Parameter-Efficient Domain Adversarial Training for Lightweight Framewise Phoneme Classification

Sohini Saha **Leslie M. Collins** **Boyla O. Mainsah**
{sohini.saha,leslie.collins,boyla.mainsah}@duke.edu
Duke University, Durham, NC, USA

Abstract

Lightweight neural network models deployed on edge devices are susceptible to performance degradation under unseen acoustic conditions due to domain mismatch between training and deployment environments. This is particularly relevant in causal low latency applications such as cochlear implants (CIs), where limited temporal context and strict computational constraints restrict model robustness. Domain adversarial learning can improve generalization by encouraging representations that are both task-discriminative and domain-invariant; however, jointly learning these representations becomes difficult in limited capacity models. We introduce an adapter-based adversarial learning framework for robust and parameter-efficient domain adaptation. The proposed approach incorporates an adapter module into a pretrained causal backbone. We first establish a robust pretrained baseline network and subsequently perform adversarial adaptation under mismatched acoustic conditions. We evaluate the proposed framework on framewise phonetic group classification to guide phoneme-based time-frequency mask estimation. Empirical results demonstrate consistent improvements in phoneme classification over both standard supervised training and conventional adversarial learning, particularly under low signal-to-noise ratio levels, resulting in statistically significant improvements in objective intelligibility measures of CI vocoded speech.

1 Introduction

Lightweight causal neural networks are used in real-time edge speech processing applications to satisfy low latency and limited computational resource requirements (Mou and Milanova, 2024; Wahab et al., 2025). However, limited capacity models operating under strict constraints of memory, power consumption and inference latency are susceptible to significant performance degradation under unseen acoustic conditions caused by train/test domain mismatch (Taori et al., 2020; Saha et al., 2026b). This challenge is particularly pronounced in real-time speech classification tasks, where limited temporal context and reduced model capacity hinder the learning of robust and domain-invariant feature representations in dynamic acoustic environments.

Domain-adversarial learning has emerged as an effective strategy for improving model robustness under distribution shifts by encouraging representations that remain discriminative for task classification while suppressing domain-specific information (Szegedy et al., 2014; Goodfellow et al., 2015; Shafahi et al., 2019; Madry et al., 2018; Chen and Ji, 2022; Kim et al., 2023; Ganin et al., 2016). However, conventional adversarial training introduces substantial optimization complexity and computational overhead due to iterative adversarial sample generation and additional adversarial optimization constraints (Madry et al., 2018; Pang et al., 2019; Gupta et al., 2020). Prior work has further shown that synthetic adversarial examples consume significantly more representational capacity than natural data, and even over-parameterized deep networks may exhibit insufficient model capacity under adversarial training (Jiang and Chiang, 2024). Consequently, there remains a need for parameter-efficient adversarial adaptation strategies that preserve robustness in low model capacity settings.

In this work, we focus on the domain mismatch problem in the context of a cochlear implant (CI), a medical device that restores auditory perception in individuals with severe to profound sensorineural hearing loss by converting acoustic signals into electrical stimulation delivered through an electrode array along the cochlea (Carlyon and Goehring, 2021). Although CI users typically achieve good speech understanding in quiet conditions, speech intelligibility deteriorates substantially in dynamic noisy and reverberant environments (Hazrati and Loizou, 2012; Kokkinakis et al., 2011). While current CI devices incorporate tradi-

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 4631–4648
November 6-10, 2026 ©2026 Association for Computational Linguistics

tional signal processing solutions, such as beamforming and signal-to-noise based noise reduction, for noise suppression, these solutions do not directly address reverberation or combinations of reverberation and noise and have limited robustness.

Studies have shown that ideal time-frequency (T-F) masking for suppressing distortion-dominated regions improve speech intelligibility for hearing impaired listeners in adverse acoustic conditions (Wang, 2005; Roman and Woodruff, 2013; Healy et al., 2019; Koning et al., 2015; Kokkinakis and Stohl, 2021). Prior approaches investigated for blind T-F mask estimation either use traditional signal processing method that require future signal information or room-specific tuning (Hazrati et al., 2013) or high capacity neural network architectures (Borjigin et al., 2024). CI sound processing must remain causal with end-to-end latencies within tolerable audiovisual asynchrony (Hay-McCutcheon et al., 2009).

Lightweight phoneme-based speech enhancement models targeted for CI applications incorporate framewise predictions of phonetic information to guide T-F mask estimation for CI processed speech(Chu et al., 2023; Saha et al., 2026c). The effectiveness of phoneme-based mask estimation is inherently tied to the robustness of the underlying phoneme classifier, making the latter a key limiting factor in downstream speech enhancement performance. Framewise phoneme classification is highly susceptible to acoustic degradation because phonemes with similar spectral and temporal characteristics are frequently confused under adverse conditions (Ladefoged and Johnson, 2014; Saha et al., 2026a,b).

Previous studies have attempted to improve robustness of phoneme-based speech enhancement by grouping phonemes into broader categories, such as manner of articulation (MOA) classes (Lu et al., 2023; Ziegler et al., 2012; Scanlon et al., 2007; Saha et al., 2026b). While such phoneme grouping strategies can reduce inter-class confusion, they do not explicitly address acoustic domain mismatch. To address these limitations, we propose an adversarial adaptation framework that alleviates the representational and computational limitations of conventional adversarial learning in lightweight causal networks. Our main contributions are summarized as follows:

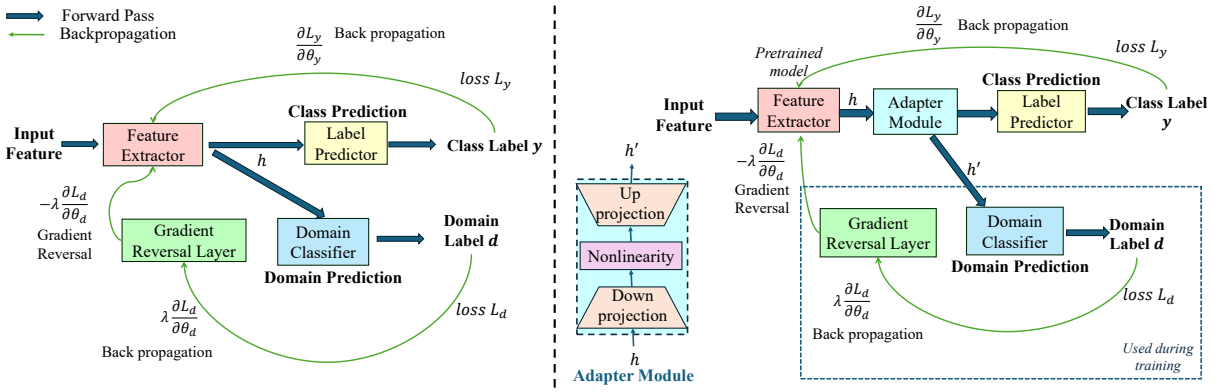
- We introduce a parameter-efficient adapter based adversarial learning framework for lightweight causal recurrent networks. Our approach integrates a lightweight bottleneck adapter module into a pretrained classification backbone and jointly optimizes both the backbone and adapter using domain adversarial learning.
- We demonstrate that pretraining plays a critical role in successful adversarial adaptation on noisy reverberant test data. By first learning stable phonetic representations under reverberant only conditions, the network is better able to capture invariant phonetic structure before adversarial adaptation to noisy reverberant domains. Interestingly, this selective domain mismatch during pretraining leads to improved robustness across domains, highlighting the counterintuitive benefits of single domain pretraining conditions.
- We show that our adapter-based adversarial adaptation method improves the learning of phoneme-discriminative and domain-invariant representations when compared to conventional adversarial learning, leading to improved downstream speech enhancement performance on CI vocoded speech.

2 Related Work

2.1 Domain Adversarial Training

Conventional supervised learning assumes that training and test samples are drawn from the same distribution. However, in real-world speech applications, this assumption is frequently violated due to variations in acoustic environments, speakers, recording devices, background noise and reverberation, leading to significant performance degradation under unseen conditions (Szegedy et al., 2014; Goodfellow et al., 2015; Shafahi et al., 2019; Madry et al., 2018; Bera and Agarwal, 2025). Domain-adversarial learning addresses this issue by reducing the dependence of learned representations on domain-specific cues, thereby improving robustness under distribution mismatch (Ganin et al., 2016). Such approaches have been successfully applied in computer vision, natural language processing (NLP) and speech recognition to improve robustness under distribution shift (Yoo and Qi, 2021; Altinisik et al., 2023; Alzantot et al., 2018; Kim et al., 2023).

A typical domain-adversarial framework consists of three components, illustrated in Figure 1a:



(a) Adversarial Training

(b) Adversarial Learning with Adapter Module

Figure 1: **(a) Conventional adversarial training framework** consisting of a feature extractor (pink), label predictor (yellow), and domain classifier (blue) with parameters θ_d and domain label d, connected through a gradient reversal layer (GRL) (green), which multiplies gradients by a negative scalar λ during backpropagation. The label predictor is parameterized by θ_y with the class labels denoted by y. Training minimizes the label prediction loss (L_y) and the domain classification loss (L_d). **(b) Our adapter-based adversarial training framework**, where a lightweight bottleneck adapter module (light blue, expanded view, left) is inserted into the pretrained backbone to refine intermediate representations h before (phoneme or manner of articulation) classification. The domain classifier (dark blue) with parameters θ_d receives the adapted features h'. The adapted features are simultaneously optimized for label prediction and domain confusion through adversarial learning.

a feature extractor, a task classifier and an auxiliary domain classifier. The task classifier predicts the primary labels, while the domain classifier attempts to identify the domain of the extracted features. During backpropagation, a gradient reversal layer reverses gradients from the domain classifier (Ganin et al., 2016), forcing the feature extractor to simultaneously minimize task loss and maximize domain classification loss. Consequently, the learned representations preserve task-relevant information while suppressing domain-specific cues, improving generalization across domains.

Conventional domain adversarial learning can be challenging to optimize in lightweight architectures with limited representational capacity. Compact recurrent networks must simultaneously preserve task-discriminative information while suppressing domain-specific variability, making robust representation learning challenging under severe domain mismatch. Moreover, full-model fine-tuning on new domains can lead to catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999) and unstable adaptation, particularly in low-resource settings where models are prone to overfitting and poor cross-domain generalization (Lee et al., 2017).

2.2 Adapter Modules for Fine-Tuning

Adapter modules have emerged as a parameter-efficient alternative to full model fine-tuning. Instead of updating all parameters of a pretrained model, adapters introduce a small number of task-specific trainable parameters while keeping most of the backbone network shared across tasks. Prior work has explored task-specific and multi-task adapter learning strategies. Bapna and Firat (2019) proposed adapter layers for neural machine translation by inserting lightweight task-specific modules into a pretrained model, while Stickland and Murray (2019); Pfeiffer et al. (2021) investigated multi-task adapter training to leverage shared representations across tasks during transfer learning. Similarly, Ahmat et al. (2023) used adapters for cross-lingual adaptation to construct language-pair-specific alignment subspaces while preserving the original semantic space of the pretrained model. Most existing approaches, however, focus on transformer-based architectures and keep the pretrained backbone fixed during adaptation.

Other studies have combined adapters with adversarial or robustness oriented objectives. Rebuffi et al. (2023) showed that lightweight adapters can improve adversarial robustness in large vision transformers using domain-specific trainable parameters, while Nguyen and Le (2024) combined adapters, adversarial training and mixup augmentation for robust NLP classification. Similarly, (Wu et al., 2025) introduced robustness feature adapters that improve adversarial generalization

through feature-space perturbation learning. However, these approaches primarily focus on robustness against synthetically generated adversarial samples in large transformer-based vision and language models, and keep the pretrained backbone weights fixed during adaptation.

In contrast, this work investigates adapter-based adversarial adaptation for offline training of lightweight networks designed for real-time edge deployment, which, to the best of our knowledge, has not been previously explored. Unlike conventional adapter fine-tuning approaches that freeze the pretrained backbone or allocate separate adapters for different domains, we jointly optimize both the backbone network and lightweight adapters within a unified domain-adversarial framework. This achieves parameter-efficient adaptation while learning acoustically invariant feature representations under distribution shifts.

3 Methodology

Our adapter-based adversarial adaptation framework for robust classification in lightweight causal recurrent neural networks is illustrated in Figure 1b. We first describe the pretrained classifier and adapter integration strategy, and the adversarial domain adaptation framework used during training.

3.1 Classifier and Adapter Integration

A classification network is first pretrained using supervised learning. Given an input feature sequence x_i, the network predicts the posterior probability of each class using the cross-entropy loss, $\mathcal{L}_{CE}$:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{|y|}\mathbf{1}(y_i = j)\log P(y = j \mid f_\Theta(x_i)), \tag{1}$$

where N denotes the number of training samples; $f_\Theta(x_i)$ represents the feature representation produced by the network with parameters Θ; y_i denotes the ground-truth class label; and $|y|$ is the total number of output classes.

After pretraining, an adapter module is inserted into the intermediate feature extraction stage of the network, as illustrated in Figure 1b. The adapter employs a bottleneck architecture consisting of a down-projection layer, nonlinear activation and a up-projection layer with a residual connection. Given an intermediate representation h, the adapted representation is computed as:

$$h' = h + W_{\text{up}}\big(\sigma(W_{\text{down}}h)\big), \tag{2}$$

where W_{down} and W_{up} denote the down- and up-projection matrices, respectively; and $\sigma(\cdot)$ denotes a nonlinear activation function. The adapter introduces only a small number of additional trainable parameters while enabling the network to learn domain-adaptive feature representations with minimal computational overhead.

3.2 Domain Adversarial Training for Lightweight Adaptation

To improve robustness across domains, the proposed framework incorporates adversarial domain adaptation during training, as illustrated in Figure 1b. An auxiliary domain discriminator is introduced to predict the domain associated with each input representation.

The domain discriminator operates on the adapted feature representation h' generated by the backbone network and the adapter module. The gradient reversal layer (GRL) is inserted between the feature extractor and the discriminator. During backpropagation, the GRL reverses gradients from the domain classifier, encouraging the feature extractor and adapter modules to learn representations that remain informative for the primary classification task while suppressing domain-specific information. The domain classification loss $\mathcal{L}_{dom}$ is defined as:

$$\mathcal{L}_{dom} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{|d|}\mathbf{1}(d_i = j)\log P(d = j \mid h'_i), \tag{3}$$

where d_i denotes the domain label associated with sample i; $|d|$ denotes the total number of domains; and h'_i represents the adapted feature representation for sample i.

The overall training objective combines the task classification and adversarial domain losses:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_{\text{adv}}\mathcal{L}_{dom}, \tag{4}$$

where λ_{adv} controls the contribution of the adversarial objective. During optimization, the domain discriminator minimizes $\mathcal{L}_{dom}$, while the GRL encourages the feature extractor and adapter modules to maximize it. The domain discriminator is used only during training and removed during inference.

Backbone	Adversarial /Adapter	Hidden Dim Size	Bottleneck Dim Size	# Parameters (Ph, MOA)	MACs (M) (Ph, MOA)	Size (MB) (Ph, MOA)
GRU+A	No ($Base$)	117	-	124996, 121338	359.07, 355.45	0.50, 0.49
GRU+A	Yes (Adv)	117	-	124996, 121338	359.07, 355.45	0.50, 0.49
GRU+A	**Yes (Adv, AM)***	**117**	**32**	**132633 , 128975**	366.63, 363.00	0.53, 0.52
LSTM+A	No ($Base$)	117	-	146524, 142866	380.95, 377.33	0.59, 0.57
LSTM+A	Yes (Adv, AM)*	117	32	154161, 150,503	388.50, 384.88	0.62, 0.60
GRU+A	No ($Base$)	120	-	130720, 126969	370.80, 367.08	0.52, 0.51
GRU+A	Yes (Adv, AM)*	120	32	138552, 134801	378.54, 374.82	0.55, 0.54
GRU+A	Yes (Adv, AM)*	117	16	128873, 125215	362.85, 359.22	0.52, 0.50
GRU+A	Yes (Adv, AM)*	117	64	140153, 136495	374.18, 370.55	0.56, 0.55

*Model specification excludes the domain discriminator (6,942 parameters), which is removed during inference.

Table 1: Model specifications for phoneme (Ph) and manner of articulation (MOA) classifiers with alternative architecture designs. GRU, Gated recurrent unit; LSTM, Long short-term memory; A, Attention; $Base$, Baseline; Adv, Adversarial training; AM, Adapter module; Bottleneck Dim Size, Dimension size of adapter module bottleneck; Hidden Dim Size: Dimension size of the hidden layer in the backbone model.

Consequently, the final model during inference retains nearly the same computational complexity as the baseline classifier, with only the addition of a lightweight adapter module.

4 Experimental Design

The adapter-based adversarial framework is evaluated on a causal frame-wise phoneme classification task for phoneme-based speech enhancement in CIs, where classifier outputs are used to guide downstream T-F mask estimation. Models were implemented in PyTorch and trained on a NVIDIA A5000 GPU.

4.1 Datasets

Training Data. The training speech dataset consists of an 8000-utterance subset (≈28 hrs) of the LibriSpeech corpus (Panayotov et al., 2015). Noise signals are selected from the WHAM! (Wichern et al., 2019) and Bern Cocktail Party (Fischer et al., 2020) datasets, containing real world environmental and babble noise recordings. Anechoic speech and noisy speech generated at SNR levels from -5 to 20 dB are convolved with room impulse responses (RIRs) from the BUT Reverb Database (Szöke et al., 2018) to create reverberant and noisy-reverberant conditions.

Test Data. Evaluation is performed using the HINT (Nilsson et al., 1994) and CUNY (Boothroyd et al., 1985) speech corpora, which are commonly used in CI research. The HINT corpus consists of 240 phonetically balanced sentences with controlled vocabulary and relatively uniform sentence lengths, enabling reliable and repeatable intelligibility assessment of speech intelligibility in noise. The CUNY corpus contains 720 syntactically natural and meaningful sentences spoken by both male and female talkers, providing greater variability in speaker characteristics and sentence content.

To evaluate performance under realistic acoustic conditions, the speech corpora were combined with real-world RIRs and environmental noise recordings representative of everyday listening environments encountered by CI users. These conditions include binaural (left- and right-channel) RIR recordings (Table A.5) from the Aachen Impulse Response dataset (Jeub et al., 2009), and public noise recordings (interiors of a cafeteria, restaurant and a busy subway station) from the DEMAND dataset (Thiemann et al., 2013). Additionally, we generated two-talker babble using the test partition of the TIMIT speech corpus (Garofolo et al., 1993). The TIMIT test set utterances were first grouped according to speaker identity. For each babble recording, two distinct speakers were randomly selected without replacement, ensuring that a speaker appeared only once within a given mixture. Test SNRs range from -5 to 20 dB. The resulting test conditions therefore capture variability in reverberation, environmental noise, speaker characteristics, and inter-channel differences.

4.2 Feature Extraction

Speech features are extracted following the Advanced Combination Encoder strategy (Vandali et al., 2000) implemented in the Nucleus Matlab CI Toolbox (Swanson and Mauch, 2006). Signals are segmented into 8 ms frames with 2 ms overlap, transformed using short-time Fourier transform (STFT) followed by a log operation, producing 65-dimensional log-compressed spectral features. Global mean-variance normalization is applied prior to classification and mask estimation.

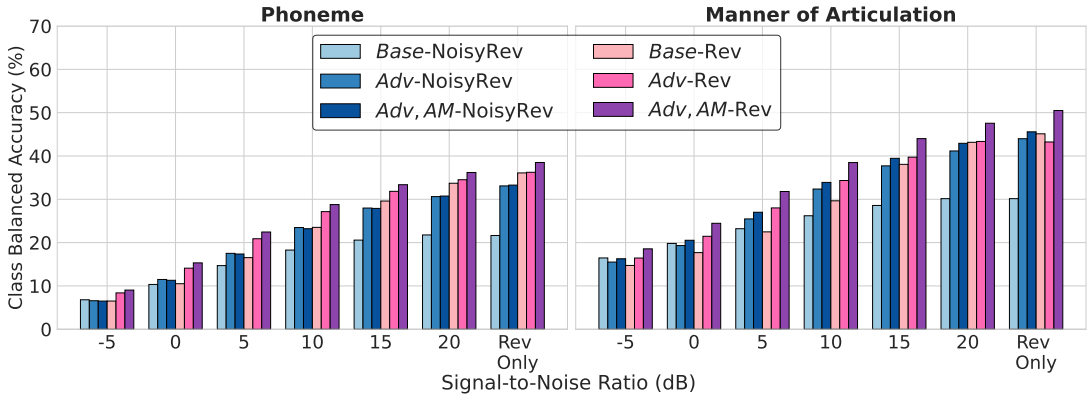


Figure 2: Frame-wise phoneme and manner of articulation classification accuracy under noisy-reverberant conditions. The evaluated models include the baseline classifier ($Base$) pretrained using either reverberant only (Rev) or noisy-reverberant (NoisyRev) data, the conventional adversarially trained model (Adv), and the proposed adapter-based adversarially trained model (Adv, AM), both initialized from the corresponding pretrained baseline.

4.3 Model Architecture and Training

The backbone classifier is a lightweight causal recurrent network for CI applications adapted from (Saha et al., 2026c), consisting of a single-layer unidirectional gated recurrent unit (GRU) or long short-term memory (LSTM) network with 117 hidden units followed by a multi-head attention module. The network predicts frame-wise phoneme posteriors over 39 phoneme classes plus a silence class (Ladefoged and Johnson, 2014). Phoneme labels are generated using Kaldi forced alignment (Povey et al., 2011) and mapped to CI time bins. MOA classification was also evaluated (Appendix Table A.1).

The baseline classifier model is pretrained using reverberant only or noisy reverberant speech and adapted using the proposed adversarial framework. During adaptation, a bottleneck adapter module with dimension 32 and ReLU activation is inserted after the attention layer. For the GRU model, additional model specifications investigated include the number of hidden dimensions, adapter bottleneck size and adversarial weight factor. Model complexity (number of parameters, model size and multiply-accumulate operations) is reported in Table 1.

Adversarial adaptation incorporates the domain classification objective (Equation 3) with $\lambda_{\text{adv}} = 0.1$. Two domains are considered during training: reverberant only speech (domain 0) and noisy-reverberant speech (domain 1). Classification models were trained using stochastic gradient descent with learning rate 1×10^{-5} and momentum 0.9. Training data were partitioned into disjoint training and validation sets. Training used 2-second segments with early stopping when the validation loss was unchanged for 10 consecutive epochs.

4.4 Speech Enhancement

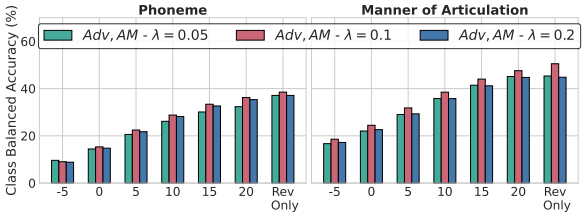
For downstream speech enhancement, the classifier output is integrated into the Omni-Expert (OE) mask estimation framework (Saha et al., 2026c). The OE model consists of a single expert backbone that applies learned task-specific feature transformations to input features based on phoneme information. The OE model estimates phoneme-specific masks, which are probabilistically combined into a composite mask based on predictions from a phoneme group classifier. The OE mask estimation procedure is detailed in Appendix A.4).

5 Results

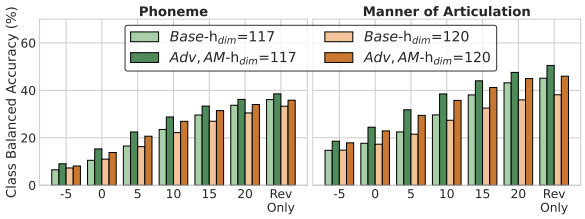
5.1 Impact of Adapter-Based Adversarial Learning and Pretraining Conditions

Figure 2 compares frame-wise classification performance for both phoneme and MOA classifiers under noisy-reverberant conditions.

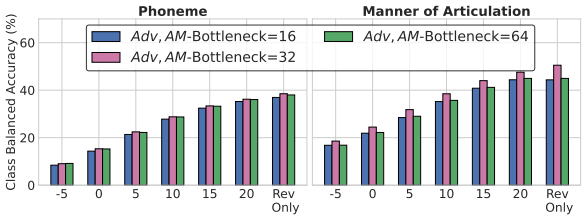
Adversarial Adaptation. The results show that both adversarial training approaches consistently improved classification robustness under diverse acoustic conditions when compared to the baseline ($Base$) models. Our adapter-based adversarially trained (Adv, AM) models significantly outperformed conventional adversarially trained (Adv) models. This improvement suggests that the adapter module more effectively refines learned intermediate feature representations during adversarial adaptation, enabling the network to preserve



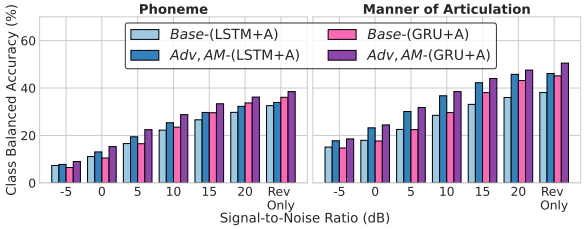
(a) Effect of the adversarial domain classification loss weighting factor, λ_{adv}.



(b) Impact of GRU hidden dimension size (h_{dim}).



(c) Impact of adapter bottleneck dimension size.



(d) Comparison of gated recurrent unit (GRU) and long short-term memory (LSTM) attention (+A) backbone architectures.

Figure 3: Comparison of architectural design choices for frame-wise phoneme and manner of articulation classification under noisy reverberant conditions. The adapter based adversarial trained model (Adv, AM) is initialized from a baseline network ($Base$) pretrained on reverberant-only data.

discriminative phonetic structure in the presence of noise and reverberation.

Pretraining Conditions. The pretraining condition also strongly influences the effectiveness of adversarial adaptation. The mismatched reverberant-only pretrained $Base$ models consistently outperformed those pretrained on matched noisy-reverberant data. In particular, reverberant-only pretraining followed by adapter based adversarial adaptation achieves the highest performance across nearly all SNR conditions.

These findings suggest that separating phonetic representation learning from robustness adaptation is beneficial. Pretraining on reverberant only speech enables the network to learn stable phonetic structure with reduced environmental interference, while subsequent adversarial adaptation introduces robustness to acoustic variability in a more controlled manner. Interestingly, the reverberant-only pretrained $Base$ model for phoneme classification remains competitive at higher SNR levels (>10 dB), where it outperforms both Adv and Adv, AM models trained on matched noisy-reverberant data. This indicates that selective single domain pretraining can itself produce robust phonetic representations, highlighting the counterintuitive benefits of controlled domain mismatch during pretraining.

5.2 Impact of Adversarial and Architectural Design Choices

Adversarial Loss Weighting. Figure 3a shows results of the impact of the adversarial domain loss weight λ_{adv} in GRU+A classifier models. Phoneme classification performance remains relatively stable across different adversarial weights, suggesting that phonetic representations are robust to varying levels of domain regularization. In contrast, MOA classification exhibits greater sensitivity to λ_{adv}, indicating that broader phonetic categories are more strongly influenced by the degree of enforced domain invariance.

Backbone Hidden Representation Size. The effect of GRU hidden dimension size is shown in Figure 3b. The Adv, AM framework consistently improves phoneme and MOA classification performance across different hidden dimensions (117 and 120), demonstrating robustness across varying model capacities. Interestingly, the best overall performance is achieved at the lower hidden dimension of 117, indicating that the proposed framework can effectively learn robust representations without requiring larger recurrent models.

Adapter Bottleneck Dimension. The bottleneck dimension determines the size of the intermediate adapter representation and consequently controls the adaptation capacity of the module. Figure 3c shows the impact of different adapter bottleneck dimensions. Smaller bottleneck dimensions (16) reduce parameter overhead but limit transformation flexibility, whereas larger dimensions (64) increase adaptation capacity at the cost of additional parameters. Overall, a bottleneck dimension of 32 provided the best trade-off between classification performance and computational efficiency.

Metric	Model	Signal-to-Noise Ratio of Noisy-Reverberant Speech						
		-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB	Rev Only
SRMR-CI	NoisyRev	0.832 (±0.004)	0.935 (±0.008)	1.055 (±0.013)	1.132 (±0.016)	1.167 (±0.018)	1.181 (±0.019)	1.413 (±0.022)
	Base	1.171 (±0.017)	1.417 (±0.019)	1.645 (±0.023)	1.754 (±0.024)	1.783 (±0.024)	1.785 (±0.024)	2.101 (±0.027)
	Adv, AM	**1.360 (±0.021)**	**1.608 (±0.025)**	**1.829 (±0.027)**	**1.943 (±0.028)**	**1.973 (±0.027)**	**1.964 (±0.026)**	**2.124 (±0.029)**
	Oracle	1.793 (±0.025)	1.953 (±0.025)	2.036 (±0.023)	2.059 (±0.022)	2.059 (±0.022)	2.055 (±0.022)	2.051 (±0.023)
	IRM	2.081 (±0.023)	2.186 (±0.022)	2.240 (±0.022)	2.261 (±0.022)	2.269 (±0.022)	2.274 (±0.022)	2.241 (±0.022)
STOI	NoisyRev	0.499 (±0.003)	0.563 (±0.003)	0.617 (±0.003)	0.656 (±0.003)	0.680 (±0.003)	0.691 (±0.003)	0.699 (±0.001)
	Base	0.546 (±0.005)	0.633 (±0.004)	0.701 (±0.003)	0.746 (±0.003)	0.771 (±0.003)	0.783 (±0.002)	**0.792 (±0.001)**
	Adv, AM	**0.552 (±0.005)**	**0.641 (±0.004)**	**0.708 (±0.004)**	**0.751 (±0.003)**	**0.775 (±0.003)**	**0.787 (±0.002)**	0.790 (±0.001)
	Oracle	0.694 (±0.004)	0.737 (±0.004)	0.774 (±0.003)	0.800 (±0.003)	0.816 (±0.002)	0.824 (±0.002)	0.829 (±0.001)
	IRM	0.959 (±0.001)	0.968 (±0.001)	0.970 (±0.001)	0.970 (±0.001)	0.969 (±0.001)	0.969 (±0.001)	0.969 (±0.000)

Table 2: Mean ± 95% confidence intervals of objective speech intelligibility metrics: speech-to-reverberation modulation energy ratio for cochlear implant users (SRMR-CI) and short-time objective intelligibility (STOI) for noisy reverberant (NoisyRev) speech enhanced using the ideal ratio mask (IRM) and the Omni-Expert (OE) model with phoneme predictions. Phoneme classifiers include the baseline model (*Base*) trained on reverberant-only data and the adapter-based adversarial trained model (*Adv*, *AM*). The oracle denotes the OE model with known/oracle phoneme information. Bold indicates highest performing non-oracle model.

Recurrent Architectures. Figure 3d compares GRU-based and LSTM-based architectures within the adversarial adapter framework. Both architectures benefit from adversarial adaptation, with the GRU+A configuration achieving higher classification performance at a lower computational complexity (Table 1). These results suggest that GRU-based recurrent modeling offers a more efficient architecture for lightweight real-time CI applications.

Model Complexity. Incorporating adversarial learning and adapter modules increases the parameter count only modestly relative to the corresponding *Base* models (Table 1). The reported parameter counts exclude the domain discriminator, which is only used during training and removed during inference. This allows the deployed model to remain lightweight.

Overall, these findings demonstrate that the adapter-based adversarial framework remains effective across multiple architectural configurations while maintaining a compact model suitable for low-latency CI deployment.

5.3 Impact on Objective Speech Intelligibility Metrics

We then assessed how improved robustness in phoneme group classification influences the downstream phoneme-based speech enhancement task. The mask estimation expert in the OE model is trained using matched noisy-reverberant conditions (Saha et al., 2026b). Two phoneme classifier configurations were evaluated: the reverberant only pretrained *Base* model and the counterpart *Adv*, *AM* model. MOA based results are provided in the Appendix Table A.2.

Mean speech intelligibility performance and 95% confidence intervals were estimated using 5,000-sample bootstrap resampling at the sample level, preserving the repeated-measures structure. Table 2 compares the speech-to-reverberation modulation energy ratio for cochlear implant users

(SRMR-CI) and short-time objective intelligibility (STOI) scores of noisy-reverberant speech enhanced using the phoneme-based OE model. Statistical test results are detailed in Appendix A.7.

The Adv, AM phoneme classifier model achieves statistically significant improvements in both SRMR-CI and STOI from the $Base$ classifier model, particularly at lower SNR levels. These results demonstrate that improved robustness in phoneme classification leads to more accurate downstream T-F mask estimation, with the largest gains observed under severe noisy-reverberant conditions.

6 Conclusion

This work introduced adapter-based adversarial training for robust framewise phoneme classification under noisy and reverberant acoustic conditions. By integrating a lightweight adapter module into a pretrained causal recurrent network during adversarial training, the model learns representations that remain discriminative for classification while reducing sensitivity to acoustic domain mismatch. The resulting architecture maintains a compact model size suitable for real-time CI applications.

Empirical findings demonstrated that staged pretraining on reverberant-only speech followed by adversarial adaptation improves robustness across a wide range of SNR conditions. Improvements in phoneme classification robustness further translated to measurable gains in downstream speech enhancement performance, as reflected by objective speech intelligibility measures. Although MOA classifiers achieved higher classification accuracy overall, phoneme-based models yielded larger improvements in speech enhancement performance, suggesting that finer-grained phoneme information is more beneficial for robust T-F mask estimation and noise suppression.

Overall, the results demonstrate that adapter-based adversarial learning provides an effective and parameter-efficient strategy for improving robust speech processing models in lightweight causal systems, particularly under challenging low-SNR acoustic conditions.

Limitations

The proposed framework was evaluated using real-world noisy and reverberant acoustic recordings, simulated CI-vocoded speech and objective speech intelligibility metrics. While this evaluation protocol is standard in offline CI speech enhancement research, future work will include human listening studies to provide a more comprehensive assessment of how improvements in model performance translate to speech perception outcomes in CI users. The current study considers a limited set of acoustic domains. Evaluating performance under broader acoustic variability and speaker diversity to reflect unseen real-world environments is also important.

Although on-device deployment was not implemented, the lightweight neural networks used in this work are within specifications (<10^6 parameters, <1.5Mb) for low-power microcontrollers targeted for assisted hearing technologies (Itani et al., 2025; Tri Lê et al., 2026). The adapter module introduces minimal computational overhead. For real-time optimization, model compression techniques, such as quantization and pruning, may be necessary for deployment on resource-constrained embedded CI hardware.

Societal Impact

Improving speech intelligibility in noisy and reverberant environments has the potential to significantly enhance communication accessibility and quality of life for CI users. Robust low-latency speech enhancement systems will support improved social interaction, educational participation, and overall listening comfort of CI users in everyday acoustic environments.

Robustness and generalization across diverse acoustic conditions remain important considerations for equitable deployment. Models trained on limited datasets may not generalize well across different speakers, accents, languages, or recording conditions. Future work should therefore prioritize broader and more diverse data collection and evaluation to ensure reliable and inclusive performance across user populations.

Beyond CI applications, the adapter-based adversarial framework may also be beneficial for other edge-deployed speech and audio systems operating with limited computational resources under acoustic domain mismatch. Potential applications include hearing aids, wearable assistive devices, mobile speech enhancement, embedded automatic speech recognition, and low-power Internet-of-Things audio processing systems, where robustness, low latency and parameter efficiency are critical requirements.

Acknowledgments

This work was supported by a grant from the National Institutes of Health (R56DC020267-01A1).

References

Ahtamjan Ahmat, Yating Yang, Bo Ma, Rui Dong, Kaiwen Lu, and Lei Wang. 2023. WAD-X: Improving Zero-shot Cross-lingual Transfer via Adapter-based Word Alignment. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(9).

Enes Altinisik, Hassan Sajjad, Husrev Sencar, Safa Messaoud, and Sanjay Chawla. 2023. Impact of Adversarial Training on Robustness and Generalizability of Language Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7828–7840, Toronto, Canada. Association for Computational Linguistics.

Moustafa Alzantot, Bharathan Balaji, and Mani Srivastava. 2018. Did you hear that? Adversarial Examples Against Automatic Speech Recognition. *arXiv e-prints*, arXiv:1801.00554.

Ankur Bapna and Orhan Firat. 2019. Simple, Scalable Adaptation for Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Anup Bera and Aanchal Agarwal. 2025. Bias Detection and Mitigation Framework for ASR System. In *7th International Conference on Signal Processing and Information Communications*, pages 13–27, Cham. Springer Nature Switzerland.

Arthur Boothroyd, Laurie Hanin, and Theresa Hnath. 1985. A sentence test of speech perception: Reliability, set equivalence, and short term learning.

Agudemu Borjigin, Kostas Kokkinakis, Hari M Bharadwaj, and Joshua S Stohl. 2024. Deep learning restores speech intelligibility in multi-talker interference for cochlear implant users. *Scientific Reports*, 14(1):13241.

Robert P. Carlyon and Tobias Goehring. 2021. Cochlear Implant Research and Development in the Twenty-first Century: A Critical Update. *Journal of the Association for Research in Otolaryngology*, 22(5):481–508.

Hanjie Chen and Yangfeng Ji. 2022. Adversarial Training for Improving Model Robustness? Look at Both Prediction and Interpretation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10463–10472. AAAI Press.

Kevin Chu, Leslie Collins, and Boyla Mainsah. 2023. Suppressing reverberation in cochlear implant stimulus patterns using time-frequency masks based on phoneme groups. *Proceedings of Meetings on Acoustics*, 50(1):050002.

Tim Fischer, Marco Caversaccio, and Wilhelm Wimmer. 2020. Multichannel acoustic source and image dataset for the cocktail party effect in hearing aid and implant users. *Scientific Data*, 7(1):440.

Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(59):1–35.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, David S Pallett, and Nancy L Dahlgren. 1993. DARPA TIMIT::acoustic-phonetic continuous speech corpus CD-ROM, NIST speech disc 1-1.1.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*.

Sidharth Gupta, Parijat Dube, and Ashish Verma. 2020. Improving the affordability of robustness training for dnns. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3383–3392.

Marcia J. Hay-McCutcheon, David B. Pisoni, and Kristopher K. Hunt. 2009. Audiovisual asynchrony detection and speech perception in hearing-impaired listeners with cochlear implants: A preliminary analysis. *International Journal of Audiology*, 48(6):321–333.

Oldooz Hazrati, Jaewook Lee, and Philipos C. Loizou. 2013. Blind binary masking for reverberation suppression in cochlear implants. *The Journal of the Acoustical Society of America*, 133(3).

Oldooz Hazrati and Philipos C. Loizou. 2012. The combined effects of reverberation and noise on speech intelligibility by cochlear implant listeners. *International Journal of Audiology*, 51(6):437–443.

Eric W. Healy, Jordan L. Vasko, and DeLiang Wang. 2019. The optimal threshold for removing noise from speech is similar across normal and impaired hearing—a time-frequency masking study. *The Journal of the Acoustical Society of America*, 145(6).

Malek Itani, Tuochao Chen, Arun Raghavan, Gavriel Kohlberg, and Shyamnath Gollakota. 2025. Wireless hearables with programmable speech AI accelerators. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking*, pages 863–877.

Marco Jeub, Magnus Schafer, and Peter Vary. 2009. A binaural room impulse response database for the evaluation of dereverberation algorithms. *16th International Conference on Digital Signal Processing*.

Yixuan Jiang and Hsiao-Dong Chiang. 2024. Geometry-Aware Weight Perturbation for Adversarial Training. *Electronics*, 13(17).

Dohee Kim, Daeyeol Shim, and Joon-Hyuk Chang. 2023. General-purpose Adversarial Training for Enhanced Automatic Speech Recognition Model Generalization. In *Interspeech 2023*, pages 889–893.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*.

Kostas Kokkinakis, Oldooz Hazrati, and P. Loizou. 2011. A channel-selection criterion for suppressing reverberation in cochlear implants. *Journal of the Acoustical Society of America*, 129(5).

Kostas Kokkinakis and Joshua S. Stohl. 2021. Optimized gain functions in ideal time-frequency masks and their application to dereverberation for cochlear implants. *JASA Express Letters*, 1(8).

Koning, Madhu, and Wouters. 2015. Ideal Time–Frequency Masking Algorithms Lead to Different Speech Intelligibility and Quality in Normal-Hearing and Cochlear Implant Listeners. *IEEE Transactions on Biomedical Engineering*, 62(1).

P. Ladefoged and K. Johnson. 2014. *A Course in Phonetics*. Cengage Learning.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully Character-Level Neural Machine Translation without Explicit Segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Heiner Löllmann, Emre Yilmaz, Marco Jeub, and Peter Vary. 2010. An improved algorithm for blind reverberation time estimation. In *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 1–4.

Yen-Ju Lu, Chia-Yu Chang, Cheng Yu, Ching-Feng Liu, Jeih-weih Hung, Shinji Watanabe, and Yu Tsao. 2023. Improving Speech Enhancement Performance by Leveraging Contextual Broad Phonetic Class Information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2738–2750.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

Afsana Mou and Mariofanna Milanova. 2024. Performance Analysis of Deep Learning Model-Compression Techniques for Audio Classification on Edge Devices. *Sci*, 6(2):21.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *International Conference on Machine Learning*, pages 807–814.

Patrick A Naylor and Nikolay D Gaubitch. 2010. *Speech dereverberation*. Springer Science & Business Media.

Tuc Van Nguyen and Thai Le. 2024. Adapters Mixup: Mixing Parameter-Efficient Adapters to Enhance the Adversarial Robustness of Fine-tuned Pre-trained Text Classifiers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21183–21203, Miami, Florida, USA. Association for Computational Linguistics.

Michael Nilsson, Sigfrid D Soli, and Jean A Sullivan. 1994. Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, 95(2).

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. 2019. Improving Adversarial Robustness via Promoting Ensemble Diversity. In *International Conference on Machine Learning*, pages 4970–4979.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, and 1 others. 2011. The Kaldi Speech Recognition Toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding*.

Sylvestre-Alvise Rebuffi, Francesco Croce, and Sven Gowal. 2023. Revisiting Adapters with Adversarial Training. In *International Conference on Learning Representations*.

Nicoleta Roman and John Woodruff. 2013. Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold. *The Journal of the Acoustical Society of America*, 133(3).

Sohini Saha, Leslie Collins, and Boyla Mainsah. 2026a. Leveraging Acoustic Domain Mismatch During Training for Robust Speech Enhancement Models for Cochlear Implants. In *IEEE Spoken Language Technology Workshop*.

Sohini Saha, Leslie Collins, and Boyla Mainsah. 2026b. Leveraging Articulatory Phonetic Groupings for Robust Speech Enhancement for Cochlear Implants in Complex Acoustic Environments. In *Findings of the 5th Asia-Pacific Chapter of the Association for Computational Linguistics & the 15th International Joint Conference on Natural Language Processing: AACL-IJCNLP*.

Sohini Saha, Mezisashe Ojuba, Leslie Collins, and Boyla Mainsah. 2026c. The Omni-Expert: A Computationally Efficient Approach to Achieve a Mixture of Experts in a Single Expert Model. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 124574–124617. Curran Associates, Inc.

Patricia Scanlon, Daniel P. W. Ellis, and Richard B. Reilly. 2007. Using Broad Phonetic Group Experts for Improved Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):803–812.

Manfred R Schroeder. 1965. New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(3):409–412.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial Training for Free! In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5986–5995. PMLR.

B Swanson and H Mauch. 2006. Nucleus Matlab Toolbox 4.20 software user manual. *Cochlear Ltd*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing Properties of Neural Networks. In *International Conference on Learning Representations*.

Igor Szöke, Miroslav Skácel, Ladislav Mošner, Jakub Paliesek, and J. Černocký. 2018. Building and evaluation of a real room impulse response dataset. *IEEE Journal on Selected Topics in Signal Processing*, 13(4).

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. Measuring Robustness to Natural Distribution Shifts in Image Classification. *Advances in Neural Information Processing Systems*, 33:18583–18599.

Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. 2013. The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings . In *Proceedings of Meetings on Acoustics*, volume 19. AIP Publishing.

Minh Tri Lê, Pierre Wolinski, and Julyan Arbel. 2026. Efficient neural networks for tiny machine learning: A comprehensive review. *ACM Transactions on Intelligent Systems and Technology*, 17(4):1–41.

Andrew E. Vandali, Lesley A. Whitford, Kerrie Plant, Clark, and Graeme M. 2000. Speech Perception as a Function of Electrical Stimulation Rate: Using the Nucleus 24 Cochlear Implant System. *Ear and Hearing*, 21:608–624.

Fazal E. Wahab, Zhongfu Ye, Nasir Saleem, Sami Bourouis, and Amir Hussain. 2025. Lightweight Adaptive Deep Learning for Efficient Real-Time Speech Enhancement on Edge Devices. *IEEE Transactions on Consumer Electronics*, 71(4):12086–12095.

DeLiang Wang. 2005. *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis*. Springer US.

Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. 2019. WHAM!: Extending Speech Separation to Noisy Environments. *Interspeech*.

Norbert Wiener. 1949. Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications.

Quanwei Wu, Jun Guo, Wei Wang, and Yi Wang. 2025. Robustness Feature Adapter for Efficient Adversarial Training. *ArXiv*, abs/2508.17680.

Jin Yong Yoo and Yanjun Qi. 2021. Towards Improving Adversarial Training of NLP Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 945–956, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stefan Ziegler, Bogdan Ludusan, and Guillaume Gravier. 2012. Using broad phonetic classes to guide search in automatic speech recognition. In *INTERSPEECH-Annual Conference of the International Speech Communication Association*.

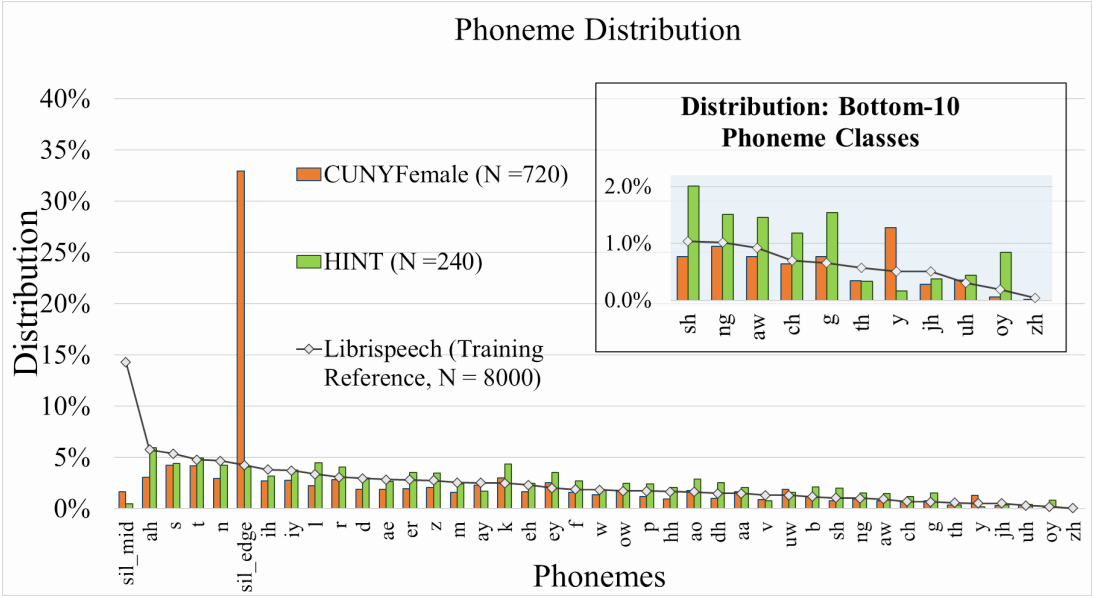


Figure A.1: Distribution of phoneme classes in the training and test speech datasets sorted by the frequency of phonemes in the training dataset. The silence class (*sil*) is divided into two categories: silences occurring at the beginning and end of an utterance, denoted by *sil_edge*; and silences occurring in the middle of an utterance, denoted by *sil_mid*. N is the number of speech files in each dataset.

A Appendix

A.1 Phoneme Distribution Across Speech Datasets

Figure A.1 illustrates the highly imbalanced phoneme class distributions in the training and test speech datasets, with classes sorted according to their frequency in the training set. The silence class (*sil*) is divided into two categories: silences occurring at the beginning and end of an utterance (*sil_edge*) and silences occurring within an utterance (*sil_mid*). In the training dataset, approximately 15% of the samples correspond to *sil_mid*, while the phoneme distribution is heavily concentrated around vowel classes. In contrast, the test datasets, particularly CUNY, contain more than 30% *sil_edge* samples. Such disparities introduce a significant class imbalance between training and testing conditions, making robust phoneme classification challenging. As a result, the classifier becomes disproportionately biased toward silence (non-phoneme) classes compared to actual phoneme classes.

A.2 Manner of Articulation (MOA)

Phoneme Groups. The phoneme classes consist of 39 standard American English phonemes (Ladefoged and Johnson, 2014) and an additional "non-phoneme" class to account for silent segments in clean speech. The MOA groups consist of 8 classes, with an additional silent class. MOA groups are summarized in Table A.1.

Manner of Articulation (MOA)	
Group	**Phonemes**
Vowel	aa, ae, ah, ao, aw, ay, eh, er, ey, ih, iy, ow, oy, uh, uw
Stop	b, d, g, k, p, t
Fricative	dh, f, s, sh, th, v, z, zh
Affricate	ch, jh
Nasal	m, n, ng
Liquid	l, r
Semivowel	w, y
Aspirate	hh
Silence	non-phoneme

Table A.1: Phoneme groupings by manner of articulation.

Results for MOA-based Mask Estimation in Noisy Reverberant Test Conditions. Table A.2 show the objective intelligibility scores for noisy-reverberant speech and speech enhanced using MOA-based masking techniques with $Base$ and Adv, AM models as classifiers.

A.3 Analysis of Framewise Phoneme Classification

Figure A.2 shows example framewise true and predicted phoneme probabilities for the baseline pre-

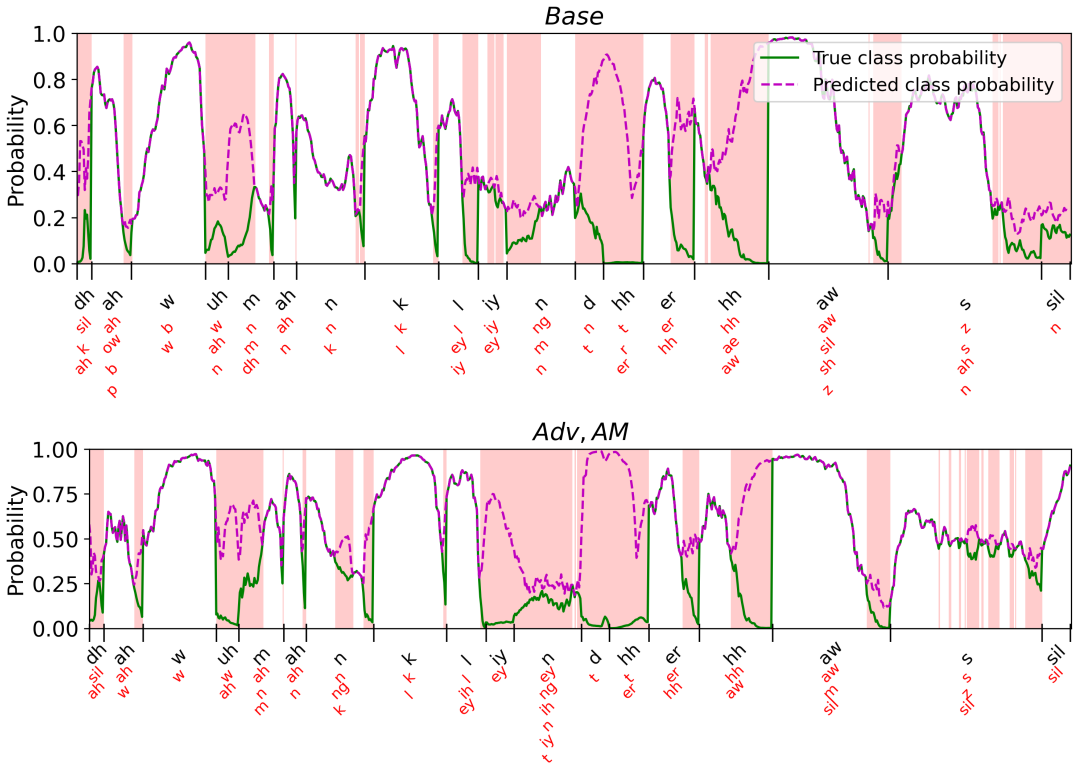


Figure A.2: Framewise probabilities of true and predicted phoneme classes for speech (from the HINT speech dataset) convolved with the Church$_1$ room impulse response. True class labels are shown in black, while unique predicted classes corresponding to each true label are shown in red. Results are presented for the baseline classifier ($Base$) trained on reverberant only data and the adapter-based adversarially trained classifier (Adv, AM).

trained classifier ($Base$) and the adapter-based adversarially trained classifier (Adv, AM). In general, Adv, AM training reduces the number of misclassifications compared to the $Base$ model. Classification errors primarily occur around phoneme transition boundaries and among phonemes in the same MOA group. Since phonemes within the same MOA group exhibit similar spectro-temporal characteristics, such confusions are expected to be less detrimental to downstream T-F mask estimation and speech enhancement performance.

A.4 Mask Estimation

An ideal ratio mask (IRM) is computed according to (Wiener, 1949):

$$M(t,f) = \left(\frac{|S(t,f)|^2}{|S(t,f)|^2 + |N(t,f)|^2} \right)^{0.5}, \quad (5)$$

$$\hat{S}(t,f) = M(t,f)X(t,f), \quad (6)$$

where $M(t,f)$ represents the IRM; $S(t,f)$ represents the clean signal; $N(t,f)$ represents the noise; $X(t,f)$ represents the degraded speech signal; and $\hat{S}(t,f)$ represents the enhanced signal after applying the mask. IRM values range from [0, 1].

The Omni-Expert (OE) model is a phoneme-based model for mask estimation (Saha et al., 2026c). The model architecture consists of a phoneme group classifier and a single expert model for mask estimation that is structurally identical to the phoneme independent model. Input features are transformed based on a phoneme group-specific transformation:

$$\mathbf{z}_n = \mathbf{a}_n \odot \mathbf{x} + \mathbf{b}_n, \quad (7)$$

where $\mathbf{z}_n$ is the transformed feature vector; $\mathbf{x}$ is the input feature vector; $\mathbf{a}_n$ and $\mathbf{b}_n$ are the scale and shifting factor vectors, respectively, for the phoneme group n; and $\odot$ is element-wise multiplication.

A composite mask is estimated accordingly:

$$\hat{M}(t,:) = \sum_n p_n(t) E(\mathbf{z}_n(t)), \quad (8)$$

where $p_n(t)$ is the framewise probability of phoneme group n at time frame index t; $\mathbf{z}_n(t)$ is the transformed feature for phoneme group n; and $E(\mathbf{z}_n(t))$ is the corresponding phoneme group specific mask estimated from the single expert E.

The phoneme group-specific transformation parameters $(\mathbf{a}_n, \mathbf{b}_n)$ are predicted by two separate multilayer perceptrons (MLPs) with a one-hot phoneme encoding as input. Both MLPs have an input size based on the number of phoneme group classes and an output size of 65 (matching the input feature size). The scale MLP uses ReLU activation, while the shift MLP uses LeakyReLU to introduce nonlinearities (Nair and Hinton, 2010). The transformation parameters are precomputed after training and stored for all phoneme group classes, effectively reducing the process during inference to a lookup table.

Model Training. Mask estimation models were trained to minimize the signal loss function:

$$L = \frac{1}{TF} \sum_{t,f} \left(\hat{M}(t,f)X(t,f) - M(t,f)X(t,f)\right)^2, \tag{9}$$

where $X(t,f)$ represents the magnitude spectrum of the distorted speech signal; $\hat{M}(t,f)$ represents the estimated mask; $M(t,f)$ represents the ideal mask; and T and F represent the number of time and frequency bins, respectively.

The expert network in the OE model was initialized using weights from a pretrained phoneme independent model. Models were trained using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 1e-3 and momentum coefficients of 0.9 and 0.999. Training batches consisted of 16 utterances segmented into 2-second chunks to accommodate memory constraints. Training was terminated when no improvement was observed for 10 consecutive epochs.

A.5 Mask Estimation Results Under Reverberant Only Conditions

We also investigated whether adapter-based adversarial adaptation with reverberant-only pretraining preserves performance in the original (i.e., matched) acoustic domain in pretraining. All models are pretrained using reverberant-only speech and evaluated under reverberant-only test conditions. Both frame-wise classification performance and downstream mask estimation results are analyzed.

Speech utterances used for testing were obtained from the Hearing In Noise Test (HINT) (Nilsson et al., 1994) and the City University of New York (CUNY) Female corpus (Boothroyd et al., 1985), which are widely used in audiology and CI listening studies. Reverberant test conditions were generated using binaural RIRs from the Aachen Impulse Response database (Jeub et al., 2009). The selected environments include office, meeting, lecture, booth, and stairway rooms with $\text{RT}_{60} < 1$ s, as well as highly reverberant church environments with RT_{60} values ranging from 4-9 s. Additional details of the simulated acoustic environments are provided in Table A.5.

Classification Performance. The class-balanced accuracy results are summarized in Table A.3. Adapter-based adversarial learning maintains, and in most cases improves, phoneme group classification performance when evaluated under reverberant conditions. These results indicate that the adversarial adaptation process improves robustness without degrading previously learned phoneme representations. Importantly, despite jointly optimizing both the pretrained backbone and adapter modules, the model does not exhibit noticeable catastrophic forgetting, which is a common concern in fine-tuning based adaptation methods. This suggests that adapter-based adversarial training generalizes effectively across acoustic conditions while preserving knowledge learned during pretraining.

Downstream Mask Estimation Performance. The speech enhancement results are summarized in Table A.4. The Adv, AM framework preserves the speech intelligibility gains achieved through adversarial adaptation even when evaluated under the same reverberant conditions used during pretraining. These results indicate that the improved phonetic robustness learned during adaptation does not compromise performance in the original domain and continues to provide effective guidance for downstream T-F mask estimation.

A.6 Room Impulse Response (RIR) Characteristics

Table A.5 lists characteristics of recorded RIRs of six rooms in the Aachen Impulse Response database (Jeub et al., 2009) used for testing. RIRs were selected from both right and left channels of an office, a lecture hall, a stairway, a meeting room, a studio booth and a highly reverberant church. Five different source-to-receiver distances were considered for the church. For the stairway and the church, RIRs were selected at an azimuth of 90 degrees, where the source and receiver are directly facing each other. RIRs were filtered using an anti-aliasing filter and then downsampled from 48 to 16 kHz before convolution with anechoic

Metric	Model	Signal-to-Noise Ratio of Noisy-Reverberant Speech						
		-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB	Rev Only
SRMR-CI	NoisyRev	0.499 (±0.003)	0.563 (±0.003)	0.617 (±0.003)	0.656 (±0.003)	0.680 (±0.003)	0.691 (±0.003)	0.699 (±0.001)
	$Base$	1.186 (±0.022)	1.383 (±0.020)	1.546 (±0.018)	1.620 (±0.018)	1.638 (±0.018)	1.633 (±0.018)	2.101 (±0.018)
	Adv, AM	1.368 (±0.018)	1.588 (±0.024)	1.752 (±0.027)	1.813 (±0.027)	1.816 (±0.025)	1.803 (±0.024)	2.086 (±0.023)
	Oracle	1.823 (±0.019)	1.914 (±0.019)	1.964 (±0.020)	1.973 (±0.020)	1.967 (±0.020)	1.964 (±0.020)	2.084 (±0.020)
	IRM	2.081 (±0.023)	2.186 (±0.022)	2.240 (±0.022)	2.261 (±0.022)	2.269 (±0.022)	2.274 (±0.022)	2.241 (±0.022)
STOI	NoisyRev	0.499 (±)	0.563 (±)	0.617 (±)	0.656 (±)	0.680 (±)	0.691 (±)	0.709 (±0.001)
	$Base$	0.534 (±0.005)	0.621 (±0.004)	0.688 (±0.004)	0.732 (±0.003)	0.758 (±0.003)	0.770 (±0.002)	0.777 (±0.001)
	Adv, AM	0.377 (±0.011)	0.490 (±0.010)	0.576 (±0.008)	0.631 (±0.007)	0.664 (±0.006)	0.679 (±0.005)	0.803 (±0.001)
	Oracle	0.669 (±0.004)	0.717 (±0.004)	0.758 (±0.003)	0.787 (±0.003)	0.805 (±0.003)	0.813 (±0.002)	0.819 (±0.001)
	IRM	0.959 (±0.001)	0.968 (±0.001)	0.970 (±0.001)	0.970 (±0.001)	0.969 (±0.001)	0.969 (±0.001)	0.969 (±0.000)

Table A.2: Mean ± 95% confidence intervals of objective speech intelligibility metrics: speech-to-reverberation modulation energy ratio for cochlear implant users (SRMR-CI) and short-time objective intelligibility (STOI) for noisy reverberant (NoisyRev) speech enhanced using the ideal ratio mask (IRM) and the Omni-Expert (OE) model with manner of articulation (MOA) group predictions. MOA classifiers include the baseline model ($Base$) trained on reverberant only-data and the adapter-based adversarial trained model (Adv, AM). The oracle denotes the OE model with known/oracle MOA information. Bold indicates highest performing non-oracle model.

speech stimuli. Reverberation times (RT_{60}s) were calculated using the Schroeder method (Schroeder, 1965) using the code provided by (Löllmann et al., 2010). The direct-to-reverberant ratios (DRRs) of the recorded RIRs were calculated using (Naylor and Gaubitch, 2010).

A.7 Statistical Analysis

Statistical analyses for speech enhancement performance (SRMR-CI) in the noisy-reverberant training condition were performed in Python using the `statsmodels` and `SciPy` packages.

A linear mixed-effects (LME) model was fit to account for the repeated-measures structure of the data. SRMR-CI scores were modeled as a function of *Speech Unit* (Phoneme vs. MOA), *Classifier* ($Base$ vs. Adv, AM), *SNR*, and their interactions. A random intercept for sample identity was included to account for repeated measurements obtained from the same underlying speech/noise/reverberation sample across methods and SNRs. The LME model was fit using maximum likelihood, and statistical significance of the fixed effects and their interactions was assessed from the fitted mixed-effects model.

Following an omnibus analysis, paired post-hoc comparisons were performed within each SNR and experimental condition. Because all methods were evaluated on the same 3,840 samples, comparisons were paired at the sample level. For each comparison, the mean paired difference, paired t-statistic, raw p-value, and Cohen's d_z were calculated. Confidence intervals for the paired differences were estimated using bootstrap resampling of complete samples, thereby preserving the within-sample pairing across methods. Multiple comparisons were controlled using the Holm correction for the post-hoc analysis. A comparison was considered statistically significant when the multiplicity-corrected p-value was below 0.05.

The post-hoc analysis showed a consistent advantage of the Adv, AM classifier over the $Base$

Class Balanced Accuracy (%)				
Rooms	**Phoneme**		**MOA**	
	Base	**Adv, AM**	**Base**	**Adv, AM**
RIRs <1 s				
office	43.7%	**46.2%**	57.3%	**61.2%**
lecture	39.1%	**40.6%**	51.1%	**54.9%**
stairway	42.3%	**45.1%**	53.8%	**59.8%**
meeting	43.7%	**46.7%**	59.9%	**64.2%**
booth	40.2%	**44.7%**	58.6%	**61.6%**
Highly Reverberant RIRs (4-9 s)				
Church_1	39.3%	**43.9%**	52.9%	**59.9%**
Church_3	34.0%	**37.8%**	42.4%	**49.2%**
Church_4	29.9%	**31.9%**	36.4%	**41.2%**
Church_5	26.2%	**26.3%**	32.3%	**34.6%**
Church_7	**22.9%**	21.7%	29.6%	**29.9%**
Aggregate	36.1%	**38.5%**	47.4%	**51.7%**

Table A.3: Frame-wise class balanced accuracy (%) under reverberant only training and testing conditions. Results are reported for phoneme and manner of articulation (MOA) classification using the baseline ($Base$) and our adapter-based adversarial (Adv, AM) models across different room environments.

	Rev	**Base**	**Adv,AM**	**Oracle**	**IRM**
Phoneme					
SRMR-CI	1.683 ± 0.008	2.316 ± 0.007	2.319 ± 0.007	2.382 ± 0.006	2.300 ± 0.005
STOI	0.749 ± 0.002	0.828 ± 0.001	0.827 ± 0.001	0.847 ± 0.001	0.971 ± 0.000
Manner of Articulation					
SRMR-CI	1.683 ± 0.008	2.261 ± 0.007	2.246 ± 0.008	2.372 ± 0.006	2.300 ± 0.005
STOI	0.749 ± 0.002	0.821 ± 0.001	0.819 ± 0.001	0.841 ± 0.001	0.971 ± 0.000

Table A.4: Objective intelligibility metrics: speech-to-reverberation modulation energy ratio for cochlear implant users (SRMR-CI) and short-time objective intelligibility (STOI) for reverberant only speech (Rev) enhanced using the ideal ratio mask (IRM) and the omni-expert (OE) model. Phoneme classifier is the adapter-based adversarial trained network (Adv, AM) and the baseline model ($Base$) while oracle denotes OE with known/oracle phoneme information. Reported results are aggregated across all 10 room impulse responses (RIRs) listed in Table A.5.

classifier for phoneme across all evaluated SNRs. The mean improvement ranged from $\Delta = 0.179$ to 0.191, with Cohen's d_z values of approximately 0.90-1.0 and all comparisons remaining significant after correction. Specifically, the mean differences were 0.190 at -5dB, 0.191 at 0dB, 0.184 at 5dB, 0.190 at 10dB, 0.190 at 15dB, and 0.179 at 20 dB. These results indicate that the benefit of the Adv, AM classifier is consistent across SNRs when the phoneme-based model is trained under noisy-reverberant conditions.

The speech unit representation comparison under noisy-reverberant training showed that the relative performance of MOA and phoneme depended on the classifier and SNR. With the Adv, AM classifier, phoneme yielded higher scores than MOA from 0 to 20 dB, with differences increasing from 0.020 at 0 dB to 0.161 at 20 dB. Overall, the results demonstrate a significant interaction between speech unit representation and classifier choice.

Room	Dimensions (L × W × H) (m)	Source–Receiver Distance (m)	RT_{60} (s)	DRR (dB)
Office	5.0 × 6.4 × 2.9	3.0	0.6 (right); 0.6 (left)	0.4 (right); 0.2 (left)
Lecture	10.8 × 10.9 × 3.15	5.56	0.9 (right); 0.8 (left)	-0.1 (right); 0.1 (left)
Stairway	7.0 × 5.2 × 2.9	3.0	0.9 (right); 1.0 (left)	1.6 (right); 1.9 (left)
Meeting	8.0 × 5.0 × 3.1	2.80	0.3 (right); 0.3 (left)	4.6 (right); 6.5 (left)
Booth	3.0 × 1.8 × 2.2	1.5	0.3 (right); 0.3 (left)	7.8 (right); 7.1 (left)
$Church_1$	19.0 × 30.0 × 9.0	1.0	4.4 (right); 4.7 (left)	9.4 (right); 7.8 (left)
$Church_3$	19.0 × 30.0 × 9.0	3.0	5.3 (right); 5.4 (left)	2.5 (right); 0.7 (left)
$Church_4$	19.0 × 30.0 × 9.0	5.0	6.5 (right); 6.6 (left)	-0.6 (right); -1.5 (left)
$Church_5$	19.0 × 30.0 × 9.0	10.0	8.4 (right); 8.5 (left)	-5.7 (right); -6.0 (left)
$Church_7$	19.0 × 30.0 × 9.0	20.0	7.0 (right); 7.0 (left)	-1.1 (right); -3.8 (left)

Table A.5: Room impulse response characteristics of test room conditions from Aachen Impulse Response (AIR) dataset. The recorded RIRs used in the testing set were binaural, recorded from right and left channels of a dummy head. RT_{60} (s): reverberation time; DRR: direct-to-reverberant ratio. $Church_{number}$ indicate different RIR number.