise; the AAFM deeply fuses these representations with audiogram priors; the BFM decouples interaural interaction cues across the two channels; the auxiliary QLM jointly regularizes the score prediction through MTL; and the explicit System ID embedding neutralizes biases induced by heterogeneous HA processing. The raincloud plots in Fig. 7 provide a complementary view on the *Unseen* set: MOS-Net and Quality-Net exhibit broad error dispersion, indicative of unstable predictions in binaural scenarios, whereas BiSQAFusion produces the most compact distribution, with the majority of samples and the density peak confined within a ± 0.05 error tolerance band, powerfully verifying its absolute superiority in predictive accuracy.

To evaluate the overall efficiency of BiSQAFusion, this study utilized the FlopCountAnalysis library to quantify the parameter size and computational complexity of relevant frontier baselines (Table 11). Although Quality-Net is the most lightweight, its shallow architecture fails to meet high accuracy requirements, while MOS-Net, despite a moderate budget, remains inadequate on complex binaural degraded speech. In contrast, BiSQAFusion achieves the highest predictive accuracy with parameter scale and computational overhead substantially below those of SQINet-HLlevel (18.11 G) and MTSE-LA (9.11 G), and its 2.67 G FLOPs are also lower than MOS-Net (7.44 G), the second-lowest computational cost among the HA-oriented models, after the HASA-Net series. A fair comparison with the strongest baseline, HASA-Net⁺, requires distinguishing the two complexity dimensions: BiSQAFusion has a comparable parameter count (1.17 M vs. 0.71 M, $1.6\times$) but approximately $5\times$ the FLOPs (2.67 G vs. 0.53 G), owing to the convolutional and BiLSTM stages introduced for deep non-linear fusion. We therefore do not claim superior efficiency over HASA-Net⁺. Instead, BiSQAFusion

trades this additional compute for a $+0.040$ PCC / $+0.035$ SRCC gain on the CPC1 *Unseen* set, and the acceptability of this trade depends on the deployment scenario. Crucially, BiSQAFusion demonstrates architectural flexibility. When configured for unilateral processing (Monaural mode), its parameter size is merely 0.51 M (1.13 G FLOPs), placing it in the same weight class as the HASA-Net series. Even under this lightweight configuration, the MAE on the CPC1 *Unseen* test set increases by only 0.003, maintaining an extremely competitive performance level. This validates the architecture’s potential for resource-constrained deployment. Overall, these results suggest that minimal complexity alone is not a sufficient criterion for HA speech quality assessment models: for accuracy-critical offline scenarios the full binaural configuration is preferable, whereas the monaural mode offers a deployment-oriented operating point with a negligible accuracy penalty.

6. Conclusion

This paper proposes BiSQAFusion, a systematic multi-level fusion network for the non-intrusive quality assessment of binaural HA speech. By deeply integrating SSL acoustic representations, audiogram priors, interaction cues, and HA configuration identifiers, the proposed model effectively addresses two key challenges in this field: the lack of a dedicated architecture for comprehensive binaural quality assessment and limited robustness across diverse HA processing conditions. Architecturally, BiSQAFusion consists of the SSL-Encoder, HL-Encoder, AAFM, BFM, QLM, and SPM. The dual-encoder design learns compact and discriminative representations from speech and audiograms, thereby facilitating cross-modal fusion. AAFM combines gated FiLM conditioning with CNN-BiLSTM modeling to bridge the gap between dynamic acoustic representations and static hearing loss priors. To transcend monaural limitations, a lightweight BFM explicitly models interaural interaction cues through a perceptually motivated, parameter-free decoupling-reconstruction logic. In addition, System ID embedding alleviates

prediction bias caused by diverse HA processing strategies, while QLM further improves score regression through auxiliary multi-task optimization. Extensive experiments confirm that BiSQA-Fusion significantly outperforms SOTA methods in terms of prediction accuracy and generalization, while remaining robust to diverse binaural hearing-loss patterns and HA configurations. In future work, we plan to enhance acoustic representation by continuing to explore cutting-edge large-scale SSL models. Additionally, we aim to employ knowledge distillation techniques to reduce model complexity, enabling deployment on resource-constrained HA devices. Furthermore, once a dataset of subjective binaural speech quality ratings for HAs becomes publicly available, we will conduct more in-depth validation of the proposed network to improve its clinical applicability.

CRediT authorship contribution statement

Yang Yang: Writing – original draft, Methodology; **Ruiyu Liang:** Supervision, Project administration, Funding acquisition; **Ye Ni:** Formal analysis, Data curation; **Cairong Zou:** Visualization, Validation, Software; **Song Sun:** Investigation, Conceptualization; **Jing Yan:** Software, Data curation; **Wei Zhou:** Validation, Conceptualization; **Weiping Ding:** Resources, Investigation; **Xiaoshuai Hao:** Writing – review & editing, Supervision.

Declaration of competing interest

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

Acknowledgments

This work was supported in part by the [National Natural Science Foundation of China](#) under grant number 62601396 and 61871213.

Data availability

The data that has been used is confidential.

References

- [1] S. Chadha, K. Kamenov, A. Cieza, The world report on hearing, 2021, Bull. World Health Organ. 99 (4) (2021) 242.
- [2] Y. Ni, R. Liang, X. Hao, J. Cheng, Q. Wang, C. Huang, C. Zou, W. Zhou, W. Ding, B.W. Schuller, Affine modulation-based audiogram fusion network for joint noise reduction and hearing loss compensation, Inf. Fusion 127 (2026) 103726.
- [3] J.B.B. Nielsen, J. Nielsen, J. Larsen, Perception-based personalization of hearing aids using Gaussian processes and active learning, IEEE/ACM Trans. Audio Speech Lang. Process. 23 (1) (2014) 162–173.
- [4] R. Liang, M. Ju, F. Kong, Y. Xie, G. Tang, A Non-Intrusive speech quality evaluation algorithm for hearing aids via an auxiliary training task, Appl. Acoust. 206 (2023) 109312.
- [5] F. Drakopoulos, S. Verhulst, A neural-network framework for the design of individualised hearing-loss compensation, IEEE/ACM Trans. Audio Speech Lang. Process. 31 (2023) 2395–2409.
- [6] Q. Liang, Y. Shen, T. Chen, L. Zhang, S. Zhao, ADTMOS - Synthesized speech quality assessment based on audio distortion tokens, IEEE Trans. Audio Speech Lang. Process. 33 (2025) 1493–1507.
- [7] K. Shen, D. Yan, J. Hu, Z. Ye, Non-intrusive speech quality assessment: a survey, Neurocomputing 580 (2024) 127471.
- [8] J.M. Kates, K.H. Arehart, The hearing-aid speech quality index (HASQI), J. Audio Eng. Soc. 58 (5) (2010) 363–381.
- [9] J.M. Kates, K.H. Arehart, The hearing-aid speech quality index (HASQI) version 2, J. Audio Eng. Soc. 62 (3) (2014) 99–117.
- [10] J.M. Kates, K.H. Arehart, The hearing-aid audio quality index (HAAQI), IEEE/ACM Trans. Audio Speech Lang. Process. 24 (2) (2016) 354–365.
- [11] R. Huber, V. Parsa, S. Scollie, Predicting the perceived sound quality of frequency-compressed speech, Plos One 9 (11) (2014) e110260.
- [12] R. Liang, Y. Xie, J. Cheng, C. Pang, B. Schuller, A non-Invasive speech quality evaluation algorithm for hearing aids with multi-Head self-Attention and audiogram-Based features, IEEE/ACM Trans. Audio Speech Lang. Process. 32 (2024) 2166–2176.
- [13] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, F. Wei, WavLM: large-Scale self-Supervised pre-Training for full stack speech processing, IEEE J. Sel. Top. Signal Process. 16 (6) (2022) 1505–1518.

- [14] A. Pasad, B. Shi, K. Livescu, Comparative layer-Wise analysis of self-Supervised speech models, in: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.
- [15] H.-T. Chiang, Y.-C. Wu, C. Yu, T. Toda, H.-M. Wang, Y.-C. Hu, Y. Tsao, HASA-Net: A non-intrusive hearing-aid speech assessment network, in: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2021, pp. 907–913.
- [16] Y. Yang, R. Liang, Y. Ni, Y. Xie, C. Zou, B.W. Schuller, A non-Intrusive speech quality evaluation framework for hearing aids based on speech label assistance and multi-Task learning strategy, IEEE Trans. Audio Speech Lang. Process. 33 (2025) 3332–3346.
- [17] H.-T. Chiang, S.-W. Fu, H.-M. Wang, Y. Tsao, J.H.L. Hansen, Multi-objective non-intrusive hearing-aid speech assessment model, J. Acoust. Soc. Am. 156 (5) (2024) 3574–3587.
- [18] R. Guo, P. Jiang, R. Liang, Y. Xie, C. Zou, A non-intrusive speech quality evaluation method based on the audiogram and weighted frequency information for hearing aid, IEICE Trans. Fundam. Electron. Commun. Comput. Sci. 106 (1) (2023) 64–68.
- [19] D.A.M.G. Wisnu, S. Rini, R.E. Zezario, H.-M. Wang, Y. Tsao, HAAQI-Net: A non-Intrusive neural music audio quality assessment model for hearing aids, IEEE Trans. Audio Speech Lang. Process. 33 (2025) 1877–1892.
- [20] M. Schäfer, M. Bahram, P. Vary, An extension of the PEAQ measure by a binaural hearing model, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 8164–8168.
- [21] J. Reimes, DNN-Based Speech quality assessment for binaural signals, in: 2022 International Workshop on Acoustic Signal Enhancement (IWAENC), 2022, pp. 1–5.
- [22] J.-H. Fleßner, R. Huber, S.D. Ewert, Assessment and prediction of binaural aspects of audio quality, J. Audio Eng. Soc. 65 (11) (2017) 929.
- [23] J.-H. Fleßner, T. Biberger, S.D. Ewert, Subjective and objective assessment of monaural and binaural aspects of audio quality, IEEE/ACM Trans. Audio Speech Lang. Process. 27 (7) (2019) 1112–1125.
- [24] A.H. Andersen, J.M. de Haan, Z.-H. Tan, J. Jensen, Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions, Speech Commun. 102 (2018) 1–13.
- [25] J. Barker, M. Akeroyd, T.J. Cox, J.F. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, et al., The 1st clarity prediction challenge: a machine learning challenge for hearing aid intelligibility prediction, in: Interspeech, 2022, pp. 3508–3512.
- [26] J. Barker, M.A. Akeroyd, W. Bailey, T.J. Cox, J.F. Culling, J. Firth, S. Graetzer, G. Naylor, The 2nd clarity prediction challenge: a machine learning challenge for hearing aid intelligibility prediction, in: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 11551–11555.
- [27] R.E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, Y. Tsao, Deep learning-Based non-Intrusive multi-Objective speech assessment model with cross-Domain features, IEEE/ACM Trans. Audio Speech Lang. Process. 31 (2023) 54–70.
- [28] A.A. Kressner, D.V. Anderson, C.J. Rozell, Robustness of the hearing aid speech quality index (HASQI), in: 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011, pp. 209–212.
- [29] A.A. Kressner, D.V. Anderson, C.J. Rozell, Evaluating the generalization of the hearing aid speech quality index (HASQI), IEEE Trans. Audio Speech Lang. Process. 21 (2) (2013) 407–415.
- [30] D. Suelzle, V. Parsa, T.H. Falk, On a reference-free speech quality estimator for hearing aids, J. Acoust. Soc. Am. 133 (5) (2013) EL412–EL418.
- [31] H. Salehi, V. Parsa, On nonintrusive speech quality estimation for hearing aids, in: 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015, pp. 1–5.
- [32] H. Salehi, D. Suelzle, P. Folkeard, V. Parsa, Learning-Based reference-Free speech quality measures for hearing aid applications, IEEE/ACM Trans. Audio Speech Lang. Process. 26 (12) (2018) 2277–2288.
- [33] S.-W. Fu, Y. Tsao, H.-T. Hwang, H.-M. Wang, Quality-Net: an end-to-end non-intrusive speech quality assessment model based on BLSTM, in: Proc. Annu. Conf. Int. Speech Commun. Assoc., INTERSPEECH, 2018, pp. 1873–1877.
- [34] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, H.-M. Wang, Mosnet: deep learning based objective assessment for voice conversion, in: Proc. Annu. Conf. Int. Speech Commun. Assoc., INTERSPEECH, 2019, pp. 1541–1545.
- [35] J.M. Kates, K.H. Arehart, The hearing-aid speech perception index (HASPI) version 2, Speech Commun. 131 (2021) 35–46.
- [36] E. Perez, F. Strub, H. De Vries, V. Dumoulin, A. Courville, Film: visual reasoning with a general conditioning layer, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018, pp. 32–38.
- [37] J. Kim, J. Kong, J. Son, Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech, in: International Conference on Machine Learning, PMLR, 2021, pp. 5530–5540.
- [38] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1501–1510.
- [39] B.C.J. Moore, B.R. Glasberg, Modeling binaural loudness, J. Acoust. Soc. Am. 121 (3) (2007) 1604–1612.
- [40] A. Bronkhorst, R. Plomp, Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing, J. Acoust. Soc. Am. 92 (6) (1992) 3132–3139.
- [41] A.W. Bronkhorst, The cocktail-party problem revisited: early processing and selection of multi-talker speech, Atten. Percept. Psychophys. 77 (5) (2015) 1465–1487.

- [42] C.K.A. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matushevych, R. Aichner, A. Aazami, S. Braun, et al., The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results, *arXiv:2005.13981* (2020).
- [43] A. Varga, H.J.M. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems, *Speech Commun.* 12 (3) (1993) 247–251.
- [44] H. Schröter, T. Rosenkranz, A.N. Escalante-B, A. Maier, Deepfilternet: Perceptually motivated real-time speech enhancement, *arXiv:2305.08227* (2023).
- [45] J. Cheng, R. Liang, L. Zhao, C. Huang, B.W. Schuller, Speech denoising and compensation for hearing aids using an FT-CRN-Based metric GAN, *IEEE Signal Process. Lett.* 30 (2023) 374–378.
- [46] J.M. Kates, K.H. Arehart, An overview of the HASPI and HASQI metrics for predicting speech intelligibility and speech quality for normal hearing, hearing loss, and hearing aids, *Hear. Res.* 426 (2022) 108608.
- [47] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, *Adv. Neural Inf. Process. Syst.* 33 (2020) 12449–12460.
- [48] W.-N. Hsu, B. Bolte, Y.-H.H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, HuBERT: self-Supervised speech representation learning by masked prediction of hidden units, *IEEE/ACM Trans. Audio Speech Lang. Process.* 29 (2021) 3451–3460.
- [49] R.H. Margolis, G.L. Saly, Asymmetric hearing loss: definition, validation, and prevalence, *Otol. Neurotol.* 29 (4) (2008) 422–431.