

Document Version

Final published version

Licence

CC BY

Citation (APA)

Olalere, F., van der Heijden, K., Stronks, H. C., Briaire, J., Frijns, J. H. M., & Güçlütürk, Y. (2026). Speech separation for hearing-impaired children in the classroom. *The Journal of the Acoustical Society of America*, 160(3), 2440-2454.
<https://doi.org/10.1121/10.0046649>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

SEPTEMBER 22 2026

Speech separation for hearing-impaired children in the classroom

Feyisayo Olalere ; Kiki van der Heijden; H. Christiaan Stronks; Jeroen Briare; Johan H. M. Frijns; Yagmur Güçlütürk



J. Acoust. Soc. Am. 160, 2440–2454 (2026)

<https://doi.org/10.1121/10.0046649>



Articles You May Be Interested In

Phase-preserving gray box model for robust sonar signal separation

J. Acoust. Soc. Am. (April 2025)

A two-stage deep learning algorithm for talker-independent speaker separation in reverberant conditions

J. Acoust. Soc. Am. (September 2020)

Enhancing speech separation performance utilizing various wavelet coefficients

J. Acoust. Soc. Am. (July 2025)



LEARN MORE

Advance your science and career as a member of the
Acoustical Society of America

Speech separation for hearing-impaired children in the classroom

Feyisayo Olalere,^{1,a)}  Kiki van der Heijden,^{1,2} H. Christiaan Stronks,³ Jeroen Braire,³ Johan H. M. Frijns,^{3,4,5} and Yagmur Güçlütürk¹

¹*Radboud University, Donders Institute for Brain, Cognition, and Behaviour, Nijmegen, The Netherlands*

²*Mortimer B. Zuckerman Mind, Brain, Behavior Institute, Columbia University, New York, New York 10027, USA*

³*Department of Otorhinolaryngology, Leiden University Medical Center, Leiden, The Netherlands*

⁴*Leiden Institute for Brain and Cognition, Leiden, The Netherlands*

⁵*Department of Bioelectronics, Delft University of Technology, Delft, The Netherlands*

ABSTRACT:

Classroom environments pose significant challenges for hearing-impaired children, where background noise, simultaneous talkers, and reverberation degrade speech perception. Most deep learning-based speech separation algorithms are developed for adult voices in simplified conditions, neglecting both the higher spectral similarity of children's voices and the acoustic complexity of classrooms. We address this gap using MIMO-TasNet, a compact, low-latency, multi-channel architecture suited for real-time processing in bilateral hearing aids or cochlear implants. We simulated naturalistic classroom conditions with moving talkers, both child–child and child–adult pairs across varying noise and distance settings. We compared three training strategies: adult speech only, classroom-specific data, and finetuning adult-trained models with limited classroom data. Results show that binaural spatial cues enable adult-trained models to perform well in clean classroom conditions, even on overlapping child talkers. Classroom-specific training improved separation quality, while finetuning with only half the data achieved superior performance, confirming data-efficient adaptation benefits. Training with diffuse babble noise enhanced robustness across conditions, and models preserved spatial awareness while generalizing to unseen talker–listener distances. Combining spatially aware architectures with targeted adaptation strategies is an important step toward enhancing computational speech separation for children, which can help guide the deployment of practical on-device assistive technologies for classroom speech-in-noise challenges.

© 2026 Acoustical Society of America. <https://doi.org/10.1121/10.0046649>

(Received 30 December 2025; revised 9 August 2026; accepted 30 August 2026; published online 22 September 2026)

[Editor: James F. Lynch]

Pages: 2440–2454

I. INTRODUCTION

Classroom environments pose significant challenges for speech perception, especially for younger children (Klatte *et al.*, 2013). These challenges are even more pronounced for children with hearing loss in classrooms designed for typical-hearing students (Grempe and Easterbrooks, 2018; Neuman *et al.*, 2010). Younger children need more favorable signal-to-noise ratios (SNRs) than adults to understand speech in reverberant conditions (Neuman *et al.*, 2010), and modest background noise can disrupt their learning (McMillan and Saffran, 2016). Children with hearing loss, whether mild, unilateral, or using hearing devices, requires better listening conditions to achieve comparable performance (Ching *et al.*, 2018; Lewis *et al.*, 2016). For those with auditory processing difficulties, noisy classrooms can further interfere with learning and participation (Jutras *et al.*, 2019). Overlapping talkers, reverberation, background noise, and talker-to-listener distance all contribute to the communication barriers faced by hearing-impaired children (Attia *et al.*, 2025; Valente *et al.*, 2012).

Hearing technologies like hearing aids and cochlear implants are essential assistive devices commonly used by hearing-impaired children (Henry *et al.*, 2023). While these devices offer substantial benefits in quiet, controlled settings, their performance often degrades in challenging real-world environments such as classrooms (Gazibegovic *et al.*, 2024; Healy and Yoho, 2016; Qin and Oxenham, 2003). Advanced hearing aids use technologies like beamforming to enhance speech from a speaker positioned directly in front or from the side of the listener (Gazibegovic *et al.*, 2024; McCreery *et al.*, 2012). Although these systems have shown promising results in adults (de Souza *et al.*, 2024; Henry *et al.*, 2023), their effectiveness is limited in school settings, where children often move unpredictably, speak from varying positions, and produce less structured, more spontaneous speech than adults (Wolfe *et al.*, 2017, 2022).

Given these limitations, there is growing interest in supplementary approaches that enhance the speech signal for use in hearing aids and cochlear implants (Johnstone *et al.*, 2018). One such approach involves wireless remote microphone (WRM) systems (Gabova *et al.*, 2024a). In this setup, the talker wears or places a microphone nearby, which transmits audio directly to the listener's hearing aid. WRMs have

^{a)}Email: feyisayo.olalere@donders.ru.nl

been shown to improve speech understanding, particularly during teacher-led instruction, by improving the SNR (Chen *et al.*, 2021; Thompson *et al.*, 2020). However, this solution is less practical for children, who must also understand their peers during discussions and social interactions. Since peers typically do not wear WRMs, these systems fall short of addressing the full scope of children's communicative needs in classrooms (Gabova *et al.*, 2024b; Gustafson *et al.*, 2021).

Recent advancements in artificial intelligence (AI), particularly in deep neural network (DNN)-based speech separation, have opened up new possibilities for improving speech intelligibility in noisy and complex auditory environments (Lee *et al.*, 2024; Lutati *et al.*, 2022; Subakan *et al.*, 2021; Zhao *et al.*, 2024). When used as a front-end for hearing assistive devices (Han *et al.*, 2020), these models can isolate individual speech components from mixtures, allowing clearer access to the target speaker. In this approach, sound is captured by hearing aid microphones and processed by DNNs in real-time to separate speech from noise or interfering speech, and the enhanced signal is returned to the listener. This method has the potential to eliminate the need for multiple WRMs, which is impractical in peer-to-peer classroom settings. Moreover, DNN-based models can adapt to dynamic listening scenarios, tracking speakers even as they move (Han *et al.*, 2021). However, most state-of-the-art speech separation models have been developed and evaluated primarily on adult speech, with little to no testing on children's voices, despite known acoustic and linguistic differences [e.g., higher fundamental frequency (F0), and shorter vocal tracts] (Lee *et al.*, 1999) that affect model performance in related domains like speech recognition.

A consistent finding across speech recognition and enhancement studies is that models trained on adult voices generalize poorly to children's speech (Attia *et al.*, 2025; Shivakumar and Georgiou, 2020; Southwell *et al.*, 2024). This performance gap is attributed to fundamental acoustic differences, such as higher fundamental frequencies, shorter vocal tracts, and greater intra-speaker variability in children's speech. Importantly, prior studies that document this mismatch have almost exclusively relied on monaural recordings, where enhancement depends heavily on spectral distinctiveness. In such settings, the adult-to-child spectral mismatch produces a marked drop in performance. The potential for binaural cues, such as interaural phase differences and interaural level differences (ILDs), to mitigate this gap remains underexplored.

Binaural hearing has long been recognized as critical for parsing multi-talker environments, particularly in noisy or reverberant spaces (Bronkhorst, 2000; Litovsky *et al.*, 2006). Our prior work also demonstrated that binaural models can leverage spatial cues to separate spectrally similar adult talkers (Olalere *et al.*, 2025). We hypothesize that, since spatial cues can reduce reliance on spectral distinctiveness, they could thereby bridge part of the adult-to-child mismatch generally observed when using adult-trained speech enhancement models for children-related speech. We ask whether binaural information can also help an

adult-trained speech separation model generalize to children's speech. At the same time, we recognize that mismatches in noise type, such as adult versus Child-Babble or having two child talkers overlap, where spectral similarity is greatest, are likely to remain challenging.

To address these challenges, this study evaluates speech separation in naturalistic classroom scenarios using MIMO-TasNet (Han *et al.*, 2021). MIMO-TasNet is a compact, low-latency binaural architecture suitable for hearing aids and can process binaural input and preserve spatial information over time. We simulated dynamic classroom environments containing moving talkers and background babble using binaural room impulse responses (BRIRs) from children's ears to evaluate both speech separation and spatial tracking. We tested three training strategies across both clean and babble-noise conditions: (1) training on adult-only speech mixtures, (2) training on full classroom mixtures (child-child and child-adult), and (3) finetuning an adult-trained model on 50% of the classroom data. Our study make four primary contributions:

- We demonstrate that binaural cues substantially mitigate the adult-to-child generalization gap in clean conditions, contrasting with the severe mismatches typically documented in monaural speech tasks (Shivakumar and Georgiou, 2020; Southwell *et al.*, 2024).
- We show that in noisy environments, training specifically with child-like babble material is essential to maintain performance.
- We establish finetuning as a highly efficient strategy, outperforming full classroom training while utilizing only half the dataset.
- We evaluate spatial preservation via direction-of-arrival (DoA) error and demonstrate robust generalization to unseen talker distances, highlighting the possibility of these models for the future deployment of hearing-assistive technologies in dynamic classroom environments.

The remainder of this paper is structured as follows: in Sec. II, we describe the method, which contains the problem definition, and the simulation of overlapping speech in naturalistic classroom settings, including spatial modeling of moving talkers, model training procedures, and evaluation metrics. In Sec. III, we present the results of our evaluations, highlighting the performance gap between adult-trained models and classroom overlapping speech scenarios. We also show the improvements gained through finetuning and robust data training. Section IV discusses the broader implications of our findings for speech enhancement in classroom environments, particularly for children using hearing aids. Finally, Sec. V concludes the paper and outlines directions for future work.

II. METHODS

A. Problem definition

The goal of this study is to perform speech separation in naturalistic classroom scenarios involving two simultaneously

moving talkers. These talkers are either two children or a child and an adult, simulating typical classroom scenarios involving overlapping speech between peers or between a student and a teacher.

Formally, given a binaural mixture signal $y_{s_1+s_2} \in \mathbb{R}^{C \times T}$, where C is the number of channels (here $C = 2$, representing sound at the left and right ears) and T is the number of time samples, the task is to estimate the individual speech signals $\hat{s}_1$ and $\hat{s}_2$ corresponding to each active talker. We use binaural modeling to replicate the signals that would be captured by the microphones of a child with bilateral hearing aids, where spatial cues from each ear are crucial for separating talkers. In our scenario, the talkers move during the utterance, producing dynamically changing spatial cues that the model must track to maintain separation performance. This motion tracking is a central aim of our setup, as it reflects naturalistic classroom listening conditions in which talkers frequently change position.

The input waveform y may also include additive Child-Babble noise in the background, making the separation task more challenging and reflective of classroom acoustics. The separation model is trained to output time-domain waveforms for each speaker with no interference from the other speaker or background noise. This setup enables the model to not only separate overlapping speech, but also to preserve spatial distinctions associated with each moving source.

B. Simulation of overlapping speech for classroom conditions

To capture the reverberant and spatial characteristics typical of classroom environments, we developed a spatialization pipeline for generating training and evaluation data (Fig. 1). This pipeline consists of five main components, which are explained below in detail:

1. Simulation of room impulse responses (RIRs)
2. Application of head-related impulse responses (HRIRs)
3. Generation of BRIRs

4. Modeling of talkers' movement trajectories
5. Synthesis of the classroom speech data

1. RIRs

To simulate naturalistic reverberant classroom acoustics, we generated RIRs that capture direct sound, early reflections, and reverberation or echo. These RIRs were used to spatialize source signals in simulated classroom environments with varying geometry, reverberation, and source–listener distances. We used the Pyroomacoustics Python package (Scheibler *et al.*, 2018), which implements the image source method to model sound propagation in rectangular (shoebox) rooms.

A total of 30 classrooms were simulated, with dimensions randomly sampled from a range of $8.5 \times 8.5 \times 3$ m to $10 \times 10 \times 3.5$ m (length $\times$ width $\times$ height), reflecting typical U.S. classroom sizes (American National Standard Institute, 2010; Crandell and Smaldino, 2000). Reverberation times (T_{60}) were drawn uniformly from 0.2 s to 0.7 s in 0.1-s increments to model variability in acoustic conditions found in real educational settings (Acoustic Frontiers, 2025; Larson Davis, 2025).

Listener positions were randomized on a 1-m spaced grid in each room, ensuring a minimum distance of 1 m from any wall to avoid boundary effects. For each simulation, one listener position was randomly selected from this grid to introduce spatial variability across the dataset. At each of these positions, we simulated RIRs as captured by a virtual multi-microphone array. This array consisted of six DPA-4060 omnidirectional microphones arranged in orthogonal pairs along a 10-cm-diameter circle, with an Earthworks M30 or M50 microphone placed at the center.

Furthermore, we simulated varying talker–listener distances by defining circular trajectories around each listener at radii of 1.0 m, 1.5 m, and 2.0 m. Along each circle, 72 talker positions were defined in 5° steps, resulting in 2160

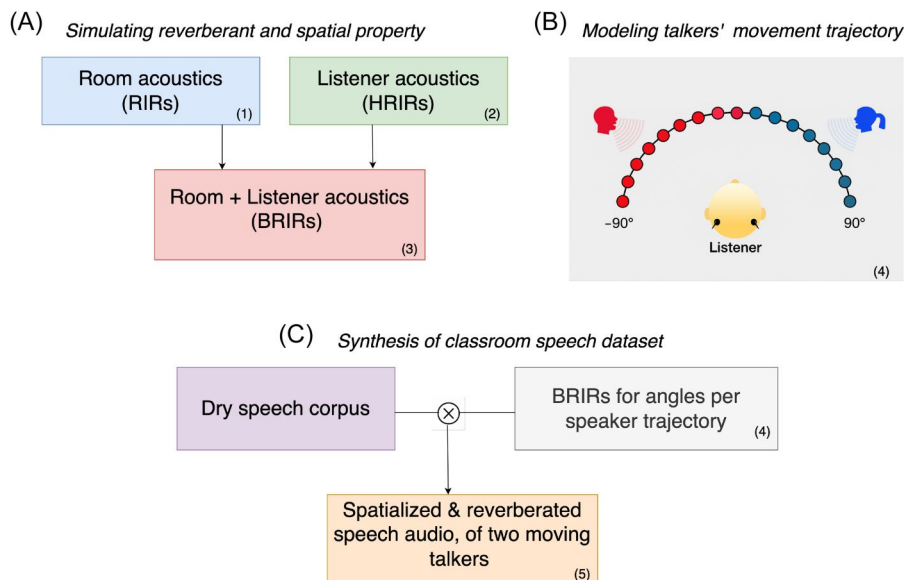


FIG. 1. Simulation pipeline illustrating the generation of reverberant and spatialized speech mixtures for classroom conditions. The process includes simulating room and listener acoustic properties (A), modeling talkers' movement trajectories (B), and synthesizing classroom speech mixtures (C). Nos. 1–5 correspond to the steps itemized in Sec. II B.

RIRs per distance setting (72 positions $\times$ 30 rooms). We chose to restrict the talker trajectories to the horizontal plane to isolate the model's ability to track dynamic azimuth shifts, without the confound of elevation-dependent spectral cues or variable temporal overlap. Importantly, the RIRs corresponding to 1.5 m and 2.0 m radii were used only at test time, allowing us to evaluate the model's ability to generalize to talker–listener distances not seen during training.

2. HRIRs

HRIRs encode the direction-dependent filtering effects of the outer ear, head, and torso on incoming sounds, providing essential cues for spatial hearing (Gardner and Martin, 1994). Unlike generic or KEMAR-based HRIRs (Burkhard and Sachs, 1975), listener-specific HRIRs enable more accurate binaural rendering. Because our focus is on supporting hearing-impaired children, all mixtures in the main experiments were rendered from the child–listener's perspective. For this, we used the CHASAR HRIR dataset (Braren and Fels, 2021), which includes measurements from children aged 5–10 years with microphones embedded in silicone hearing-aid domes, sampled on a 5° azimuth and elevation grid.

The CHASAR dataset was recorded with sources placed 1 m from the receiver. Training on this fixed distance allowed us to control for variability in spatial cues and to explicitly assess generalization to unseen distances (1.5 m and 2.0 m), which were simulated by applying inverse-square law amplitude corrections. While this approximation does not model distance-dependent spectral effects (e.g., air absorption or early reflections), it provides a reasonable proxy for energy-level differences in far-field conditions (Blauert, 1997).

3. BRIRs

BRIRs combine room-specific acoustic characteristics captured by RIRs with listener-specific spatial filtering effects modeled by HRIRs (Garí *et al.*, 2019). Convolution of a single-channel sound signal with a BRIR yields a spatialized, reverberant binaural signal that reflects both the room environment and the listener's anatomical filtering.

To generate these two-channel BRIRs, we adopted the methodology detailed in the study by Amengual Garí *et al.* (2021). This approach leverages the spatial decomposition method (Tervo *et al.*, 2013) to efficiently convert the multi-channel RIRs (generated from our virtual 7-microphone array, as described in Sec. II B 1) into directional information, which is then convolved with the listener-specific HRIRs. This process projects the spatial sound field, captured by the microphone array, onto the two ear channels of a binaural listener. This multi-stage process was applied to all simulated RIRs at each listener–talker distance, resulting in a total of 2160 BRIRs per distance condition (72 directions $\times$ 30 rooms), each being a two-channel (left and right ear) impulse response.

4. Modeling movement trajectories

To simulate spatially dynamic speech sources, we adapted the moving speaker simulation approach detailed in the study by Han *et al.* (2021). In this framework, each talker moves along a trajectory constrained to the frontal azimuthal plane, spanning from -90° to 90° with a 5° angular resolution, consistent with the spatial resolution of our BRIRs.

Each talker's trajectory is dynamically determined over the 2.4-s utterance length. They begin at a randomly selected azimuth angle, and their movement direction (clockwise or counterclockwise) is randomly assigned. Based on the assigned movement direction and their angular velocity, the movement pattern is generated. Their angular velocity is sampled uniformly between 8° and 15° per second. This angular velocity dictates how many 5° steps the talker traverses, resulting in discrete changes to the talker's spatial position (azimuth angle) over time.

This movement simulation was achieved by segmenting the 2.4-s clean speech utterance into short, time-varying blocks. For each block, a specific BRIR was selected from our pre-generated set, corresponding to the talker's angular position at that precise time within the trajectory. The selection process ensures that the sound for each segment is convolved with the BRIR associated with the talker's discrete spatial location during that segment. To minimize audible artifacts at the transitions between these segments, a short cross-fading window (5 ms) is applied between consecutive segments. This method generates spatially time-varying binaural signals with dynamic interaural cues, including ILDs and interaural time differences (ITDs), reflecting the speaker's changing location over time.

A total of 56 000 trajectory pairs were generated, split into 40 000 pairs for training, 10 000 for validation, and 6000 for testing. The full code for trajectory simulation is available here (<https://github.com/sayo20/Speech-Enhancement-for-Moving-Speakers-in-Classroom-Environment>).

5. Synthesis of the classroom speech data

We used two distinct datasets for this study: (1) a classroom speech dataset with overlapping child–child and child–adult talker pairs, and (2) an adult-only dataset used to establish a baseline for generalization performance. In both datasets, spatialized two-talker mixtures were created by convolving clean speech utterances with BRIRs sampled along predefined movement trajectories (see Sec. II B 4). All waveforms were resampled to 16 kHz, and each utterance was 2.4 s in duration, matching the simulated trajectory length. See summary in Table I and details below.

For the classroom dataset, we used the MyST corpus (Ward *et al.*, 2019) to source child utterances from speakers aged 8–11. We followed the official train/dev/test splits to ensure speaker disjointness across subsets. Low-quality segments labeled as NO_SIGNAL, SILENCE, DISCARD, SIDE_SPEECH, ECHO, or BREATH were removed. Adult speech used in mixed-age scenes was drawn from the

TABLE I. Summary of the datasets created in this study. Both datasets contain spatialized two-talker mixtures with and without background babble noise.

Dataset	Speaker types	Speech corpora	Babble corpora
Classroom	Child–child, child–adult	MyST, LibriSpeech	CSLU
Adult-only	Adult–adult	LibriSpeech	WSJ

LibriSpeech corpus (Panayotov *et al.*, 2015). We simulated two overlapping speech types: (1) two moving children, and (2) a moving child and a moving adult pair, both with and without background Child-Babble. A diffuse babble noise field was generated using the CSLU corpus (Cole *et al.*, 1998), with overlapping utterances (70% overlap), spatialized using BRIRs from 3–8 random locations following Francel and McDermott (2022).

The adult-only dataset was constructed solely from LibriSpeech utterances, following the mixture and movement framework described above. Two moving adult speakers were selected per mixture, and an additional adult babble stream was optionally added using utterances from the WSJ corpus (Paul and Baker, 1992). The adult babble diffuse noise field followed the same overlap and spatialization method described above for generating the children’s babble noise field.

Across both datasets, we generated 40 000 training mixtures, 10 000 for validation, and 6000 for testing. The SNRs for the speech mixtures were sampled between 0 and 5 dB, while speech-to-babble SNRs ranged from –2.5 to 15 dB.

To highlight acoustic differences between adult and child speech, Fig. 2 presents annotated sample spectrograms of example utterances. As shown, child speech exhibits a higher fundamental frequency (F_0) due to the faster vibration rate of smaller vocal folds, producing more widely separated harmonics (Lee *et al.*, 1999). Furthermore, the child’s shorter vocal tract elevates formant frequencies and widens their spacing, resulting in a broader distribution of spectral energy compared to adult speech (Vorperian and Kent, 2007). The child signal also displays greater pitch modulation and less stable harmonic structure over time (Assmann and Katz, 2000). In contrast, adult speech shows lower and more closely spaced formants, and a more stable harmonic

structure over time (Assmann and Katz, 2000; Lee *et al.*, 1999).

These differences in vocal tract characteristics, F_0 , and spectral shape are known to affect both perceptual intelligibility and the behavior of enhancement models. Furthermore, prior work has shown that children’s voices tend to be more spectrally similar to one another than adult voices (Gerosa *et al.*, 2006; Lee *et al.*, 1999; Schuster *et al.*, 2014). This spectral overlap reduces the distinctiveness between speakers in a mixture, making source separation more challenging. In contrast, adult voices exhibit more inter-speaker spectral variability, providing models with stronger cues to differentiate speakers. This spectral similarity among children may help explain why models trained on adult speech could fail to generalize to child–child mixtures, where fewer acoustic cues are available for speaker discrimination.

C. Binaural speech separation model

We adopt a MIMO-TasNet-based architecture, as described by Han *et al.* (2021), which is designed for real-time binaural speech separation and achieves a latency as low as 2 ms. The model comprises two main components: a separation module and an enhancement module (Fig. 3). The DoA estimation module, which shares architectural similarities with the enhancement module, is exclusively used for *post hoc* evaluation and operates on the enhanced outputs to classify the talker’s trajectory. The entire model has 1.9×10^6 parameters.

The separation module processes the binaural mixture, leveraging both spectral and spatial cues, including interaural phase differences and ILDs, which are crucial for source localization. These cues, combined with learned representations from an encoder, are passed through a temporal convolutional network (TCN) to estimate time-domain source masks. The output consists of initial binaural waveform estimates corresponding to the two speakers in the mixture.

These initial estimates are then passed through the enhancement module, which shares an identical TCN-based architecture with the separation module but operates with independent parameters. This module takes as input both the noisy binaural speech mixture (y) and the initial estimated

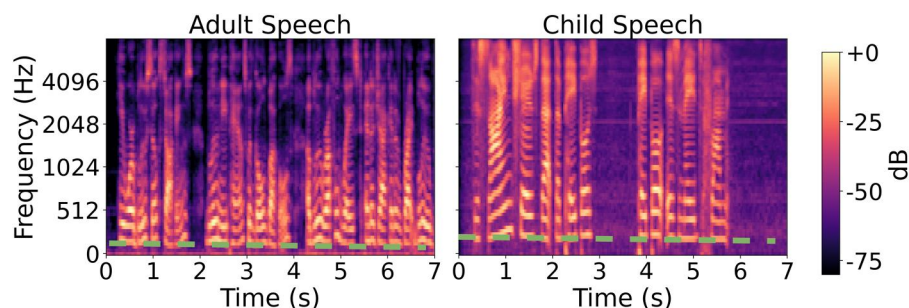


FIG. 2. Spectrogram of utterances from an adult speaker (left) and a child speaker (right). The child’s speech shows higher fundamental frequency (F_0) and greater pitch variation (visible in the more widely spaced and fluctuating harmonic structure), with energy distributed across a wider frequency range. In contrast, the adult’s speech has a lower F_0 and denser harmonic spacing. The green dashed lines indicate estimated F_0 .

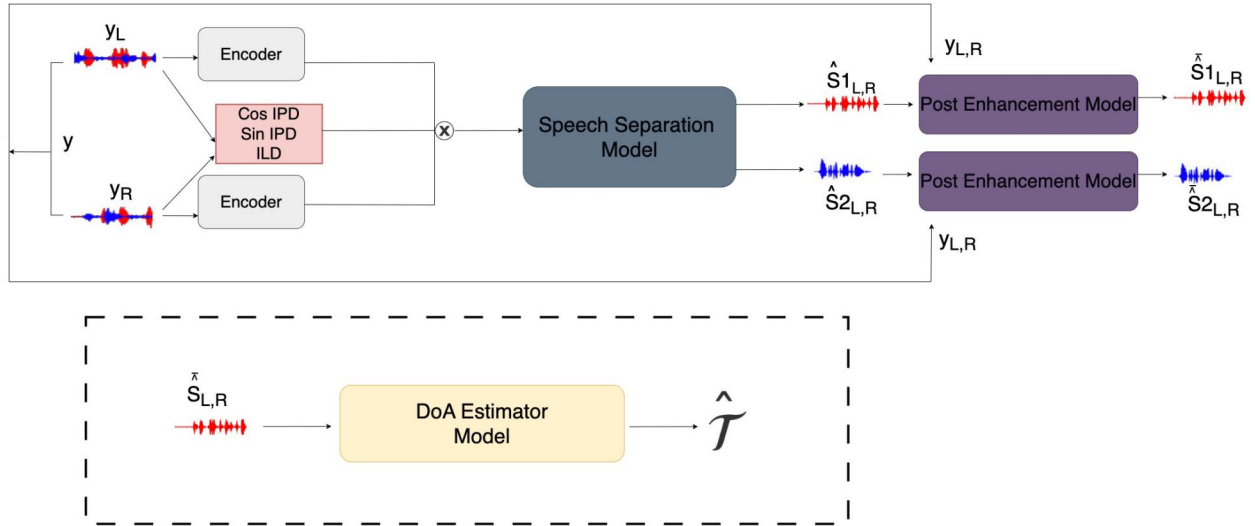


FIG. 3. Overall architecture of the proposed binaural speech separation and enhancement system (Han *et al.*, 2021). The model processes left and right ear mixed signals (y_L, y_R) to output separated ($\hat{s}_{1,L,R}, \hat{s}_{2,L,R}$) and subsequently enhanced speech signals ($\bar{s}_{1,L,R}, \bar{s}_{2,L,R}$) for two sources. To evaluate whether spatial cues are preserved after enhancement, each enhanced binaural signal is further passed through a dedicated DoA estimation module, which predicts the estimated speaker trajectories ($\hat{\mathcal{T}}$). The DoA estimator is used exclusively for evaluation and is trained independently, ensuring it does not influence the separation or enhancement stages.

speech signals ($\hat{s}$). Its purpose is to refine and produce a cleaner version of the separated speech signals.

During training, we optimized the SNR loss between the estimated and reference signals, encouraging accurate waveform reconstruction while preserving spatial cues. The SNR is defined as

$$\text{SNR}(s, \hat{s}) = 10 \log_{10} \frac{\|s\|^2}{\|s - \hat{s}\|^2}, \quad (1)$$

where (s) and ($\hat{s}$) denote the reference and estimated speech signals, respectively.

To correctly match each separated speaker output to its corresponding reference signal, we applied permutation-invariant training (Yu *et al.*, 2017). Specifically, we minimized the loss over all possible speaker permutations

$$\mathcal{L} = \min_{\pi \in P} \sum_{c=1}^C [\text{SNR}(s_{\pi(c)}^L, \hat{s}_c^L) + \text{SNR}(s_{\pi(c)}^R, \hat{s}_c^R)], \quad (2)$$

where (P) is the set of all ($C!$) permutations, and C is the number of speakers. The same permutation (π) is applied to both left and right channels to maintain consistent speaker ordering across ears.

To evaluate whether the enhanced outputs successfully retained spatial information about the moving speakers, a DoA estimation module was employed. This DoA model was trained independently, only after the separation and enhancement modules had been completed, ensuring its sole purpose was to assess the preservation of spatial cues. Its architecture adopts the same TCN-based structure as the enhancement module but is trained as a classifier over discretized azimuth angles, following the approach described by Han *et al.* (2021) (Fig. 3). The stereo signals from the

enhancement output, $\hat{s}_{1,L,R}$ and $\hat{s}_{2,L,R}$, are encoded, passed through temporal convolutional layers, and mapped to per-frame DoA class probabilities. Let $\hat{\Theta}(t) \in \{\theta_1, \dots, \theta_K\}$ denote the predicted DoA class at time t . Each frame is labeled with a single ground-truth azimuth, and training is performed with a cross-entropy loss. During inference, these predictions are aggregated in fixed-length chunks, producing an estimated trajectory $\hat{\mathcal{T}} = \{\hat{\theta}_1, \dots, \hat{\theta}_T\}$. By training the DoA estimator separately in this way, we ensure that its performance directly reflects how well the enhancement module preserved directional cues, rather than any confounding effects from joint optimization with the separation stage.

The model was implemented in the PyTorch framework (Paszke *et al.*, 2019) and trained for 100 epochs using the Adam optimizer with an initial learning rate of 1×10^{-3} . An exponential learning rate scheduler (ExponentialLR) was applied with a decay factor of 0.98 every two epochs during the main training phase. During subsequent finetuning, the decay was applied every 5 epochs.

D. Computational experiments

All models in this study share the same MIMO-TasNet architecture described in Sec. II C, but differ in the training data composition and noise conditions. Importantly, all experiments were rendered from the child-listener perspective using CHASAR BRIRs, ensuring that the simulated mixtures reflect the signals received by a child with bilateral hearing aids. To avoid ambiguity, models are named according to (1) the type of speech data used for training (adult vs classroom) and (2) whether babble noise was present during training (clean vs babble).

Adult-Clean and Adult-Babble were trained exclusively on the adult dataset (adult-adult overlapping talkers), without and with background babble noise, respectively. These

conditions serve as baselines and mirror common practice in prior work (Chen *et al.*, 2018; Gu *et al.*, 2019, 2020; Han *et al.*, 2021; Wang and Wang, 2019), where models are trained primarily on adult speech. Evaluating these models on the classroom dataset allows us to assess their ability to generalize to child-dominant acoustic scenes.

Classroom-Clean and Classroom-Babble were trained entirely on the classroom dataset described in Sec. II B 5, including child–child and child–adult overlapping talkers, again without and with background babble noise, respectively. These models reflect a child-aware training regimen that explicitly targets the acoustic variability present in children’s speech and naturalistic classroom environments.

Finetuned-Clean and Finetuned-Babble were initialized from the corresponding adult models and then finetuned using only 50% of the classroom training data. These configurations test whether prior knowledge from adult speech separation can accelerate learning or improve generalization to children’s speech with reduced data requirements.

All models were evaluated on the classroom test set, referred to as Class-Clean (with no babble) and Class-Babble (with a babble background), allowing us to compare their performance in naturalistic school-like environments. We report the SNR and the SNR improvement (SNRi) to quantify the similarity between the separated output and the reference clean speech. We also report the estimated DoA error, which measures the mean absolute angular deviation between the model’s estimated talker position and the true position. Lower DoA is indicative of a more accurate spatial preservation. We further examine how training with or without babble noise influences robustness under noisy evaluation conditions.

III. RESULTS

The results are organized to address three main questions. First, we ask whether models trained exclusively on adult talkers can generalize to child-listener classroom conditions. Within this analysis, we consider two dimensions of generalization: (1) noise robustness, comparing performance on clean versus babble mixtures, and (2) voice-type effects, comparing adult–child versus child–child overlaps. Second, we evaluate whether training directly on classroom-specific data or finetuning an adult-trained model improves separation and localization performance in the classroom environment. Finally, we test the robustness of all models to additional factors, including the presence of babble during training and generalization to unseen source–listener distances.

A. MIMO-model trained on adult-only data

A common issue in speech enhancement is the observable drop in performance when algorithms trained on adult voices are applied to children’s speech. Existing studies typically use a single microphone. In the current study, we utilize recordings from two microphones placed on a bilateral hearing aid in order to leverage the inherent spatial cues that

exist in the audio scenes, as well as introduce precomputed cues computed from both microphones. Our first objective is to determine if additional spatial and bilateral cues mitigate the performance drop typically observed when using adult-trained models on children’s speech.

1. Noise generalization

Table II summarizes how the model trained on adult-only data with and without babble performs on two different test sets.

When the model trained on the Adult-Clean data is evaluated on the adult-clean test data, we observe an SNRi of 11.37 dB. When the same model is evaluated on the Class-Clean test data, we observe a SNRi of 10.81 dB. While this reduction is small, it was statistically significant (Mann–Whitney $U = 7.80 \times 10^7$, $p < 10^{-29}$, $r = -0.084$). Furthermore, we see that the Adult-Clean model was better able to track the movement in the Adult-Clean test data (DoA error = 3.82°) in comparison to the movements in the Class-Clean test data (DoA error = 4.90°; Mann–Whitney $U = 6.3 \times 10^7$, $p < 10^{-61}$, $r = 0.123$).

When the model trained on the Adult-Babble data is evaluated on the Adult-Babble test data, we observe an SNRi of 9.92 dB. When the same model is evaluated on the Class-Babble test data, we observe an SNRi of 8.50 dB. The difference observed was also statistically significant (Mann–Whitney $U = 8.91 \times 10^7$, $p < 10^{-221}$, $r = -0.237$). Although overall speech separation performance (SNRi) dropped, indicating a clear domain shift challenge, the DoA error demonstrated a conflicting trend. The model showed a statistically significant, albeit small, improvement in tracking movements in the Class-Babble dataset (10.81° vs 11.37°; Mann–Whitney $U = 8.13 \times 10^7$, $p < 10^{-68}$, $r = -0.13$, where the negative r reflects the reduction in error). This result suggests that, while the spectral characteristics of the Child-Babble interfere with mask estimation (reducing SNRi), the spatial cues required for DoA estimation remain robust, or are even slightly clearer, in the simulated classroom environment.

TABLE II. Performance of adult-trained models on in-domain (adult test sets) and out-of-domain (classroom test sets). The Adult-Clean model was trained without babble noise and tested on Adult-Clean and Classroom-Clean sets, while the Adult-Babble model was trained with babble noise and tested on Adult-Babble and Classroom-Babble sets.

Test data	Adult-Clean model	
	SNRi (dB) ↑	DoA (°) ↓
Adult-Clean	11.37	3.82
Class-Clean	10.81	4.90
	Adult-Babble model	
	SNRi (dB) ↑	DoA (°) ↓
Adult-Babble	9.92	11.37
Class-Babble	8.50	10.81

TABLE III. Performance of Adult-Clean and Adult-Babble models on classroom mixtures, partitioned by overlap type (adult-child vs child-child). The Adult-Clean model is evaluated on the Classroom-Clean test set, while the Adult-Babble model is evaluated on the Classroom-Babble test set.

Overlap type	Adult-Clean model	
	SNRi (dB) ↑	DoA (°) ↓
Adult-Child	11.38	4.56
Child-Child	10.25	5.24
	Adult-Babble model	
	SNRi (dB) ↑	DoA (°) ↓
Adult-Child	8.92	10.59
Child-Child	8.08	11.03

Taken together, we observed that introducing background babble leads to a higher mismatch, likely because the properties of babble noise differ between adults' and children's voices.

2. Voice-type generalization

As described in Sec. II B 5, our classroom dataset is composed of adult-child talker pairs and child-child talker pairs. We evaluate the performance of the Adult-Clean and Adult-Babble models across these two speaker configurations (Table III).

We observed a performance drop across both adult-child and child-child pairings in both clean and babble conditions. For the Adult-Clean model, SNRi decreased from 11.38 dB in adult-child mixtures to 10.25 dB in child-child mixtures (Mann-Whitney $U = 2.03 \times 10^7$, $p < 10^{-32}$, $r = -0.125$). Similarly, the model had a better DoA error of 4.56° in adult-child mixtures than 5.24° in child-child mixtures (Mann-Whitney $U = 1.63 \times 10^7$, $p < 10^{-19}$, $r = 0.093$). This reduction, even in the absence of background noise, shows that the spectral mismatch between children's voices still makes separation more difficult for adult-trained models.

TABLE IV. Performance of models trained without and with babble noise, evaluated on classroom mixtures with and without babble. All models were trained and evaluated from the child-listener perspective.

Models trained without babble noise						
Test data	Adult-Clean		Classroom-Clean		Finetuned-Clean	
	SNRi (dB) ↑	DoA (°) ↓	SNRi (dB) ↑	DoA (°) ↓	SNRi (dB) ↑	DoA (°) ↓
Class-Clean	10.81	4.90	10.80	4.68	11.65	4.11
Class-Babble	7.77	13.85	8.01	13.05	8.00	12.58
Models trained with babble noise						
	Adult-Babble		Classroom-Babble		Finetuned-Babble	
	SNRi (dB) ↑	DoA (°) ↓	SNRi (dB) ↑	DoA (°) ↓	SNRi (dB) ↑	DoA (°) ↓
Class-Clean	9.76	6.36	10.48	5.74	11.05	4.99
Class-Babble	8.50	10.81	10.62	6.78	11.02	6.55

The same pattern is observed under babble noise. For the Adult-Babble model, SNRi dropped from 8.92 dB (adult-child) to 8.08 dB (child-child; Mann-Whitney $U = 2.02 \times 10^7$, $p < 10^{-29}$, $r = -0.120$). However, with the introduction of children babble, the model performed similarly in the localization of the adult-child pairs (DoA error = 10.59°) and the child-child pairs (DoA error = 11.03°), as the difference was not statistically significant (Mann-Whitney $U = 1.83 \times 10^7$, $p = 0.06$, $r = -0.20$).

Taken together, we observe that, when both talkers are children, the combined effects of spectral similarity and environmental noise impose a significant challenge on adult-trained models. Nevertheless, we see that the model still maintains its ability to separate and track the child talkers when it can leverage the binaural audio and spatial cues.

B. MIMO-model trained on adult data vs MIMO-model trained on classroom data

In this section, we test whether direct exposure to domain-specific data improves performance and robustness in the classroom, especially with child-child talkers. To investigate this, we evaluated the effect of training the model directly on the classroom dataset.

1. Performance in clean classroom conditions

We first compared the Adult-Clean and Classroom-Clean models on the Class-Clean test data [Table IV; Fig. 4(A)]. The two models achieved nearly identical overall performance (SNRi = 10.81 vs 10.80 dB, DoA = 4.90° vs 4.68°). This suggests that, in the absence of babble, the spatial cues are sufficient to allow adult-trained models to generalize remarkably well to classroom-related speech mixtures.

However, when we zoom into the different talker pairs in the Class-Clean data, a consistent gap remained between the model performance on adult-child and child-child talkers. The Adult-Clean model had an SNRi of 11.38 dB and DoA error of 4.56° on the adult-child pair and 10.25 dB and

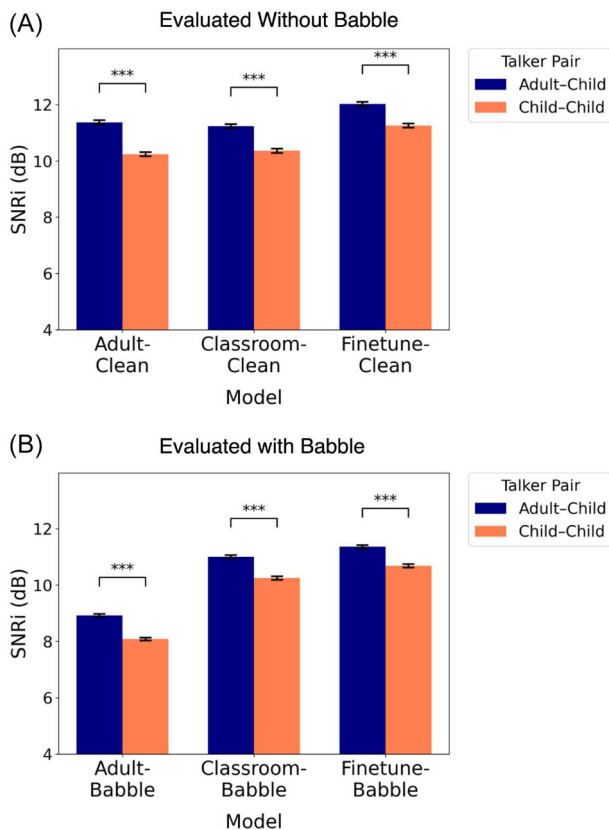


FIG. 4. Speech separation performance of MIMO models trained on different datasets and evaluated under varying noise conditions. Bars represent mean SNRi for models trained on Adult, Class, and Finetuned splits, evaluated on the Class dataset with Adult-Child (blue) and Child-Child (orange) talker pairs. (A) Performance of models trained and evaluated in babble-free conditions. (B) Performance of models trained and evaluated in conditions with background babble. Error bars indicate SEM. *** $p < 0.001$ (Mann-Whitney U test).

DoA error of 5.24° on the child-child pair (drop in SNRi = 1.13 dB, increase in DoA error = 0.68° ; Mann-Whitney $U = 2.03 \times 10^7$, $p < 10^{-32}$, $r = -0.125$, FDR-corrected). While the Classroom-Clean model had a SNRi of 11.24 dB and DoA error of 4.89° on the adult-child pair and 10.37 dB and DoA error of 4.47° on the child-child pair (drop in SNRi = 0.87 dB, decrease in DoA error = 0.42° ; Mann-Whitney $U = 1.96 \times 10^7$, $p < 10^{-16}$, $r = -0.089$, FDR-corrected).

Taken together, we observed that, while domain-specific training provides a slight improvement for the most challenging case (child-child mixtures), it does not offer a substantial advantage over the Adult-Clean model in clean classroom conditions.

2. Performance in noisy classroom conditions

In babble conditions, domain-specific training conferred a clearer advantage [Table IV; Fig. 4(B)].

When evaluated on the Class-Babble test data, the Classroom-Babble model achieved an SNRi of 10.62 dB and a DoA of 6.78° , which is a substantial improvement over the Adult-Babble model's performance (SNRi: 8.50 dB, DoA

error: 10.81° ; Mann-Whitney $U = 5.16 \times 10^7$, $p < 10^{-60}$, $r = -0.28$, FDR-corrected). This demonstrates that direct exposure to the specific acoustic properties of babble noise and children's voices is critical for robust performance in noisy environments.

The Classroom-Babble model also performed better on the child-child talker pairs, with an SNRi of 10.25 dB, and DoA error of 6.94° . While the Adult-Babble model achieved an SNRi of 8.08 dB, and DoA error of 10.78° on the child-child talker pairs. Furthermore, we observed a smaller performance drop between the talker pairs for the Classroom-Babble model. The Adult-Babble model showed a 0.84 dB drop in SNRi between adult-child and child-child talkers (Mann-Whitney $U = 2.02 \times 10^7$, $p < 10^{-30}$, $r = -0.12$, FDR-corrected), while the drop in the Classroom-Babble model was 0.76 dB (Mann-Whitney $U = 1.96 \times 10^7$, $p < 10^{-17}$, $r = -0.088$, FDR-corrected) meaning the difference between the talker pairs within this model was negligible [Fig. 4(B)].

Taken together, these results show that, while using bin-aural audio and providing spatial cues can alleviate a significant portion of the performance loss seen with adult-trained models, they still suffer in noisy environments.

C. MIMO-model trained on adult data finetuned on classroom dataset

While full-domain training on classroom data is beneficial, particularly in noisy conditions, acquiring a large, high-quality children's speech corpus is a well-known challenge. This makes children's speech a low-resource domain. Given that the adult-trained model already shows strong generalization in clean conditions due to spatial cues, it suggests that the model's foundational knowledge of speech separation is transferable. Therefore, we investigate a more data-efficient approach: finetuning the robust adult-trained model with a limited amount of classroom data. This allows us to determine whether we can achieve performance comparable to a fully trained classroom model without the need for a large, specialized dataset.

1. Comparison of Classroom-Clean and Finetuned-Clean models

The Finetuned-Clean model (Table IV) showed higher overall performance across both metrics compared to the Classroom-Clean model, achieving an SNRi of 11.65 dB (vs 10.80 dB) and a lower DoA error of 4.11° (vs 4.68°), despite being trained on only half the amount of classroom scenes (Mann-Whitney $U = 6.39 \times 10^7$, $p < 10^{-50}$, $r = -0.11$, FDR-corrected).

Furthermore, finetuning showed a similar drop in performance between the adult-child and the child-child talker pairs. The Finetuned-Clean model exhibited an SNRi drop of 0.77 dB between the adult-child and child-child pairings (Mann-Whitney $U = 1.9 \times 10^6$, $p < 10^{-15}$, $r = -0.08$, FDR-corrected), compared to the 0.84 dB drop observed in the Classroom-Clean model [Fig. 4(A)]. While the

Finetuned-Clean model yields marginally better SNRi and DoA values, the performance of both models is comparable under these conditions.

These findings demonstrate that finetuning effectively transfers spectral priors learned from adult speech representations to the children's classroom domain. This strategy is viable in clean conditions, as it leverages the model's existing knowledge of speech acoustics without the added noise complexity.

2. Comparison of Classroom-Babble and Finetuned-Babble models

On the Class-Babble test data (Table IV), the Finetuned-Babble model achieved an SNRi of 11.02 dB and a DoA error of 6.55°, which is marginally higher than the Classroom-Babble model's 10.62 dB SNRi and 6.78° DoA error (Mann-Whitney $U = 6.79 \times 10^7$, $p < 10^{-14}$, $r = -0.06$, FDR-corrected). As with the Finetuned-Clean model, the Finetuned-Babble model was trained with only half of the Classroom scenes, as was used in the Classroom-Babble model.

The Finetuned-Babble model also performed slightly better when handling the child-child talker pairs (SNRi = 10.68 dB, DoA error = 6.55°), compared to the fully trained Classroom-Babble model (SNRi = 10.25 dB, DoA error = 6.94°). Additionally, the Finetuned-Babble model had a smaller 0.68 dB SNRi drop between the adult-child and child-child talker pairs (Mann-Whitney $U = 1.9 \times 10^6$, $p < 10^{-14}$, $r = -0.08$, FDR-corrected), compared to the 0.76 dB drop observed in the Classroom-Babble model [Fig. 4(B)]. Similar to what was observed with the Finetuned-Clean model, the Finetuned-Babble model shows comparable performance to the full training done with the Classroom-Babble model.

These results collectively demonstrate that finetuning an adult-trained model with a limited amount of domain-specific data is not just a viable alternative but could be a better data-efficient strategy for speech separation in a noisy classroom.

D. Robustness to unseen training conditions

To assess the model's robustness and generalization capabilities, we evaluated its performance on acoustic conditions on which it was not explicitly trained. We investigate two key scenarios: (1) Noise generalization: We first examine how models trained with and without babble noise perform when exposed to the opposite condition. (2) Distance generalization: We then take the worst and best performing model trained on the noisy condition (Adult-Babble, Finetuned-Babble) and test its ability to generalize to unseen talker distances.

1. Noise generalization; impact of babble during training

The results (Table IV; Fig. 5) showed a clear asymmetry in the model's robustness. Models trained without babble noise, Adult-Clean, Classroom-Clean, and Finetuned-Clean

demonstrated significant performance degradation when tested on noisy mixtures (Table IV). For example, Finetuned-Clean's SNRi dropped from an SNRi of 11.65 dB and a DoA error of 4.11° on clean mixtures to an SNRi of only 8.00 dB and a DoA error of 12.58° on noisy mixtures [Fig. 5(A); Mann-Whitney $U = 4.3 \times 10^7$, $p < 10^{-60}$, $r = -0.02$, FDR-corrected], highlighting the vulnerability of the clean-trained models to unseen noise, which is common in a classroom environment.

In contrast, models trained with babble noise showed better generalization. These models not only performed robustly in noisy conditions but also transferred well to clean environments. The Finetuned-Babble model demonstrated the strongest generalization. It maintained a high SNRi of 11.05 dB and a DoA error of 4.99° in clean conditions, and an SNRi of 11.02 dB and DoA error of 6.55° in noisy conditions. The Classroom-Babble and Adult-Babble models also showed good generalization, with performance on clean and noisy data being closely matched (10.48 dB and 10.62 dB for Classroom-Babble; 9.76 dB and 8.50 dB for Adult-Babble).

2. Generalization to unseen talker distances

To assess the models' ability to track talkers in different spatial configurations, we evaluated the Adult-Babble and

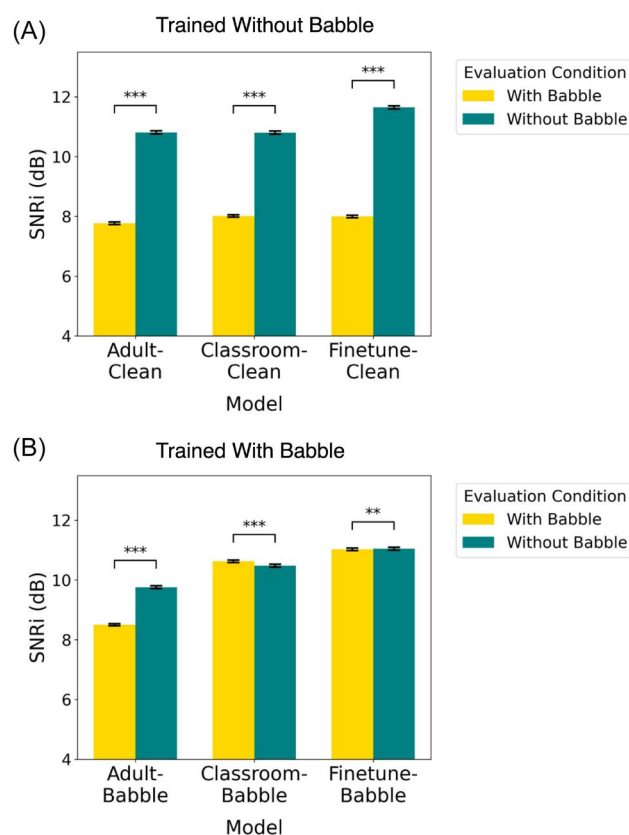


FIG. 5. SNR performance of three models under clean and noisy evaluation conditions. (A) Models trained without babble noise. (B) Models trained with babble noise. All models were evaluated on classroom data with and without background babble. Teal bars represent evaluations without babble. Yellow bars represent evaluations with babble. Error bars indicate SEM. *** $p < 0.01$; **** $p < 0.001$ (Mann-Whitney U test).

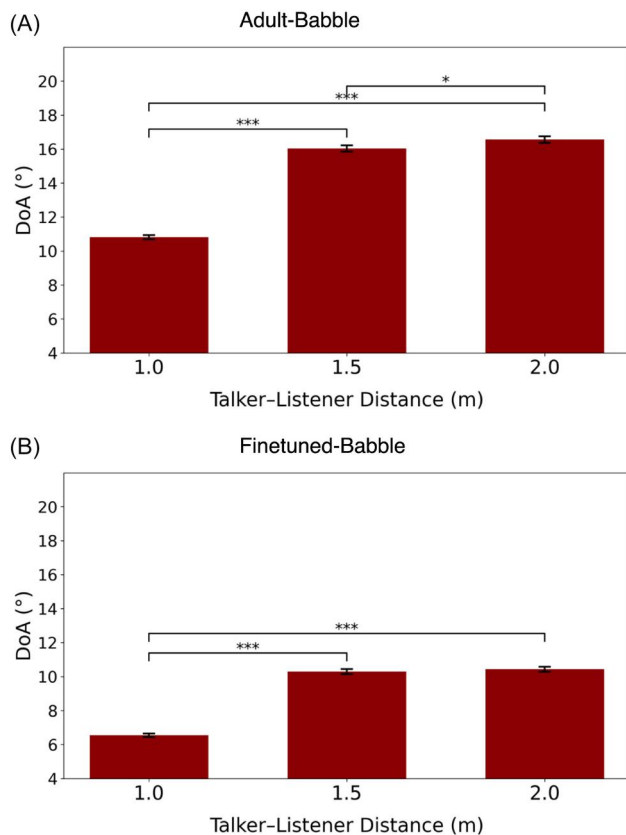


FIG. 6. Average DoA achieved by the Adult-Babble (A) and Finetuned-Babble (B) when evaluated on the Class-Babble test set at three talker-listener distances: 1.0 m, 1.5 m, and 2.0 m. Error bars indicate SEM. Asterisks denote statistically significant differences between distances: * $p < 0.05$, *** $p < 0.001$ (Mann-Whitney U test, FDR-corrected).

Finetuned-Babble models on the Class-Babble test set with talkers at increased distances of 1.5 m and 2.0 m, as they were only trained on 1.0 m configurations. This setup was made to simulate some variability in the speaker-listener distances.

As shown in Fig. 6, both the Adult-Babble and Finetuned-Babble models showed a decline in performance as talker distance increased. However, the Finetuned-Babble model demonstrated more robustness to the unseen talker distances. The Finetuned-Babble model's DoA error increased by 3.74° between 1.0 m and 1.5 m, and by 3.89° between 1.0 m and 2.0 m. In contrast, the Adult-Babble model's error increased by 5.22° and 5.75° , respectively.

IV. DISCUSSION

This study explored how DNN-based speech separation can be adapted to classroom environments and children's speech to provide effective speech enhancement for hearing-impaired children. Our approach leverages binaural audio, which consists of simulated yet naturalistic acoustic scenes with two moving talkers. Our findings demonstrate that, while the model trained solely on binaural adult speech recordings can generalize to children's speech in quiet conditions, its performance significantly degrades in the

presence of children's babble noise. Crucially, we show that training on the classroom dataset and, even more effectively, finetuning a pre-trained model are essential strategies for achieving robust performance in noisy classroom settings. The success of these adaptation methods can have direct and important implications for developing adaptive hearing aids and assistive listening devices that can better serve children in dynamic educational settings.

Prior work has consistently shown that models trained on adult speech generalize poorly to children's speech, primarily due to key spectral differences, such as higher fundamental frequencies and less stable harmonics (Attia *et al.*, 2025; Potamianos *et al.*, 1997; Shivakumar and Georgiou, 2020). These studies, however, are almost exclusively based on monaural recordings, where separation relies heavily on spectral distinctiveness of the talkers, e.g., the dissimilarity between adult and female talkers (Wang *et al.*, 2018, 2020). We hypothesized that a binaural approach could mitigate this challenge. Based on our previous work (Olaire *et al.*, 2025), which showed that a binaural model can effectively leverage spatial cues (e.g., ITDs and ILDs) to resolve spectrally similar adult talkers, we hypothesized that this spatial mechanism could be robust enough to overcome the spectral challenges inherent in child-talker separation. Our results confirm that, in clean conditions, the performance drop between the Adult-Only model (trained exclusively on adult pairs) and the Classroom model (trained on mixed adult-child and child-child pairs) is minimal. This finding is significant as it suggests that, in the absence of noise, the spatial cues alone are sufficient to reduce the spectral mismatch barrier observed in speech tasks using monaural recordings. While monaural separation relies on talker-specific spectral features like pitch and harmonics (Wang *et al.*, 2018), binaural models exploit relative time and level differences between the ears (Blauert, 1997). Because these spatial features depend primarily on room and head geometry rather than talker-specific features, the adult model could generalize effectively to children's speech in quiet environments without prior exposure to child speech training data.

This spatial advantage leveraged by the adult-trained model in clean conditions, however, diminishes in the presence of babble noise, which is a defining characteristic of real-world classroom environments. Our results suggest that the acoustic nature of the noise itself, rather than the target voice mismatch alone, poses the primary barrier. In clean conditions, binaural models effectively separate target talkers by exploiting localized spatial features (Blauert, 1997; Bronkhorst, 2000). In contrast, background babble in a classroom setting is a diffused noise source, meaning it originates from multiple randomized locations, which degrades the quality of the interaural spatial cues (Faller and Merimaa, 2004). When the spatial features become corrupted by noise, the models may depend more heavily on spectral structures to estimate separation masks (Olaire *et al.*, 2025). Hence, the model trained exclusively on adult acoustic scenes lacks the spectral priors needed to separate target child speech from background Child-Babble. These

findings demonstrate that diffuse noise actively impairs spatial separation, underscoring the necessity of incorporating domain-specific noise during training. To address the resulting need for domain-specific noise training in this low-resource domain, we investigated finetuning on Classroom data as an efficient strategy (Karunanayake *et al.*, 2019; Shivakumar and Georgiou, 2020). The Finetuned-Babble model performed better than the Adult-Babble model on the Class data. This demonstrates that the spectral priors learned from adult speech and adult babble provide a strong foundation for rapid adaptation to children's speech in both clean and noisy environments, using only 50% of the target data. This hybrid approach is valuable for low-resource domains and also offers practical benefits, such as reducing training time from 17 h to just 8 h in our case. We therefore recommend prioritizing resource-efficient adaptation methods, such as finetuning with limited additional data, while in parallel expanding high-quality datasets of children's speech to enable effective adaptation. While our finetuning strategy was simple, more sophisticated methods, such as discriminative layer-wise tuning or parameter-efficient transfer methods (Howard and Ruder, 2018; Hu *et al.*, 2022), could unlock even greater improvements.

Beyond speech separation, our models preserved the spatial locations of the moving talkers in noise, as evidenced by lower DoA tracking errors. Accurate sound localization helps listeners orient their attention and track talkers in multi-speaker scenes (Bronkhorst, 2000; Litovsky *et al.*, 2006). Interestingly, the Classroom-Babble and Finetuned-Babble models maintained relatively low trajectory tracking errors (DoA). This is likely attributed to the sparse temporal and spectral structure of speech (Cooke, 2006). While background babble reduces overall interaural coherence across time, speech signals contain brief time-frequency glimpses where the target talker's SNR remains high. In these brief intervals, the model can still extract reliable ITDs/ILDs from the direct-path sound. Additionally, low-frequency ITDs remain relatively stable in diffuse noise, which could provide consistent phase cues for localization (Faller and Merimaa, 2004). Also, training with babble noise as data augmentation likely forces the model to rely on these localized phase cues rather than spectral features that are easily corrupted by background noise (Adavanne *et al.*, 2019; Chakrabarty and Habets, 2019). This would therefore explain why the babble-trained models outperformed clean-trained models across both clean and noisy conditions while the clean-trained models failed to generalize to noisy tracking scenarios.

When translating these objective improvements to practical benefits, not all observed model differences are equally meaningful. In clean conditions, the Finetuned-Clean model achieved a marginal 0.85 dB SNRi gain and a 0.57° reduction in DoA error compared to the Classroom-Clean model, both corresponding to small effect sizes ($r = -0.11$ and $r = 0.06$, respectively). In practice, the 0.57° DoA difference falls below the human minimum audible angle of 1°–2° under ideal midline conditions (Mills, 1958), while the

0.85 dB SNRi difference is likewise minimal. The spatial tracking and separation capabilities of the two models, therefore, appear practically equivalent in quiet settings. In contrast to quiet conditions, the Classroom-Babble model showed a 4.03° lower DoA error ($r = 0.33$) alongside a 2.12 dB higher SNRi ($r = -0.28$) compared to the Adult-Babble baseline, representing moderate effect sizes that could be functionally beneficial. In complex auditory environments, minimizing directional errors preserves the integrity of binaural cues needed to segregate target speech from background noise (Bronkhorst, 2000). Such tracking improvements could be leveraged in assistive devices to provide better directional estimates needed to steer downstream neural beamformers in hearing aids. Furthermore, SNRi gains have also been conceptually linked to reduced listening effort (Pichora-Fuller *et al.*, 2016). While these gains do not translate linearly to intelligibility observed in human listening tests, prior work (Thoidis and Goehring, 2024) showed that higher SNRi corresponded to better speech understanding in hearing-impaired adult listeners. Whether these relationships hold for children in classroom environments remains an open question, as children require higher SNRs than adults to achieve equivalent comprehension (Howard *et al.*, 2010). Behavioral testing with child listeners remains a necessary next step to validate whether these objective gains produce meaningful real-world benefits.

We imposed some methodological constraints to help with understanding and interpreting the results. First, we relied on simulated RIRs/BRIRs and child HRIRs measured at a fixed 1 m distance, rather than real recordings. The first reason for this is that there is no sufficiently large children's BRIR corpus available, especially with the addition of the classroom constraint. By simulating under this constraint, we do not sufficiently capture the acoustic scattering that classroom furniture would produce in a real room (Bistafa and Bradley, 2000; Savioja and Svensson, 2015). On the other hand, a simulation-based approach provides a reproducible ground truth for spatial tracking and lets us control for head-movement variability that would be difficult to isolate in live classroom recordings. Another constraint we introduced was restricting talker trajectories to the horizontal plane and using fixed utterance durations. Restricting trajectories to the horizontal plane allowed us to focus on azimuth tracking within the classroom, such as a teacher walking in front of a whiteboard, without introducing elevation-dependent spectral variations, while fixed durations ensured we evaluated continuous tracking performance under constant signal overlap. As a result, we do not account for expected vertical head tilts, median-plane localization, or the elevation-dependent spectral cues introduced by the pinnae (Blauert, 1997), nor do we model the variable temporal overlap of natural classroom speech. Consequently, our findings address front-facing tracking scenarios rather than the fuller range of head and body movements or intermittent speech overlap typical of classroom interactions.

Finally, our findings on the model's robustness to unseen talker distances are encouraging. Despite being trained only at 1 m, the Finetuned-Babble model maintained

functional separation and spatial tracking at 1.5 m and 2.0 m, suggesting that it had internalized scale-invariant spatial patterns rather than distance-specific artifacts of its training configuration. We simulated these distances using an inverse-square law amplitude correction. Using this method only accounts for the overall energy loss at a distance and fails to model how a real room introduces early reflections or spectral filtering (Kuttruff, 2016). In a real classroom, moving farther from a speaker means the direct voice is increasingly overtaken by room reverberation. These echoes can disrupt the spatial cues needed for binaural release from masking, which normally assists listeners in separating a target talker from background babble (Bronkhorst, 2000). Our results, therefore, demonstrate robustness to the signal-level changes associated with distance; whether this extends to the fuller acoustic profile of a real room at those distances remains to be tested. Future work can expand our simulation pipeline to account for more complex scenarios, such as multiple simultaneous talkers and a broader range of talker distances.

V. CONCLUSION

This study demonstrates that binaural spatial cues can substantially reduce the performance gap when adult-trained speech separation models are applied to children's speech in classroom environments. While adult-trained models generalize reasonably well in clean conditions, their performance degrades sharply in the presence of realistic children's babble noise. We showed that domain-specific classroom training improves noise robustness and that finetuning adult-trained models with limited classroom data is even more effective, offering both higher performance and reduced training costs. Crucially, finetuned models also preserved spatial localization cues, enabling accurate talker tracking under challenging, noisy conditions.

ACKNOWLEDGMENTS

This work is part of the INTENSE consortium, which has received funding from NWO Cross-over Grant No. 17619. We also acknowledge the use of AI-assisted tools, such as Grammarly and Google Gemini, for grammar checking and enhancing the clarity of the manuscript. We take full responsibility for the work.

AUTHOR DECLARATIONS

Conflict of Interest

All authors have no competing interests to declare.

DATA AVAILABILITY

The speech corpora used in this study (MyST, CSLU, WSJ) are subject to their respective distributors' licensing terms and cannot be redistributed directly, while LibriSpeech and the HRIRs (CHASAR) used are publicly available corpora. However, we make available the trained

checkpoints for all models trained in this work, the simulated RIRs and BRIRs, and the scripts needed to simulate moving-talker mixtures using any speech corpus. Code and instructions are available on our GitHub repository: <https://github.com/sayo20/Speech-Enhancement-for-Moving-Speakers-in-Classroom-Environment>.

- Acoustic Frontiers (2025). "Reverberation time in room acoustics," <https://acousticfrontiers.com/blogs/articles/reverberation-time-in-room-acoustics> (Last viewed April 15, 2025).
- Adavanne, S., Politis, A., Nikunen, J., and Virtanen, T. (2019). "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Top. Signal Process.* **13**(1), 34–48.
- Amengual Garí, S. V., Arend, J. M., Calamia, P. T., and Robinson, P. W. (2021). "Optimizations of the spatial decomposition method for binaural reproduction," *J. Audio Eng. Soc.* **68**(12), 959–976.
- American National Standard Institute (2010). "Acoustical performance criteria, design requirements and guidelines for schools: 1. Permanent schools," S12.60.
- Assmann, P. F., and Katz, W. F. (2000). "Time-varying spectral change in the vowels of children and adults," *J. Acoust. Soc. Am.* **108**(4), 1856–1866.
- Attia, A. A., Demszky, D., Ògúnremí, T., Liu, J., and Espy-Wilson, C. (2025). "CPT-boosted wav2vec2. 0: Towards noise robust speech recognition for classroom environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5.
- Bistafa, S. R., and Bradley, J. S. (2000). "Predicting reverberation times in a simulated classroom," *J. Acoust. Soc. Am.* **108**(4), 1721–1731.
- Blauert, J. (1997). *Spatial Hearing: The Psychophysics of Human Sound Localization* (Massachusetts Institute of Technology, Cambridge, MA).
- Braren, H. S., and Fels, J. (2021). "Towards child-appropriate virtual acoustic environments: A database of high-resolution HRTF measurements and 3D-scans of children," *Int. J. Environ. Res. Public Health* **19**(1), 324.
- Bronkhorst, A. W. (2000). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acust. Acust.* **86**(1), 117–128.
- Burkhard, M., and Sachs, R. (1975). "Anthropometric manikin for acoustic research," *J. Acoust. Soc. Am.* **58**(1), 214–222.
- Chakrabarty, S., and Habets, E. A. (2019). "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE J. Sel. Top. Signal Process.* **13**(1), 8–21.
- Chen, J., Wang, Z., Dong, R., Fu, X., Wang, Y., and Wang, S. (2021). "Effects of wireless remote microphone on speech recognition in noise for hearing aid users in China," *Front. Neurosci.* **15**, 643205.
- Chen, Z., Xiao, X., Yoshioka, T., Erdogan, H., Li, J., and Gong, Y. (2018). "Multi-channel overlapped speech recognition with location guided speech extraction network," in *IEEE Spoken Language Technology Workshop*, pp. 558–565.
- Ching, T. Y., Zhang, V. W., Flynn, C., Burns, L., Button, L., Hou, S., McGhie, K., and Van Buynder, P. (2018). "Factors influencing speech perception in noise for 5-year-old children using hearing aids or cochlear implants," *Int. J. Audiol.* **57**(suppl 2), S70–S80.
- Cole, R. A., Noel, M., and Noel, V. (1998). "The CSLU speaker recognition corpus," in *International Conference on Spoken Language Processing*, Vol. 98, pp. 3167–3170.
- Cooke, M. (2006). "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.* **119**(3), 1562–1573.
- Crandell, C. C., and Smaldino, J. J. (2000). "Classroom acoustics for children with normal hearing and with hearing impairment," *Lang. Speech Hear. Serv. Sch.* **31**(4), 362–370.
- de Souza, L. M., Costa, M. H., and Borges, R. C. (2024). "Envelope-based multichannel noise reduction for cochlear implant applications," *IEEE/ACM Trans. Audio Speech Lang. Process.* **32**, 1873.
- Faller, C., and Merimaa, J. (2004). "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.* **116**(5), 3075–3089.
- Franci, A., and McDermott, J. H. (2022). "Deep neural network models of sound localization reveal how perception is adapted to real-world environments," *Nat. Hum. Behav.* **6**(1), 111–133.

- Gabova, K., Furstova, J., and Tavel, P. (2024a). "Wireless microphones for students who are deaf and hard of hearing: Czech teachers' perspective," *Res. Spec. Educ. Needs* **24**(1), 68–79.
- Gabova, K., Meier, Z., and Tavel, P. (2024b). "Parents' experiences of remote microphone systems for children with hearing loss," *Disabil. Rehabil. Assist. Technol.* **19**(3), 831–840.
- Gardner, B., and Martin, K. (1994). "HRFT measurements of a Kemar dummy-head microphone," in Vision and Modeling Group, Media Laboratory, Massachusetts Institute of technology.
- Gari, S. V. A., Brimijoin, W. O., Hassager, H. G., and Robinson, P. W. (2019). "Flexible binaural resynthesis of room impulse responses for augmented reality research," in *EAA Spatial Audio Signal Processing Symposium*, pp. 161–166.
- Gazibegovic, D., Bohnert, A., and Laessig, A. K. (2024). "Hearing aid apps: Are they safe, practical and beneficial for children and teens in challenging situations?," *Eur. Arch. Otorhinolaryngol.* **281**, 6021.
- Gerosa, M., Lee, S., Giuliani, D., and Narayanan, S. (2006). "Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition," in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 1, pp. I–I.
- Gremp, M. A., and Easterbrooks, S. R. (2018). "A descriptive analysis of noise in classrooms across the US and Canada for children who are deaf and hard of hearing," *Volta Rev.* **117**, 5–31.
- Gu, R., Chen, L., Zhang, S.-X., Zheng, J., Xu, Y., Yu, M., Su, D., Zou, Y., and Yu, D. (2019). "Neural spatial filter: Target speaker speech separation assisted with directional information," in *Interspeech*, pp. 4290–4294.
- Gu, R., Zhang, S.-X., Chen, L., Xu, Y., Yu, M., Su, D., Zou, Y., and Yu, D. (2020). "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7319–7323.
- Gustafson, S. J., Ricketts, T. A., and Picou, E. M. (2021). "Individual differences offer insight into clinical recommendations for directional and remote microphone technology use in children," *J. Speech Lang. Hear. Res.* **64**(2), 635–650.
- Han, C., Luo, Y., and Mesgarani, N. (2020). "Real-time binaural speech separation with preserved spatial cues," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6404–6408.
- Han, C., Luo, Y., and Mesgarani, N. (2021). "Binaural speech separation of moving speakers with preserved spatial cues," in *Interspeech*, pp. 3505–3509.
- Healy, E. W., and Yoho, S. E. (2016). "Difficulty understanding speech in noise by the hearing impaired: Underlying causes and technological solutions," in *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 89–92.
- Henry, F., Glavin, M., and Jones, E. (2023). "Noise reduction in cochlear implant signal processing: A review and recent developments," *IEEE Rev. Biomed. Eng.* **16**, 319–331.
- Howard, C. S., Munro, K. J., and Plack, C. J. (2010). "Listening effort at signal-to-noise ratios that are typical of the school classroom," *Int. J. Audiol.* **49**(12), 928–932.
- Howard, J., and Ruder, S. (2018). "Universal language model finetuning for text classification," *arXiv:1801.06146*.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). "LoRA: Low-rank adaptation of large language models," *arXiv:2106.09685*.
- Johnstone, P. M., Mills, K. E., Humphrey, E., Yeager, K. R., Jones, E., McElligott, K., Pierce, A., Agrawal, S., Froeling, C., and Little, J. P. (2018). "Using microphone technology to improve speech perception in noise in children with cochlear implants," *J. Am. Acad. Audiol.* **29**(9), 814–825.
- Jutras, B., Lafontaine, L., East, M.-P., and Noël, M. (2019). "Listening in noise training in children with auditory processing disorder: Exploring group and individual data," *Disabil. Rehabil.* **41**(24), 2918–2926.
- Karunanayake, Y., Thayasivam, U., and Ranathunga, S. (2019). "Transfer learning based free-form speech command classification for low-resource languages," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 288–294.
- Klatte, M., Bergström, K., and Lachmann, T. (2013). "Does noise affect learning? A short review on noise effects on cognitive performance in children," *Front. Psychol.* **4**, 578.
- Kuttruff, H. (2016). *Room Acoustics* (CRC, Boca Raton, FL).
- Larson Davis (2025). "Reverberation time in room acoustics," <https://www.larsondavis.com/learn/building-acoustics/Reverberation-Time-in-Room-Acoustics> (Last viewed April 15, 2025).
- Lee, S., Potamianos, A., and Narayanan, S. (1999). "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *J. Acoust. Soc. Am.* **105**(3), 1455–1468.
- Lee, Y., Choi, S., Kim, B.-Y., Wang, Z.-Q., and Watanabe, S. (2024). "Boosting unknown-number speaker separation with transformer decoder-based attractor," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 446–450.
- Lewis, D., Schmid, K., O'Leary, S., Spalding, J., Heinrichs-Graham, E., and High, R. (2016). "Effects of noise on speech recognition and listening effort in children with normal hearing and children with mild bilateral or unilateral hearing loss," *J. Speech Lang. Hear. Res.* **59**(5), 1218–1232.
- Litovsky, R. Y., Johnstone, P. M., Godar, S., Agrawal, S., Parkinson, A., Peters, R., and Lake, J. (2006). "Bilateral cochlear implants in children: Localization acuity measured with minimum audible angle," *Ear Hear.* **27**(1), 43–59.
- Lutati, S., Nachmani, E., and Wolf, L. (2022). "SepIt: Approaching a single channel speech separation bound," *arXiv:2205.11801*.
- McCreery, R. W., Venediktov, R. A., Coleman, J. J., and Leech, H. M. (2012). "An evidence-based systematic review of directional microphones and digital noise reduction hearing aids in school-age children with hearing loss," *Am. J. Audiol.* **21**(2), 295–312.
- McMillan, B. T., and Saffran, J. R. (2016). "Learning in complex environments: The effects of background speech on early word learning," *Child Dev.* **87**(6), 1841–1855.
- Mills, A. W. (1958). "On the minimum audible angle," *J. Acoust. Soc. Am.* **30**(4), 237–246.
- Neuman, A. C., Wroblewski, M., Hajicek, J., and Rubinstein, A. (2010). "Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults," *Ear Hear.* **31**(3), 336–344.
- Olalere, F., van den Heijden, K., Stronks, H. C., Briare, J., Frijns, J. H., and van Gerven, M. (2025). "Leveraging spatial cues from cochlear implant microphones to efficiently enhance speech separation in naturalistic listening scenes," *Sci. Rep.* **16**(1), 2255.
- Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). "LibriSpeech: An ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. et al. (2019). "PyTorch: An imperative style, high-performance deep learning library," *arXiv:1912.01703*.
- Paul, D. B., and Baker, J. M. (1992). "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the 2nd International Conference on Spoken Language Processing*, pp. 899–902.
- Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., Lemke, U., Lunner, T., Matthén, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., and Wingfield, A. (2016). "Hearing impairment and cognitive energy: The framework for understanding effortful listening (fuel)," *Ear Hear.* **37**, 5S–27S.
- Potamianos, A., Narayanan, S. S., and Lee, S. (1997). "Automatic speech recognition for children," in *Eurospeech*, Vol. 97, pp. 2371–2374.
- Qin, M. K., and Oxenham, A. J. (2003). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," *J. Acoust. Soc. Am.* **114**(1), 446–454.
- Savioja, L., and Svensson, U. P. (2015). "Overview of geometrical room acoustic modeling techniques," *J. Acoust. Soc. Am.* **138**(2), 708–730.
- Scheibler, R., Bezzam, E., and Dokmanić, I. (2018). "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 351–355.
- Schuster, S., Pancoast, S., Ganjoo, M., Frank, M. C., and Jurafsky, D. (2014). "Speaker-independent detection of child-directed speech," in *IEEE Spoken Language Technology Workshop*, pp. 366–371.
- Shivakumar, P. G., and Georgiou, P. (2020). "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," *Comput. Speech Lang.* **63**, 101077.
- Southwell, R., Ward, W., Trinh, V. A., Clevenger, Ch., Clevenger, C., Watts, E., Reitman, J., D'Mello, S., and Whitehill, J. (2024). "Automatic speech recognition tuned for child speech in the classroom," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 12291–12295.

- Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., and Zhong, J. (2021). "Attention is all you need in speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 21–25.
- Tervo, S., Pätynen, J., Kuusinen, A., and Lokki, T. (2013). "Spatial decomposition method for room impulse responses," *J. Audio Eng. Soc.* **61**(1), 17–28.
- Thoidis, I., and Goehring, T. (2024). "Using deep learning to improve the intelligibility of a target speaker in noisy multi-talker environments for people with normal hearing and hearing loss," *J. Acoust. Soc. Am.* **156**(1), 706–724.
- Thompson, E. C., Benítez-Barrera, C. R., Angley, G. P., Woynaroski, T., and Tharpe, A. M. (2020). "Remote microphone system use in the homes of children with hearing loss: Impact on caregiver communication and child vocalizations," *J. Speech Lang. Hear. Res.* **63**(2), 633–642.
- Valente, D. L., Plevinsky, H. M., Franco, J. M., Heinrichs-Graham, E. C., and Lewis, D. E. (2012). "Experimental investigation of the effects of the acoustical conditions in a simulated classroom on speech recognition and learning in children," *J. Acoust. Soc. Am.* **131**(1), 232–246.
- Vorperian, H. K., and Kent, R. D. (2007). "Vowel acoustic space development in children: A synthesis of acoustic and anatomic data," *J. Speech Lang. Hear. Res.* **50**(6), 1510–1545.
- Wang, X., Du, J., Cristia, A., Sun, L., and Lee, C.-H. (2020). "A study of child speech extraction using joint speech enhancement and separation in realistic conditions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7304–7308.
- Wang, X., Du, J., Sun, L., Wang, Q., and Lee, C.-H. (2018). "A progressive deep learning approach to child speech separation," in *11th International Symposium on Chinese Spoken Language Processing*, IEEE, pp. 76–80.
- Wang, Z.-Q., and Wang, D. (2019). "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Trans. Audio Speech Lang. Process.* **27**(2), 457–468.
- Ward, W., Cole, R., and Pradhan, S. (2019). *My Science Tutor and the Myst Corpus* (Boulder Learning, Carbondale, CO).
- Wolfe, J., Duke, M., Miller, S., Schafer, E., Jones, C., Rakita, L., Dunn, A., Browning, S., and Neumann, S. (2022). "Evaluation of potential benefits and limitations of noise-management technologies for children with hearing aids," *J. Am. Acad. Audiol.* **33**(2), 66–74.
- Wolfe, J., Duke, M., Schafer, E., Jones, C., and Rakita, L. (2017). "Evaluation of adaptive noise management technologies for school-age children with hearing loss," *J. Am. Acad. Audiol.* **28**(5), 415–435.
- Yu, D., Kolbaek, M., Tan, Z.-H., and Jensen, J. (2017). "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 241–245.
- Zhao, S., Ma, Y., Ni, C., Zhang, C., Wang, H., Nguyen, T. H., Zhou, K., Yip, J. Q., Ng, D., and Ma, B. (2024). "Mossformer2: Combining transformer and RNN-free recurrent network for enhanced time-domain monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 10356–10360.