Neuro-MSBG: An End-to-End Neural Model for Hearing Loss Simulation

Hui-Guan Yuan*†, Ryandhimas E. Zezario*, Shafique Ahmed*, Hsin-Min Wang*, Kai-Lung Hua†‡, Yu Tsao*

*Academia Sinica, Taipei, Taiwan
†National Taiwan University of Science and Technology, Taipei, Taiwan
‡Microsoft, Taipei, Taiwan

***Abstract*—Hearing loss simulation models are essential for hearing aid deployment. However, existing models have high computational complexity and latency, which limits real-time applications, and lack direct integration with speech processing systems. To address these issues, we propose Neuro-MSBG, a lightweight end-to-end model with a personalized audiogram encoder for effective time-frequency modeling. Experiments show that Neuro-MSBG supports parallel inference and retains the intelligibility and perceptual quality of the original MSBG, with a Spearman's rank correlation coefficient (SRCC) of 0.9247 for Short-Time Objective Intelligibility (STOI) and 0.8671 for Perceptual Evaluation of Speech Quality (PESQ). Neuro-MSBG reduces simulation runtime by 46 times (from 0.970 seconds to 0.021 seconds for a 1 second input), further demonstrating its efficiency and practicality.**

***Index Terms*—hearing loss model, mamba, differentiable framework, audiogram, real-time inference**

I. Introduction

Hearing loss simulation models aim to simulate how hearing impairment affects sound processing in the auditory system and have become essential tools in both research and evaluation. For example, the *Clarity Challenge* [1] uses the Moore, Stone, Baer, and Glasberg (MSBG) model [2]–[5] to simulate individual perceptual conditions based on audiograms. Similarly, the *Cadenza Challenge* [6] and the *Clarity Challenge* adopt perceptually grounded metrics such as the hearing-aid speech perception index (HASPI) [7], the hearing-aid speech quality index (HASQI) [8], and the hearing-aid audio quality index (HAAQI) [9], which embed auditory processing to assess quality and intelligibility under hearing loss conditions.

Existing hearing loss models are generally divided into two categories: physiological models and engineering-oriented models. Physiological models, such as the model proposed in [10] and the transmission-line (TL) cochlear model [11], are designed to accurately model cochlear mechanics, but their complexity limits real-time integration. In contrast, engineering-oriented models, such as the Hohmann filterbank [12], the Auditory Toolbox [13], and MSBG, balance deployment practicality with perceptual accuracy and provide computational stability, making them well suited for real-time speech processing. Among engineering-oriented models, MSBG [2] is currently the most widely used. It simulates sensorineural hearing loss based on audiograms and reproduces key perceptual effects. Despite its widespread adoption in both academic and practical applications, MSBG has two major limitations: (i) it does not support parallel processing, which reduces its applicability to real-time speech systems; and (ii) it relies on multiple filtering stages, which introduces variable delays. These limitations restrict the integration of MSBG into end-to-end learning frameworks and reduce its effectiveness in real-time or large-scale speech processing.

Recent studies [14], [15] have attempted to simplify both physiological and engineering-oriented models for real-time applications. For example, CoNNear [14] simplifies the TL cochlear model and supports real-time simulation of auditory nerve responses. Similarly, P Leer et al. [16] trained neural models to emulate the Verhulst auditory periphery model for varying hearing-loss profiles. However, the lack of waveform-level output generation prevents waveform-level supervision and reduces its applicability to speech processing and hearing aid systems. For engineering-oriented models, the Wakayama University Hearing Impairment Simulator (WHIS) [15] addresses the latency and computational cost issues associated with MSBG. WHIS first computes the cochlear excitation pattern of a target hearing-impaired individual using a Gammachirp filterbank, and then dynamically generates a time-varying minimum-phase filter based on this pattern to transform normal-hearing speech into its hearing-loss-simulated counterpart. This method reduces the processing time for one second of speech to approximately 10 milliseconds while maintaining near-perfect temporal alignment with the original waveform. Despite its efficiency, WHIS still relies on frame-by-frame computation of infinite impulse response (IIR) filter coefficients and dynamic gain selection, and has not been optimized for vectorized or parallel processing. Therefore, its integration into an end-to-end deep learning framework remains challenging, and joint optimization with compensation models remains an open research direction.

With the increasing use of differentiable hearing loss models in hearing aid compensation, optimizing their design and performance has become a research focus. In physiological modeling, the auditory nerve responses generated by CoNNear have been used as loss functions to guide the training of compensation models [17]–[19]. This approach attempts to make the neural responses of hearing-impaired people when receiving compensated speech similar to the neural responses of normal-hearing people when listening to the original

signal. In engineering-oriented models, Tu et al. proposed the Differentiable Hearing Aid Speech Processing (DHASP) framework [20], which reimplements the auditory processing pipeline in HASPI [7] and uses differentiable modules for backpropagation. Tu et al. also introduced a differentiable version of the MSBG model and applied it to the training of hearing aid algorithms [21]. These engineering-oriented approaches typically consist of differentiable finite impulse response (FIR) filters and audio processing steps designed to approximate auditory mechanisms.

Despite some progress in differentiable engineering-oriented hearing loss models, most efforts have focused on magnitude-domain simulation, with limited attention paid to the role of phase information. Meanwhile, recent advances in speech enhancement have highlighted the importance of phase modeling for perceptual quality. For instance, MP-SENet [22] adopts a joint enhancement strategy for both magnitude and phase spectra, achieving significantly better performance than traditional magnitude-only methods and highlighting the importance of incorporating phase modeling. Inspired by these findings, we investigate the role of phase information in hearing loss simulation and propose Neuro-MSBG, an end-to-end fully differentiable hearing loss model. Neuro-MSBG outputs simulated audio in the waveform domain, which can be directly integrated with modern speech enhancement systems that rely on waveform-based losses and evaluation metrics (e.g., mean squared error (MSE), short-time objective intelligibility (STOI) [23], and perceptual evaluation of speech quality (PESQ) [24]). It also supports noisy speech input, further enhancing its practical applicability. Experimental results show that the addition of phase processing significantly improves the fidelity of MSBG hearing loss simulation, highlighting the importance of phase modeling in replicating authentic auditory perception. The main contributions of our model are as follows:

- **Parallelizable and lightweight simulation:** Neuro-MSBG achieves parallel inference, reducing the simulation time for one second of audio from 0.970 seconds in the original MSBG to 0.021 seconds, a **46× speedup**.
- **Seamless integration with end-to-end speech systems:** By resolving the delay issues inherent in the original MSBG, Neuro-MSBG can be integrated into modern speech compensator training pipelines.
- **Phase-aware modeling:** By incorporating phase information, Neuro-MSBG maintains the intelligibility and perceptual quality of the original MSBG, achieving a Spearman's rank correlation coefficient (SRCC) of 0.9247 for STOI and 0.8671 for PESQ.

At the end of this paper, we also demonstrate the preliminary integration of Neuro-MSBG into a speech compensator pipeline, thus confirming its practicality as a differentiable hearing loss simulation module. The remainder of this paper is organized as follows. Section II presents the proposed method. Section III describes the experimental setup and results. Finally, Section IV presents conclusions.

II. Methodology

This section introduces the model architecture and training criteria of Neuro-MSBG. As shown in Fig. 1, the model takes normal speech signals and monaural audiograms as input. The audiogram is transformed into personalized hearing features through the Audiogram Encoder, while the speech signal is converted into the time-frequency domain through Short-Time Fourier Transform (STFT) to obtain magnitude and phase features. These three types of features, including personalized hearing features, magnitude features, and phase features, are concatenated and then input into the Neural Network Block (NN Block). The network then branches into two decoders: the Magnitude Mask Decoder and the Phase Decoder, which respectively predict the magnitude and phase shifts associated with hearing loss. Finally, the predicted magnitude and phase are combined to reconstruct the speech signal perceived by the hearing-impaired listener through inverse STFT.

A. Neuro-MSBG

Our model adopts an architecture based on MP-SENet [22] and the advanced SE-Mamba framework [25]. This design is inspired by our experimental findings that phase information is critical for accurate hearing loss simulation (see Section III for details). To evaluate how well different neural modules capture spectral and temporal cues, we replace the original attention-based components with alternatives such as LSTM, Transformer, CNN, and Mamba blocks. Given Mamba's efficiency in modeling long sequences and its low latency, we also control the number of parameters to ensure that the model remains lightweight and effective.

For audiogram integration, previous methods typically concatenate the audiogram representation along the frequency dimension. In contrast, we found that treating the audiogram as an additional input channel in addition to magnitude and phase produces more stable and effective results. This channel-based integration may enable the model to receive more consistent, spatially aligned conditioning across layers, thereby improving its ability to modulate internal feature representations.

1) Audiogram Encoder: To incorporate personalized hearing profiles, we design a lightweight Audiogram Encoder that transforms the audiogram $\mathbf{a} \in \mathrm{R}^{B\times 8}$ into a frequency-aligned representation $\mathbf{a}_{\text{enc}} \in \mathrm{R}^{B\times F}$, where B denotes the batch size, and $F = 201$ matches the STFT resolution. The transformation process is defined as:

$$\mathbf{a}_{\text{enc}} = \mathbf{W} \cdot \text{Flatten}\left(\text{AvgPool}\left(\sigma\left(\text{Conv}(\mathbf{a})\right)\right)\right), \qquad (1)$$

where Conv is a 1D convolution layer, σ denotes a ReLU activation, and $\mathbf{W}$ is a linear projection matrix.

The encoded vector is then broadcast along the time axis and concatenated with the magnitude and phase features to form the three-channel input $\mathrm{R}^{B\times 3\times T\times F}$ of the DenseEncoder and NN Block. This channel-based integration enables the model to consistently inject hearing-profile information in both time and frequency dimensions.

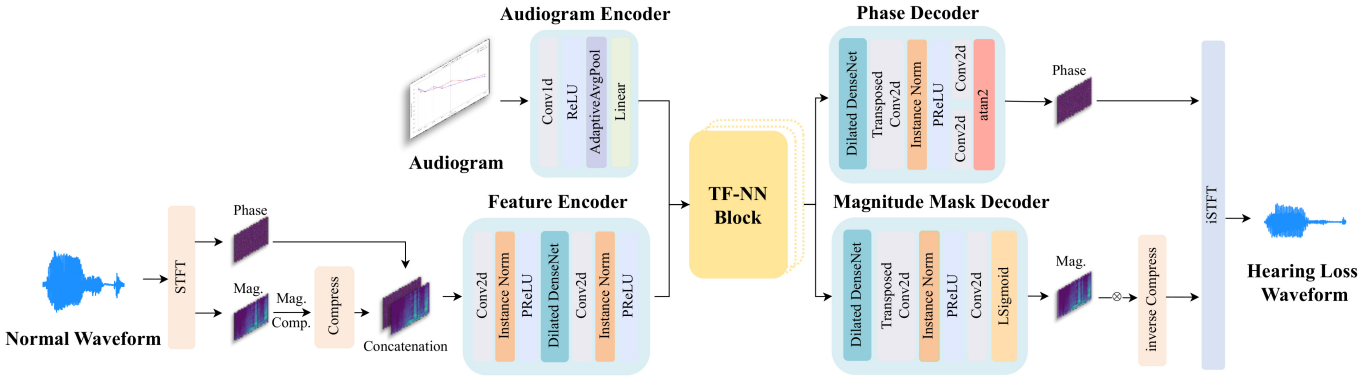

Fig. 1. **Overview of the proposed Neuro-MSBG framework.**

2) Neural Network Blocks: To capture the temporal and spectral structure of hearing-loss-affected speech, we design and compare multiple NN Blocks, each of which adopts a dual-path architecture to process the time and frequency dimensions separately. The input tensor is first rearranged and reshaped for temporal modeling, followed by a similar process for frequency modeling. Each path contains residual connections and a ConvTranspose1d layer to restore the original shape to ensure compatibility with subsequent modules.

In this unified framework, we replace the core time-frequency module with one of the following four alternatives:

- **Mamba Block**: Combined with bidirectional Mamba modules, time and frequency are modeled separately to provide efficient long-range dependency modeling.
- **Transformer Block**: Transformer encoder layers are applied to both axes, and global attention is used to capture contextual information.
- **LSTM Block**: Bidirectional LSTM is used to model sequential patterns, and then linear projection is performed to maintain dimensional consistency.
- **CNN Block**: One-dimensional convolution is used to extract local features, followed by channel expansion, activation, and residual fusion.

This dual-axis design forms a flexible and stable framework for the simulation of hearing loss. It also allows for systematic comparison across architectures, demonstrating Mamba's potential for low-latency, high-fidelity speech modeling.

B. Training Criteria

We adopt a multi-objective loss to jointly supervise spectral accuracy, phase consistency, and time-domain fidelity. The **magnitude loss** $\mathcal{L}_{\text{Mag}}$ is defined as the MSE between the predicted magnitude $\hat{m}$ and the ground-truth magnitude m:

$$\mathcal{L}_{\text{Mag}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{m}_i - m_i\|^2 , \tag{2}$$

where N is the number of training samples. The **phase loss** $\mathcal{L}_{\text{Pha}}$ is inspired by the anti-wrapping strategy proposed in [26] and consists of three components: the instantaneous phase loss $\mathcal{L}_{\text{IP}}$, the group delay loss $\mathcal{L}_{\text{GD}}$, and the integrated absolute frequency loss $\mathcal{L}_{\text{IAF}}$, defined as:

$$\mathcal{L}_{\text{IP}} = \mathbb{E}_{p,\hat{p}}\left[\|f_{\text{aw}}\left(p-\hat{p}\right)\|_1\right], \tag{3}$$

$$\mathcal{L}_{\text{GD}} = \mathbb{E}_{p,\hat{p}}\left[\|f_{\text{aw}}\left(\Delta_{\text{F}}(p-\hat{p})\right)\|_1\right], \tag{4}$$

$$\mathcal{L}_{\text{IAF}} = \mathbb{E}_{p,\hat{p}}\left[\|f_{\text{aw}}\left(\Delta_{\text{T}}(p-\hat{p})\right)\|_1\right], \tag{5}$$

$$\mathcal{L}_{\text{Pha}} = \mathcal{L}_{\text{IP}} + \mathcal{L}_{\text{GD}} + \mathcal{L}_{\text{IAF}}, \tag{6}$$

where $f_{\text{aw}}(\cdot)$ denotes the anti-wrapping function used to mitigate 2π discontinuity, $\Delta_{\text{F}}(\cdot)$ and $\Delta_{\text{T}}(\cdot)$ represent the first-order differences of the phase error along the *frequency* axis and *time* axis, respectively. The **complex loss** $\mathcal{L}_{\text{Com}}$ measures the MSE between the predicted complex spectrogram $\hat{c}$ and the ground-truth complex spectrogram c (including both real and imaginary parts):

$$\mathcal{L}_{\text{Com}} = 2 \cdot \frac{1}{N}\sum_{i=1}^{N} \|\hat{c}_i - c_i\|^2 . \tag{7}$$

The **time-domain loss** $\mathcal{L}_{\text{Time}}$ is calculated as the L_1 distance between the predicted waveform $\hat{x}$ and the reference waveform x to preserve temporal fidelity:

$$\mathcal{L}_{\text{Time}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{x}_i - x_i\|_1 . \tag{8}$$

The total training loss is a weighted sum of all components:

$$\mathcal{L}_{\text{Total}} = \lambda_{\text{Mag}}\mathcal{L}_{\text{Mag}} + \lambda_{\text{Pha}}\mathcal{L}_{\text{Pha}} + \lambda_{\text{Com}}\mathcal{L}_{\text{Com}} + \lambda_{\text{Time}}\mathcal{L}_{\text{Time}}, \tag{9}$$

where each λ is a tunable scalar weight used to balance the contribution of each loss term.

III. Experiment

The dataset comprises 12,396 clean utterances (11,572 for training and 824 for testing) from VoiceBank [27] and their noisy counterparts from VoiceBank–DEMAND [28], created by mixing each clean utterance with one randomly selected noise type from 10 DEMAND recordings [29] and one signal-to-noise-ratio (SNR) level. To ensure generalization, 8 noise types are used for training and 2 disjoint types for testing, with SNR levels drawn from {0, 5, 10, 15} dB for training and

TABLE I
PERFORMANCE COMPARISON BETWEEN MONOLITHIC MODELS AND OUR MODULAR NEURO-MSBG FRAMEWORK WITH DIFFERENT TF-NN BLOCKS.

Model Type	Architecture	STOI LCC	STOI SRCC	STOI MSE	PESQ LCC	PESQ SRCC	PESQ MSE	RAW MSE
Monolithic	Transformer	0.7123	0.6932	0.0028	0.6026	0.6250	0.1677	0.0870
	CNN	0.8156	0.8037	0.0019	0.6058	0.6433	0.1455	0.0670
	LSTM	0.5004	0.5064	0.0137	0.3291	0.4193	0.2242	0.1054
Neuro-MSBG TF-NN Block Replacement	Transformer	0.7271	0.7593	0.0012	0.5326	0.5823	0.4910	0.0529
	CNN	0.8234	0.8591	0.0009	0.7374	0.7580	0.1606	0.0670
	LSTM	0.8443	0.8999	0.0006	0.8226	0.8312	0.0801	0.0691
	Mamba	**0.8475**	**0.9247**	**0.0006**	**0.8519**	**0.8671**	**0.0782**	**0.0669**

$\{2.5, 7.5, 12.5, 17.5\}$ dB for testing. Each utterance is paired with two monaural audiograms, representing different levels of hearing loss, ranging from mild to severe. Consequently, the training set is expanded to $11,572 \times 2 \times 2 = 46,288$ samples, and the test set is expanded to $824 \times 2 \times 2 = 3,296$ samples. Speech data and audiograms are disjoint across training and testing splits, ensuring that the model is evaluated on unseen utterances and unseen hearing-loss profiles.

Specifically, this study conducted a series of experiments covering five main aspects: (i) comparison of Neuro-MSBG with different architectures and monolithic baselines (Table I); (ii) runtime evaluation (Table II); (iii) validation of the necessity of phase prediction (Table III); (iv) comparison of audiogram integration strategies (Table IV); and (v) application of Neuro-MSBG in a speech compensator (Table V). In all quantitative tables, **bold** highlights the best-performing model and underline marks the second best. All experiments were performed on a single NVIDIA RTX 3090. The models were trained for 200 epochs with a batch size of 3, an initial learning rate of 0.0005, and the AdamW optimizer.

A. Data Preparation

We use the MSBG model to simulate hearing loss by applying an ear-specific gain curve to each input, resulting in single-channel impaired speech. However, due to the multi-stage filtering in MSBG, the output may exhibit unpredictable delay relative to the original signal. Such misalignment is undesirable for downstream tasks, such as speech enhancement or intelligibility assessment.

To estimate and correct this delay, we generate an auxiliary reference signal along with the main audio during the MSBG process [1]. This auxiliary signal is a silent waveform with the same length and sampling rate as the input, used solely for delay tracking. A unit impulse is inserted into this signal, which is defined as:

$$\delta[n-k] = \begin{cases} 1, & \text{if } n = k \\ 0, & \text{otherwise} \end{cases}, \tag{10}$$

where n denotes the discrete time index, and $k = \frac{F_s}{2}$ is the sample position of the impulse, where F_s represents the sampling rate in Hz.

The impulse is inserted at the midpoint of the auxiliary reference signal. After MSBG simulation, we compare the pre/post impulse positions to estimate the delay introduced by MSBG. We then use the estimated delay to time-align the original normal-hearing input, the clean reference, and the impaired output to ensure accurate evaluation through metrics such as STOI and PESQ that are highly sensitive to timing errors. Each training instance consists of: 1) the single-ear impaired speech, 2) the aligned clean reference, 3) the aligned normal-hearing input, and 4) the associated 8-dimensional audiogram vector. Fig. 2 shows the waveform alignment before and after the delay-compensation shift.

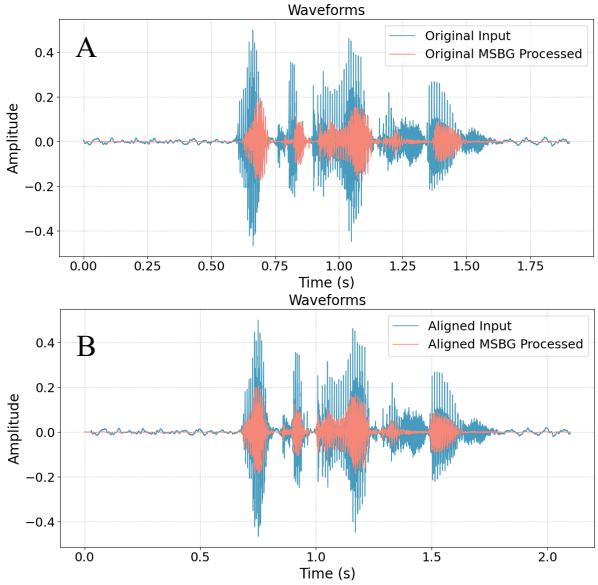

Fig. 2. **Waveform alignment before and after shifting.** In A, the MSBG-processed signal (red) shows a clear delay relative to the original input (blue). In B, the waveforms are time-aligned using impulse-based method, allowing for a fair and accurate assessment of the effect of MSBG.

B. Experimental Results

In the early stages of our experiment, we used monolithic models with unified architectures such as CNN, LSTM, and Transformer for hearing simulation. However, due to limited performance, we subsequently adopted the Neuro-MSBG framework, replacing only the TF-NN block with different architectures, including CNN, LSTM, Transformer, and Mamba. The results are shown in Table I. To ensure a

TABLE II
COMPARISON OF RUNTIME OF MSBG AND NEURO-MSBG ON DIFFERENT DEVICES. WE MEASURED THE INFERENCE TIME REQUIRED TO PROCESS A 1-SECOND, 44.1 KHZ AUDIO SIGNAL USING AN INTEL XEON GOLD 6152 CPU AND AN NVIDIA RTX 3090 GPU.

Model	CPU	GPU	Param
MSBG	0.970	NA	NA
Neuro-MSBG (Mamba)	NA	0.021 s	1.45M
Neuro-MSBG (LSTM)	0.617 s	0.016 s	1.47M
Neuro-MSBG (CNN)	0.592 s	0.016 s	1.45M
CoNNear	0.025 s	0.099 s	11.7M

fair comparison across different architectures, the number of parameters of all models was set to be roughly the same. From Table I, we observe that the Neuro-MSBG variants consistently outperform the monolithic baselines on unseen test data. Among them, Neuro-MSBG with Mamba achieves the best performance across all evaluation metrics.

Next, we compare Neuro-MSBG with an existing neural network-based model, CoNNear, in terms of model size and inference time. Although the two models have different goals—CoNNear simulates physiologically grounded auditory nerve responses, while our framework focuses on perceptual signal transformation—the comparison is appropriate given their common goal of real-time hearing loss modeling. As shown in Table II, CoNNear has approximately 11.7 million parameters, while Neuro-MSBG has only 1.45 million parameters. In addition to its lightweight architecture, Neuro-MSBG supports noisy input conditions and accommodates a wide range of audiogram configurations, providing additional advantages for practical applications involving diverse acoustic environments and personalized hearing loss profiles.

Table II also shows the inference time of different models. MSBG does not support parallel processing and cannot be executed on GPU; therefore, the corresponding GPU column is marked as NA. In contrast, Neuro-MSBG (Mamba) leverages a CUDA-accelerated selective scan kernel for core operations, which currently only supports GPU execution. Therefore, only the inference time on GPU is reported. In terms of inference time, Neuro-MSBG (Mamba) achieves about 46× speedup on GPU over MSBG on CPU. For CPU-executable variants such as Neuro-MSBG (LSTM), the inference time is 0.016 seconds, which is 60× faster than MSBG's 0.970 seconds. In addition, we also implemented and evaluated the CoNNear model. Despite its larger parameter size (11.7 million), it shows fast inference in our computation environment, with a GPU runtime of 0.099 seconds and a notably fast CPU runtime of 0.025 seconds. The faster CPU runtime of CoNNear than the GPU version is likely due to the batch size of 1 used in this experiment, which limits the benefits of GPU parallelism.

C. Ablation Study

Previous approaches to modeling hearing loss typically predict only the magnitude spectrum while reusing the input phase for waveform reconstruction. We initially adopted this approach; however, our empirical analysis revealed that phase information plays a crucial role in MSBG-based simulation. We conducted an ablation study using Neuro-MSBG (Mamba). As shown in Table III, predicting both magnitude and phase substantially outperforms magnitude-only prediction in all metrics (STOI MSE, PESQ MSE, and waveform-level MSE). Specifically, the STOI MSE decreased from 0.0041 to 0.0006, indicating a notable improvement in intelligibility, while the PESQ MSE decreased from 2.3579 to 0.0782, reflecting improved perceptual quality. At the waveform level, the MSE decreased from 0.0986 to 0.0669, confirming that phase modeling is critical for both perceptual fidelity and accurate signal reconstruction.

TABLE III
PERFORMANCE COMPARISON OF NEURO-MSBG MODELS WITH AND WITHOUT PHASE PREDICTION.

Setting	STOI MSE	PESQ MSE	RAW MSE
Magnitude Only	0.0041	2.3579	0.0986
Magnitude + Phase	**0.0006**	**0.0782**	**0.0669**

Furthermore, in many speech-related applications involving audiograms, a common practice is to concatenate the audiogram vector with the audio features before feeding them into the model. We initially adopted this simple approach, but found that it could not effectively capture the relationship between hearing profiles and spectral features. To address this issue, we introduced a lightweight Audiogram Encoder that transforms the 8-dimensional audiogram vector into a frequency-aligned representation. This representation is appended as a third channel along with the magnitude and phase features. As shown in Table IV, incorporating the Audiogram Encoder leads to consistent reduction in STOI MSE, PESQ MSE, and waveform-level MSE, demonstrating its effectiveness in integrating personalized hearing profiles to improve hearing loss modeling.

D. Qualitative Evaluation

To evaluate the model's performance in reconstructing hearing-loss-affected speech, we compare the log-magnitude spectrograms of speech outputs of seven models and ground-truth speech (Fig. 3). Among them, Neuro-MSBG (Mamba) produces the most accurate reconstruction, preserving the harmonic structure and high-frequency energy. Neuro-MSBG with CNN, LSTM, and Transformer Blocks also preserve key spectral features but exhibit some energy imbalance or mild distortion. In contrast, the baseline models introduce more obvious artifacts: the CNN and LSTM variants lack clarity and high-frequency content, while the Transformer variant has difficulty reconstructing an accurate spectrogram.

E. Training a Compensator with Neuro-MSBG

To advance end-to-end hearing loss compensation, recent work (e.g., NeuroAMP [30]) has integrated audiogram-aware processing directly into neural networks, replacing traditional modular pipelines with personalized, data-driven amplification. Inspired by this direction, we propose a complementary

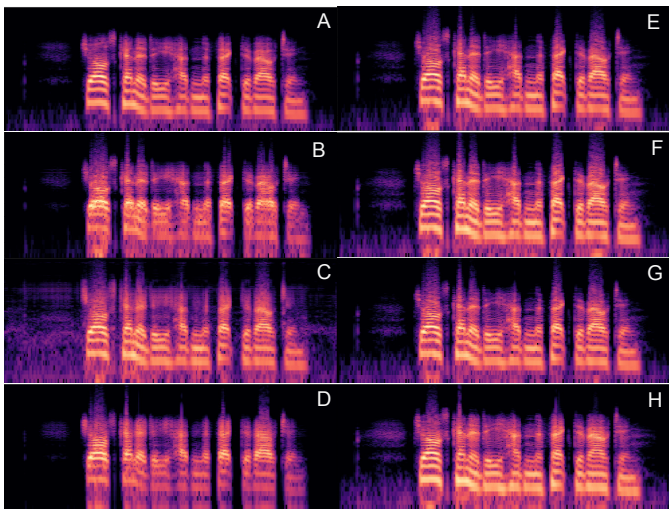

Fig. 3. Log-magnitude spectrograms of speech outputs of seven models and ground-truth speech: (A) ground truth, (B), (C), and (D) outputs of baseline CNN, LSTM, and Transformer models, and (E), (F), (G), and (H) outputs of Neuro-MSBG models using Mamba, CNN, LSTM, and Transformer Blocks.

TABLE IV
EFFECTS OF THE AUDIOGRAM ENCODER IN THE NEURO-MSBG MODEL.

Setting	STOI MSE	PESQ MSE	RAW MSE
w/o Audiogram Encoder	0.0013	0.1152	0.0670
w/ Audiogram Encoder	**0.0006**	**0.0782**	**0.0669**

approach that connects a trainable compensator to a frozen, perceptually grounded simulator (Neuro-MSBG), enabling the compensator to shape its input to match the individual's hearing profiles.

Neuro-MSBG is lightweight, differentiable, and does not require clean reference alignment, making it suitable for integration into an end-to-end hearing loss compensation system. Building on this feature, we explore a new use case: connecting a pre-trained Neuro-MSBG model to a trainable compensator to achieve personalized hearing enhancement, as illustrated in Fig. 4. The training and test sets are from VoiceBank [27].

The goal of the compensator is to transform an input waveform into a personalized, compensated version. This compensated waveform is then passed into the frozen Neuro-MSBG model, with the training objective of closely matching the final output with the original clean speech. This design enables the compensator to function as a personalized module that adjusts the audio to each user's hearing condition. It is important to note that we did not fine-tune Neuro-MSBG, as our main objective was to initially verify the feasibility and effectiveness of integrating Neuro-MSBG into the training pipeline. The compensator adopts the same architecture as Neuro-MSBG, but with a key modification in the magnitude path: the original masking-based magnitude encoder is replaced by a mapping strategy designed to enhance or restore lost information. This adjustment better aligns the model with

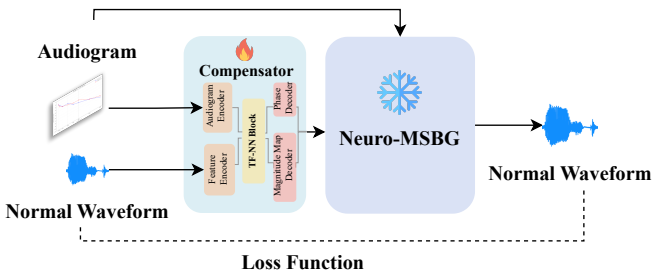

Fig. 4. **Training framework of a personalized speech compensator using Neuro-MSBG as a fixed hearing loss simulator.** The left module (trainable, labeled "Compensator") is optimized to minimize the loss between the Neuro-MSBG output and the original clean waveform. The right module ("Neuro-MSBG") is frozen during training and used only for perceptual feedback.

TABLE V
STATISTICAL COMPARISON OF HASPI SCORES BEFORE AND AFTER APPLYING THE PROPOSED COMPENSATOR.

Metric	Original	Compensated	Change (Δ)
HASPI (mean $\pm$ std)	0.428 ± 0.360	0.616 ± 0.352	$+0.187$
t-test (paired)	$t = -24.113$, $p < 0.00001$		
Wilcoxon signed-rank	$W = 292426.0$, $p < 0.00001$		
Cohen's d (paired)	$d = 0.594$		

the goal of compensation, enabling the model to generate gain-adjusted outputs that enhance speech intelligibility for hearing-impaired users.

We evaluate the effectiveness of the proposed compensator using the HASPI metric to assess the improvement in perceptual speech intelligibility. As shown in Table V, the compensator significantly improves the average HASPI score from 0.428 to 0.616 ($\Delta = +0.187$). This improvement is statistically significant, supported by both the paired t-test ($t = -24.113$, $p < 0.00001$) and the Wilcoxon signed-rank test ($W = 292426.0$, $p < 0.00001$). The observed effect size is moderately large ($d = 0.594$), indicating a substantial improvement in perceptual quality across samples.

IV. CONCLUSION

This paper introduces Neuro-MSBG, a lightweight and fully differentiable hearing loss simulation model that addresses key limitations of traditional approaches, including delay issues, limited integration flexibility, and lack of parallel processing capabilities. Unlike conventional models that require clean reference signal alignment, Neuro-MSBG can be seamlessly integrated into end-to-end training pipelines and avoids timing mismatches that may affect evaluation metrics such as STOI and PESQ. Its parallelizable architecture and low-latency design make it well suited for scalable speech processing applications. In particular, the Mamba-based Neuro-MSBG achieves 46× speedup over the original MSBG, reducing the simulation time for one second of audio from 0.970 seconds to 0.021 seconds through parallel inference. Meanwhile, the LSTM-based variant achieves an inference time of 0.016 seconds, which is 60× faster than MSBG. Experimental results fur-

ther demonstrate that jointly predicting magnitude and phase can significantly improve speech intelligibility and perceptual quality, with SRCC of 0.9247 for STOI and 0.8671 for PESQ. In addition, the proposed Audiogram Encoder can effectively transform audiogram vectors into frequency-aligned features, outperforming the simple concatenation method and more accurately modeling individual hearing profiles.

REFERENCES

[1] Simone Graetzer, Jon Barker, Trevor J. Cox, Michael Akeroyd, John F. Culling, Graham Naylor, Eszter Porter, and Rhoddy Viveros Muñoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Proc. INTERSPEECH*, 2021, pp. 686–690.

[2] Brian C. J. Moore, Brian R. Glasberg, and Thomas Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *Journal of the Audio Engineering Society*, vol. 45, pp. 224–240, 1997.

[3] Thomas Baer and Brian C. J. Moore, "Effects of spectral smearing on the intelligibility of sentences in noise," *The Journal of the Acoustical Society of America*, vol. 94, pp. 1229–1241, 1993.

[4] Thomas Baer and Brian C. J. Moore, "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech," *The Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2050–2062, 1993.

[5] Brian C. J. Moore and Brian R. Glasberg, "Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech," *The Journal of the Acoustical Society of America*, vol. 94, no. 4, pp. 2050–2062, 1993.

[6] Gerardo Roa Dabike, Jon Barker, John F. Culling, et al., "The ICASSP SP Cadenza challenge: Music demixing/remixing for hearing aids," *arXiv preprint arXiv:2310.03480*, 2023.

[7] Kathryn H. Arehart James M. Kates, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, 2014.

[8] James Kates and Kathryn Arehart, "The hearing-aid speech quality index (HASQI)," *AES: Journal of the Audio Engineering Society*, vol. 58, pp. 363–381, 2010.

[9] J. M. Kates and K. H. Arehart, "The hearing-aid audio quality index (HAAQI)," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 354–365, 2016.

[10] Muhammad S. A. Zilany and Ian C. Bruce, "Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery," *The Journal of the Acoustical Society of America*, vol. 120, no. 3, pp. 1446–1466, 2006.

[11] Sarah Verhulst, Alessandro Altoè, and Viacheslav Vasilkov, "Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss," *Hearing Research*, vol. 360, pp. 55–75, 2018.

[12] Volker Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acta Acustica united with Acustica*, vol. 88, pp. 433–442, 2002.

[13] Malcolm Slaney, "Auditory toolbox version 2: A MATLAB toolbox for auditory modeling work," Tech. Rep. 1998-010, Interval Research Corporation, 1998.

[14] Arthur Van Den Broucke, Deepak Baby, and Sarah Verhulst, "Hearing-impaired bio-inspired cochlear models for real-time auditory applications," in *Proc. INTERSPEECH*, 2020, pp. 2842–2846.

[15] Toshio Irino, "Hearing impairment simulator based on auditory excitation pattern playback: WHIS," *IEEE Access*, vol. 11, pp. 78419–78430, 2023.

[16] Peter Leer, Jesper Jensen, Zheng-Hua Tan, Jan Østergaard, and Lars Bramsløw, "How to train your ears: Auditory-model emulation for large-dynamic-range inputs and mild-to-severe hearing losses," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 32, pp. 2006–2020.

[17] Fotios Drakopoulos and Sarah Verhulst, "A neural-network framework for the design of individualised hearing-loss compensation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2395–2409, 2023.

[18] Fotios Drakopoulos and Sarah Verhulst, "A differentiable optimisation framework for the design of individualised dnn-based hearing-aid strategies," in *Proc. ICASSP*, 2022, pp. 351–355.

[19] Fotios Drakopoulos, Arthur Van Den Broucke, and Sarah Verhulst, "A dnn-based hearing-aid strategy for real-time processing: One size fits all," in *Proc. ICASSP*, 2023, pp. 1–5.

[20] Zehai Tu, Ning Ma, and Jon Barker, "Dhasp: Differentiable hearing aid speech processing," in *Proc. ICASSP*, 2021, pp. 296–300.

[21] Zehai Tu, Ning Ma, and Jon Barker, "Optimising hearing aid fittings for speech in noise with a differentiable hearing loss model," in *Proc. INTERSPEECH*, 2021, pp. 691–695.

[22] Ye-Xin Lu, Yang Ai, and Zhen-Hua Ling, "MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra," in *Proc. INTERSPEECH*, 2023, pp. 3834–3838.

[23] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[24] Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, pp. 749–752.

[25] Rong Chao, Wen-Huang Cheng, Moreno La Quatra, Sabato Marco Siniscalchi, Chao-Han Huck Yang, Szu-Wei Fu, and Yu Tsao, "An investigation of incorporating mamba for speech enhancement," in *Proc. SLT*, 2024, pp. 302–308.

[26] Yuxuan Ai and Zhen-Hua Ling, "Neural speech phase prediction based on parallel estimation architecture and anti-wrapping losses," in *Proc. ICASSP*, 2023, pp. 1–5.

[27] Christophe Veaux, Junichi Yamagishi, and Simon King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. O-COCOSDA/CASLRE*, 2013, pp. 1–4.

[28] Cassia Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and tts models," 2017, University of Edinburgh, Centre for Speech Technology Research.

[29] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "The diverse environments multichannel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Meetings on Acoustic*, 2013, pp. 1–6.

[30] Shafique Ahmed, Ryandhimas E. Zezario, Hui-Guan Yuan, Amir Hussain, Hsin-Min Wang, Wei-Ho Chung, and Yu Tsao, "Neuroamp: A novel end-to-end general purpose deep neural amplifier for personalized hearing aids," *arXiv preprint arXiv:2502.10822*, 2025.