Abstract

This work proposes and validates a differentiable, end-to-end trainable hearing-loss simulation model designed to overcome the computational limitations of the traditional MSBG auditory model (Brian C. J. Moore & Brian R. Glasberg). Although the MSBG model reliably reproduces perceptual characteristics of sensorineural hearing loss, its cascaded filtering and nonlinear compression introduce high latency and poor parallelizability, making real-time integration with modern neural speech systems challenging.

To address these issues, we introduce HL-Mamba, a neural architecture that replaces cascaded filtering with a differentiable signal-mapping pipeline. HL-Mamba incorporates Audiogram Encoding and time—frequency modeling to jointly learn amplitude attenuation and phase distortion, supported by a feed-forward structure that enables efficient parallel inference. A lightweight Audiogram Encoder embeds clinical audiograms via frequency alignment, allowing individualized degeneration patterns. Experiments further confirm the importance of phase modeling, which improves similarity to Moore—Glasberg responses and enhances perceived naturalness.

Overall, this work aims to (1) reduce simulation latency for real-time speech processing, (2) improve perceptual realism and clarity of degraded speech, and (3) provide a differentiable hearing-loss module for speech enhancement and compensation frameworks. Results demonstrate that the proposed Neuro-MSBG model preserves core Moore Glasberg perceptual behavior while achieving higher efficiency and perceptual similarity, showing strong integration potential in end-to-end speech pipelines.

Keywords: Hearing loss simulation, Differentiable auditory modeling, End-to-end speech processing, Phase-aware degradation

Chapter 1 Introduction

Hearing loss simulators aim to reproduce how impaired auditory perception alters sound processing along the human auditory pathway. As their modeling fidelity and computational usability increase, these systems have become foundational tools for research, evaluation, and algorithm benchmarking. This trend is evident in initiatives such as the *Clarity Challenge* [10], which adopts the Moore—Stone—Baer—Glasberg (MSBG) model [19, 3, 4, 18] to generate individualized listening conditions from audiograms. Challenges including *Cadenza* [21] and *Clarity* further emphasize perceptual evaluation, incorporating HASPI [13], HASQI [15], and HAAQI [14] as objective intelligibility and quality metrics grounded in auditory physiology.

Existing simulation approaches may be broadly divided into physiological and engineering-oriented methodologies. Physiological models, such as Zilany et al. [30] and the transmission line cochlear representation of Verhulst et al. [29], provide biologically interpretable insights but often impose heavy computational costs, limiting real-time deployment. In contrast, engineering-focused solutions seek a practical balance between perceptual accuracy and compute efficiency, as seen in the Hohmann filterbank [11], the Auditory Toolbox [22], and MSBG [19]. Among these, MSBG remains dominant due to its audiogram-driven perceptual alignment. However, its sequential multi-stage filtering blocks parallelization, and its non-uniform latency complicates waveform alignment—both of which hinder integration into end-to-end learning frameworks.

To reduce computational overhead, recent work [5, 12] has moved toward neural approximations and streamlined auditory simulation. CoNNear [5] provides a neural surrogate for cochlear mechanics, facilitating real-time auditory nerve response generation, while Tan et al. [16] approximate the Verhulst auditory periphery across hearing-loss conditions. Although effective at the neural representation level, these methods do not output waveform-domain signals, which limits their suitability for time-domain supervision and downstream enhancement tasks. On the engineering side, the WHIS simulator [12] accelerates MSBG through a Gammachirp filterbank and a time-varying minimum-phase reconstruction, reducing one second of speech to roughly 10 ms of processing and maintaining near-perfect

temporal alignment. Yet, WHIS still relies on frame-based IIR coefficient estimation and dynamic gain selection, lacks vectorization and parallel inference, and remains difficult to embed within differentiable training pipelines.

As differentiable auditory simulation gains traction in hearing-aid optimization, improving learnability and computational efficiency has emerged as a key research objective. CoNNear-derived auditory nerve responses have been incorporated as loss terms for training compensation networks [8, 7, 9], encouraging restored speech to approximate the neural behaviour of a normal listener. Engineering-oriented work has likewise progressed: Tu et al. [25] introduced the DHASP framework, enabling backpropagation through the HASPI processing pipeline, and later developed a differentiable MSBG implementation for hearing-aid optimization [26], demonstrating that physiologically motivated simulation can be integrated into learning-based compensation.

Despite these developments, most differentiable simulators remain focused on magnitude-only processing. Phase information—known to influence perceived quality in speech enhancement [17]—is rarely modeled in hearing loss simulation. Motivated by evidence that phase-aware enhancement improves perceptual realism, we propose **Neuro-MSBG**, a fully differentiable, end-to-end simulation model that generates impaired speech directly in the time domain. Waveform-domain operation allows seamless compatibility with loss functions such as MSE, STOI [23], and PESQ [20], and supports both clean and noisy input speech. Our experiments show that incorporating phase significantly enhances perceptual correspondence with MSBG, confirming the importance of phase-aware auditory modeling.

- **Efficient and parallelizable:** One second of speech is simulated in 0.021 s, a $46\times$ improvement over MSBG.
- **Latency-stable and alignment-preserving:** Eliminates irregular delay and directly supports end-to-end learning pipelines.
- **Phase-aware modeling:** Improves perceptual consistency, yielding SRCC scores of 0.9247 (STOI) and 0.8671 (PESQ).

We additionally demonstrate its integration with a differentiable compensation net-

work, illustrating practical feasibility for hearing aid algorithm development. Section II details the proposed model, Section III presents experimental results, and Section IV concludes the work.

Chapter 2 Related Work

Hearing loss simulation has been widely studied from two major perspectives: (1) physiologically grounded auditory models that replicate cochlear mechanics and neural transduction, and (2) engineering-oriented models that emphasize computational efficiency and practical deployment. This chapter reviews representative works in both categories and discusses their advantages, limitations, and relevance to the development of HL-Mamba.

2.1 Physiological Models of Hearing Loss

Physiological models aim to mimic the real auditory periphery, including basilar membrane vibration, auditory nerve firing patterns, and nonlinear cochlear dynamics. Among them, the auditory periphery model of Zilany et al. [30] and the transmission-line (TL) cochlear model proposed by Verhulst et al. [29] are two of the most influential frameworks. These models provide biologically realistic output that is valuable for analyzing neural encoding and psychoacoustic perception under sensorineural hearing loss. However, their numerical complexity and multi-stage computation make them computationally expensive, leading to slow simulation speed, limited scalability, and difficulty in real-time deployment.

To improve usability, several efforts have attempted to approximate cochlear dynamics with neural architectures. CoNNear [5] accelerates TL-model inference using convolutional networks, enabling real-time simulation of auditory nerve responses. Tan et al. [16] further demonstrated that deep learning can emulate the Verhulst auditory periphery model for diverse hearing profiles. Despite these advances, physiological models generally do not produce waveform outputs and therefore cannot be directly incorporated into speech enhancement, hearing aid learning frameworks, or perceptual objective metrics such as STOI or PESQ. As a result, their impact remains more prominent in auditory neuroscience than in deployable audio signal processing.

2.2 Engineering-Oriented Models of Hearing Loss

Engineering approaches aim to provide perceptually grounded yet computationally efficient hearing loss simulation. Classic examples include the Hohmann auditory filterbank [11], the Auditory Toolbox [22], and the Moore, Stone, Baer, and Glasberg (MSBG) model [19]. MSBG remains the most widely used model in research and evaluation due to its ability to reproduce spectral smearing, loudness recruitment, and reduced frequency selectivity based on audiograms. The Clarity [10] and Cadenza [21] challenges both adopt MSBG-generated data as perceptually aligned impaired speech, together with evaluation metrics such as HASPI, HASQI, and HAAQI [13, 15, 14].

However, MSBG suffers from two primary limitations: (1) lack of parallel computing support, which hinders real-time application and GPU scaling, and (2) variable latency caused by multi-stage filtering, making end-to-end learning integration difficult. To mitigate these constraints, Irino et al. proposed WHIS [12], which constructs an excitation-based minimum-phase filter for fast waveform transformation. WHIS reduces one-second processing to 10 ms while preserving temporal alignment. Nonetheless, it still requires frame-wise IIR estimation, lacks efficient vectorization, and has not yet been fully integrated into neural frameworks.

Recent differentiable auditory models further enable end-to-end optimization. Tu et al. introduced DHASP [25] and a differentiable MSBG variant [26], demonstrating that auditory simulation modules can participate in gradient-based training. However, existing engineering-oriented models focus mainly on magnitude, while phase—critical for perceptual realism—remains under-modeled. This motivates the development of HL-Mamba, which incorporates phase modeling, supports parallel execution, and enables seamless integration into waveform-based hearing compensation pipelines.

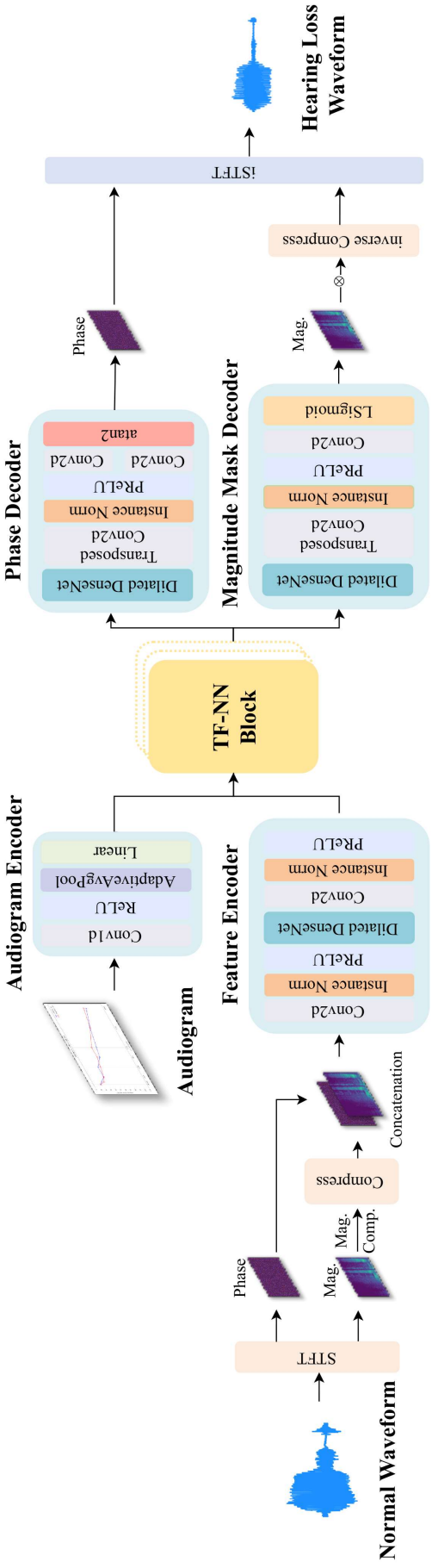

Figure 2.1: **Overall architecture of the proposed HL-Mamba framework.**

Chapter 3 Methodology

This chapter presents the model architecture and training objectives of the proposed HL-Mamba framework. As illustrated in Fig. 1, the model takes normal-hearing speech signals and monaural audiograms as inputs.

The audiogram is first processed by the Audiogram Encoder to generate personalized hearing-condition features. Meanwhile, the speech signal is transformed into the time-frequency domain via Short-Time Fourier Transform (STFT), yielding its magnitude and phase components. These three types of features, namely the personalized hearing features, the magnitude features, and the phase features, are then concatenated and fed into the Neural Network Block (NN Block).

Subsequently, the network splits into two decoding branches: the Magnitude Mask Decoder and the Phase Decoder, which respectively estimate the magnitude and phase modifications associated with hearing loss. Finally, the predicted magnitude and phase are combined and converted back to the time domain using inverse STFT, producing the speech signal as it would be perceived by a listener with hearing impairment.

3.1 HL-Mamba

The overall architecture of HL-Mamba is built upon the MP-SENet [17] and the SE-Mamba framework [6], which together offer efficient and expressive mechanisms for feature interaction. This design choice is motivated by our empirical observation that phase information plays a crucial role in accurately simulating hearing loss effects (see Section III for further analysis). To systematically assess how different neural modules capture spectral and temporal cues, we replace the original attention-based components with several alternatives, including LSTM, Transformer, CNN, and Mamba blocks. In view of Mamba's strength in modeling long sequences with low latency, we also carefully control the model size so that the resulting architecture remains lightweight while preserving strong performance.

Regarding audiogram integration, prior approaches often append audiogram representations along the frequency axis. In contrast, our experiments indicate that treating the audiogram as an additional input channel, alongside the magnitude and phase channels, leads to more stable and effective learning. This channel-based conditioning strategy provides the model with a spatially aligned, layer-wise consistent way of injecting hearing-profile information, thereby enhancing its capability to modulate internal feature representations.

Audiogram Encoder

To incorporate individualized hearing profiles, we design a lightweight Audiogram Encoder that maps the audiogram $\mathbf{a} \in \mathrm{R}^{B\times 8}$ to 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 frequency resolution. The transformation is formulated as:

$$\mathbf{a}_{\text{enc}} = \mathbf{W} \cdot \text{Flatten}\left(\text{AvgPool}\left(\sigma\left(\text{Conv}(\mathbf{a})\right)\right)\right), \tag{3.1}$$

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 dimension and concatenated with the magnitude and phase features, forming a three-channel input tensor of shape $\mathrm{R}^{B\times 3\times T\times F}$ to the DenseEncoder and NN Block. This channel-based integration ensures that the hearing-profile information is injected consistently across both time and frequency axes, facilitating effective conditioning throughout the network.

Neural Network Blocks

To effectively capture the temporal and spectral characteristics of speech affected by hearing loss, we develop and compare several NN Blocks, each following a dual-path design that separately models the time and frequency dimensions. Concretely, the input tensor is first rearranged and reshaped for temporal modeling; a similar operation is then performed for frequency modeling. Each path is equipped with residual connections and a

ConvTranspose1d layer to restore the original tensor shape, ensuring seamless interaction with subsequent modules.

Within this unified framework, the core time-frequency modeling module is instantiated using one of the following four alternatives:

- **Mamba Block**: Bidirectional Mamba modules are used to model time and frequency independently, providing efficient long-range dependency modeling across both dimensions.
- **Transformer Block**: Transformer encoder layers are applied along each axis, using global attention to capture rich contextual information.
- **LSTM Block**: Bidirectional LSTMs are employed to model sequential dynamics, followed by linear projection layers to maintain dimensional compatibility.
- **CNN Block**: One-dimensional convolutions are used to extract local patterns, followed by channel expansion, non-linear activation, and residual fusion to enhance feature expressiveness.

This dual-axis architecture yields a flexible and robust framework for hearing loss simulation. It also enables a controlled comparison of different neural architectures under the same interface, highlighting the potential of Mamba for low-latency, high-fidelity speech modeling.

3.2 Training Criteria

To guide the learning of HL-Mamba, we adopt a multi-objective loss function that jointly enforces spectral accuracy, phase consistency, and time-domain fidelity.

First, the **magnitude loss** $\mathcal{L}_{\text{Mag}}$ is defined as the mean squared error (MSE) between the predicted magnitude $\hat{m}$ and the ground-truth magnitude m:

$$\mathcal{L}_{\text{Mag}} = \frac{1}{N} \sum_{i=1}^{N} \|\hat{m}_i - m_i\|^2 , \tag{3.2}$$

where N denotes the number of training samples.

The **phase loss** $\mathcal{L}_{\text{Pha}}$ is derived from the anti-wrapping strategy in [2] 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 $\mathcal{L}_{\text{IAF}}$. They are defined as:

$$\mathcal{L}_{\text{IP}} = \mathbb{E}_{p,\hat{p}} \left[\| f_{\text{aw}} \left(p - \hat{p} \right) \|_1 \right], \tag{3.3}$$

$$\mathcal{L}_{\text{GD}} = \mathbb{E}_{p,\hat{p}} \left[\| f_{\text{aw}} \left(\Delta_{\text{F}}(p - \hat{p}) \right) \|_1 \right], \tag{3.4}$$

$$\mathcal{L}_{\text{IAF}} = \mathbb{E}_{p,\hat{p}} \left[\| f_{\text{aw}} \left(\Delta_{\text{T}}(p - \hat{p}) \right) \|_1 \right], \tag{3.5}$$

$$\mathcal{L}_{\text{Pha}} = \mathcal{L}_{\text{IP}} + \mathcal{L}_{\text{GD}} + \mathcal{L}_{\text{IAF}}, \tag{3.6}$$

where $f_{\text{aw}}(\cdot)$ denotes the anti-wrapping function used to alleviate 2π discontinuities, and $\Delta_{\text{F}}(\cdot)$ and $\Delta_{\text{T}}(\cdot)$ represent the first-order phase differences along the *frequency* and *time* axes, respectively.

The **complex loss** $\mathcal{L}_{\text{Com}}$ measures the MSE between the predicted complex spectrogram $\hat{c}$ and the reference complex spectrogram c (including both real and imaginary parts):

$$\mathcal{L}_{\text{Com}} = 2 \cdot \frac{1}{N} \sum_{i=1}^{N} \| \hat{c}_i - c_i \|^2 . \tag{3.7}$$

To further preserve waveform-level fidelity, we introduce the **time-domain loss** $\mathcal{L}_{\text{Time}}$, computed as the L_1 distance between the predicted waveform $\hat{x}$ and the target waveform x:

$$\mathcal{L}_{\text{Time}} = \frac{1}{N} \sum_{i=1}^{N} \| \hat{x}_i - x_i \|_1 . \tag{3.8}$$

The overall training objective is a weighted sum of the above components:

$$\mathcal{L}_{\text{Total}} = \lambda_{\text{Mag}} \mathcal{L}_{\text{Mag}} + \lambda_{\text{Pha}} \mathcal{L}_{\text{Pha}} + \lambda_{\text{Com}} \mathcal{L}_{\text{Com}} + \lambda_{\text{Time}} \mathcal{L}_{\text{Time}}, \tag{3.9}$$

where each λ is a tunable scalar that controls the relative contribution of the corresponding loss term.

Chapter 4 Experiments and Results

The dataset consists of 12,396 clean utterances from VoiceBank [28], where 11,572 are used for training and 824 for testing. Their noisy counterparts originate from the VoiceBank–DEMAND corpus [27], created by mixing each clean utterance with a randomly selected noise recording from the 10 DEMAND environments [24] and a specified signal-to-noise ratio (SNR). To promote generalization, 8 noise types are allocated to the training split and 2 disjoint noise types are used exclusively for testing. SNR values are drawn from {0, 5, 10, 15} dB during training and from {2.5, 7.5, 12.5, 17.5} dB for testing. Each utterance is associated with two monaural audiograms that represent different degrees of hearing loss, spanning mild to severe impairment.

As a result, the size of the training set expands to $11,572 \times 2 \times 2 = 46,288$ samples, while the test set expands to $824 \times 2 \times 2 = 3,296$ samples. There is no overlap in speech content or audiograms between the training and testing splits, ensuring that evaluation is performed on previously unseen utterances and unseen hearing-loss profiles.

This work conducts a comprehensive set of experiments across five major dimensions: (i) comparison of HL-Mamba variants with different architectures and monolithic baselines (Table 4.1); (ii) evaluation of inference latency (Table 4.3); (iii) analysis of the importance of phase modeling (Table 4.2); (iv) investigation of audiogram integration strategies (Table 4.4); and (v) application of HL-Mamba within a speech compensation framework (Table 4.5). Across all quantitative results, **bold** denotes the top-performing system, while <u>underline</u> indicates the runner-up.

All experiments were executed on a single NVIDIA RTX 3090 GPU. Model training was conducted for 200 epochs with a batch size of 3, using an initial learning rate of 0.0005 and the AdamW optimizer.

4.1 MSBG-Based Alignment

We employ the MSBG model to emulate hearing loss by applying an ear-dependent gain profile to each input signal, producing a single-channel impaired output. However, the multi-stage filtering operations in MSBG may introduce an unknown latency relative to the original waveform. This temporal shift is undesirable for subsequent tasks such as speech enhancement or intelligibility evaluation, where accurate temporal alignment is critical.

To estimate and correct for this latency, we follow the procedure described in [10] and generate an auxiliary reference signal together with the main audio during the MSBG simulation. This reference is a silent waveform that has the same duration and sampling rate as the input, and it is used exclusively for delay monitoring. A unit impulse is embedded into this reference signal, defined as:

$$\delta[n-k] = \begin{cases} 1, & \text{if } n = k \\ 0, & \text{otherwise} \end{cases}, \tag{4.1}$$

where n denotes the discrete time index, and $k = \frac{F_s}{2}$ specifies the impulse location, with F_s being the sampling rate in Hz.

The impulse is placed at the center of the auxiliary reference signal. After passing through the MSBG processing, we compute the shift between the original and processed impulse positions to infer the delay introduced by MSBG. This estimated delay is then applied to synchronize the normal-hearing input, the clean reference, and the impaired output, enabling reliable computation of highly time-sensitive metrics such as STOI and PESQ.

Each training example therefore contains: 1) the impaired single-ear speech, 2) the aligned clean reference, 3) the aligned normal-hearing input, and 4) the corresponding 8-dimensional audiogram vector. Figure 4.1 illustrates the waveform alignment before and after applying the delay compensation.

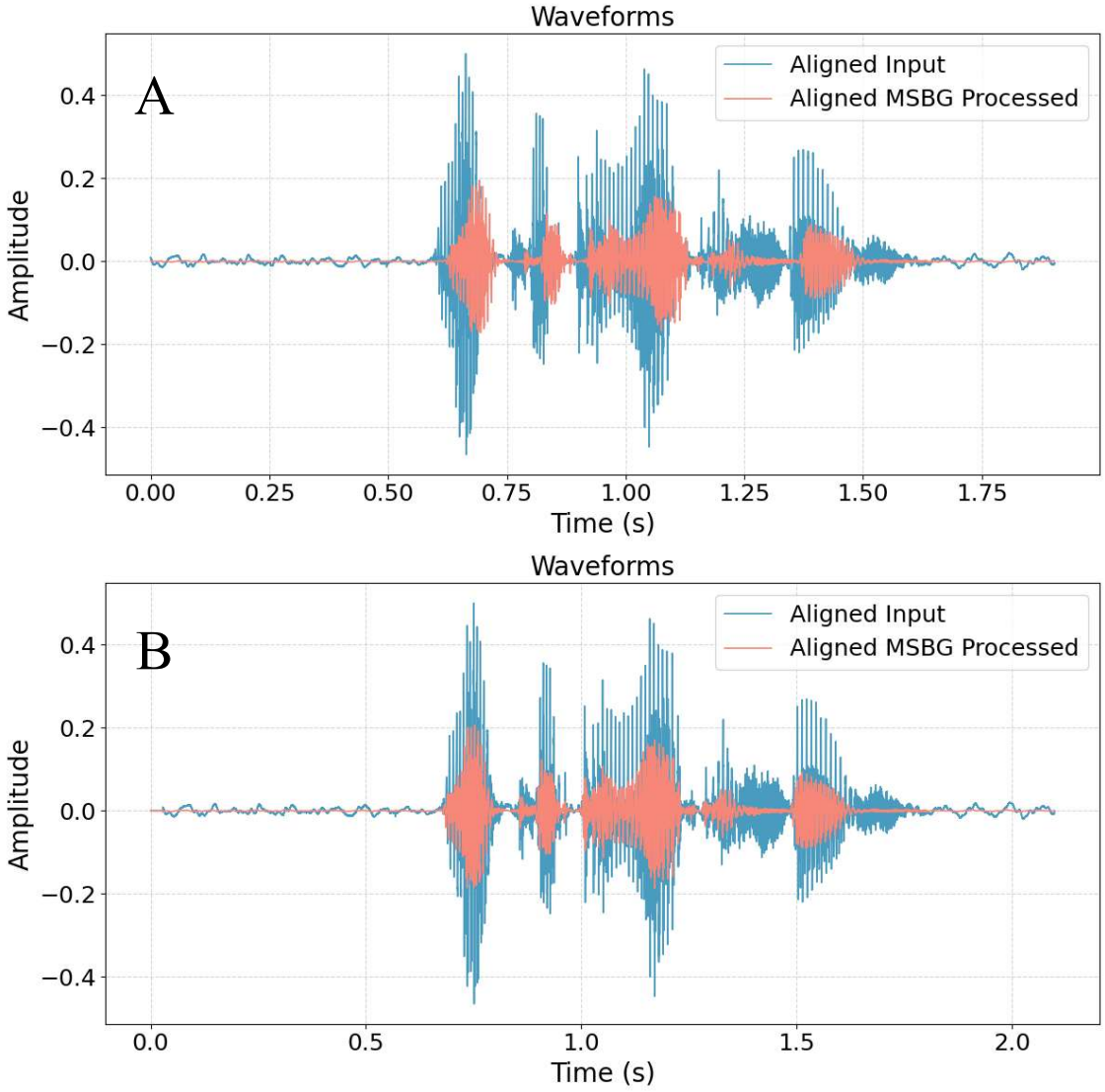

Figure 4.1: Waveform alignment before and after correction. (A) The MSBG output exhibits a delay relative to the input signal. (B) The proposed impulse-based correction compensates for the delay and restores temporal alignment for accurate evaluation.

4.2 Main Performance Comparison

Table 4.1: Comparison of the performance of monolithic models and HL-Mamba variants 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
HL-Mamba 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**

In the initial phase of our study, we employed monolithic models with single, uniform architectures such as CNN, LSTM, and Transformer to simulate hearing loss. Due to their limited performance, we transitioned to the HL-Mamba framework, modifying only the TF-NN block while testing multiple architectures, including CNN, LSTM, Transformer, and Mamba. The results in Table 4.1 present the outcomes of this comparison. To ensure fairness across models, we controlled the total parameter count so that all architectures had approximately the same number of learnable parameters. As indicated in Table 4.1, the HL-Mamba variants consistently surpass the monolithic baselines on unseen test sets. Among these variants, the Mamba-based implementation of HL-Mamba demonstrates the highest performance across all evaluation metrics. We further compare HL-Mamba with an existing neural model, CoNNear, by examining model complexity and inference latency. Although the two approaches serve different purposes—CoNNear focuses on predicting physiologically accurate auditory nerve responses, whereas HL-Mamba is designed for perceptually oriented signal transformation—the comparison remains reasonable because both pursue real-time hearing loss simulation. According to Table 4.3, CoNNear contains roughly 11.7 million parameters, while HL-Mamba has only 1.45 million. Beyond being more compact, HL-Mamba can process noisy speech inputs and flexibly handle a diverse range of audiogram configurations, offering advantages for real-world applications with varying acoustic conditions and personalized hearing requirements.

Table 4.3 also reports inference latency for each model. MSBG does not support parallelism and therefore cannot be executed on GPU, so its GPU runtime is listed as NA. In contrast, HL-Mamba (Mamba) relies on a CUDA-accelerated selective scan kernel for its core computations, which currently runs only on GPU; therefore, only GPU inference time is measured. In terms of speed, HL-Mamba (Mamba) achieves roughly a 46$\times$ acceleration on GPU compared to MSBG on CPU. For CPU-compatible variants such as HL-Mamba (LSTM), the inference time is 0.016 seconds, corresponding to a 60$\times$ speedup over MSBG's 0.970 seconds. We additionally implemented CoNNear for comparison. Despite its relatively large parameter size of 11.7 million, it yields fast inference on our hardware, with GPU computation taking 0.099 seconds and notably faster CPU computation taking 0.025 seconds. The faster CPU runtime compared to GPU is likely due to the batch size of 1 adopted in this experiment, which limits the computational benefits of

GPU parallelization.

4.3 Ablation Study

Previous studies on hearing loss modeling typically estimate only the magnitude spectrum and reuse the input phase when reconstructing the waveform. We initially followed this strategy as well, but our experiments showed that phase information is highly influential in MSBG-based simulation. To examine this effect, we performed an ablation study using HL-Mamba (Mamba). As reported in Table 4.2, simultaneously predicting magnitude and phase yields substantial improvements over magnitude-only prediction across all evaluation criteria (STOI MSE, PESQ MSE, and waveform MSE). Notably, STOI MSE decreases from 0.0041 to 0.0006, indicating improved intelligibility, while PESQ MSE drops from 2.3579 to 0.0782, signifying better perceptual quality. At the waveform level, the MSE decreases from 0.0986 to 0.0669, demonstrating that phase estimation is essential for perceptual fidelity as well as accurate reconstruction.

In addition, many speech-related systems that incorporate audiometric information adopt a straightforward strategy of concatenating the audiogram vector with the spectral features before feeding them to the model. We initially applied this method, but observed that it could not sufficiently represent the relationship between hearing profiles and acoustic cues. To overcome this limitation, we designed a lightweight Audiogram Encoder that maps the 8-dimensional audiogram vector to a frequency-aligned representation, which is then added as a separate channel alongside the magnitude and phase features. According to the results in Table 4.4, the proposed encoder consistently reduces STOI MSE, PESQ MSE, and waveform-level MSE, indicating that more effective integration of individualized hearing characteristics leads to improved hearing loss simulation.

Table 4.2: **Phase prediction is essential.** Magnitude-only severely degrades perceptual accuracy.

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

Table 4.3: Comparison of runtime of MSBG and HL-Mamba 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
HL-Mamba (Mamba)	NA	0.021 s	1.45M
HL-Mamba (LSTM)	0.617 s	0.016 s	1.47M
HL-Mamba (CNN)	0.592 s	0.016 s	1.45M
CoNNear	0.025 s	0.099 s	11.7M

Table 4.4: **Audiogram Encoder improves hearing—profile conditioning.**

Setting	STOI MSE	PESQ MSE	RAW MSE
Without Encoder	0.0013	0.1152	0.0670
With Encoder	**0.0006**	**0.0782**	**0.0669**

4.4 Qualitative Evaluation

To assess how well each model reconstructs speech affected by hearing loss, we perform a qualitative comparison of the log-magnitude spectrograms generated by seven different models against the ground-truth reference (Fig. 4.2). Among all models, HL-Mamba (Mamba) produces spectrograms that most closely match the reference, effectively retaining harmonic structures and high-frequency components. The HL-Mamba variants with CNN, LSTM, and Transformer blocks also capture important spectral characteristics, though they exhibit slight energy imbalance or mild distortion. In contrast, the baseline models introduce more noticeable artifacts: the CNN and LSTM versions show reduced clarity and insufficient high-frequency detail, while the Transformer baseline struggles to generate an accurate spectral representation.

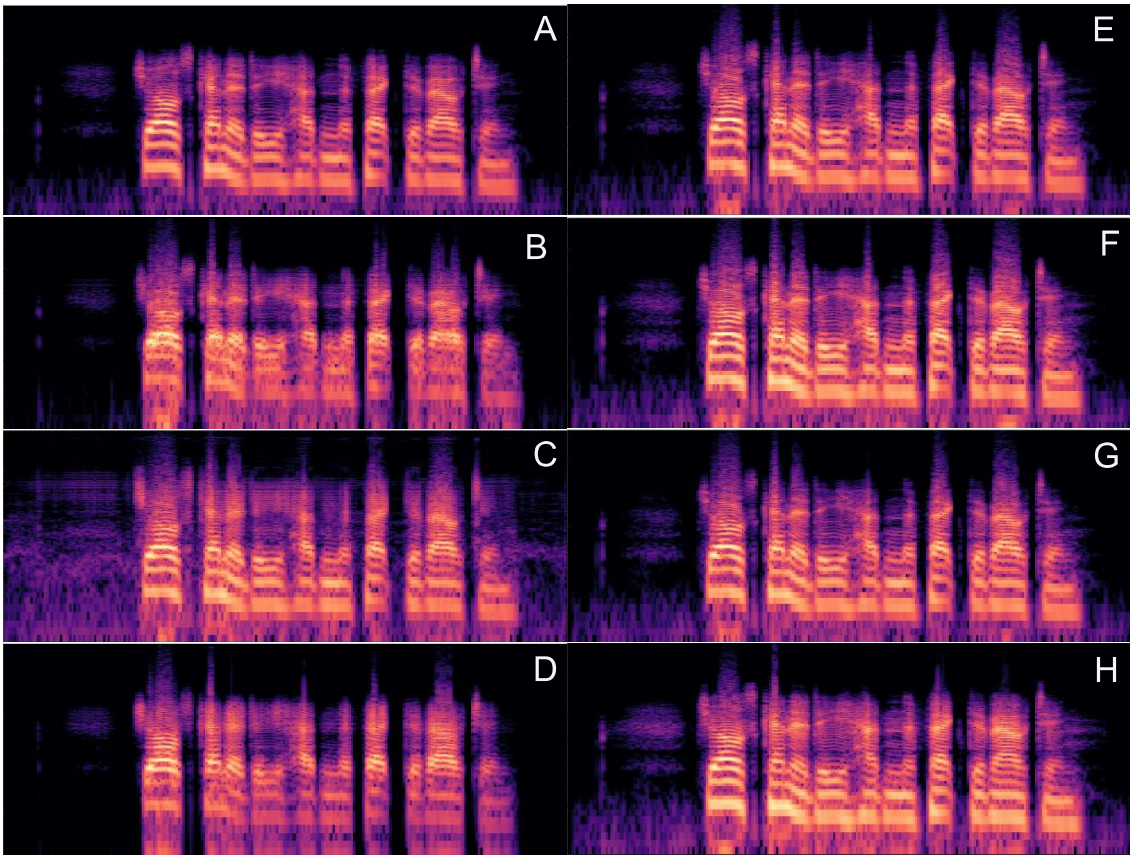

Figure 4.2: **Spectrogram comparison.** (A) Ground truth (B—D) Baseline CNN/LSTM/Transformer (E—H) HL-Mamba with Mamba/CNN/LSTM/Transformer Mamba reconstruction best preserves harmonic structure + high-frequency detail.

4.5 Compensator Training with HL-Mamba

To facilitate end-to-end hearing loss compensation, recent approaches such as NeuroAMP [1] have incorporated audiogram-aware processing directly into neural architectures, effectively replacing conventional modular pipelines with personalized, data-driven amplification. Motivated by this trend, we propose a complementary strategy in which a trainable compensator is connected to a frozen, perceptually informed simulator (HL-Mamba). In this setup, the compensator learns to modify its input so that the resulting output from the simulator matches each listener's hearing profile.

HL-Mamba is lightweight, differentiable, and does not rely on time-domain alignment with clean references, characteristics that make it well-suited for building an end-to-end hearing loss compensation pipeline. Leveraging these properties, we explore a novel application: integrating a pre-trained HL-Mamba model with a trainable compensator to perform personalized hearing enhancement, as illustrated in Fig. 4.3. Both training and evaluation are conducted using the VoiceBank dataset [28].

The compensator is designed to convert the input waveform into a personalized, compensated signal. This processed signal is then fed into the frozen HL-Mamba model, and the system is trained such that the final output closely matches the clean speech. In this configuration, the compensator acts as an individualized module that adapts the signal according to a user' s hearing characteristics. Importantly, HL-Mamba itself remains fixed during training; our intention is to first confirm that the integration of HL-Mamba into the optimization loop is feasible and effective. The compensator follows the HL-Mamba architecture with one major modification in the magnitude path: instead of the original masking-based magnitude encoder, we employ a mapping-based strategy aimed at restoring or enhancing missing spectral information. This modification aligns the network with the objective of signal compensation by enabling the generation of gain-adjusted outputs that improve intelligibility for hearing-impaired listeners.

To quantify the impact of the proposed approach, we evaluate the system using the HASPI metric, which measures perceptual speech intelligibility. As presented in Table 4.5, the compensator produces a notable increase in average HASPI score from 0.428

to 0.616 ($\Delta = +0.187$). The improvement is statistically significant according to both a paired t-test ($t = -24.113$, $p < 0.00001$) and a Wilcoxon signed-rank test ($W = 292426.0$, $p < 0.00001$). The resulting effect size ($d = 0.594$) is moderately large, indicating meaningful perceptual gains across samples.

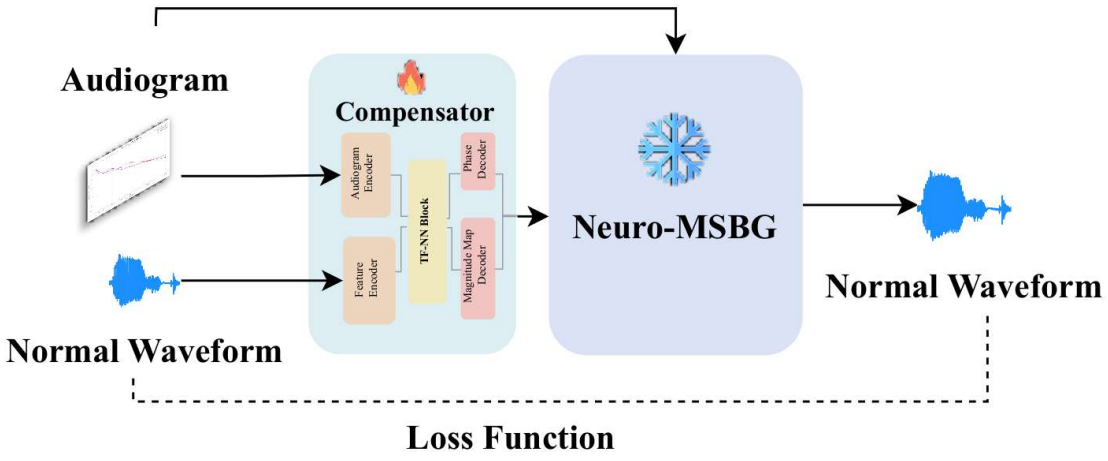

Figure 4.3: **Trainable compensator with HL-Mamba.** Simulator remains frozen—only compensator updates to improve intelligibility.

Table 4.5: **HASPI improvement using compensator.**

	Original	**Compensated**	**Gain**
HASPI	0.428	**0.616**	+0.187
t-test		$p < 10^{-5}$	
Wilcoxon		$p < 10^{-5}$	
Effect Size d		0.594	

Results confirm that HL-Mamba enables learnable, profile—aware enhancement.

Chapter 5 Conclusions

Neuro-MSBG is proposed as a differentiable simulation framework for hearing loss, aimed at resolving several limitations inherent in traditional models, such as high delay variability, limited compatibility with learning-based pipelines, and the lack of scalable parallel execution. Unlike conventional approaches that rely on explicit alignment to a clean reference signal, Neuro-MSBG can be embedded directly into end-to-end architectures without inducing temporal desynchronization, thereby preventing degradation in objective metrics such as STOI and PESQ. Its highly parallelizable structure, together with low latency, enables deployment at scale and supports real-time speech processing applications.

Among the developed variants, the Mamba-based model demonstrates substantial computational advantages, running approximately $46\times$ faster than the original MSBG implementation and reducing the processing time of one second of audio from 0.970 s to 0.021 s. The LSTM-based version further improves efficiency, achieving an inference time of only 0.016 s, corresponding to roughly a $60\times$ speedup. Experimental evidence also indicates that jointly estimating magnitude and phase leads to notable improvements in simulated speech intelligibility and perceived quality, reaching SRCC scores of 0.9247 for STOI and 0.8671 for PESQ.

In addition, we introduce an Audiogram Encoder that transforms audiogram representations into frequency-aligned latent features. This design surpasses simple vector concatenation strategies and offers a more faithful representation of individualized hearing characteristics.

References

[1] 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.

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

[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*, 94:1229—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*, 95(4):2050—2062, 1993.

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

[6] 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*, pages 302–308, 2024.

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

[8] 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*, 31:2395–2409, 2023.

[9] 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*, pages 1–5, 2023.

[10] 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*, pages 686–690, 2021.

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

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

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

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

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

[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.*, 32:2006—2020.

[17] 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*, pages 3834–3838, 2023.

[18] 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*, 94(4):2050—2062, 1993.

[19] 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*, 45:224–240, 1997.

[20] 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*, pages 749–752, 2001.

[21] 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.

[22] Malcolm Slaney. Auditory toolbox version 2: A MATLAB toolbox for auditory modeling work. Technical Report 1998-010, Interval Research Corporation, 1998.

[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*, 19(7):2125–2136, 2011.

[24] 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*, pages 1–6, 2013.

[25] Zehai Tu, Ning Ma, and Jon Barker. Dhasp: Differentiable hearing aid speech processing. In *Proc. ICASSP*, pages 296–300, 2021.