Leveraging Articulatory Phonetic Groupings for Robust Speech Enhancement for Cochlear Implants in Complex Acoustic Environments

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

Abstract

Cochlear implant (CI) users often experience reduced speech understanding in noisy and reverberant environments. Our recent work has shown the benefits of exploiting knowledge of speech content to drive speech enhancement for CI applications by estimating time-frequency masks based on predictions from a phoneme classifier. However, phonemes with similar temporal and spectral characteristics are likely to be confused, leading to misclassifications that negatively impact the downstream phoneme-based task. Also, machine learning models are generally vulnerable to performance degradation in unseen acoustic conditions. Using a lightweight and causal phoneme-based speech enhancement model developed for CI applications, we demonstrate the utility of alternative phoneme groups to improve model robustness in diverse acoustic settings. Results from objective measures of speech enhancement show that manner of articulation (MOA) groups provide a balance between fine-grained but error-prone phonemes and limited acoustic specificity with voicing groups. In particular, a counterintuitive finding is that MOA-based models trained in reverberant conditions are particularly robust under mismatched train-test conditions. Our findings highlight the importance of linguistically motivated phonetic grouping and selective training strategies for developing robust speech enhancement models for CI applications while maintaining lightweight model complexity for edge deployment.

1 Introduction

Cochlear implants (CIs) restore sound perception to individuals with moderate to severe sensorineural hearing loss by converting sound into a sequence of electrical pulses delivered via electrodes along the cochlea (Carlyon and Goehring, 2021). Although most CI users typically have good speech understanding in quiet listening conditions, they struggle to understand speech in complex acoustic environments that include background noise and reverberation (Hazrati and Loizou, 2012; Kokkinakis et al., 2011).

Front-end noise reduction strategies available in current CI devices, such as beamforming and signal-to-noise ratio (SNR)-based noise reduction (Buechner et al., 2014; Dawson et al., 2011; Henry et al., 2021), are mostly beneficial when noise and speech sources are spatially separated and the noise is stationary. However, solutions for noise reduction are not explicitly designed to address reverberation that results from delayed and attenuated reflections of speech in enclosed spaces (Naylor and Gaubitch, 2010). Unlike additive noise, reverberant reflections are similar to target speech, limiting the effectiveness of conventional noise suppression techniques. In real-world acoustic settings, noise and reverberation frequently co-occur, and solutions developed to address either acoustic distortion type in isolation are not necessarily effective in mitigating combined noise and reverberation.

A common approach for speech enhancement is time-frequency (T-F) masking, where a gain matrix (or mask) is applied to a T-F representation of the degraded speech to separate segments dominated by speech and acoustic distortions. An ideal mask is calculated based on a measure of the distortion of a degraded speech signal relative to its clean counterpart. Studies have shown ideal T-F masks for denoising and dereverberation improve speech intelligibility for hearing-impaired listeners, including CI users (Roman and Woodruff, 2013; Kokkinakis and Stohl, 2021; Healy et al., 2019; Koning et al., 2015; Wang, 2005; Hazrati et al., 2013). However, the clean signal is unavailable in practice, thus the T-F mask must be estimated from the degraded speech signal (Kokkinakis et al., 2011).

Research on speech enhancement models for CI users focuses on machine learning (ML) methods

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 4503–4529
November 6-10, 2026 ©2026 Association for Computational Linguistics

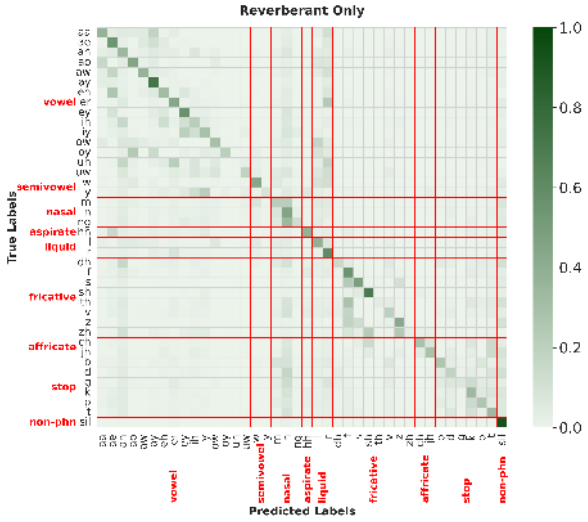


Figure 1: Confusion matrix of phoneme predictions with a phoneme classifier across reverberant room conditions in test datasets. Phonemes are annotated by manner of articulation groups.

for direct estimation of T-F masks from distorted speech signals (Carlyon and Goehring, 2021; Yuliani et al., 2021), though no ML-based speech enhancement models are currently deployed in CI devices. T-F mask estimation models that exploit predicted phoneme content have been shown to improve speech enhancement performance (Chazan et al., 2021, 2016; Wang et al., 2016; Chu et al., 2022). However, the performance of phoneme-based mask estimation is highly dependent on the performance of the phoneme recognition model as phonemes with similar temporal and spectral characteristics are prone to mutual confusion (Saha et al., 2026b).

The goal of phoneme-based speech enhancement is not necessarily to distinguish phoneme identity but to provide acoustically coherent adaptation for mask estimation. In articulatory phonetics, manner of articulation (MOA) and voicing grouping of phonemes reflect similarities in speech production mechanisms, which generally correspond to similarities in time-frequency energy distributions (Ladefoged and Johnson, 2014). As illustrated in Figure 1, phoneme prediction errors frequently occur within the same MOA group. Errors within the same MOA group are likely to be less detrimental for mask estimation because acoustically similar phonemes tend to require similar T-F masks for enhancement. This suggests that alternative phoneme groupings based on acoustic similarity may provide more robust conditioning signals for phoneme-based mask estimation while reducing sensitivity to phoneme classification errors.

In addition, ML models for speech enhancement are sensitive to shifts in data distribution between training and testing, particularly under unseen combinations of noise and reverberation encountered in real-world acoustic environments (Taori et al., 2020; Saha et al., 2026a). Understanding how alternative phoneme grouping strategies affect model robustness under matched and mismatched train/test acoustic conditions is therefore important to assess the effectiveness of phoneme-based mask estimation for CIs in real-word settings.

Our focus is not large-capacity general-purpose models for speech enhancement, but lightweight models for frame-by-frame speech enhancement targeted for CI deployed settings with low latency, memory, and power constraints. Unlike bidirectional speech enhancement models that can implicitly learn sequential phoneme structure from long-context representations, causal lightweight models benefit substantially from explicit phoneme conditioning. Therefore, we focus on comparisons with counterpart causal lightweight baselines with comparable model complexity.

In this work, we investigate how linguistically motivated phonetic representations can provide robust conditioning signals for speech enhancement under challenging acoustic conditions. Rather than treating phoneme grouping solely as an alternative classification scheme, we examine how different levels of phonetic abstraction influence the robustness of T-F mask estimation. In particular, we show that MOA groupings provide an effective balance between phoneme-level specificity and robustness to phoneme classification errors, improving speech enhancement performance without increasing model complexity. We further evaluate robustness under matched and mismatched train/test acoustic conditions and show that MOA-based models provide the most robustness to distribution shifts than phoneme- and VUV-based alternatives.

2 Related Work

Research on speech enhancement model in CIs now utilizes deep neural networks (DNNs) that offer improved robustness over linear models in noisy and reverberant environments (Borjigin et al., 2024; Huang et al., 2024; Mamun and Hansen, 2024; Gajecki et al., 2023; Gaultier and Goehring,

2024). Phoneme-based mask estimation models, where separate models are trained for different phonemes, have shown improved performance in speech denoising for automated speech processing applications (Wang et al., 2016; Chazan et al., 2016). Some studies have used a mixture-of-experts (MoE) architecture for phoneme-based speech enhancement (Chazan et al., 2021; Chu et al., 2022), consisting of multiple experts to estimate phoneme-specific masks that are probabilistically combined based on predictions from a phoneme classifier. Chu et al. (2022) developed a MoE model for speech dereverberation in CIs that relies exclusively on causal features extracted from within the CI signal processing pipeline, thus real-time feasible.

However, MoE models pose challenges when considering deployment on resource-constrained edge devices like CIs. Due to the relatively high error rates with framewise phoneme classification, Chu et al. (2022) used a dense MoE for phoneme-based mask estimation; thus, all experts are activated during inference. We recently introduced the Omni-Expert (OE) model (Saha et al., 2026c) as a computationally efficient alternative to the dense MoE model in (Chu et al., 2022). In the OE model, phoneme-specific transformations are applied in parallel to input features to achieve phoneme-specific expert specialization for mask estimation within a single expert network.

The performance of phoneme-based speech enhancement is inherently tied to the accuracy of phoneme classification. Prior work in automatic speech recognition has explored reducing phoneme confusion through phoneme grouping strategies based on contextual information (Lu et al., 2023), decision-tree clustering (Ziegler et al., 2012) and mutual information driven feature selection for predetermined phoneme groups (Scanlon et al., 2007). These prior works suggest further investigation into how phoneme grouping affects the performance of speech enhancement models tailored for CIs under adverse acoustic conditions. However, the impact of phoneme grouping on downstream speech enhancement robustness, particularly under noisy and reverberant train-test mismatch relevant to CI applications, remains largely unexplored.

Unlike prior work that conditions speech enhancement on individual phoneme identities, we investigate whether linguistically motivated phoneme groupings can improve robustness of lightweight phoneme-conditioned enhancement models under adverse acoustic conditions. In this work, we conduct a sensitivity analysis to explore how model performance and robustness varies with alternative phoneme grouping strategies under matched and mismatched train-test acoustic conditions.

3 Methodology

Models were implemented in PyTorch and trained on a NVIDIA A5000 GPU.

3.1 Datasets

Training. Speech utterances were a randomly selected 8000-utterance subset (≈28 hours) of the 100-hour LibriSpeech corpus (Panayotov et al., 2015). Noise signals used for training were selected from the WSJ0 Hipster Ambient Mixtures (WHAM!) (Wichern et al., 2019) and the Bern Cocktail Party (Fischer et al., 2020) datasets. The WHAM! dataset includes noise audio collected from restaurants, cafes, bars and parks in urban locations. Bern Cocktail Party audio were recorded on microphones mounted on and around a CI processor dummy on the ear of a human or head simulator. The train partition of the Bern Cocktail Party dataset includes babble noise with 2, 4, 6, 8, 10, 15, 20, and 30 speakers. Noise was added to speech utterances from the LibriSpeech corpus at SNR levels of -5, 0, 5, 10, 15, and 20 dB. Reverberant and noisy reverberant speech signals were obtained by convolving anechoic and noisy speech, respectively, with recorded room impulse responses (RIRs) from the Brno University of Technology@FIT Reverb Database (Szöke et al., 2018).

Testing. Speech utterances used for testing were from speech datasets that are widely used in audiology research and CI listening studies: Hearing In Noise Test (HINT) (Nilsson et al., 1994) and the City University of New York (CUNY) (Boothroyd et al., 1985) Female datasets. HINT consists of phonetically balanced sentences with controlled vocabulary designed for reliable intelligibility assessment in noise, while the CUNY corpus contains meaningful conversational sentences from both male and female speakers. Real-world noise and reverberation conditions representative of everyday listening environments were simulated using recorded binaural RIRs and environmental noise datasets. Three different test acoustic conditions were developed:

- *Reverberant only:* Left and right channel recorded RIRs were selected from diverse rooms in the Aachen Impulse Response (AIR) dataset (Jeub

et al., 2009), which include booth, office, lecture, stairway, meeting rooms with reverberation times (RT_{60}) < 1s, and highly reverberant churches with RT_{60} ranging from 4-9s. Additional details of test room characteristics are provided in Appendix A1. The reverberant-only test condition explicitly evaluates performance in the absence of additive noise.

- *Noisy only:* Noisy speech signals were generated by adding noise from the DEMAND dataset (Thiemann et al., 2013) and two-talker babble generated from the TIMIT dataset (Garofolo et al., 1993) to clean speech. From the DEMAND dataset, public noise was selected, which includes the interiors of a cafeteria, restaurant and a busy subway station. Two-talker babble was generated from the TIMIT test partition by randomly selecting two distinct speakers without replacement for each mixture. Utterances from each speaker were concatenated and looped to form 60s streams, which were summed and amplitude-normalized. A total of 200 unique babble recordings were generated from distinct speaker pairs among the approximately 168 test speakers. Noise was added at SNR levels of -5, 0, 5, 10, 15, 20 dB.

- *Noisy reverberant speech:* Noisy speech was convolved with left and right channels recorded RIRs from diverse rooms (stairway and church) from the AIR database.

3.2 Feature Extraction

T-F features were extracted from speech signals following the Advanced Combination Encoder (ACE) strategy (Vandali et al., 2000) implemented in the Nucleus Matlab CI Toolbox (Swanson and Mauch, 2006). The acoustic signal is segmented into 8 ms frames with a 2 ms overlap and processed via short-time Fourier transform, yielding 65 frequency bins. The power spectrum was computed and then log-compressed to reduce dynamic range. Feature normalization was applied using the global mean and variance calculated from the training set. The normalized log-compressed spectral features were used as inputs to both the phoneme classification and mask estimation models.

3.3 Model Architecture

For real-time feasibility in CIs, we use a lightweight recurrent neural network developed for CI applications (Saha et al., 2026c). The base architecture consists of a single layer of 117 unidirectional gated recurrent units (GRU) with a multi-head attention (A) layer. The outputs of the GRU network following layer normalization are fed into the multi-head attention module, a residual connection around the attention block, followed by layer normalization. The attention outputs are fused with the original GRU hidden state outputs via element-wise multiplication so that the final representation retains both the sequential encoding and the learned contextual weighting. Model complexity is detailed in Table A2.

3.4 Gating Network for Phoneme Group Classification

The GRU+A model is followed by a fully connected layer for phoneme group classification.

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 and voiced/unvoiced (VUV) groups consist of 8 and 2 classes, respectively, with an additional silent class. MOA and VUV phoneme groups are summarized in Table 1.

MOA groupings reflect similarities in speech production mechanisms that often correspond to shared spectro-temporal acoustic energy patterns. For speech enhancement, MOA groupings may provide acoustically coherent conditioning signals while reducing sensitivity to phoneme classification errors. In contrast, coarse VUV grouping may not provide sufficient acoustic structure to inform mask estimation.

Model Training. Phoneme labels were generated from the direct-path signal using forced alignment based on the LibriSpeech recipe in the Kaldi Speech Recognition Toolkit (Povey et al., 2011), and mapped into corresponding MOA/VUV-based groups. Phoneme time stamps were converted into CI-based time bins to match the temporal resolution of the mask estimation model. A one-hot vector was used to encode the phoneme label.

Model parameters were randomly initialized from a uniform distribution in the range [-0.1,0.1]. Training was performed using stochastic gradient descent with a learning rate of 1e-5 and momentum of 0.9. Each training batch consisted of a single utterance split into 2-second non-overlapping segments to accommodate memory constraints and maintain sequence continuity. Training data were partitioned into disjoint training and validation sets. The model was optimized using cross-entropy loss

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

Table 1: Phoneme groupings by manner of articulation and voicing.

and training terminated when the validation loss was unchanged for 10 consecutive epochs.

3.5 Mask Estimation

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

$$M(t,f) = \left(\frac{|S(t,f)|^2}{|S(t,f)|^2 + |N(t,f)|^2} \right)^{0.5}, \quad (1)$$

$$\widehat{S}(t,f) = M(t,f)X(t,f), \quad (2)$$

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

We trained two different types of models:

- ***Phoneme Independent (PI) model***, the conventional method for mask estimation directly from input features. The base model architecture is followed by a fully connected layer and an output layer with 65 sigmoidal units, representing the 65-dimensional IRM values.

- ***Omni-Expert (OE) model***, 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:

$$\mathbf{z}_n = \mathbf{a}_n \odot \mathbf{x} + \mathbf{b}_n, \quad (3)$$

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:

$$\hat{M}(t,:) = \sum_n p_n(t) E(\mathbf{z}_n(t)), \quad (4)$$

where $p_n(t)$ is the 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. Unlike the phoneme-based model, which learns separate shift-scale conditioning parameters for each phoneme class, the MOA/VUV-based model learns conditioning parameters shared across phonemes belonging to the same group.

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

$$L = \frac{1}{TF} \sum_{t,f} \Big(\hat{M}(t,f)X(t,f) - M(t,f)X(t,f) \Big)^2 \quad (5)$$

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 PI model was initialized using weights drawn from a uniform (-0.1,0.1) distribution. The

expert network in the OE model was initialized using pre-trained weights from the PI 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.

3.6 Performance Evaluation

We evaluated CI vocoded speech using objective metrics that are correlated with speech evaluation measures in CI users (Falk et al., 2015):

Speech-to-reverberation modulation energy ratio for CI users (SRMR-CI) (Santos and Falk, 2014), developed for reverberant speech and adapted for CI users, thus provides a more reliable speech intelligibility predictor for CI users in reverberation.

Short-time objective intelligibility (STOI) (Taal et al., 2011), with the direct path signal used as the reference signal. STOI ranges between [0, 1]. STOI was originally developed for speech in additive noise and does not evaluate silent regions of the reference signal. Late reverberation occurs in speech gaps, and therefore improvements in these regions are not captured by STOI scores. Moreover, as a bounded metric, STOI provides less margin for improvement under challenging SNR conditions. In particular, at low SNRs, where the baseline system already exhibits substantial degradation, even modest STOI gains can indicate improved preservation of speech-relevant information.

Frequency-weighted segmental speech-to-distortion ratio (fwSDR) (Ma et al., 2009), a speech quality metric that measures how close a target signal is to the reference clean speech while emphasizing perceptually important frequency regions. fwSDR ranges between [-10, 35] dB and higher values indicate better speech quality.

Cepstral distance (CD) (Kitawaki et al., 1988), which quantifies the difference between two speech spectra using their cepstral coefficients, typically delivered from linear predictive coding analysis. CD ranges between [0, 10] dB and lower values indicate better speech quality.

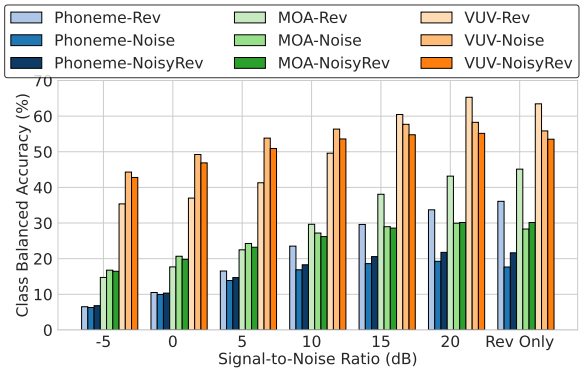


Figure 2: Framewise phoneme/manner of articulation (MOA)/ voiced-unvoiced (VUV) classification performance in noisy reverberant conditions. Models are trained using reverberant-only (Rev), noise-only (Noise), and noisy reverberant (NoisyRev) conditions.

4 Results

4.1 Phoneme Group Classification

Figure 2 shows performance of classification models for each phoneme group scheme trained in different acoustic conditions and tested on noisy reverberant speech to evaluate sensitivity to train-test mismatch. For phoneme classification, models trained exclusively on reverberant data consistently outperformed those trained on noise-containing data across SNR levels. In contrast, for MOA and VUV classification, models trained on noise-containing data performed better at lower SNRs while those trained on reverberant-only data performed better at higher SNRs, with a wider positive SNR range for MOA. These results indicate that the impact of train-test mismatch on classification performance in noisy reverberant conditions depends on the phoneme grouping scheme.

One possible explanation is that reverberant-only training encourages the classifier to learn phonetic representations that rely less on fine temporal details and more on stable articulatory-acoustic characteristics. Such characteristics may be preserved more reliably in the presence of additive noise, allowing the resulting representations to generalize to unseen noisy-reverberant conditions. In contrast, noise-containing training may encourage representations that are more closely tied to the characteristics of the training noise distribution. For MOA-based grouping, pooling acoustically related phonemes within the same articulatory category may further reduce the impact of phoneme-level classification errors induced by noise.

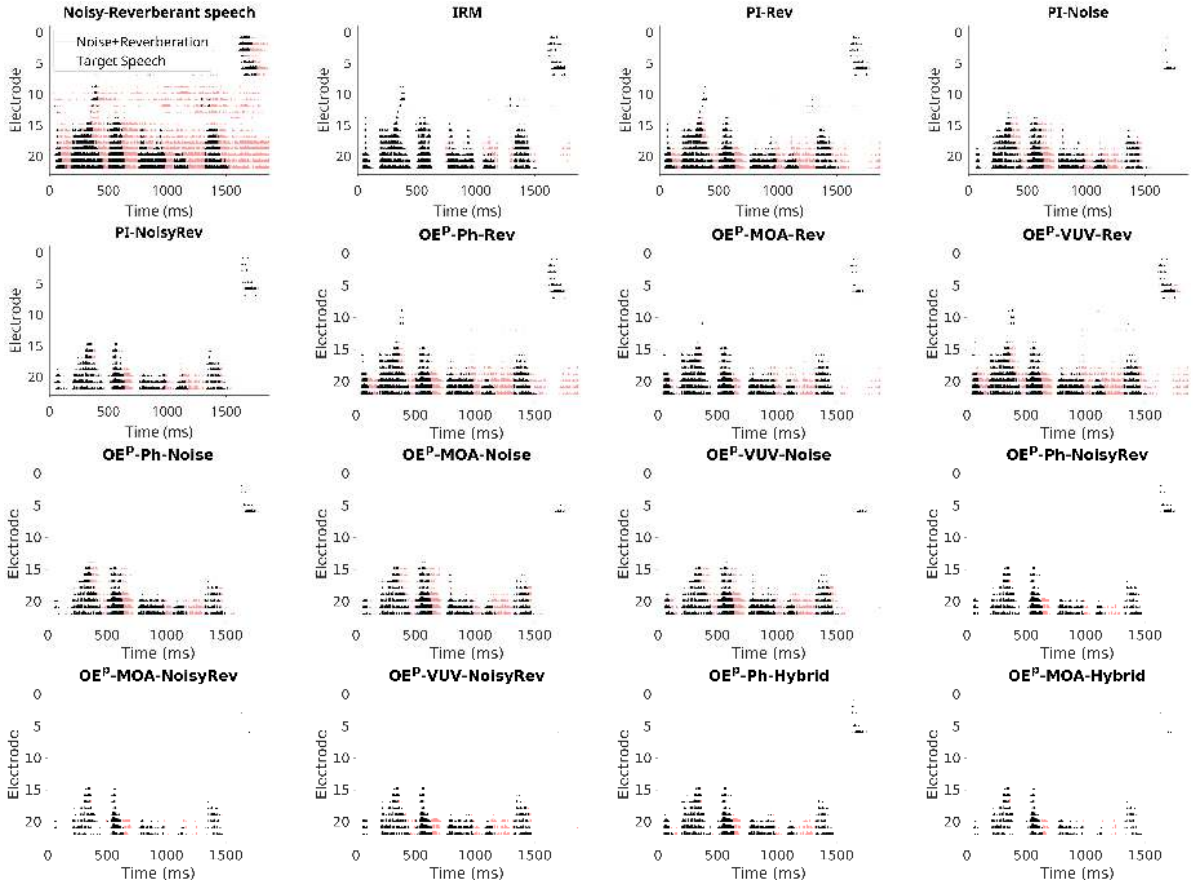


Figure 3: Electrodograms of the speech utterance *"the boy broke the wooden fence"* (HINT dataset) with noise (public noise from DEMAND dataset) added at a 5 dB signal-to-noise ratio (SNR) and convolved with a stairway room impulse response (Aachen Impulse Response dataset). Visualizations of enhanced noisy-reverberant signals with the ideal ratio mask (IRM) and estimated masks using phoneme independent (PI) models and Omni-Expert (OE) models with predicted phoneme (Ph), manner of articulation (MOA) and voiced/unvoiced (VUV) groups. Models are trained using reverberant-only (Rev), noise-only (Noise) and noisy reverberant (NoisyRev) conditions. Additionally, two hybrid training conditions of Ph and MOA based OE models with phoneme classifier trained using Rev and expert network for mask estimation trained using NoisyRev are shown.

4.2 Mask Estimation

We evaluated PI and OE models with predicted phoneme-, MOA-, and VUV-specific weights, as well as corresponding oracle OE models with ideal phoneme group classification.

Figure 3 shows example electrodograms illustrating electrical stimuli delivered to CI electrodes, analogous to acoustic spectrograms. The electrodograms are labeled with target speech and acoustic artifacts (noise and reverberation), and include an ideal mask reference to assess how applied estimated masks preserve speech while suppressing distortions. Sample electrodograms are generated using a stairway RIR, a HINT speech utterance and public noise at 5 dB SNR. Results of mean signal loss across frequency bins are provided in Figure A5.1.

Results of noisy reverberant test conditions across various SNR levels are shown in Figure 4; results for models trained and evaluated on reverberant speech and oracle conditions of the OE models across SNR levels are provided in Figure A7.1 and Figure A8.1, respectively. For SRMR-CI, MOA-based OE models achieve the highest scores amongst models trained on reverberant-only or noise-only data. In contrast, when models are trained on noisy-reverberant data, phoneme-based groupings yield the best SRMR-CI performance. STOI scores show minimal differentiation across phoneme- and MOA-based OE models in all training conditions, consistent with the known limitation of STOI not capturing reverberation-related distortions in silent speech gaps. Despite achieving the highest classification accuracy, VUV-based OE models perform comparably or lower than the PI baseline. Statistical test results are detailed in

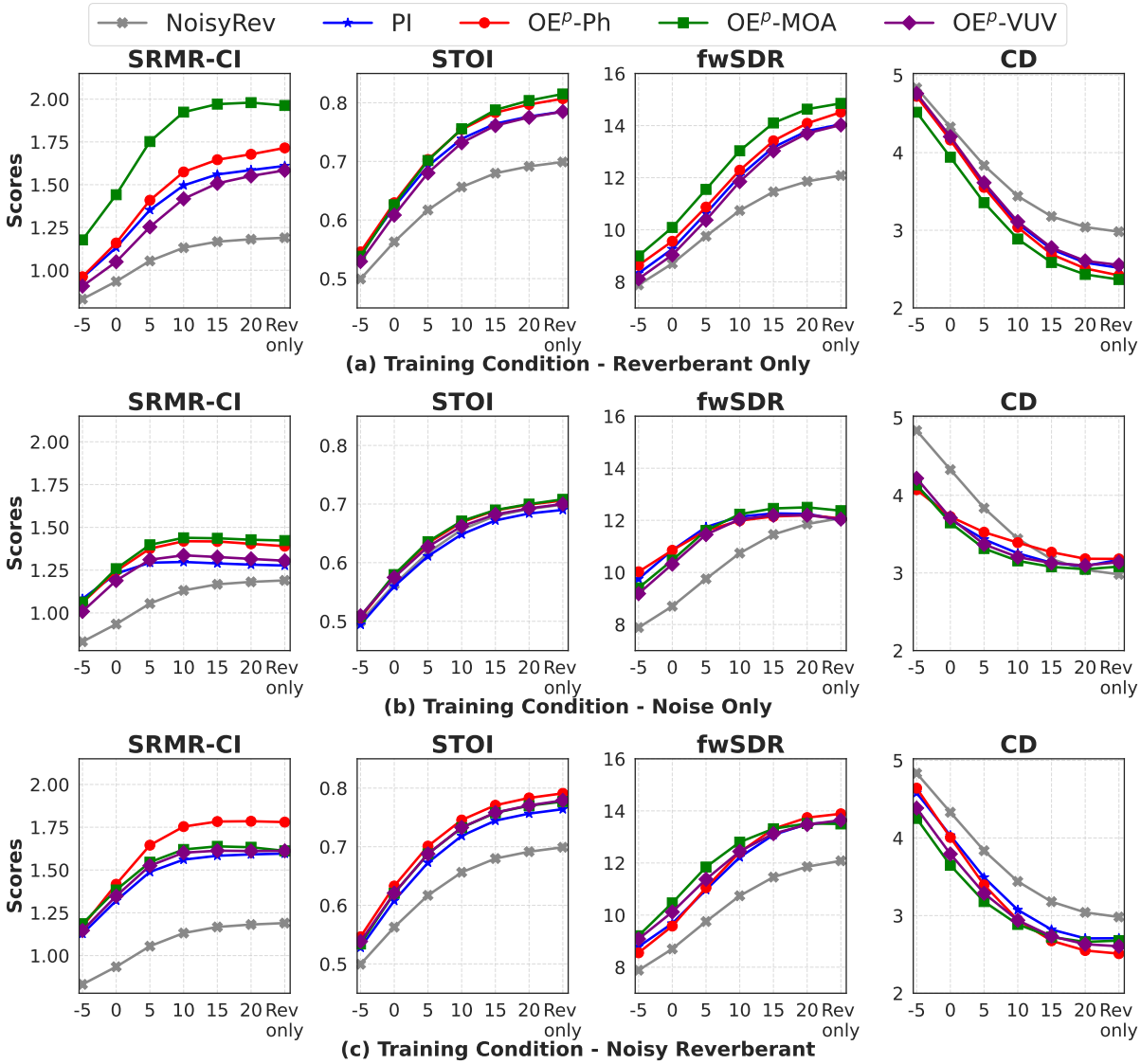


Figure 4: Impact of Acoustic Training Condition. Objective speech evaluation measures of speech-to-reverberation modulation energy ratio for cochlear implant users (SRMR-CI), short-time objective intelligibility (STOI) scores, frequency-weighted segmental speech-to-distortion ratio (fwSDR) and cepstral distance (CD) for noisy reverberant (NoisyRev) speech enhanced using phoneme independent (PI) model and Omni-expert (OE^p) models trained on **(a)** reverberant only, **(b)** noise only, and **(c)** noisy-reverberant conditions across varied reverberant rooms and noise types, aggregated across test speech datasets. Results are shown for OE models with phoneme, manner of articulation (MOA) and voiced/unvoiced (VUV) groups predicted using the gating network. Lower CD scores indicate better performance.

Appendix A.9.

For the fwSDR and CD metrics, models trained under reverberant-only data show trends that closely align with those observed for SRMR-CI and STOI. However, models trained on noise-only data exhibit relatively similar performance across groupings, with phoneme-based models performing slightly better at lower SNRs and MOA-based models showing a slight advantage at higher SNRs. Under noisy-reverberant training conditions, MOA-based models generally outperform phoneme-based models while phoneme-based models perform better at higher SNRs.

4.3 Training Data Optimization

We consider *conventional* training as using the same training data for both the gating and expert networks in the OE model. We further investigated a *hybrid* training strategy that independently selects the best performing training condition for each component in the OE model: the phoneme/MOA classifier trained on reverberant-only data (Fig-

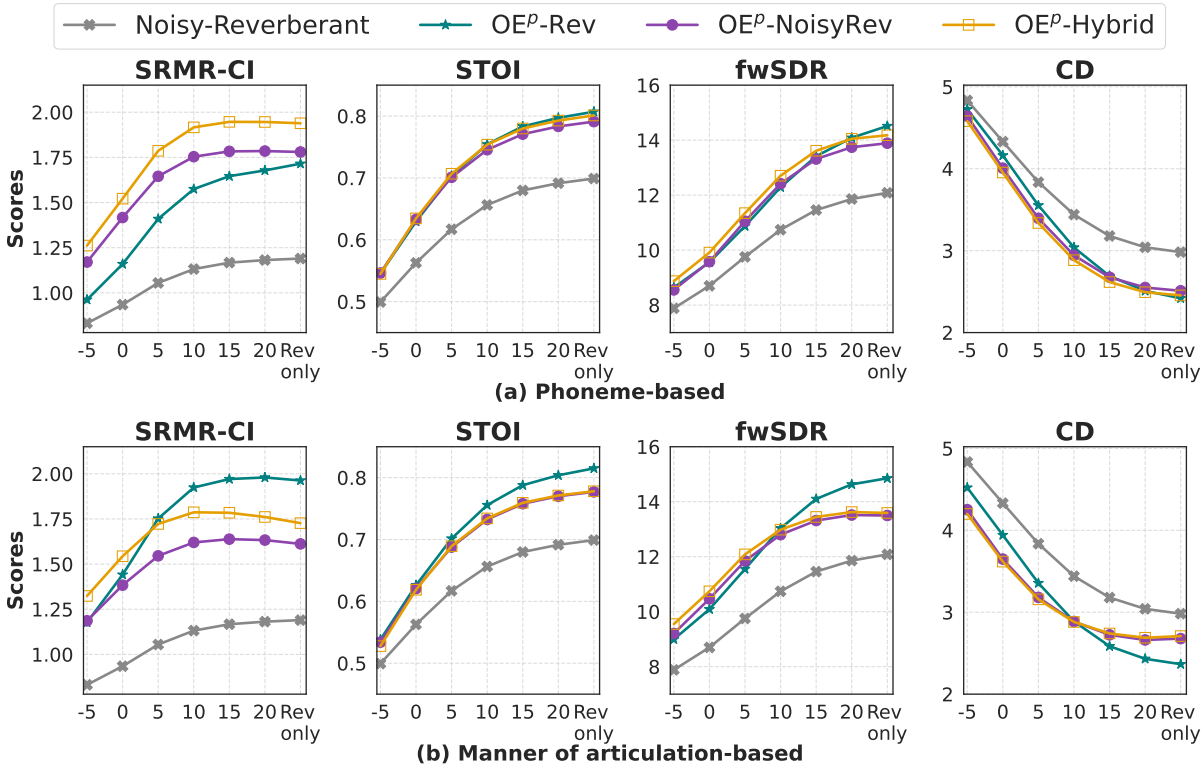


Figure 5: Impact of Hybrid Training in Omni-Expert (OE) Models. Objective speech evaluation measures of speech-to-reverberation modulation energy ratio for cochlear implant users (SRMR-CI), short-time objective intelligibility (STOI) scores, frequency-weighted segmental speech-to-distortion ratio (fwSDR) and cepstral distance (CD) for noisy reverberant (NoisyRev) speech enhanced using OE models with predicted phoneme groups (OE^p) trained conventionally on Rev and NoisyRev conditions, and hybrid training where mask estimation models are trained using NoisyRev conditions and gating networks are trained using Rev conditions. Results are shown for models with **(a)** phoneme and **(b)** manner of articulation (MOA) groups predicted using the gating network. For CD scores, lower values indicate better performance.

ure 2) and oracle mask estimation expert networks trained on noisy-reverberant data (Figure A8.1).

Figure 5 shows the performance of the different model configurations with conventional and hybrid training strategies. Phoneme- and MOA-based models with hybrid training outperform their counterpart models trained conventionally on noisy reverberant data, demonstrating that higher phoneme-group classification performance improves downstream speech enhancement. Notably, the phoneme-based model with hybrid training achieves comparable performance as the highest performing model with conventional training, i.e., the MOA-based model trained on reverberant only data, at higher SNR levels. Both phoneme- and MOA-based models with hybrid training maintain a performance edge at lower SNR levels. These results highlight the benefits of selectively introducing train-test mismatch to optimize training conditions in the different components of the OE model.

5 Conclusion

We investigated the impact of different phoneme grouping schemes and training strategies on speech enhancement models for CI users in noisy reverberant environments. Overall, our findings highlight the value of incorporating alternative phoneme grouping with selective train-test domain mismatch as viable data-centric strategies to improve speech enhancement models without increasing model complexity. These insights provide practical guidance for designing lightweight, real-time feasible and robust speech enhancement models for CI users in complex acoustic environments.

Limitations

MOA and VUV phoneme groups are based on vocal linguistics and not necessarily optimized for the downstream machine learning task of T-F mask estimation. Exploring learned phoneme grouping strategies may further enhance model performance

and robustness across diverse acoustic conditions.

Our speech enhancement evaluation was conducted using simulated CI vocoded speech and objective intelligibility metrics. Although this evaluation protocol is standard in offline CI speech enhancement research, the proposed approach needs to be validated in human listening studies to provide a more comprehensive assessment of the impact of model performance on speech perception in CI users.

Results obtained from English speech may not necessarily generalize directly to other languages with different phonetic inventories, phonological structures, or acoustic characteristics. Future work will investigate generalization across languages, speakers, acoustic environments, and real-world CI deployment conditions.

Societal Impact

This work aims to improve speech enhancement for cochlear implant users in complex acoustic environments, which can contribute to improved speech understanding and listening comfort in daily communication settings. A robust lightweight speech enhancement model may also support deployment in resource-constrained assistive hearing devices with strict latency and power requirements.

Beyond CI applications, the work connects to broader computational linguistics areas in phonetics, spoken language understanding, and linguistically informed representations for speech models. In particular, we explored phonetic categories derived from speech production as linguistically informed representations for improving robustness under noisy and degraded acoustic conditions. This perspective highlights the broader potential of incorporating phonetic and linguistic structure into spoken language technologies operating in challenging real-world environments.

Acknowledgments

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

References

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, Joshua S. Stohl, 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).

Andreas Buechner, Karl-Heinz Dyballa, Phillipp Hehrmann, Stefan Fredelake, and Thomas Lenarz. 2014. Advanced Beamformers for Cochlear Implant Users: Acute Measurement of Speech Perception in Challenging Listening Conditions. *PLOS ONE*, 9(4).

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.

Chazan, Shlomo E., S. Gannot, and J. Goldberger. 2016. A phoneme-based pre-training approach for deep neural network with application to speech enhancement. *International Workshop on Acoustic Signal Enhancement*.

Shlomo E Chazan, Jacob Goldberger, and Sharon Gannot. 2021. Speech Enhancement with Mixture of Deep Experts with Clean Clustering Pre-Training. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 716–720. IEEE.

Kevin Chu, Leslie Collins, and Boyla Mainsah. 2022. Suppressing reverberation in cochlear implant stimulus patterns using time-frequency masks based on phoneme groups. In *Proceedings of Meetings on Acoustics*, volume 50. Acoustical Society of America.

Pam W Dawson, Stefan J Mauger, and Adam A Hersbach. 2011. Clinical Evaluation of Signal-to-Noise Ratio–Based Noise Reduction in Nucleus® Cochlear Implant Recipients. *Ear and Hearing*, 32(3):382–390.

Tiago H Falk, Vijay Parsa, Joao F Santos, Kathryn Arehart, Oldooz Hazrati, Rainer Huber, James M Kates, and Susan Scollie. 2015. Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and limitations of existing tools. *IEEE Signal Processing Magazine*, 32(2):114–124.

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.

Tom Gajecki, Yichi Zhang, and Waldo Nogueira. 2023. A Deep Denoising Sound Coding Strategy for Cochlear Implants. *IEEE Transactions on Biomedical Engineering*, 70(9):2700–2709.

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.

Clément Gaultier and Tobias Goehring. 2024. Recovering speech intelligibility with deep learning and multiple microphones in noisy-reverberant situations for people using cochlear implants. *The Journal of the Acoustical Society of America*, 155(6):3833–3847.

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).

Fergal Henry, Martin Glavin, and Edward Jones. 2021. Noise Reduction in Cochlear Implant Signal Processing: A Review and Recent Developments. *IEEE Reviews in Biomedical Engineering*, 16:319–331.

Enoch Hsin-Ho Huang, Rong Chao, Yu Tsao, and Chao-Min Wu. 2024. ElectrodeNet - A Deep-Learning-Based Sound Coding Strategy for Cochlear Implants. *IEEE Transactions on Cognitive and Developmental Systems*, 16(1):346–357.

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*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Nobuhiko Kitawaki, Hiromi Nagabuchi, and Kenzo Itoh. 1988. Objective quality evaluation for low-bit-rate speech coding systems. *IEEE Journal on Selected Areas in Communications*, 6(2):242–248.

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.

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.

Jianfen Ma, Yi Hu, and Philipos C. Loizou. 2009. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *The Journal of the Acoustical Society of America*, 125(5):3387–3405.

Nursadul Mamun and John H L Hansen. 2024. Speech Enhancement for Cochlear Implant Recipients Using Deep Complex Convolution Transformer With Frequency Transformation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

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.

M. Nilsson, SD. Soli, and JA. 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)*.

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*.

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. Parameter-Efficient Domain Adversarial Learning for Lightweight Real-Time Phoneme Classification. 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.

João F Santos and Tiago H Falk. 2014. Updating the SRMR-CI Metric for Improved Intelligibility Prediction for Cochlear Implant Users. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12).

Patricia Scanlon, Daniel PW 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.

Vladislav Sovrasov. 2018. ptflops: A flops counting tool for neural networks in Pytorch framework.

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

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).

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2011. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136.

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.

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.

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

Zhong-Qiu Wang, Yan Zhao, and DeLiang Wang. 2016. Phoneme-specific speech separation. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

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.

Asri Yuliani, Mohd Amri, Endang Suryawati, Ade Ramdan, and Hilman Pardede. 2021. Speech Enhancement Using Deep Learning Methods: A Review. *Jurnal Elektronika dan Telekomunikasi*, 21:19.

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*.

A Appendix

A.1 Room Impulse Response (RIR) Characteristics

Table A1 lists characteristics of recorded RIRs of six rooms in the Aachen Impulse Response database (Jeub et al., 2009) used for testing. Recorded RIRs were selected for both right and left channels on a dummy head placed in 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 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).

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 A1: Characteristics of test room conditions from the Aachen Impulse Response (AIR) dataset. The recorded room impulse responses (RIRs) used were binaural, recorded from right and left channels of a dummy head. DRR, direct-to-reverberant ratio; RT_{60} (s), reverberation time; DRR: direct-to-reverberant ratio. $Church_{number}$ indicates a different RIR number.

A.2 Model Complexity

Table A2 summarizes complexity (number of model parameters, model size, and multiply–accumulate operations (MACs)) of the Omni-Expert model across different phoneme grouping schemes. The parameter counts and computational load were obtained using the open-source *ptflops* package (Sovrasov, 2018).

A.3 Classifier

Frame-wise class-balanced classification accuracies of phoneme groups for reverberant-only room conditions are shown in Figure A3.1.

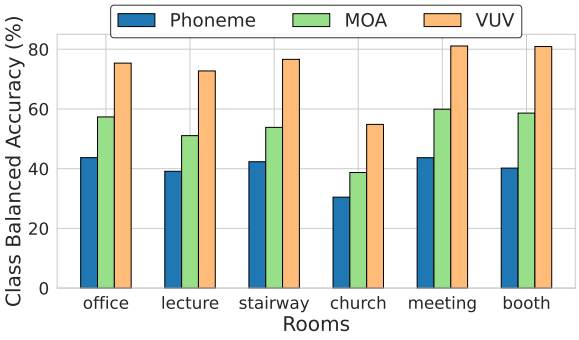


Figure A3.1: Framewise phoneme/manner of articulation (MOA)/voice and unvoiced (VUV) classification performance in reverberant room conditions. Models are trained using reverberant-only data.

Figure A3.2 and Figure A3.3 show framewise phoneme classifier predictions using all the phoneme grouping strategies applied to sample reverberant speech utterances from HINT and CUNYFemale datasets, respectively, in an office room condition. A reduction of misclassifications can be seen after using the MOA phoneme groups, with classes most likely to be misclassified occurring at the transitions..

A.4 Feature Visualization

Figure A4.1 and Figure A4.2 demonstrate the impact of phoneme grouping strategies in the t-SNE feature space across reverberant-only, noise-only and noisy reverberant training conditions with noise added at -5, 5 and 15 dB signal-to-noise (SNR).

A.5 Signal Loss

Figure A5.1 show the mean signal loss across frequency bins to illustrate the frequency-dependent effects of mask estimation for models trained under reverberant-only, noise-only and noisy reverberant conditions.

A.6 Additonal Electrodograms

Figure A6.1 and Figure A6.2 show electrodograms for reverberant-only speech generated using stairway and church RIRs, respectively, convolved with a HINT speech utterance.

Figure A6.3 show electrodograms for noisy reverberant speech generated using stairway RIR, HINT speech utterance and public noise at 5 dB. These electrodograms are for oracle conditions (predicted conditions in Figure 3).

Figure A6.4 show electrodograms for noisy reverberant speech generated using church RIR, HINT speech utterance and two-talker babble noise at 5 dB.

Number of Parameters		
Group	**Classifier**	**Mask Estimator (Expert)**
Phoneme	124,996	133,276
Manner of Articulation	121,338	129,246
Voiced and Unvoiced	120,630	128,466

(a) Number of parameters

MACs (M)		
Group	**Classifier**	**Mask Estimator (Expert)**
Phoneme	359.07	362.0
Manner of Articulation	355.45	362.0
Voiced and Unvoiced	354.74	362.0

(b) Multiply-accumulate operations (MACs, in millions)

Size (MB)		
Group	**Classifier**	**Mask Estimator (Expert)**
Phoneme	0.5	0.53
Manner of Articulation	0.49	0.52
Voiced and Unvoiced	0.48	0.51

(c) Model size, in MB

Table A2: Complexity of Omni-Expert model components across different phoneme grouping schemes.

A.7 Room-wise Mask Estimation Results - Reverberation Only Test Conditions

Table A7.1, Table A7.2, Table A7.3 and Table A7.4 show the room-wise objective intelligibility measures for reverberant only test conditions. The aggregated scores across all speech datasets and reverberant room conditions are shown in Figure A7.1.

A.8 Mask Estimation Results - Noisy Reverberant Test Conditions

Figure A8.1 illustrates the objective intelligibility measures of noisy reverberant speech and enhanced speech using masks estimated using oracle conditions.

A.9 Statistical Analysis

Statistical tests were performed to evaluate the effects of phoneme grouping technique, SNR, and training condition on SRMR-CI scores. All statistical analyses were performed in Python using the `statsmodels` and `SciPy` packages.

Mixed Effects Analysis. A linear mixed-effects (LME) model was fit with SRMR-CI score as the response variable. Fixed-effects factors included: *phoneme grouping method* (Ph, MOA, VUV); *SNR* ({-5, 0, 5, 10, 15, 20} dB); and *training condition* (Rev, Noisy, NoisyRev), together with all two-way and three-way interactions. To account for repeated measurements obtained from the same speech sample, a random intercept was included for sample identity. The LME model was fitted using maximum likelihood estimation. Convergence was achieved after optimization using the Limited-memory BFGS algorithm.

The LME analysis showed significant effects of SNR, training condition, and their interactions with phoneme grouping; all $p < 0.001$ except the *VUV* $\times$ *noisy-reverb* effect at -5dB ($p = 0.044$). The random-intercept variance was 0.067. Relative to -5dB (reference level), higher SNR levels have significantly higher SRMR-CI scores ($\beta = +0.195$ to $+0.376$, all $p < 0.001$). At -5dB across models with noise-only training, phoneme-based model did not differ significantly from MOA-based model ($\beta = -0.004$, $p = 0.411$), whereas VUV-based grouping had significantly lower scores ($\beta = -0.055$, $p < 0.001$). Significant *phoneme grouping* $\times$ *SNR* interactions were observed for Ph ($\beta = -0.010$ to -0.020) and VUV ($\beta = -0.015$ to -0.057). Reverberant-only training significantly altered the relative phoneme grouping method effects (Ph: $\beta = -0.212$; VUV: $\beta = -0.215$; both $p < 0.001$), while significant three-way interactions indicated that differences across phoneme grouping methods varied with SNR under both noisy-reverberant and reverberant-only training.

Speech-to-reverberation modulation energy ratio for cochlear implant users (SRMR-CI)					
Room	Rev	PI	OE^p-Ph	OE^p-MOA	OE^p-VUV
office	1.487 ±0.008	2.009 ±0.013	2.092 ±0.014	**2.303 ±0.014**	2.018 ±0.013
lecture	1.411 ±0.008	1.855 ±0.013	1.936 ±0.015	**2.251 ±0.014**	1.832 ±0.013
stairway	1.458 ±0.008	1.981 ±0.014	2.023 ±0.015	**2.247 ±0.015**	1.931 ±0.014
church	1.124 ±0.008	1.464 ±0.012	1.611 ±0.012	**1.834 ±0.011**	1.501 ±0.012
meeting	2.056 ±0.012	2.441 ±0.014	2.427 ±0.013	**2.469 ±0.013**	2.426 ±0.014
booth	2.764 ±0.016	**2.807 ±0.016**	2.696 ±0.016	2.652 ±0.016	2.802 ±0.016
Room	OE^k-Ph	OE^k-MOA	OE^k-VUV	IRM	DP
office	2.154 ±0.014	2.280 ±0.014	2.024 ±0.014	2.218 ±0.013	2.440 ±0.015
lecture	2.091 ±0.014	2.303 ±0.014	1.848 ±0.013	2.176 ±0.012	2.426 ±0.015
stairway	2.144 ±0.015	2.290 ±0.015	1.965 ±0.015	2.181 ±0.013	2.473 ±0.015
church	1.815 ±0.010	2.000 ±0.008	1.561 ±0.011	2.076 ±0.007	2.451 ±0.008
meeting	2.425 ±0.013	2.410 ±0.013	2.406 ±0.014	2.305 ±0.014	2.654 ±0.016
booth	2.711 ±0.016	2.594 ±0.015	2.772 ±0.016	2.578 ±0.016	2.641 ±0.017

Table A7.1: Mean ± 95% confidence interval of objective speech intelligibility scores for the reverberant signal (Rev), direct path signal (DP), and across different mask estimation methods: ideal ratio mask (IRM), phoneme independent model (PI), mask predicted by Omni-Expert with ideal phoneme/MOA/VUV knowledge (OE^k), and using predicted classifier probabilities (OE^p). Bold indicates the highest performance among the non-oracle models.

Short-time objective intelligibility (STOI)					
Room	Rev	PI	OE^p-Ph	OE^p-MOA	OE^p-VUV
office	0.721 ±0.002	0.797 ±0.001	0.820 ±0.001	**0.840 ±0.001**	0.802 ±0.001
lecture	0.692 ±0.002	0.780 ±0.002	0.805 ±0.001	**0.835 ±0.001**	0.785 ±0.001
stairway	0.736 ±0.001	0.825 ±0.001	0.844 ±0.001	**0.858 ±0.001**	0.827 ±0.001
church	0.688 ±0.003	0.750 ±0.002	0.770 ±0.002	**0.776 ±0.002**	0.753 ±0.002
meeting	0.876 ±0.001	**0.907 ±0.001**	0.910 ±0.001	0.901 ±0.001	0.905 ±0.001
booth	**0.950 ±0.001**	0.943 ±0.001	0.938 ±0.001	0.914 ±0.001	0.936 ±0.001
Room	OE^k-Ph	OE^k-MOA	OE^k-VUV	IRM	DP
office	0.833 ±0.001	0.852 ±0.001	0.807 ±0.001	0.968 ±0.001	1.000 ±0.000
lecture	0.827 ±0.001	0.850 ±0.001	0.794 ±0.001	0.971 ±0.000	1.000 ±0.000
stairway	0.861 ±0.001	0.870 ±0.001	0.833 ±0.001	0.966 ±0.001	1.000 ±0.000
church	0.810 ±0.002	0.814 ±0.002	0.773 ±0.002	0.969 ±0.000	1.000 ±0.000
meeting	0.912 ±0.001	0.909 ±0.001	0.908 ±0.001	0.951 ±0.001	1.000 ±0.000
booth	0.939 ±0.001	0.923 ±0.001	0.942 ±0.001	0.983 ±0.000	1.000 ±0.000

Table A7.2: Mean ± 95% confidence interval of objective speech intelligibility scores for the reverberant signal (Rev), direct path signal (DP), and across different mask estimation methods: ideal ratio mask (IRM), phoneme independent model (PI), mask predicted by Omni-Expert with ideal phoneme/MOA/VUV knowledge (OE^k), and using predicted classifier probabilities (OE^p). Bold indicates the highest performance among the non-oracle models.

Frequency-weighted segmental speech-to-distortion ratio (fwSDR)					
Room	Rev	PI	OE^p-Ph	OE^p-MOA	OE^p-VUV
office	12.970 ±0.235	14.592 ±0.212	15.143 ±0.198	**15.540 ±0.196**	14.664 ±0.207
lecture	12.292 ±0.251	14.524 ±0.226	15.148 ±0.220	**15.745 ±0.208**	14.609 ±0.227
stairway	12.858 ±0.248	15.579 ±0.220	16.122 ±0.210	**16.409 ±0.204**	15.649 ±0.216
church	12.607 ±0.124	13.754 ±0.120	13.813 ±0.117	**14.270 ±0.113**	13.647 ±0.121
meeting	17.423 ±0.214	18.418 ±0.179	**18.622 ±0.179**	18.077 ±0.177	18.236 ±0.178
booth	**20.834 ±0.166**	19.397 ±0.160	19.690 ±0.171	18.425 ±0.173	19.126 ±0.163
Room	OE^k-Ph	OE^k-MOA	OE^k-VUV	IRM	DP
office	15.404 ±0.193	15.624 ±0.192	14.753 ±0.205	25.920 ±0.091	35 ±0.000
lecture	15.654 ±0.214	15.702 ±0.200	14.787 ±0.226	26.023 ±0.092	35 ±0.000
stairway	16.642 ±0.206	16.427 ±0.194	15.859 ±0.216	25.853 ±0.088	35 ±0.000
church	14.800 ±0.114	15.293 ±0.109	14.020 ±0.121	25.622 ±0.043	35 ±0.000
meeting	18.476 ±0.169	18.287 ±0.175	18.185 ±0.174	24.348 ±0.093	35 ±0.000
booth	19.702 ±0.166	18.674 ±0.175	19.145 ±0.160	25.546 ±0.095	35 ±0.000

Table A7.3: Mean ± 95% confidence interval of objective speech intelligibility scores for reverberant speech (Rev), direct path speech (DP), and enhanced speech across different mask estimation methods: ideal ratio mask (IRM), phoneme independent (PI) model and Omni-Expert with ideal phoneme/MOA/VUV knowledge (OE^k), and using predicted classifier probabilities (OE^p). Bold indicates the highest performance among the non-oracle models.

The corresponding three-way effects ranged from $+0.059$ to $+0.188$ for phoneme-based model and from $+0.019$ to $+0.075$ for VUV-based model under noisy-reverberant training. Under reverberant-only training, the effects ranged from -0.057 to -0.118 for phoneme-based model and from -0.101 to -0.197 for VUV-based model.

Multiple Comparison Test. Post-hoc pairwise comparisons were performed because the LME analysis revealed higher-order interactions among phoneme grouping method, SNR, and training condition. For each comparison, the mean paired difference, paired t statistic, unadjusted p-value, and Cohen's d_z were calculated. 95% confidence intervals were estimated using 5000 bootstrap iterations.

Under reverberant-only training, MOA-based model significantly outperformed both phoneme- and VUV-based models across all SNR levels, with range of effect sizes of d_z = [1.032, 1.881] for MOA vs. Ph; and d_z = [1.115, 2.611] for MOA vs. VUV. The phoneme-based model also significantly outperformed the VUV-based model across all SNR levels, d_z = [0.404, 1.454].

Under noise-only training, MOA-based model significantly outperformed phoneme-based model across all SNR levels, with paired effect sizes ranging from small to modest, d_z = [0.041, 0.284]. The MOA-based model also significantly outperformed VUV-based model at every SNR (d_z = [0.555, 1.157]), while phoneme-based model significantly outperformed VUV-based model (d_z= [0.441, 0.891]), yielding a consistent ordering of performance with MOA $>$ Ph $>$ VUV.

Under noisy-reverberant training, the relative ordering changed with SNR. At -5 dB, MOA-based model significantly outperformed both phoneme-based (d_z = 0.082) and VUV-based (d_z = 0.209) models, while phoneme-based model also outperformed VUV-based model (d_z = 0.121). From SNR $\geq$ 0dB, phoneme-based model significantly outperformed MOA-based model, with the effect increasing from d_z = 0.194 at 0 dB to d_z = 1.241 at 20 dB. Across all SNR levels, MOA-based model significantly outperformed VUV-based model (d_z = [0.127, 0.209]), and phoneme-based model significantly outperformed VUV-based model (d_z= [0.362, 1.241]. Thus, under noisy-reverberant training, the ordering changed from MOA $>$ Ph $>$ VUV at -5dB to Ph $>$ MOA $>$ VUV for SNR levels between 0 to 20dB.

The results collectively demonstrate that the relative effectiveness of linguistic representations depends strongly on the acoustic training condition and the SNR level. MOA- and phoneme-based conditioning exhibit different sensitivities to combined noise and reverberant training condition. Under noise-only and reverberant-only training, the MOA-based model shows highest overall performance and most robustness when compared to phoneme- and VUV-based models.

Cepstral distance (CD) (Lower is better)					
Room	Rev	PI	OE^p-Ph	OE^p-MOA	OE^p-VUV
office	2.869 ±0.023	2.579 ±0.019	2.393 ±0.019	**2.274 ±0.016**	2.629 ±0.019
lecture	3.184 ±0.033	2.597 ±0.027	2.417 ±0.026	**2.334 ±0.021**	2.622 ±0.028
stairway	2.714 ±0.038	2.194 ±0.031	**2.100 ±0.027**	2.112 ±0.021	2.213 ±0.030
church	3.040 ±0.018	2.670 ±0.017	2.587 ±0.016	**2.517 ±0.014**	2.695 ±0.017
meeting	2.152 ±0.025	2.206 ±0.017	**2.109 ±0.016**	2.158 ±0.014	2.268 ±0.017
booth	**1.386 ±0.016**	1.842 ±0.018	1.671 ±0.014	1.782 ±0.013	1.892 ±0.015
Room	OE^k-Ph	OE^k-MOA	OE^k-VUV	IRM	DP
office	2.438 ±0.017	2.320 ±0.016	2.637 ±0.018	0.839 ±0.006	0 ±0.000
lecture	2.414 ±0.024	2.418 ±0.018	2.599 ±0.027	0.836 ±0.005	0 ±0.000
stairway	2.064 ±0.025	2.196 ±0.018	2.188 ±0.029	0.753 ±0.005	0 ±0.000
church	2.383 ±0.015	2.366 ±0.012	2.597 ±0.017	0.817 ±0.003	0 ±0.000
meeting	2.190 ±0.015	2.187 ±0.014	2.298 ±0.017	1.037 ±0.007	0 ±0.000
booth	1.772 ±0.014	1.791 ±0.014	1.921 ±0.016	0.777 ±0.006	0 ±0.000

Table A7.4: Mean ± 95% confidence interval of cepstral distance scores for the reverberant speech (Rev), direct path speech (DP), and enhance speech with different mask estimation methods: ideal ratio mask (IRM), phoneme independent (PI) model, Omni-Expert model with ideal phoneme/manner of articulation (MOA)/voiced-unvoiced(VUV) knowledge (OE^k) and using predicted classifier probabilities (OE^p). Bold indicates the highest performance among the non-oracle models.

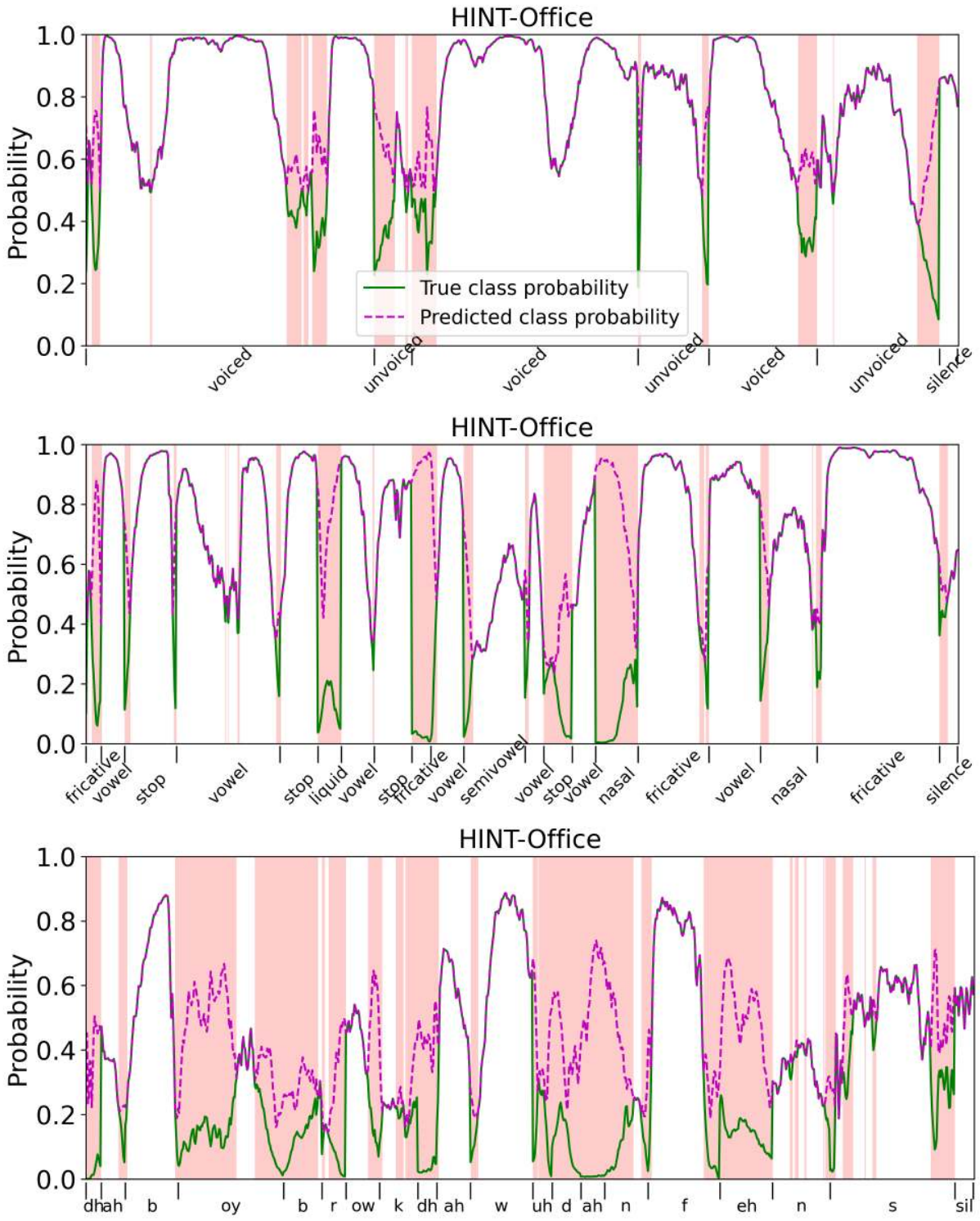


Figure A3.2: Framewise probabilities of true and predicted voiced and unvoiced (VUV), manner of articulation (MOA) and phoneme groups for reverberant HINT speech utterance *"the boy broke the wooden fence"* generated using a room impulse response from an office.

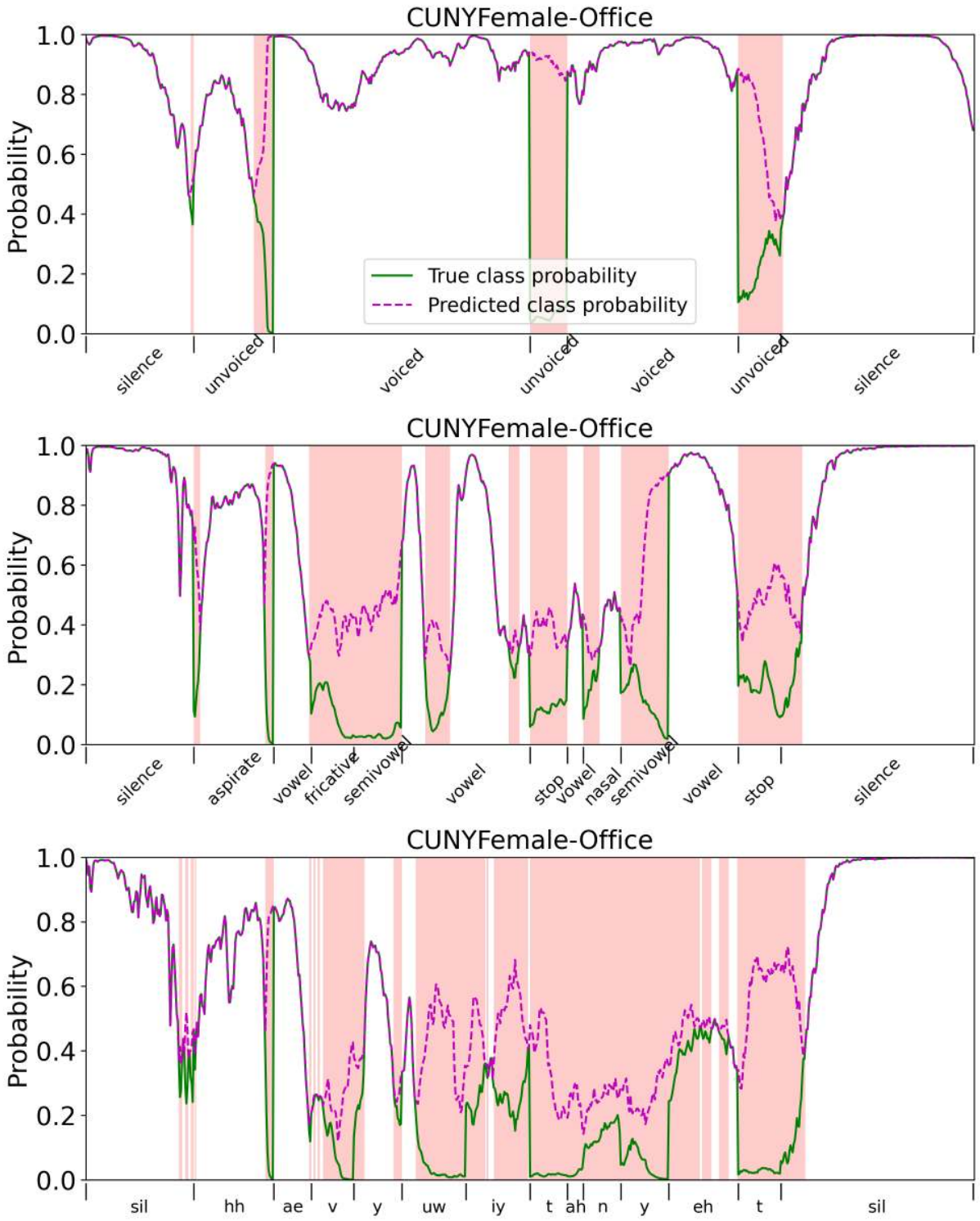


Figure A3.3: Framewise probabilities of true and predicted voiced and unvoiced (VUV), manner of articulation (MOA) and phoneme groups for reverberant CUNYFemale speech utterance *"have you eaten yet"* generated using a room impulse response from an office.

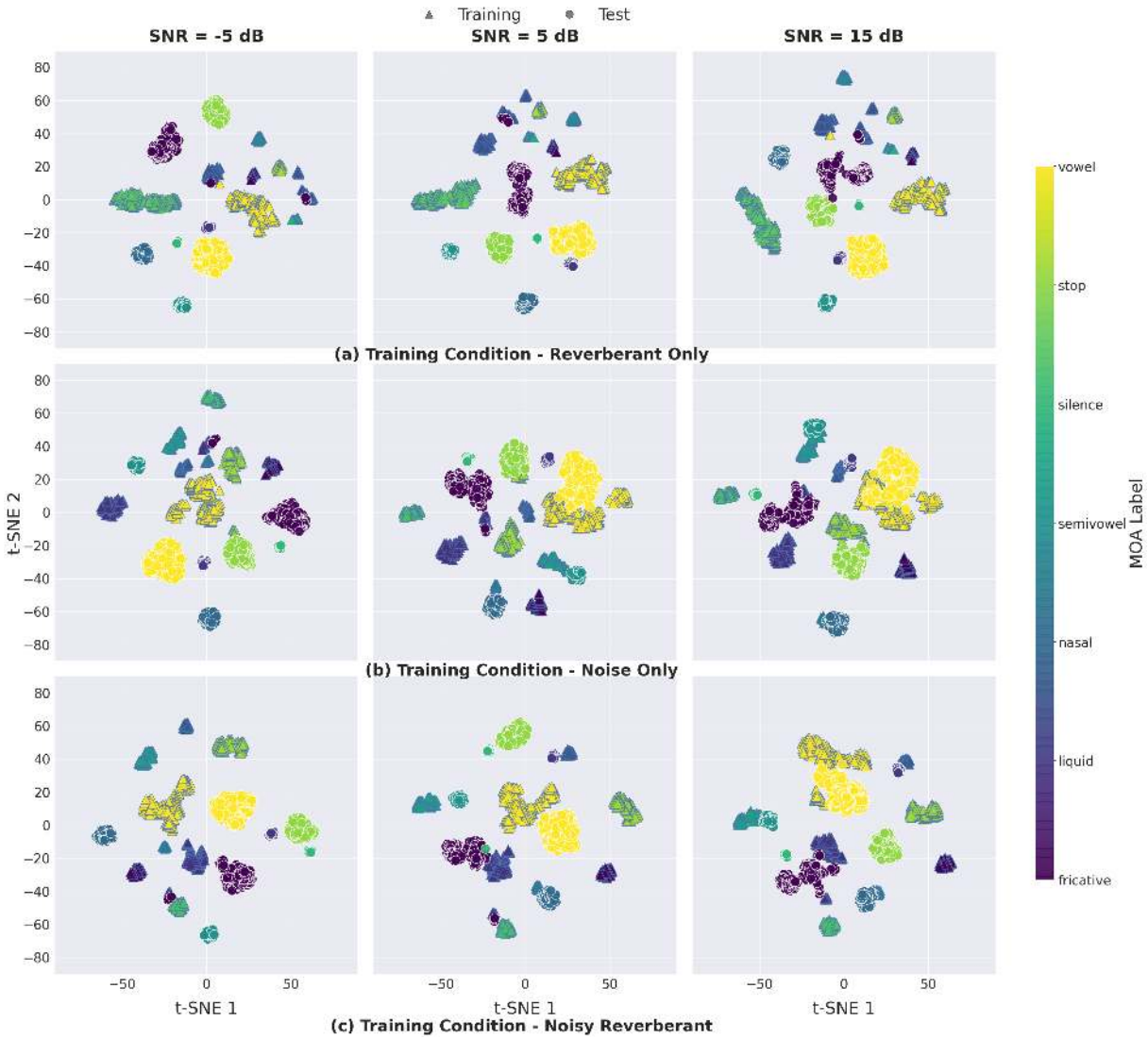


Figure A4.1: t-SNE visualization of transformed manner of articulation (MOA)-based features across training conditions for randomly sampled 1000 features from noisy reverberant dataset comprised of HINT speech, public noise and stairway room impulse response at signal-to-noise (SNR) of -5,5 and 15 dB.

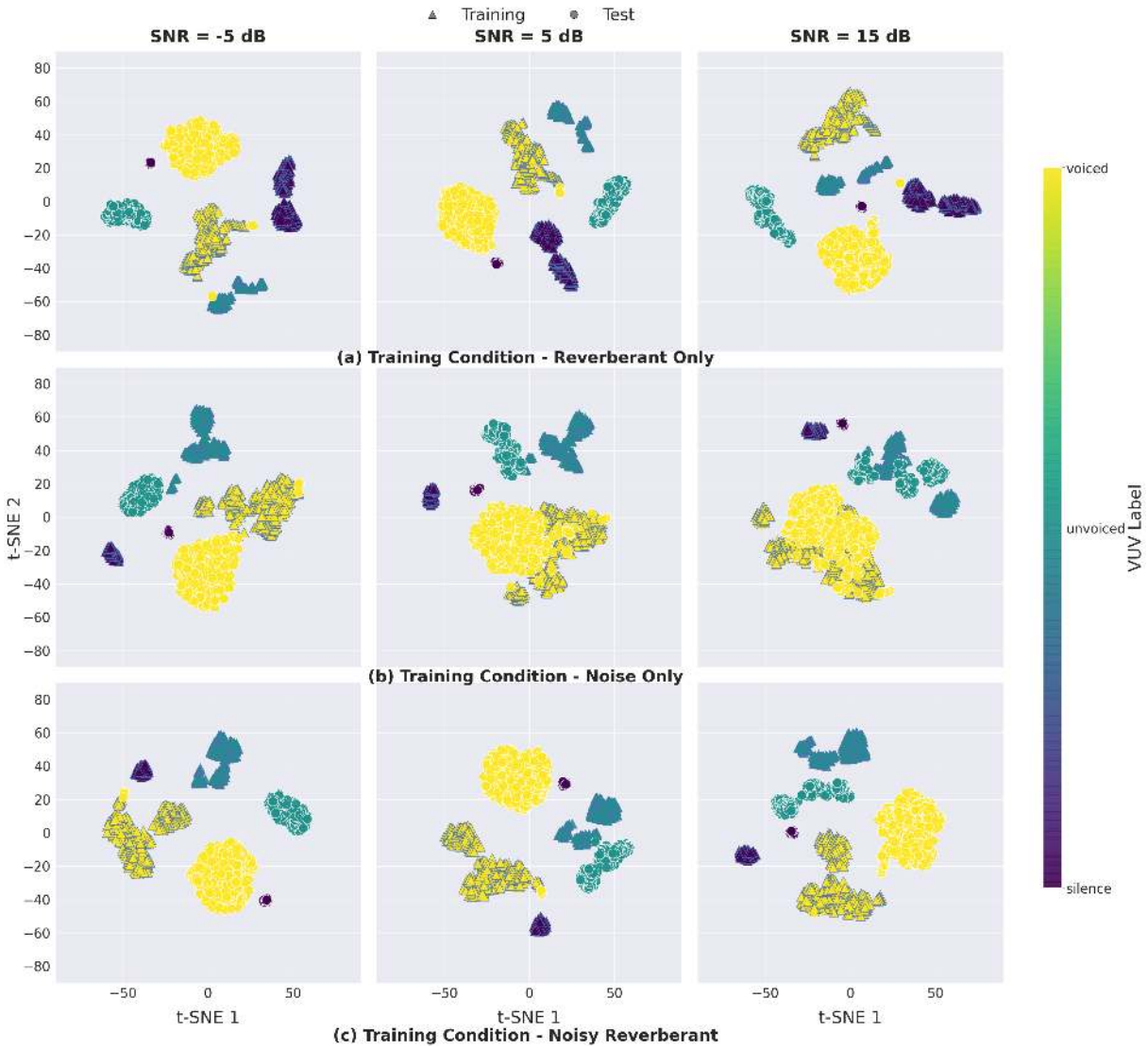
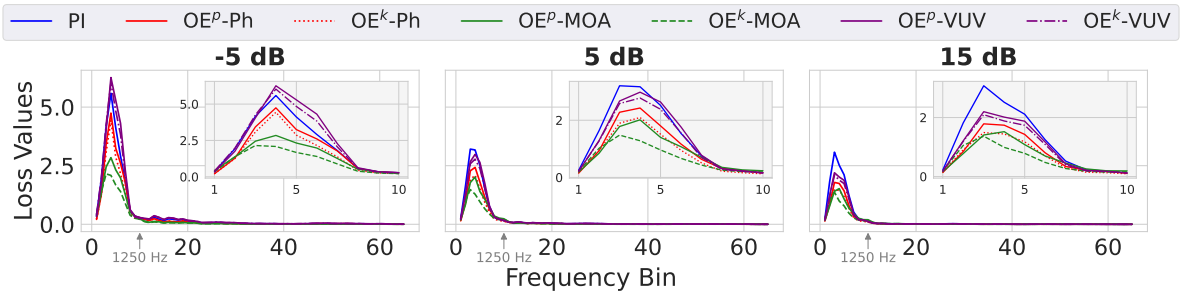
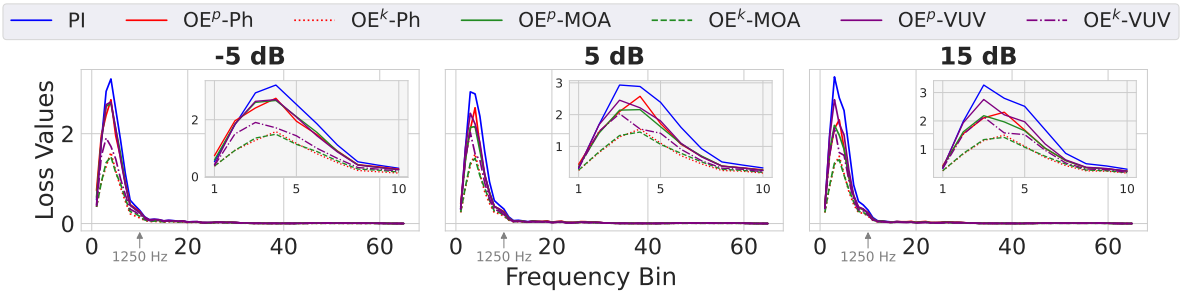


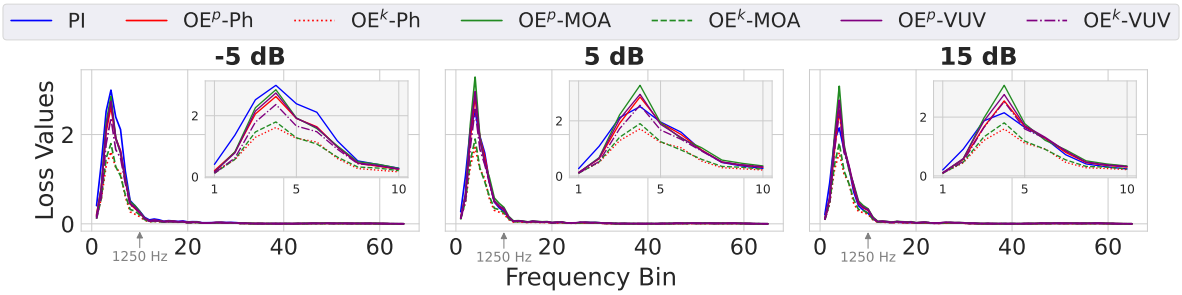
Figure A4.2: t-SNE visualization of transformed voiced/unvoiced (VUV)-based features across training conditions for randomly sampled 1000 features from noisy reverberant dataset comprised of HINT speech, public noise and stairway room impulse response at signal-to-noixe ratio (SNR) of -5,5 and 15 dB.



(a) Training Condition: Reverberation Only



(b) Training Condition: Noise Only



(c) Training Condition: Noisy Reverberant

Figure A5.1: Average signal loss across frequency bins with phoneme-independent (PI) mask estimation and mask estimation with the Omni-expert (OE) model using phonemes (Ph), manner of articulation (MOA), and voiced/unvoiced (VUV) grouping strategies. Results are shown for predicted and known conditions (OE$^{p/k}$).

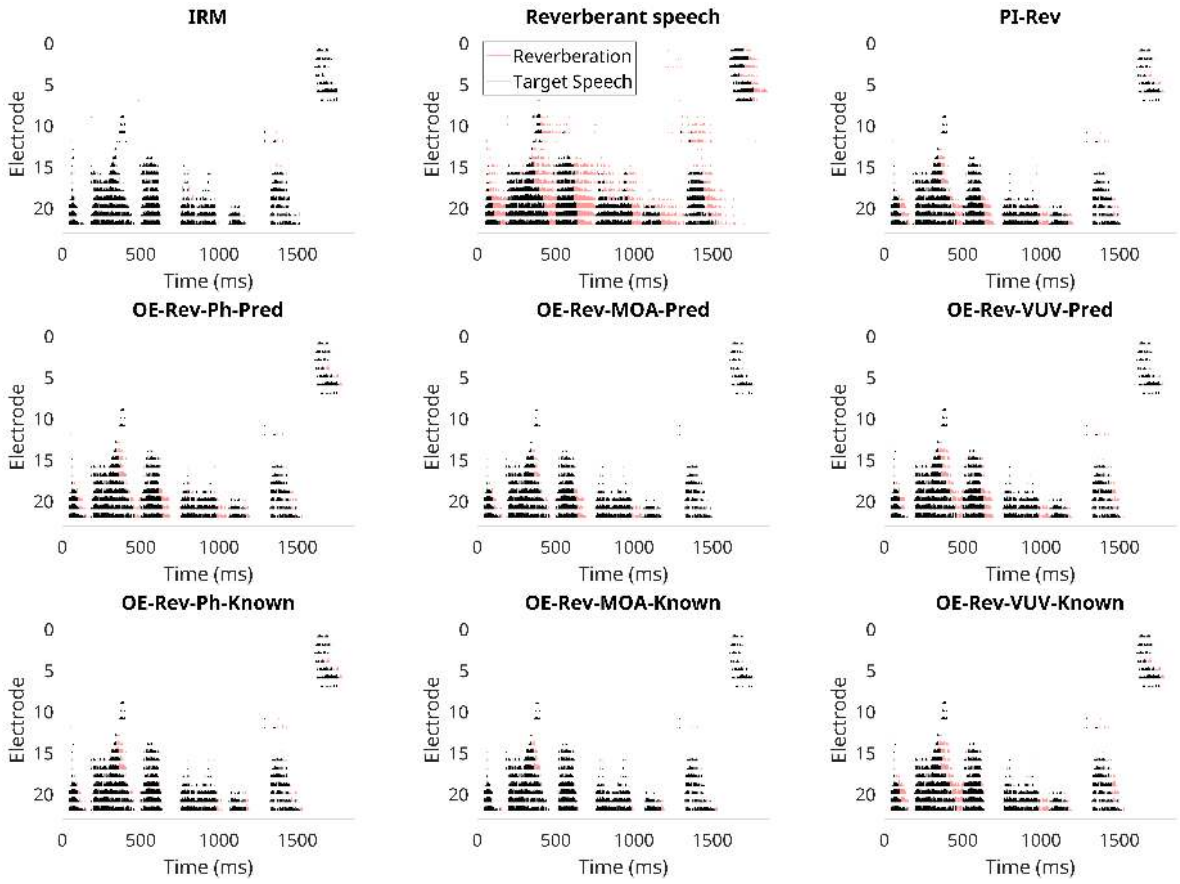


Figure A6.1: Electrodograms of the speech utterance *"the boy broke the wooden fence."* (HINT dataset) convolved with a stairway room impulse response (Aachen Impulse Response dataset). Visualizations of ideal ratio mask (IRM); unenhanced reverberant speech; enhanced reverberant speech using phoneme independent (PI) models trained using reverberant-only (Rev) data; and Omni-Expert (OE) models trained using Rev data, with predicted/known phonemes (Ph), manner of articulation (MOA) and voiced/unvoiced (VUV) groups.

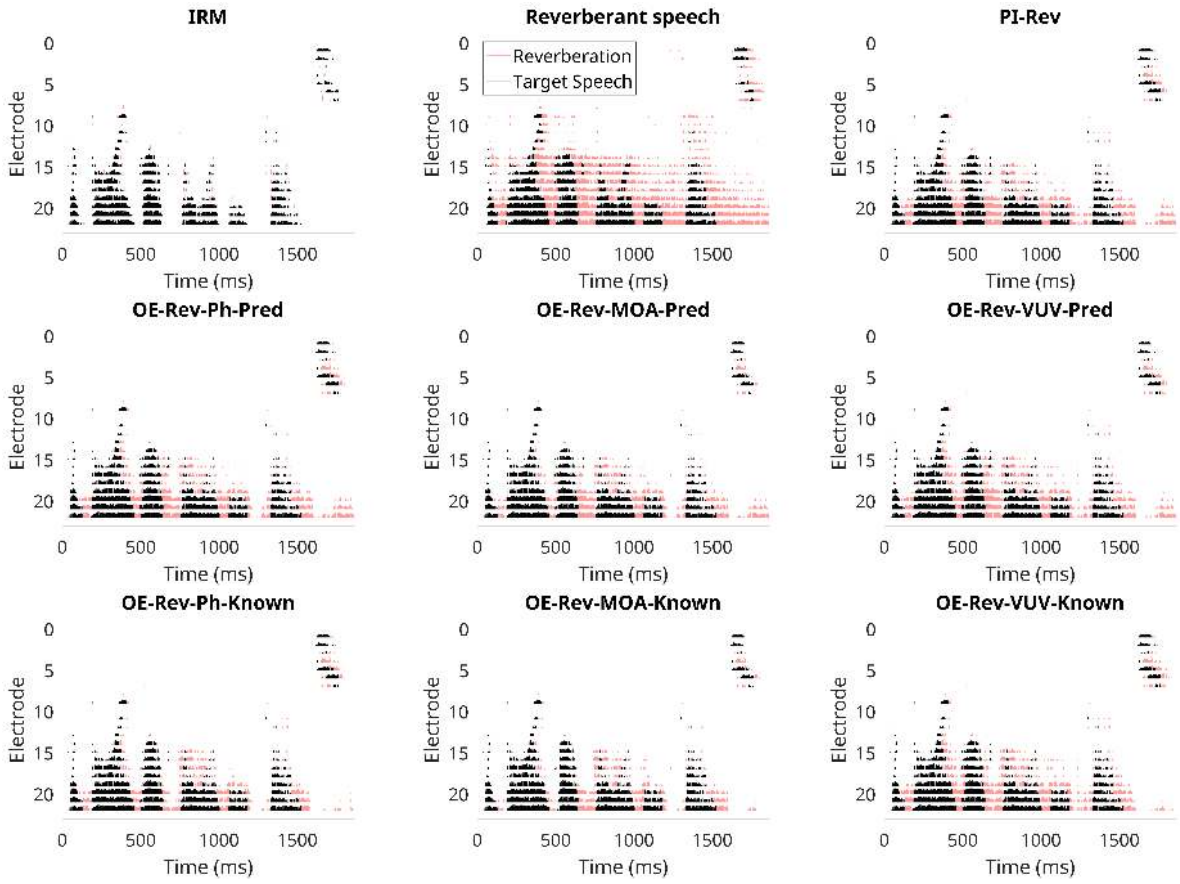


Figure A6.2: Electrodograms of the speech utterance *"the boy broke the wooden fence."* The speech utterance is from the HINT dataset. The room impulse response is that of a church. Visualizations of ideal ratio mask (IRM); unenhanced reverberant signal; enhanced reverberant signals using phoneme independent (PI) models trained using reverberant-only (Rev); and Omni-Expert (OE) models trained using reverberant-only (Rev), with predicted/known phonemes (Ph), manner of articulation (MOA) and voiced/unvoiced (VUV) groups.

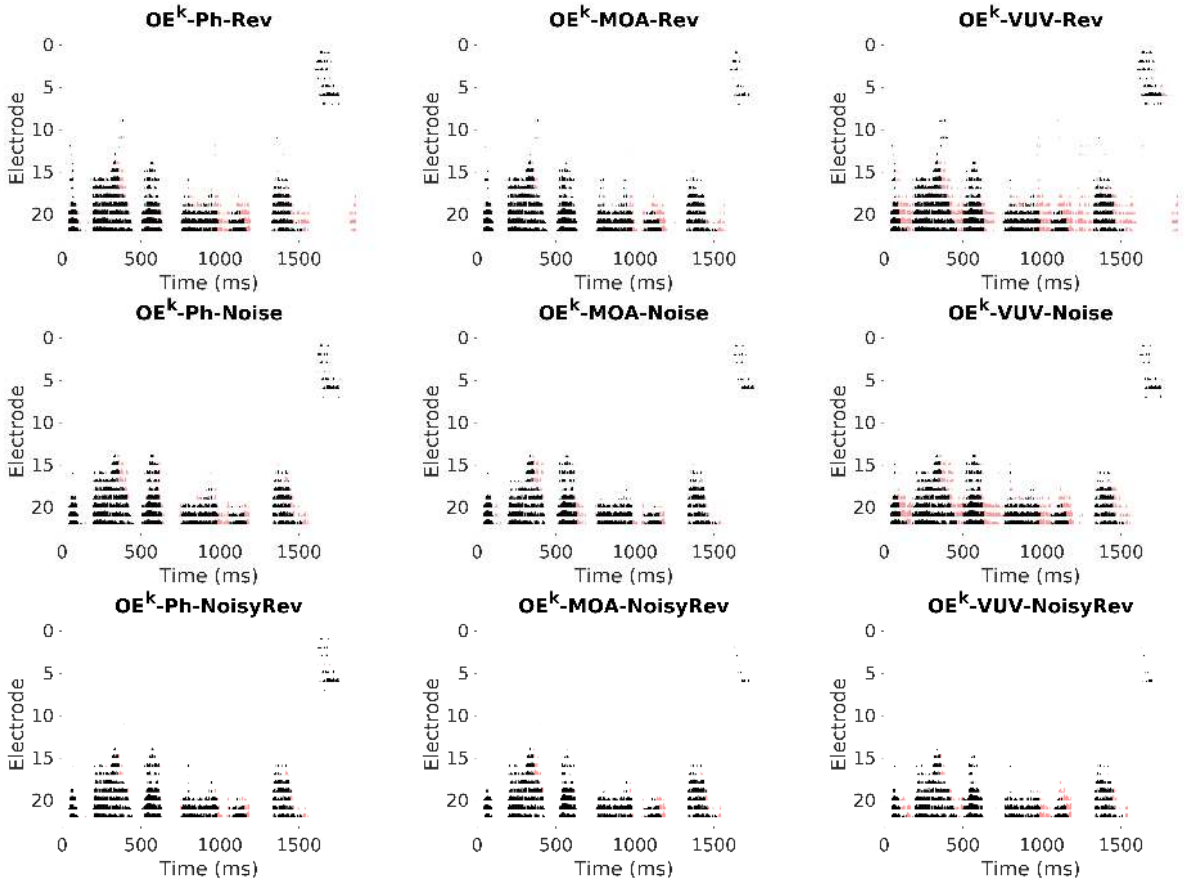


Figure A6.3: Electrodograms of the speech utterance *"the boy broke the wooden fence"* (HINT dataset) with public noise (DEMAND dataset) added at 5dB signal-to-noise ratio (SNR) and convolved with a stairway room impulse response (Aachen Impulse Response dataset). Visualizations of enhanced noisy reverberant signals using oracle Omni-Expert (OE^k) models trained on reverberant-only (Rev), noise-only (Noise), and noisy reverberant (NoisyRev) data with known phonemes (Ph), manner of articulation (MOA) and voiced/unvoiced (VUV) groups.

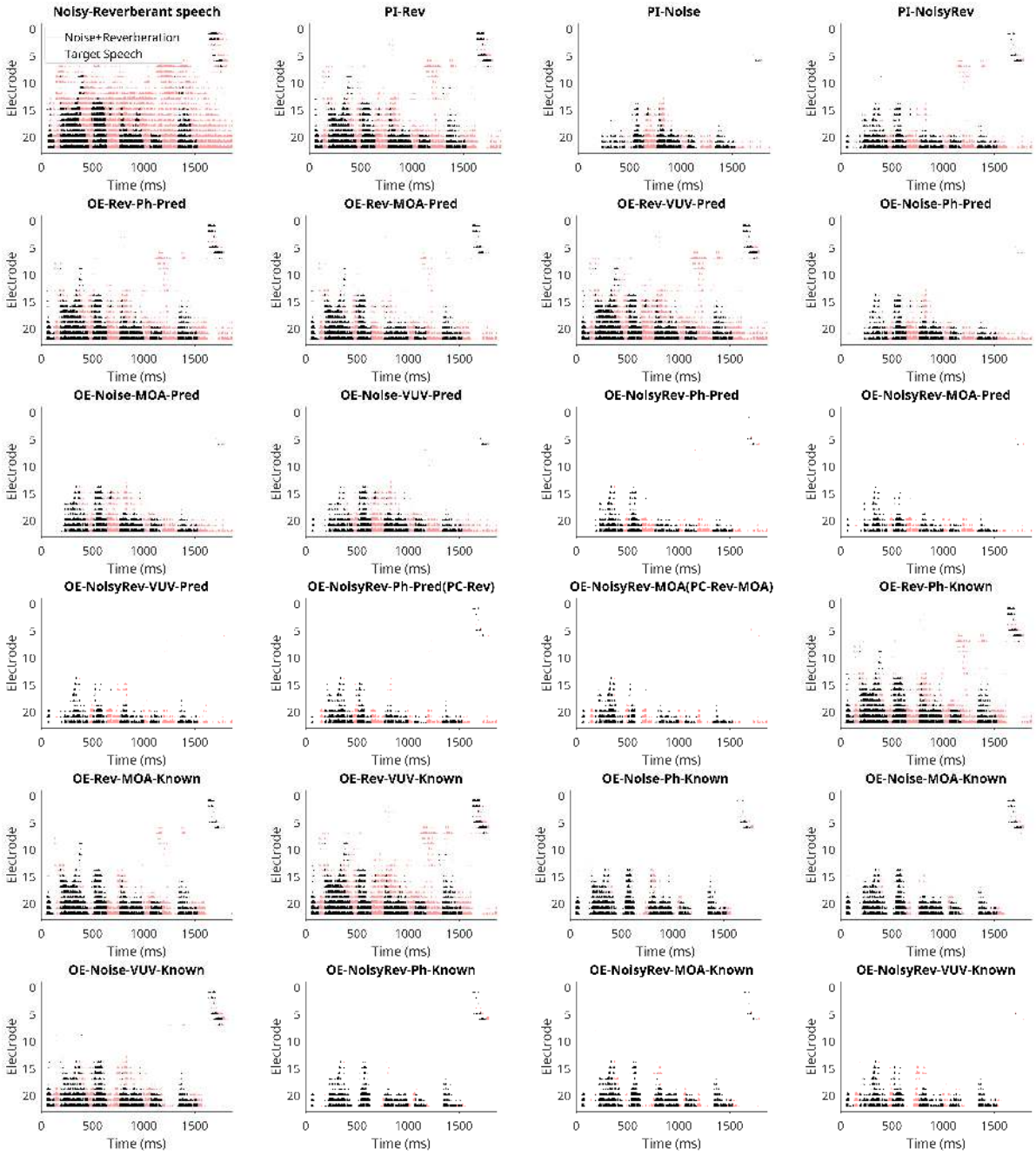


Figure A6.4: Electrodograms of the speech utterance *"the boy broke the wooden fence"* (HINT dataset) with two talker babble noise (Bern Cocktail Party dataset) added at a 5 dB signal-to-noise (SNR) and convolved with a church room impulse response (Aachen Impulse Response dataset). Visualizations of enhanced noisy reverberant signals using Omni-Expert (OE) models trained using reverberant-only (Rev), noise-only (Noise), and noisy reverberant (NoisyRev) data, with predicted/known phonemes (Ph), manner of articulation (MOA) and voiced/unvoiced (VUV) groups.

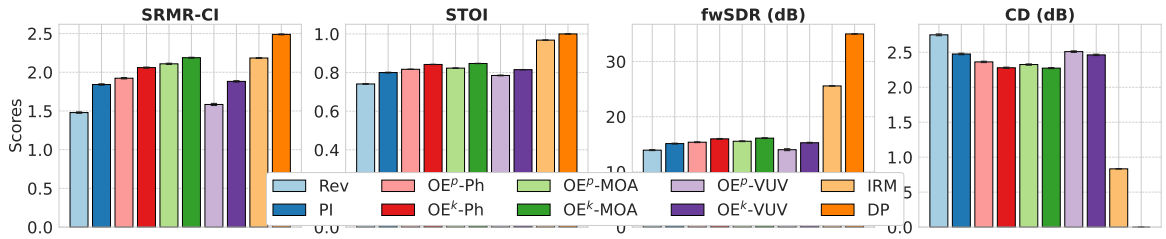


Figure A7.1: Estimated marginal means and 95% confidence intervals of speech-to-reverberation modulation energy ratio for cochlear implant users (SRMR-CI), short-time objective intelligibility (STOI) scores, frequency-weighted segmental speech-to-distortion ratio (fwSDR) and cepstral distance (CD) for direct path (DP), reverberant speech (Rev), enhanced reverberant speech after applying the ideal ratio mask (IRM) and estimated masks with the phoneme independent (PI) model, and Omni-Expert model (OE$^{p/k}$) with predicted and known phonemes (Ph), manner of articulation (MOA) and voiced/unvoiced (VUV) groups. Results aggregated across all speech datasets and reverberant room conditions. For CD scores, lower values indicate better performance.

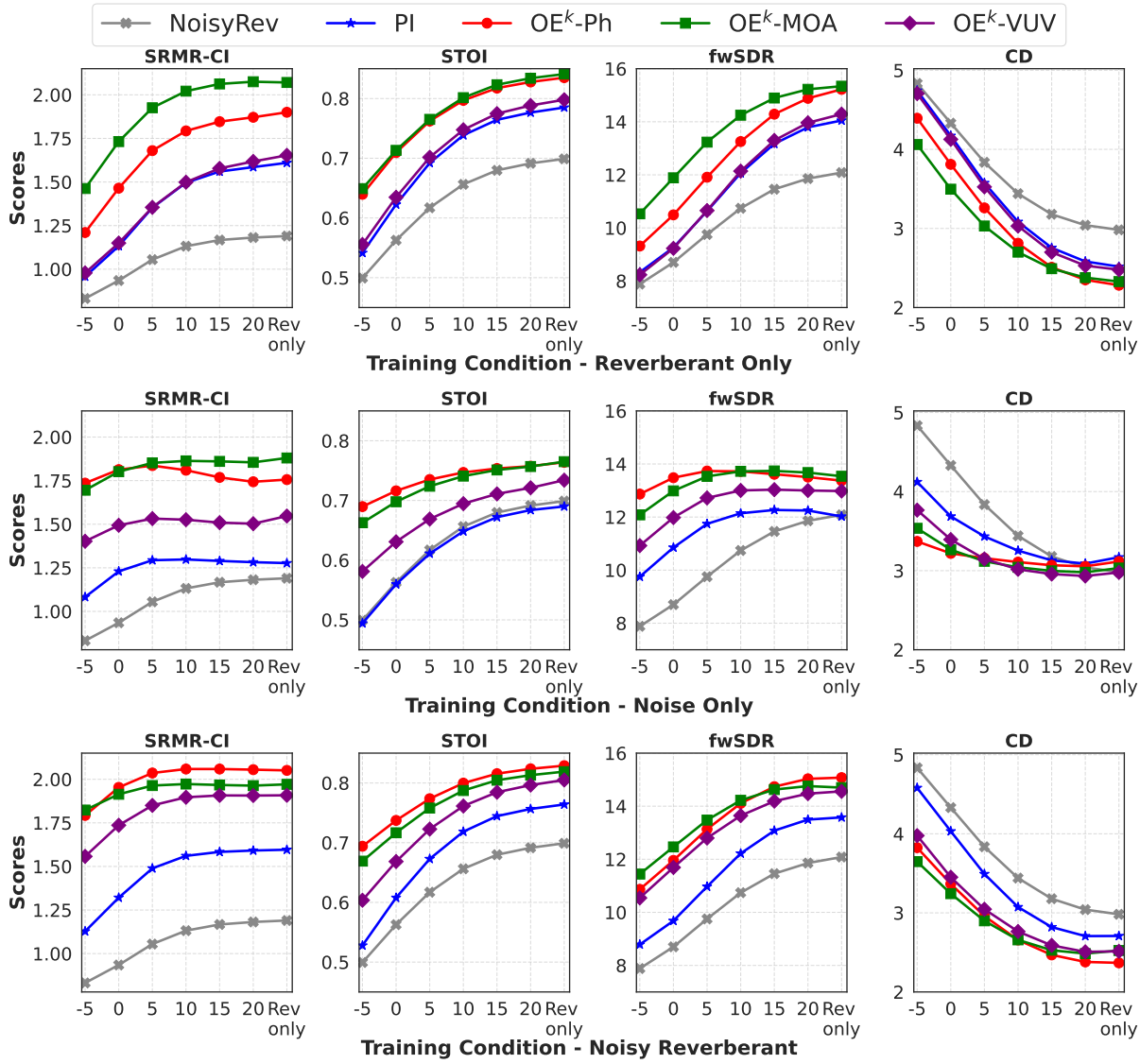


Figure A8.1: Objective speech intelligibility metrics of noisy-reverberant (NoisyRev) speech enhanced using phoneme independent model (PI) and oracle Omni-Expert (OE^k) models trained on reverberant-only, noisy-reverberant, and noise-only conditions across varied reverberant rooms and noise types, aggregated across test speech datasets. Results are shown for models with oracle phoneme, manner of articulation (MOA) and voiced/unvoiced (VUV) groups.