Single Microphone Own Voice Detection based on Simulated Transfer Functions for Hearing Aids

Mathuranathan Mayuravaani, W. Bastiaan Kleijn, Andrew Lensen, Charlotte Sørensen

***Abstract*—This paper presents a simulation-based approach to own voice detection (OVD) in hearing aids using a single microphone. While OVD can significantly improve user comfort and speech intelligibility, enabling reliable OVD with a single microphone is desirable for simplifying hardware and reducing power consumption in compact hearing devices. However, most existing solutions rely on multiple microphones or additional sensors, increasing device complexity and cost. To enable ML-based OVD without requiring costly transfer-function measurements, we propose a data augmentation strategy based on simulated acoustic transfer functions (ATFs) that expose the model to a wide range of spatial propagation conditions. A transformer-based classifier is trained using analytically generated ATFs and further fine-tuned using numerically simulated ATFs with increasing geometric realism. This hierarchical adaptation enables the model to refine its spatial understanding while maintaining generalization. Experimental results show 95.52% accuracy on simulated head-and-torso test data and 90.02% accuracy for one-second speech segments, demonstrating robustness to short durations. When evaluated on real-world hearing-aid recordings, few-shot fine-tuning using a small subset achieves 91% accuracy, demonstrating that limited real-world data can effectively adapt the simulation-trained model. These results highlight the potential of simulation-based training for enabling practical single-microphone OVD systems in hearing aids.**

***Index Terms*—own voice detection, hearing instruments, machine learning, signal processing, speech processing.**

I. Introduction

HEARING AIDS play a crucial role in improving the quality of life for individuals with hearing impairments by amplifying speech. However, a common complaint among hearing aid users is that their own voice sounds excessively loud or unnatural when using the device [1]. To mitigate this issue, hearing care professionals often reduce amplification levels to enhance comfort. This reduction, however, compromises speech clarity and audibility, making communication difficult [2]. Developing a robust method for efficiently identifying the wearer's own voice is crucial, as it enables personalized adjustments to maintain comfort and intelligibility [3].

Although many modern hearing aids incorporate multiple microphones, reliable single microphone own voice detection (OVD) remains desirable for several practical reasons. Multi-microphone processing increases hardware cost, power consumption, and calibration complexity, limiting applicability in low-cost devices [4]. In addition, some users rely on a single device due to unilateral hearing loss or personal preference. Moreover, signals may become unavailable or degraded due to occlusion, wind noise, or power-saving modes [5], making single-microphone OVD a useful fallback. For these reasons, single-microphone OVD remains a practically important and complementary alternative to multi-microphone solutions.

The majority of existing OVD methods rely on traditional signal processing techniques, such as cross-correlation, adaptive filtering, and beamforming [6]–[9]. These methods typically require multiple microphones and complex directional filtering, which increases hardware costs and makes them unsuitable for individuals with single-ear hearing loss.

Machine learning (ML) approaches, often applied in single-microphone configurations, have gained popularity in speaker identification and distance estimation tasks [10]–[13], as they can capture complex patterns in speech signals and improve adaptability. However, many ML-based approaches rely primarily on speaker-dependent vocal characteristics to distinguish the user's voice. In contrast, this work proposes a spatial-analysis-based approach that uses differences in acoustic propagation paths between the user's own voice and external speakers. Acoustic transfer functions (ATFs) are used to model the distinct propagation paths of own voice and external speech as they reach the hearing-aid microphone, and a data-driven classifier is trained to learn discriminative spatial–spectral patterns from ATF-augmented speech.

To overcome the practical difficulty of collecting large-scale measured ATFs across anatomies and device configurations, we propose a simulation-based ATF generation pipeline that progressively increases geometric realism. We first generate analytically derived ATFs based on simplified geometries to cover a wide range of spatial configurations, enabling scalable data generation with controlled variation. We then generate numerically simulated ATFs using boundary-element modeling, transitioning from a rigid sphere to a detailed head-and-torso model. These ATFs are used to augment speech signals for training the OVD classifier. We formulate OVD as a segment-level binary classification and use a transformer-based classifier to aggregate frame-level features into a single decision.

The main contributions of this work are as follows: (i) enabling scalable training of single-microphone OVD models using simulated ATFs without requiring costly measured transfer functions; (ii) employing a progressive training strategy that transitions from analytical to numerically simulated ATFs with increasing anatomical realism; and (iii) demonstrating that spatial cues embedded through simulated ATFs enable effective single-microphone OVD on real-world hearing-aid recordings. This study focuses on offline segment-level detection as a feasibility analysis; causal streaming implementation with low-latency inference is left for future work.

II. Background

OVD has been a longstanding challenge in speech processing, particularly for hearing aids. Early approaches relied on classical signal processing techniques, later evolving to incorporate multichannel methods and, more recently, ML-based solutions. One of the earliest OVD methods was proposed by Rasmussen et al. [8], who used a cross-correlation technique to exploit the symmetrical positioning of the user's mouth. Their method aimed to detect speech originating from the median plane. To improve near-field detection, they introduced the mouth-to-random-far-field (M2R) technique; however, if a straight-ahead external speaker is close enough, their voice may also exhibit near-field characteristics, posing a challenge for accurate differentiation. Another classical approach involved adaptive filtering, as proposed by Merks et al. [6], where two microphones were used to estimate a transfer function for OVD. While effective in some cases, this method was susceptible to errors in the presence of strong external speech sources. Bone-conductive sensors have also been explored [6], [14], but their effectiveness is limited by low signal-to-noise ratio (SNR) and the need for direct skull contact.

Hoang et al. [15] applied beamforming and relative ATFs for interfering speech suppression. They considered the ATF from the user's mouth to the microphones to be approximately time-invariant. With the microphones positioned close to the user's mouth, the own voice signal can be louder than both the target and interfering speech, especially at lower frequencies when the own voice is active. However, these considerations may not apply to a speaker facing directly forward or with varying loudness. Frequency-dependent variations in the response of rear microphones in head-mounted arrays have been observed and attributed to diffraction and directional propagation of own voice [16]. A similar trend appears in our free-space ATFs, where the response changes progressively with frequency (see Section III-A).

ML-based OVD methods have also been explored. Pohlhausen et al. [17] proposed a binaural OVD method utilizing near-ear microphones and a random forest classifier, distinguishing own voice from external sounds based on phase and amplitude symmetry. Similarly, Pertilä et al. [18] developed an LSTM-based online OVD system for in-ear devices using multiple microphones and accelerometers, which increase system complexity and cost. Lopez-Espejo et al. [19], [20] trained a deep residual network for OVD within a keyword spotting system, leveraging measured head-related transfer functions (HRTFs) and own voice transfer functions (OVTFs) from a hearing assistive device. However, these approaches rely on measured, device-specific transfer functions, which limits scalability because collecting such measurements across different anatomies, device placements, and acoustic conditions is costly and time-consuming. As a result, the coverage of possible acoustic propagation paths remains limited. This motivates the use of simulation-based ATF generation, which enables scalable training data generation with controlled variation across spatial configurations. Additionally, distance-based ML approaches for speaker separation have been investigated using single-channel audio [11], [13].

Acoustic modeling has played a significant role in understanding voice propagation and directivity patterns. Various analytical models [21], [22] have been employed to study speech and singing sound propagation, often utilizing simplified representations of the human head. Many of these models approximate the acoustic effects of the head and mouth using geometrical shapes, such as spheres. For instance, Blandin et al. [21] analyzed the directivity patterns around the head, assuming a vibrating plane piston model set within an infinite baffle. Their work aimed to enhance the quality of artistic performances in concert halls by investigating the influence of the mouth opening size on voice directivity, comparing both measurements and simulations. They observed that the overall features of the radiated sound could be predicted from the properties of the vibrating piston. Similarly, the impact of mouth configuration and torso shape on voice directivity was explored during singing [22]. Utilizing both piston and spherical cap models for their simulations, they found that significant changes in the mouth opening size considerably affect voice directivity. In our approach, we present a model that represents the human head as a rigid sphere with the mouth modeled as a vibrating cap on its surface, providing a physically grounded approach to detecting the wearer's own voice in a hearing aid. A point source is represented as an external speaker, allowing us to analyze how sound scatters from the rigid sphere. By jointly modeling both the wearer's voice and external sources, we capture the spatial propagation differences that distinguish them, allowing us to generate ATFs that enable us to augment the speech data to train the ML algorithm.

While previous works have applied ML for OVD and utilized transfer functions in speech modeling, to the best of our knowledge, no study has incorporated analytical and numerical ATFs for augmenting training data in ML-based OVD using a single microphone. In this work, we propose a data augmentation strategy that enhances training data diversity using synthetic ATF-augmented speech, and we evaluate a conformer encoder-based classifier [23] for segment-level own-voice versus external-speaker classification. The conformer architecture provides a flexible mechanism for aggregating frame-level speech features into a single decision, improving robustness to phonetic variability while operating in a single microphone setting.

III. Method

In this study, our objective is to classify the hearing aid wearer's own voice versus external speakers' voices. To achieve this, we employ an ML algorithm to perform the classification. Since ML models require large and diverse training datasets to generalize effectively, we augment the speech data using ATFs. Specifically, we generate training samples using both analytically derived ATFs, which introduce controlled variability, and numerically simulated ATFs, which incorporate detailed human anatomy for more realistic spatial representations. The model was first trained on analytically augmented data and subsequently fine-tuned using samples augmented from numerical simulations.

To construct these ATFs for data augmentation and analysis, we model the wearer's head as a rigid sphere that captures scattering effects shaping the acoustic paths to the hearing-aid microphone. External speech is modeled as radiation from a point source scattered by the rigid sphere, whereas own voice is modeled as radiation from a vibrating spherical cap on the same sphere, representing mouth radiation coupled with head scattering. In both cases, the resulting pressure is evaluated at the hearing-aid microphone location on the sphere surface. The same source definitions are used in the numerical simulations, while the geometry is progressively refined from a rigid sphere to detailed head and head-and-torso models to increase anatomical realism. Similar geometric approximations, such as spherical caps and pistons, have been used in previous studies [21], [22], [24] to model sound radiation and head-related scattering.

Section III-A presents the fundamental equations governing the considered geometries. Section III-B details the analytical approach for modeling sound propagation. We validate this approach by comparing analytical results with numerical simulations and also generate ATFs using simulation software, as discussed in Section III-C. Finally, the OVD algorithm is described in Section III-D.

A. Geometry and basic formulas

We consider a point source with a volume velocity $\tilde{U}_0$ located at a distance d from a rigid sphere of radius R (Fig. 1) to model the effect of an external speaker's voice reaching a hearing-aid user. Our goal is to determine the pressure field $\tilde{p}(\mathbf{r}, \theta)$ at an observation point on the sphere, corresponding to the ear location of the hearing-aid user. $r_1 = \sqrt{r^2 + d^2 - 2rd\cos\theta}$ is the distance from the point source to the observation point. In spherical coordinates, the incident pressure field due to the point source is [25], [26]:

$$\tilde{p}(r, \theta) = jk\rho_0 c\tilde{U}_0 \frac{e^{-jkr_1}}{4\pi r_1}, \tag{1}$$

where j is the imaginary unit, k is the wave number, ρ_0 is the density of air and c is the speed of sound in air.

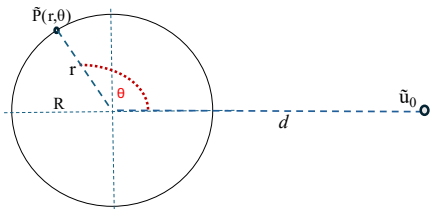


Fig. 1: Geometry of point source and rigid sphere.

This expression describes the free-field pressure at an arbitrary location before considering reflections and scattering. This expression can be expanded in terms of spherical Bessel functions j_n and Legendre functions P_n as follows [25]:

$$\tilde{p}_I(r, \theta) = \frac{k^2\rho_0 c\tilde{U}_0}{4\pi} \sum_{n=0}^{\infty} (2n+1) h_n^{(2)}(kd) j_n(kr) P_n(\cos\theta), \quad r \leq d. \tag{2}$$

When an external speaker's voice reaches the hearing-aid user, part of the sound is scattered by the head. The scattered pressure field is given by [25]:

$$\tilde{p}_s(r, \theta) = -\frac{k^2\rho_0 c\tilde{U}_0}{4\pi} \sum_{n=0}^{\infty} (2n+1) h_n^{(2)}(kd) \frac{j_n'(kR)}{h_n'^{(2)}(kR)} h_n^{(2)}(kr) P_n(\cos\theta). \tag{3}$$

The total pressure field at the observation point is obtained by summing the incident and scattered fields:

$$\tilde{p}(r, \theta) = \tilde{p}_I(r, \theta) + \tilde{p}_s(r, \theta). \tag{4}$$

This total pressure field is critical for analyzing how an external speaker's voice propagates to the hearing-aid user. In contrast to an external speaker, the hearing-aid user's own voice originates from the mouth and propagates differently. To model this, we represent the mouth as a vibrating spherical cap on a rigid sphere that vibrates with an axial velocity of $\tilde{u}_0$ (Fig. 2) [25], [27]. The near-field pressure due to the oscillating cap describes the propagation of the user's own voice from the mouth to the hearing-aid microphone. This pressure field is given by [28]:

$$\begin{aligned}\tilde{p}(r, \theta) = -jk\rho_0 c\tilde{u}_0 \Bigg(& \frac{\sin^2\alpha}{4h_0'^{(2)}(kR)} h_0^{(2)}(kr) + \frac{1-\cos^3\alpha}{2h_1'^{(2)}(kR)} h_1^{(2)}(kr)\cos\theta \\ & - \sin\alpha \sum_{n=2}^{\infty} (2n+1) \frac{\sin\alpha P_n(\cos\alpha) + \cos\alpha P_n^1(\cos\alpha)}{2(n-1)(n+2)h_n'^{(2)}(kR)} h_n^{(2)}(kr) P_n(\cos\theta) \Bigg),\end{aligned} \tag{5}$$

where α is the half-angle of the arc formed by the cap.

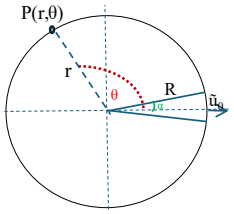


Fig. 2: Geometry of oscillating cap in a rigid sphere.

A comparison of directivity patterns at 5 kHz is presented in Fig. 3, considering a point source scattering from a rigid sphere and a vibrating spherical cap on the sphere. The results show a significant difference in the directivity pattern with respect to the on-axis direction. This effect becomes more pronounced at higher frequencies, leading to reduced pressure at the ear location in the own-voice scenario.

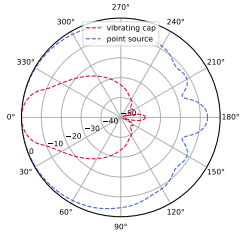


Fig. 3: Directivity pattern $20\log_{10}(D(\theta)/D(0))$ of the near-field pressure for a straight-ahead point source scattering from a 7 cm rigid sphere and a spherical cap ($\alpha = 20°$).

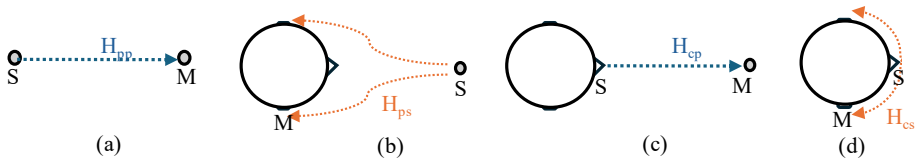


Fig. 4: Mapping the pressure between (a) a point source and a point receiver, (b) a point source and a rigid sphere, (c) a rigid sphere source with vibrating cap and a point receiver, and (d) vibrating spherical cap and location on the rigid sphere. Here, S denotes the source location and M denotes the receiver (microphone) location.

B. Analytical Modeling Approach

This section describes how analytically derived ATFs are computed and applied to generate synthetic hearing-aid microphone signals. We consider two classes: (i) external-speaker and (ii) own-voice. The two classes differ in the assumed geometry model and propagation path: external speech is modeled as radiation from a point source scattered by a rigid sphere, whereas own voice is modeled as radiation from a vibrating spherical cap mounted on the same sphere. The resulting ATFs characterize the frequency-domain mapping from the source model to the hearing-aid microphone position and are subsequently used to generate ATF-augmented speech samples for ML training (described in Section III-D).

Let $P_{out\_r}(k)$ denote the complex pressure spectrum at a reference microphone receiver location (used as a known observation in the simulation pipeline). Our objective is to generate the pressure spectrum at the hearing-aid location, denoted by $P_{f\_other}(k)$ and $P_{f\_own}(k)$, for both external and own voice conditions. The mapping is performed in two steps. First, the ATF corresponding to the reference propagation path is inverted to estimate an equivalent source spectrum. Second, the ATF from the source to the hearing-aid microphone is applied to obtain the pressure at the hearing-aid. We define the following frequency-domain transfer functions parameterized by the acoustic wavenumber $k = \omega/c$ (see Fig. 4):

- $H_{pp}(k)$: ATF from a point source to a point receiver.
- $H_{ps}(k)$: ATF from a point source to a receiver location on the rigid sphere.
- $H_{cp}(k)$: ATF from a vibrating spherical-cap source (own-voice model) to a point receiver.
- $H_{cs}(k)$: ATF from a vibrating spherical-cap source to a receiver location on the rigid sphere.

For the external-speaker class, the speech waveform is assumed to radiate from a point source. The pressure at the reference receiver $P_{out\_r}(k)$ is first mapped back to an estimate of the source pressure using the inverse ATF $H_{pp}^{-1}(k)$, and then forward-propagated to the hearing-aid microphone position using $H_{ps}(k)$ (see Fig. 5(a)):

$$P_{f\_other}(k) = H_{ps}(k)\, H_{pp}^{-1}(k)\, P_{out\_r}(k). \tag{6}$$

For the own voice class, the speech waveform is assumed to radiate from a vibrating spherical cap mounted on a rigid sphere, representing sound radiation from the wearer's mouth. The reference observation is mapped to an equivalent cap-source spectrum using $H_{cp}^{-1}(k)$ and then propagated to the hearing-aid microphone using $H_{cs}(k)$ (see Fig. 5(b)):

$$P_{f\_own}(k) = H_{cs}(k)\, H_{cp}^{-1}(k)\, P_{out\_r}(k). \tag{7}$$

These ATFs (see Fig. 6(a)) enable the simulation of how speech propagates to the hearing-aid user.

1) $H_{pp}(k)$: The ATF models free-field propagation from a point source to a point receiver (Fig. 4(a)). From the point-source pressure expression in (1), the corresponding ATF is defined as:

$$H_{pp}(k) = \frac{r_0 e^{-jkr_1}}{r_1 e^{-jkr_0}}. \tag{8}$$

Here, r_0 is a reference distance close to the source, and r_1 is the source-to-receiver distance. We use $H_{pp}^{-1}(k)$ to map the reference pressure back to the equivalent source location.

2) $H_{ps}(k)$: The ATF models propagation from a point source to a receiver location on a rigid sphere (Fig. 4(b)). It is defined as the ratio of the total pressure at the receiver on the sphere surface, obtained from (4) by setting $r = R$, to the corresponding free-field point-source pressure at distance r_0, given in (1). This reference pressure corresponds to the incident free-field pressure in the absence of the sphere, which can be expressed by (1) or equivalently by (2). With

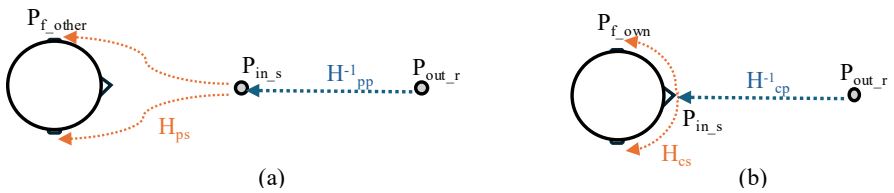
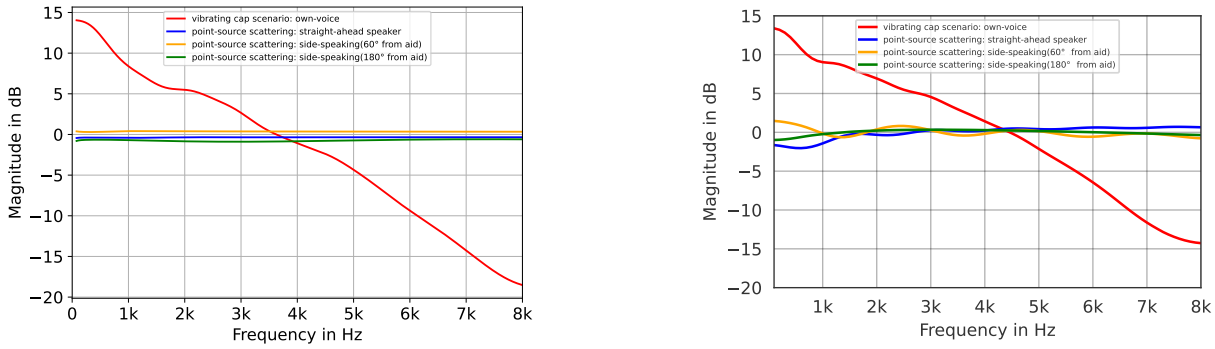


Fig. 5: Mapping the pressure at an angle on the sphere when (a) a speech signal is from a distant point source (external speaker) and (b) a speech signal is from a vibrating cap on the rigid sphere (own voice).



(a) Analytical results.

(b) Simulation results.

Fig. 6: Comparison between analytical and numerical simulation results of ATF with rigid sphere.

this formulation, $H_{ps}(k)$ captures both propagation and rigid sphere scattering:

$$H_{ps}(k) = \frac{kr_0 \sum_{n=0}^{\infty}(2n+1)h_n^{(2)}(kd)\left(j_n(kR) - \frac{j_n'(kR)}{h_n'^{(2)}(kR)}h_n^{(2)}(kR)\right)P_n(\cos\theta)}{je^{-jkr_0}}. \tag{9}$$

3) $H_{cp}(k)$: The ATF models propagation from a vibrating spherical-cap source mounted on a rigid sphere to a point receiver (Fig. 4(c)). It is derived from the spherical-cap pressure field in (5) by expressing the pressure at the receiver distance r_1 relative to that at the reference distance r_0, which is chosen close to the cap. With this formulation, $H_{cp}(k)$ captures the propagation from the vibrating spherical-cap source to the point receiver:

$$H_{cp}(k) = \frac{\left(\frac{\sin^2\alpha}{4h_0'^{(2)}(kR)}h_0^{(2)}(kr_1) + \frac{1-\cos^3\alpha}{2h_1'^{(2)}(kR)}h_1^{(2)}(kr_1) - \sin\alpha \right.\\ \left. \sum_{n=2}^{\infty}(2n+1)\frac{\sin\alpha P_n(\cos\alpha)+\cos\alpha P_n^1(\cos\alpha)}{2(n-1)(n+2)h_n'^{(2)}(kR)}h_n^{(2)}(kr_1)\right)}{\left(\frac{\sin^2\alpha}{4h_0'^{(2)}(kR)}h_0^{(2)}(kr_0) + \frac{1-\cos^3\alpha}{2h_1'^{(2)}(kR)}h_1^{(2)}(kr_0) - \sin\alpha \right.\\ \left. \sum_{n=2}^{\infty}(2n+1)\frac{\sin\alpha P_n(\cos\alpha)+\cos\alpha P_n^1(\cos\alpha)}{2(n-1)(n+2)h_n'^{(2)}(kR)}h_n^{(2)}(kr_0)\right)}. \tag{10}$$

We use $H_{cp}^{-1}(k)$ to estimate the equivalent spherical-cap source spectrum from the reference observation.

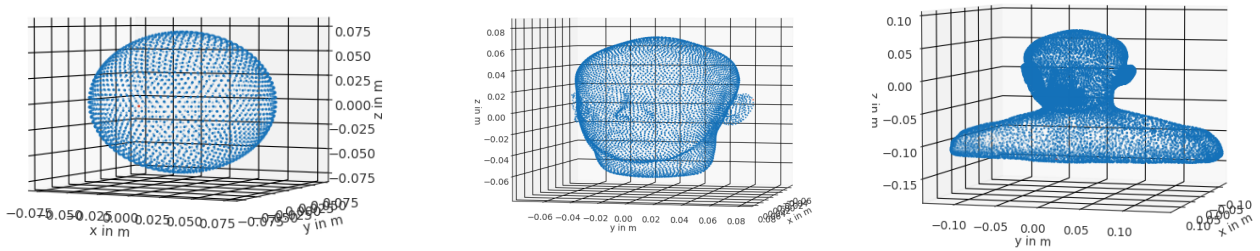
4) $H_{cs}(k)$: The ATF models propagation from a vibrating spherical-cap source to a receiver location on the rigid sphere (Fig. 4(d)), representing how the own voice is observed at the hearing-aid microphone. It is derived from the spherical-cap pressure field in (5) by evaluating the pressure on the sphere surface and expressing the response at the receiver angle θ relative to the corresponding on-axis reference response. With this formulation, $H_{cs}(k)$ captures the directional own-voice radiation from the vibrating spherical cap to the hearing-aid microphone location on the sphere surface:

$$H_{cs}(k) = \frac{\left(\frac{\sin^2\alpha}{4h_0'^{(2)}(kR)}h_0^{(2)}(kR) + \frac{1-\cos^3\alpha}{2h_1'^{(2)}(kR)}h_1^{(2)}(kR)\cos\theta - \sin\alpha \right.\\ \left. \sum_{n=2}^{\infty}(2n+1)\frac{\sin\alpha P_n(\cos\alpha)+\cos\alpha P_n^1(\cos\alpha)}{2(n-1)(n+2)h_n'^{(2)}(kR)}h_n^{(2)}(kR)P_n(\cos\theta)\right)}{\left(\frac{\sin^2\alpha}{4h_0'^{(2)}(kR)}h_0^{(2)}(kR) + \frac{1-\cos^3\alpha}{2h_1'^{(2)}(kR)}h_1^{(2)}(kR) - \sin\alpha \right.\\ \left. \sum_{n=2}^{\infty}(2n+1)\frac{\sin\alpha P_n(\cos\alpha)+\cos\alpha P_n^1(\cos\alpha)}{2(n-1)(n+2)h_n'^{(2)}(kR)}h_n^{(2)}(kR)\right)}. \tag{11}$$

C. Numerical Modeling Approach

We used *Mesh2HRTF* [29], [30], an open-source software package for the numerical calculation of HRTFs to generate ATFs from 3D mesh models. We utilized its implementation of the multilevel fast multipole method (ML-FMM-BEM), a combination of the multilevel fast multipole method (ML-FMM) [31] and boundary element method (BEM) [32]. The 3D model meshes of rigid sphere (see Fig. 7(a)) and human models (see Fig. 7(b), (c)) were created using Blender [33] and post-processed using MeshInspector [34].

The rigid sphere mesh enables direct comparison with our analytical model and provides a foundational step for progressively fine-tuning the model toward more anatomically realistic head and torso representations. In Mesh2HRTF, radiation from

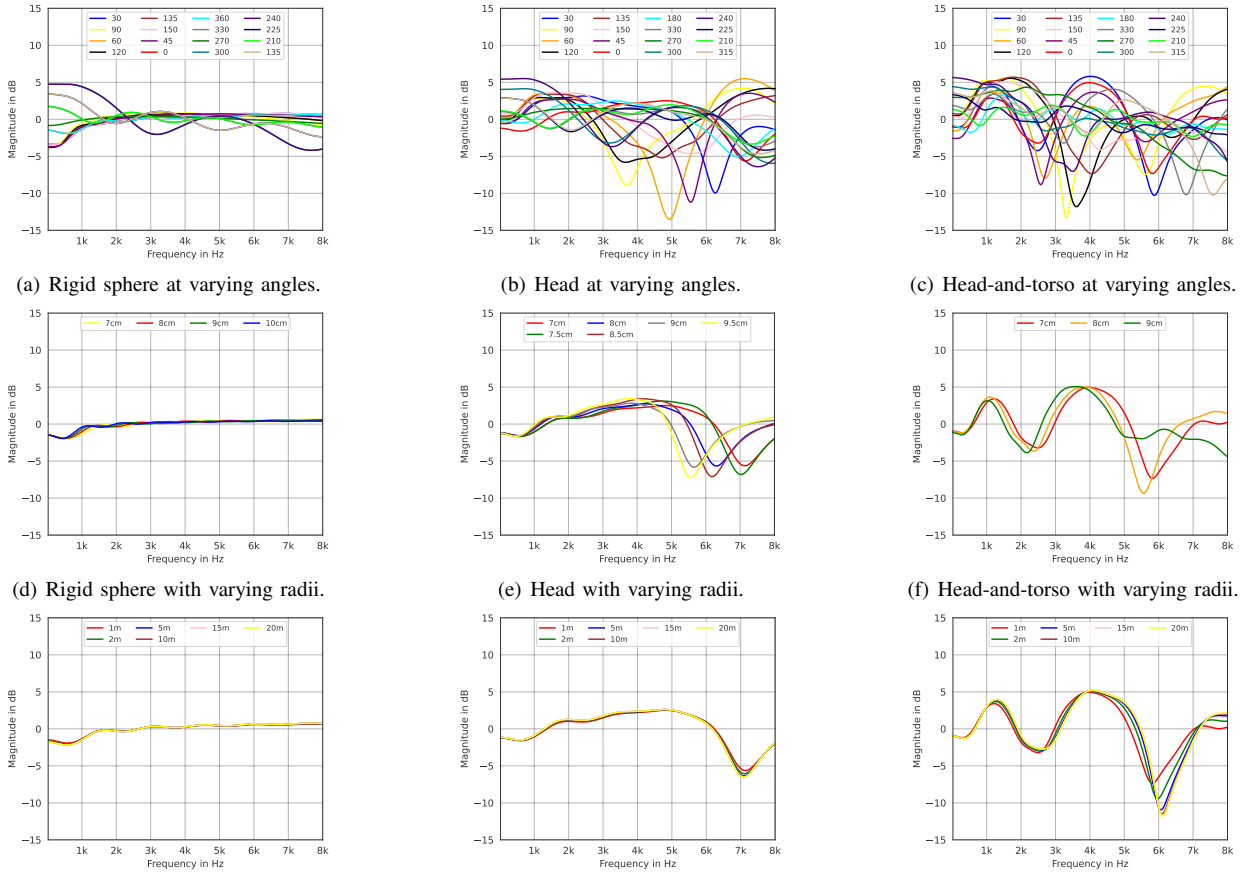


(a) 3D Rigid sphere model.

(b) 3D Human head model.

(c) 3D Human head-and-torso model.

Fig. 7: 3D model meshes created in Blender for simulation.

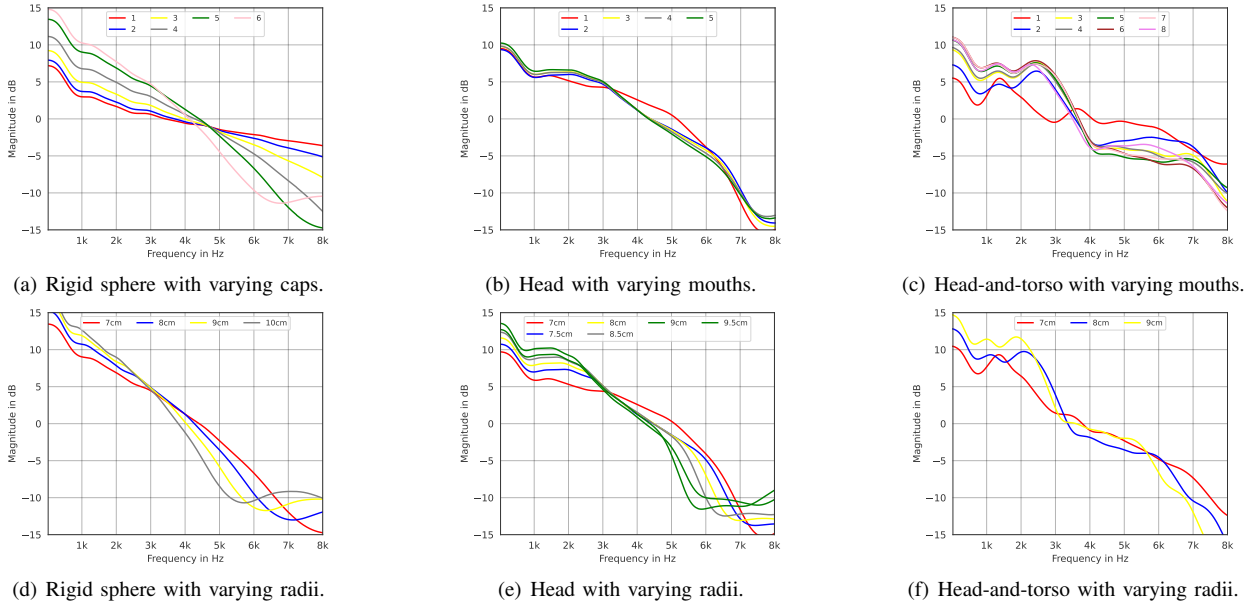


(a) Rigid sphere at varying angles. (b) Head at varying angles. (c) Head-and-torso at varying angles.

(d) Rigid sphere with varying radii. (e) Head with varying radii. (f) Head-and-torso with varying radii.

(g) Rigid sphere at varying distances. (h) Head at varying distances. (i) Head-and-torso at varying distances.

Fig. 8: ATFs for external speaker scenarios: (a)–(c) show a 3D model with a 7 cm head radius and a point source positioned 1 m away at various angles. (d)–(f) illustrate a 3D model with varying head radii and a point source positioned 1 m straight ahead. (g)–(i) depict a 3D model with a 7 cm head radius and a point source at varying distances.



(a) Rigid sphere with varying caps. (b) Head with varying mouths. (c) Head-and-torso with varying mouths.

(d) Rigid sphere with varying radii. (e) Head with varying radii. (f) Head-and-torso with varying radii.

Fig. 9: ATFs for own voice scenarios: (a)–(c) show a 3D model with a 7 cm head radius, with varying mouth angles, where the number indicates increasing angles. (d)–(f) illustrate a 3D model with varying head radii.

vibrating elements were used to model the vibrating cap for the rigid sphere and the mouth for human models, while the external speaker was represented as a point source. The source code for inspecting SOFA files was modified, and customized evaluation grids were generated to suit our specific requirements. The simulation scenarios for other speakers and own voice were generated as illustrated in Figs. 5(a) and 5(b), respectively. Figure 6 compares analytical and numerical simulation results for the rigid sphere model, showing close agreement between the two approaches.

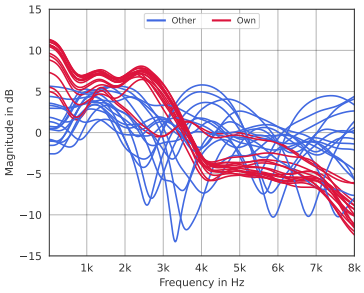


Fig. 10: Comparison of own-voice and external-speaker ATFs at various angles using the head-and-torso model.

The ATFs were generated for three models of increasing anatomical complexity: a rigid sphere, a human head, and a head-and-torso configuration. These ATFs account for variations in speaker angles, distances, and anatomical dimensions. In addition, ATFs were generated for elevation angles of $+5^\circ$ and -5° in head-and-torso models to account for slight vertical speaker offsets. Figure 8 summarizes the ATFs for external-speaker scenarios. Subplots (a)-(c) demonstrate how directional angle influences ATFs across models, with the rigid sphere showing smoother responses and the head-and-torso model exhibiting stronger spectral notches and directional dependence, especially above 3 kHz. Subplots (d)-(f) highlight the impact of changing head radius, where increases in head size lead to more pronounced high-frequency attenuation, particularly evident in the human head and head-and-torso cases. Subplots (g)–(i) show that increasing distance reduces fluctuation strength in the ATF, though structural model differences persist.

Figure 9 focuses on the ATFs for the own voice condition. Unlike external sources, these responses exhibit a steep spectral slope, particularly in the head-and-torso model, indicating strong near-field characteristics and spatial filtering. Furthermore, Fig. 10 directly compares own-voice and external-speaker ATFs using the head-and-torso model. This comparison further underscores that own-voice ATFs maintain steep spectral slopes with relatively consistent patterns, while external speaker ATFs show greater angular variability and flatter responses.

D. Own voice detection algorithm

We apply transfer functions derived from both analytical and numerical simulation models to a clean speech wave in order to replicate how a hearing aid user perceives incoming sound. To incorporate these ATFs into the speech, we first extract the Short-Time Fourier Transform (STFT) from raw audio data, and the extracted STFT is then modified by applying the ATFs, simulating how sound propagates to the hearing aid user's ear. Following this, a log-mel spectrogram is derived from the transformed STFT and used as input to the classifier.

We formulate OVD as a segment-level binary classification task, where each input segment is assigned a single label (own-voice or external-speaker). We adopt a transformer-based model with temporal gate pooling [35], which aggregates frame-level features across the segment into a single decision. This temporal aggregation improves robustness to phonetic variability and short-term spectral fluctuations while leveraging spatial–spectral cues embedded by the ATFs. We emphasize that the present work evaluates offline segment-level classification.

IV. Experimental Setup

In this section, we describe the dataset used for training and evaluation, the model architecture and hyperparameters, and the fine-tuning strategies employed.

A. Dataset

VoxCeleb1 dataset [36] was used as the source of clean speech waveforms for generating ATF-augmented training data. The dataset contains 153,516 utterances, which are randomly assigned to the own voice or external speaker class (50% each) prior to ATF-based transformations. The ATF dataset was generated using both analytically derived ATFs and numerically simulated ATFs. For the numerically simulated ATFs, each ATF class (external-speaker and own-voice) was split into training (70%), validation (10%), and testing (20%), ensuring that test ATFs are not seen during training. The detailed ATF configuration parameters and dataset sizes are provided in Table I, and the dataset split is summarized

TABLE I: Description of the ATF dataset. (*) indicates the number of variations considered for each parameter.

Model	Angles	Head radius	Distance	Mouth angle	Ear location	Total ATF	Own voice	Other speaker
Analytical approach								
Rigid sphere	0°-360° (360)	7cm-10cm (50)	0.5m-20m (100)	5°-20° (15)	90°-120° (20)	153,516	76,757	76,759
Numerical approach								
Rigid sphere	0°-360° (18)	7cm-10cm (4)	1m-20m (39)	(6)	90°	2,544	48	2,496
Human head	0°-360° (18)	7cm-9.5cm (6)	1m-20m (24)	(5)	90°	2,414	110	2,304
Head & torso	0°-360° (18)	7cm-9cm (3)	0.5m-20m (25)	(8)	90°	2,904	336	2,568

TABLE II: Distribution of ATF dataset into training, validation, and testing sets for classes 0 (Other) and 1(Own).

ATF Dataset	Training		Validation		Testing	
	1	0	1	0	1	0
Rigid sphere (Analytical)	69,180	69,181	3,452	3,452	4,125	4,126
Rigid sphere (Numerical)	33	1,747	5	250	10	499
Head model (Numerical)	77	1,612	11	231	22	461
Head & torso (Numerical)	235	1,797	33	257	68	514

in Table II. For each speech utterance, an ATF is randomly sampled from the corresponding split and applied to generate the ATF-augmented waveform.

To assess noise robustness, the MUSAN dataset [37], which includes a diverse set of noises such as music, speech, and ambient sounds, was used to augment the training data by adding background noise at various SNR levels (0, 5, 10, 15, and 20 dB). To assess generalization to unseen speech content and speakers, we evaluate the model on LibriSpeech [38], which is not used during training. OVD test samples are generated by applying the same ATF augmentation procedure using held-out ATFs from the test split. The resulting LibriSpeech test set contains 4,125 utterances per class.

We also included real-world recordings from a hearing-aid prototype to assess the model's robustness under practical acoustic conditions. Three participants provided informed consent prior to data collection. Recordings were obtained using two prototype units of the same hearing-aid device design connected to a Mackie ProFX16 audio mixer in a controlled acoustic room. For the own-voice condition, participants wore the device and read pre-defined sentences to ensure consistent speech content. For the external-speaker condition, participants continued wearing the device while a second person spoke in the same room, representing external speech in a realistic conversational scenario. The estimated SNR of the recordings was approximately 23 dB. For evaluation, 250 non-overlapping one-second segments were extracted, evenly distributed between the two classes. A randomly selected, disjoint subset of 50 segments was used for calibration to estimate statistics for domain-adaptation compensation, while the remaining segments were used exclusively for testing. Due to the limited size of the real-world dataset, this evaluation primarily serves as a proof-of-concept validation of simulation-to-real-world transfer.

In addition to real-world recordings, we evaluated the proposed approach on a publicly available measured-ATF hearing-aid dataset [19]. In this dataset, clean speech signals from the Google Speech Commands dataset [39] are convolved with ATFs measured from a hearing aid worn by a human subject, yielding both own-voice and external-speaker conditions. The measured transfer functions consist of 48 external-speaker HRTFs measured at different azimuths and one own-voice transfer function. Although the dataset provides multi-microphone recordings, only the front-microphone channel is used to match the single microphone setting considered in this work.

B. Model

We followed the standard training approach from [35], which employs a conformer-based encoder [23], and reduced the number of parameters (0.9M) to optimize computational efficiency while maintaining classification performance. The model is configured with a hidden size of 128, a feed-forward hidden size of 512, 4 attention heads, and a single encoder layer. In preprocessing, raw audio data of 16 kHz sample rate is segmented into fixed 15-second segments. The STFT is computed using a 25 ms window with a 10 ms stride, and an 80-channel log-mel spectrogram is used as input for the ML model.

C. Training methodology

The model was trained using a progressive adaptation training strategy [40], where the dataset complexity gradually increases to improve its performance in OVD. This approach follows a structured, multi-phase training process:

1) Training with analytical ATFs: Establishes a foundational understanding of sound propagation characteristics by allowing a wider range of parameter variations, including angles, distances, and head radii. For training, an AdamW optimizer with a learning rate of 1e-4 and a linear-warmup-and-cosine-decay scheduler were employed, conducting training across a maximum of 50 epochs with a batch size of 64.
2) Fine-tuning with numerical ATFs: The model is fine-tuned using simulated ATFs, progressively transitioning from a rigid sphere model $\rightarrow$ human head model $\rightarrow$ head-and-torso model, allowing it to incrementally adapt to more complex practical applications. Each fine-tuning stage was run for 5 epochs. We employed two distinct fine-tuning strategies: (1) fine-tuning only the last classification layer and (2) fine-tuning both the feed-forward and classification layers [41].

To assess robustness under noisy conditions, the model was trained using simulated speech data mixed with noise. Specifically, 70% of the training data consisted of clean speech, while the remaining 30% was randomly mixed with noise samples from the MUSAN dataset [37] at varying SNR levels of 0, 5, 10, 15, and 20 dB.

D. Test-time feature compensation

When evaluating the model on real-world hearing-aid recordings, a distribution mismatch arises because the model is trained on simulated ATF-augmented speech. To mitigate this mismatch, we apply a lightweight test-time feature compensation. Let X denote the log-Mel spectrogram extracted from a real-world speech segment. Device-dependent spectral coloration introduced by the hearing aid [42] is first compensated using a white-noise reference recording, yielding $\tilde{X} = X - \mu_W$, where μ_W is the mean log-Mel spectrum of recorded white noise. To align the feature distribution with the simulated training domain, we apply a lightweight test-time statistic matching step inspired by correlation-alignment approaches in unsupervised domain adaptation (e.g., CORAL [43]). Specifically,

$$X' = (\tilde{X} - \mu_R) \odot \frac{\sigma_S}{\sigma_R} + \mu_S, \tag{12}$$

TABLE III: Test accuracy (%) for models fine-tuned with different ATF configurations using two strategies: (i) classification layer only, and (ii) feed-forward layers with the classification layer. Evaluation is performed on the head-and-torso ATF dataset using utterances up to 15 s.

Model fine-tuned on	Test Accuracy (%) on Head & Torso ATF dataset	
	Last Layer Fine-Tuned	Feed-Forward + Last Layer Fine-Tuned
+ Sphere ATF	83.78 ± 0.24	91.46 ± 0.20
+ Head ATF	90.54 ± 0.30	93.40 ± 0.19
+ Head & torso ATF	91.04 ± 0.18	95.52 ± 0.20

where μ_R and σ_R are the per-frequency mean and standard deviation estimated from a disjoint calibration subset of the evaluation data, and μ_S and σ_S are the corresponding statistics estimated from the simulated training data. The operator $\odot$ denotes element-wise multiplication. This compensation is applied only during inference and does not require any model fine-tuning.

E. Few-shot real-world data fine-tuning

While the above compensation method enables simulation-to-real-world transfer, we also investigate few-shot model adaptation using a small set of real-world recordings. We use 50 labeled real-world segments for fine-tuning and keep the remaining segments as a held-out test set. The split is sentence-independent, with no sentence appearing in both the fine-tuning and test sets. To mitigate overfitting due to the limited data, additive background noise sampled from the MUSAN corpus is used for data augmentation, increasing the effective training set to 300 samples. During fine-tuning, the pretrained encoder parameters are frozen, while the feed-forward layers and classifier are updated for 5 epochs.

V. Results

In this section, we present and analyze the experimental results. The OVD task is formulated as a segment-level binary classification problem, where each speech segment is assigned a single label (own-voice vs. external-speaker). VoxCeleb1 contains variable-length utterances; during training we use the full available utterance duration up to a maximum of 15 s to expose the model to diverse phonetic content. The conformer encoder supports variable-length inputs via padding or cropping with attention masking and produces frame-level embeddings that are aggregated using temporal gate pooling to obtain a single segment-level prediction. At test time, we additionally evaluate the same trained model on shorter one-second segments to assess robustness under lower-latency conditions and to approximate practical real-world usage scenarios.

A. Training and progressive fine-tuning

The model trained solely on analytically simulated ATFs already demonstrated strong generalization to simulation-based test sets, achieving 86% accuracy on the head-and-torso dataset without any fine-tuning. Table III presents the test accuracy for each stage of fine-tuning. Testing was performed on the head-and-torso ATF dataset, and the test set includes variable-length utterances up to 15 seconds. When fine-tuning only the classification layer with numerically simulated ATFs from the same sphere geometry, accuracy dropped slightly to 83.78%, likely due to domain shift between analytical and numerical ATFs. This indicates that updating only the classifier is insufficient to bridge the gap between different modeling approaches. Progressively fine-tuning the classification layer alone with head-and-torso ATFs improved the accuracy to 91.04%, surpassing the original baseline. However, fine-tuning both the feed-forward and classification layers led to a substantial performance gain, achieving 95.52% accuracy on the same dataset. These findings highlight the importance of internal feature representations in adapting to complex spatial conditions, and indicate that limiting fine-tuning to the classification layer constrains the model's generalization capacity. Table IV shows the test accuracy of models evaluated on one-second utterances. The final fine-tuned model achieved an accuracy of 90.02% on the head-and-torso ATF dataset.

B. Noise augmentation

To further assess the robustness of the proposed model, we trained it with background noise augmentation using the MUSAN dataset. Table V reports the test accuracy across different noise types (Noise, Music, and Speech) and SNRs. The model shows reasonable resilience to noise, with performance decreasing gradually as SNR decreases. Furthermore, to

TABLE IV: Test accuracy (%) across different fine-tuning stages and ATF datasets, demonstrating generalization performance. The first section of the table reports results from the model trained on the analytical ATF dataset and tested on all ATF datasets. The second section presents results after progressive fine-tuning of both the feed-forward and classification layers. All models were evaluated using one-second speech segments.

Model	Test Accuracy (%)			
	Analytical ATFs	Rigid Sphere (Mesh)	Head Model (Mesh)	Head & Torso (Mesh)
Training on analytical ATFs				
Baseline (Analytical ATFs) (50 epochs)	89.59 ± 0.14	83.69 ± 0.25	86.43 ± 0.15	**85.12 ± 0.14**
Feed-Forward + classification layer fine-tuning on numerical simulation ATFs				
+ Fine-tuned on Rigid Sphere (5 epochs)	-	84.42 ± 0.27	88.74 ± 0.08	**86.87 ± 0.19**
+ Fine-tuned on Head Model (5 epochs)	-	-	92.61 ± 0.17	**88.51 ± 0.17**
+ Fine-tuned on Head & Torso (5 epochs)	-	-	-	**90.02 ± 0.29**

TABLE V: Robustness to different noise types and SNR levels. Test accuracy (%) of the model trained on VoxCeleb1 with MUSAN noise augmentation, evaluated on both VoxCeleb1 and LibriSpeech using one-second utterances. The test ATF set is based on the Mesh2HRTF head-and-torso configuration.

Noise Type	SNR (dB)	VoxCeleb1	LibriSpeech
Noise	0	75.86 ± 0.55	73.24 ± 0.49
	5	77.88 ± 0.34	75.55 ± 0.53
	10	79.45 ± 0.42	77.10 ± 0.57
	15	80.70 ± 0.35	78.28 ± 0.51
	20	82.30 ± 0.39	79.62 ± 0.49
	30	85.69 ± 0.16	82.79 ± 0.38
Music	0	71.07 ± 0.36	68.53 ± 0.42
	5	74.26 ± 0.21	71.70 ± 0.47
	10	77.62 ± 0.18	74.91 ± 0.37
	15	81.00 ± 0.28	77.67 ± 0.30
	20	83.69 ± 0.26	80.37 ± 0.45
	30	87.96 ± 0.15	84.83 ± 0.46
Speech	0	81.23 ± 0.24	75.46 ± 0.39
	5	84.19 ± 0.30	78.73 ± 0.27
	10	86.59 ± 0.33	81.62 ± 0.27
	15	88.41 ± 0.23	83.89 ± 0.49
	20	89.65 ± 0.09	85.67 ± 0.38
	30	90.67 ± 0.13	87.93 ± 0.33

assess generalization, we tested the model on the LibriSpeech dataset with added noise, a dataset not seen during training, and observed consistently high performance across all SNR levels. These results confirm the model's ability to generalize well across datasets and remain robust in diverse acoustic conditions. All evaluations were conducted on the Mesh2HRTF head-and-torso ATF test set using one-second utterances.

C. Baseline comparison and measured-ATF evaluation

To provide a direct baseline comparison and to isolate the contribution of the proposed simulation-based training strategy, we conduct three complementary evaluations. We use the ResNet-based single-microphone OVD model of López-Espejo et al. [19] as the baseline because, to the best of our knowledge, it is the only published baseline directly comparable to our setting. This setting involves offline, segment-level, single-microphone own-voice/external-speaker detection using measured hearing-aid ATFs. In all comparisons in this section, Conformer-small denotes a reduced configuration of the proposed conformer model, selected to have a parameter count similar to the ResNet baseline [19]. All models in these comparisons are trained and evaluated using one-second input segments to match the offline segment-level OVD setting.

The first evaluation compares the proposed Conformer-small model with the ResNet baseline on the measured-ATF benchmark. The second evaluation compares ResNet and Conformer-small under a controlled simulated-ATF setting, where both models are trained and evaluated using the same ATF-augmented Google Speech Commands data generated using our ATF pipeline. The third evaluation is a same-architecture comparison between Conformer-small trained only on measured ATFs and Conformer-small pretrained on simulated ATFs and then fine-tuned on measured ATFs. This final comparison isolates the benefit of the proposed simulated-ATF pretraining strategy from the effect of model architecture.

Table VI compares the performance and computational complexity of the proposed Conformer-small and the ResNet-based OVD baseline of López-Espejo et al. [19] on the measured-ATF benchmark. The ResNet baseline is trained using speech data generated with 49 measured ATF conditions, whereas the proposed system is pretrained using speech data generated with 17,966 analytically simulated ATF conditions and then partially fine-tuned using speech data generated with the same 49 measured ATF conditions. The proposed system improves own-voice, external-speaker, and overall accuracy, with a particularly large gain in external-speaker accuracy, while requiring substantially fewer MACs and lower CPU execution latency. These results indicate a better accuracy-complexity trade-off for offline segment-level OVD, which is important for hearing-aid deployment. Recent lightweight and low-power neural speech-enhancement studies [44], [45] further highlight the importance of computational efficiency for resource-constrained audio devices.

TABLE VII: OVD performance on simulated-ATF Google Speech Commands dataset [39]. Both models use the same dataset split and are sequentially fine-tuned up to the head-and-torso ATF stage.

Model	#Params (K)	Accuracy (%) on head & torso ATF data		
		Own	External	Overall
ResNet [19]	238.5	98.24	83.18	94.30
Conformer-small	307.2	**98.38**	**91.80**	**96.66**

As an additional controlled architecture comparison, Table VII compares ResNet and Conformer-small when both models are trained and evaluated on the same simulated-ATF Google Speech Commands data generated using our ATF pipeline with 17,966 analytically simulated ATF variations.

TABLE VI: Comparison between the baseline ResNet and the proposed Conformer-small architecture on the measured-ATF dataset. Performance and complexity metrics are reported per 101-frame audio segment (≈1 s), with execution latency measured on a host CPU (Intel Core i7, 2.90 GHz). MACs were estimated per input segment using framework profilers.

Model	Performance (%) on measured-ATF data				Computational Complexity		
	Own acc.	External acc.	Overall acc.	Macro-F1	#Params (K)/ Size (MiB)	Estimated MACs (M)	Execution Latency (ms)
ResNet [19]	97.52	80.81	93.15	90.75	238.5/ 0.91	1,002.49	8.40
Conformer-small	**98.77**	**89.94**	**96.46**	**95.41**	307.2 / 1.17	**17.61**	**2.98**

TABLE VIII: Same-architecture comparison for evaluating the benefit of simulated-ATF pretraining. Both systems use Conformer-small, and all trainable parameters are updated during measured-ATF training/fine-tuning. FT denotes fine-tuning.

Split protocol	Training strategy	Own acc.	External acc.	Overall acc.	Macro-Precision	Macro-Recall	Macro-F1
Speaker-disjoint	Measured only	99.10	95.69	98.21	97.95	97.40	97.67
Speaker-disjoint	Sim. pretrain + measured FT	**99.22**	**96.96**	**98.63**	**98.35**	**98.09**	**98.22**
External-ATF-disjoint	Measured only	99.16	93.87	97.51	97.67	96.52	97.07
External-ATF-disjoint	Sim. pretrain + measured FT	**99.40**	**96.84**	**98.60**	**98.62**	**98.12**	**98.36**

The results show that Conformer-small achieves comparable own-voice accuracy while providing substantially improved external-speaker and overall accuracy relative to the ResNet model. We note that the overall accuracy is influenced by class imbalance, with a larger proportion of own-voice samples, consistent with the evaluation protocol in [19].

Finally, Table VIII isolates the contribution of the proposed simulated-ATF pretraining strategy by holding the model architecture fixed. Both systems use the same Conformer-small architecture: one is trained only on measured ATFs, while the other is pretrained on simulated ATFs and subsequently fine-tuned on measured ATFs. Therefore, the performance difference reflects the benefit of simulated-ATF pretraining rather than a difference in model architecture. Simulated-ATF pretraining improves performance across both split protocols, with a more pronounced gain under the external-ATF-disjoint protocol, where external-speaker accuracy increases from 93.87% to 96.84%, indicating improved generalization to unseen ATFs.

D. Real-world hearing-aid recording evaluation

The model fine-tuned on simulated head-and-torso ATFs achieved 90.02% accuracy on the corresponding simulated test set. When evaluated on the real-world recordings using compensation (Eq. 12), the model attained 80% accuracy, demonstrating effective simulation-to-real-world transfer despite the absence of fine-tuning. The evaluation was performed on non-overlapping one-second segments. For reference, an oracle class-conditional compensation achieves 86.5% accuracy, representing an upper bound for the compensation-based approach.

Few-shot fine-tuning (see Section IV-E) improves the real-world accuracy to 91% without applying the compensation step during inference, indicating that limited model adaptation with a small set of real-world recordings can effectively reduce the simulation-to-real-world mismatch. For class-wise accuracy, the model obtains 92.0% for the external-speaker class and 90.0% for the own-voice class. The macro-precision, macro-recall, and macro-F1 scores are 91.06%, 91.0%, and 91.0%, respectively. Figure 11 shows ROC curves for the real-world hearing-aid recordings comparing test-time compensation and few-shot fine-tuning. Test-time compensation achieves an AUC of 0.80, while few-shot fine-tuning improves the detection performance to an AUC of 0.96. These results demonstrate the effectiveness of model adaptation and further support the viability of simulation-based training for OVD in hearing aids. The present work evaluates OVD in an offline, segment-level setting. A fully causal low-latency hearing-aid implementation is left for future work.

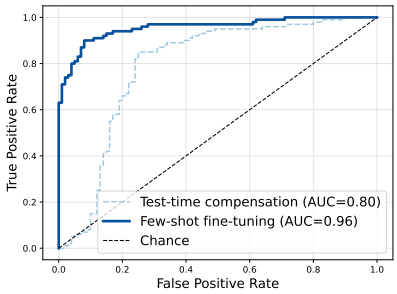


Fig. 11: ROC curves for real-world test data.

E. Analysis of acoustic features for OVD

While previous results demonstrate the model's robustness, it remains important to understand the acoustic features the model relies on for OVD. In particular, loudness differences between the wearer's voice and external speakers could be a simple cue that the model might exploit. To investigate this, we conducted experiments where the amplitude of the external speaker's audio was artificially increased to simulate closer proximity. Despite these modifications, the model's detection accuracy remained largely unchanged, achieving 89.90% for the fine-tuned model on the head-and-torso dataset. These results indicate that the model does not rely solely on amplitude cues but instead leverages spatial acoustic features to discriminate between own voice and other speakers. This is crucial for real-world use, where speaker volume may vary significantly due to differences in speaking style or environment.

Additionally, repeating the one-second utterance to fill the 15-second input window did not improve accuracy compared to zero-padding, indicating that the model does not benefit from redundant temporal information. Finally, an ablation study where the ATF input was removed during testing resulted in a substantial drop in accuracy to 49.67%, confirming that spatial cues are critical for the classification.

VI. Conclusion

This study presents a simulation-based training framework for single microphone OVD in hearing aids. A conformer-based classifier was first trained on analytically derived ATFs to expose the model to a wide range of spatial configurations, and subsequently fine-tuned using numerically simulated ATFs with increasing anatomical realism. Unlike prior approaches that rely on multi-microphone arrays, speaker-dependent cues, or costly transfer-function measurements, the proposed method leverages simulated spatial propagation cues to distinguish own voice from external speakers. The model achieved 95.52%

accuracy on simulated head-and-torso test data using full-length utterances and maintained 90.02% accuracy on one-second segments, demonstrating robustness to shorter input durations. Evaluation on the measured-ATF benchmark showed that simulated-ATF pretraining improves performance, particularly under unseen ATFs, indicating that exposure to more diverse simulated ATFs enhances the effectiveness and generalization of the model. Few-shot fine-tuning using a small set of real-world hearing-aid recordings achieved 91% accuracy, indicating that the simulation-trained model can be effectively adapted using real-world data. Although the real-world validation is limited in scale, it provides an initial feasibility demonstration that models trained on simulated ATFs can generalize to the real world. Future work will focus on reducing model size, computational cost, and energy consumption for hearing-aid deployment, informed by recent advances in lightweight and low-power neural speech-processing models [44], [45], while also targeting shorter-window detection, causal real-time operation, and validation on larger and more diverse real-world hearing-aid datasets.

ACKNOWLEDGMENTS

Funding for this work was provided by GN Store Nord A/S.

REFERENCES

[1] M. Froehlich, T. Powers, E. Branda, and J. Weber, "Perception of own voice wearing hearing aids: Why "natural" is the new normal," *Audiology Online*, 2018.

[2] E. H. Høydal, "A new own voice processing system for optimizing communication," *Hearing Review*, vol. 24, no. 11, pp. 20–22, 2017.

[3] J. Hengen, I. L. Hammarström, and S. Stenfelt, "Perception of one's own voice after hearing-aid fitting for naive hearing-aid users and hearing-aid refitting for experienced hearing-aid users," *Trends in hearing*, vol. 24, p. 2331216520932467, 2020.

[4] J. M. Kates, *Hearing Aids*. Springer, 2008.

[5] K. Chung, "Challenges and recent developments in hearing aids: Part i. speech understanding in noise, microphone technologies and noise reduction algorithms," *Trends in amplification*, vol. 8, no. 3, pp. 83–124, 2004.

[6] I. Merks, "Hearing assistance system with own voice detection," *Acoustical Society of America Journal*, vol. 135, no. 1, p. 573, 2014.

[7] O. Al-hazaimeh, S. A. Alomari, J. Alsakran, and N. Alhindawi, "Cross correlation–new based technique for speaker recognition," *Int J Acad Res*, vol. 6, pp. 232–239, 2014.

[8] K. B. Rasmussen and S. Laugesen, "Method for detection of own voice activity in a communication device," Mar. 31 2009, US Patent 7,512,245.

[9] M. Zohourian, G. Enzner, and R. Martin, "Binaural speaker localization integrated into an adaptive beamformer for hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 515–528, 2017.

[10] J. Wang, H. Liu, L. Xu, W. Yang, W. Yi, and F. Liu, "Lightweight target speaker separation network based on joint training," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, p. 53, 2023.

[11] K. Patterson, K. Wilson, S. Wisdom, and J. R. Hershey, "Distance-based sound separation," *arXiv preprint arXiv:2207.00562*, 2022.

[12] D. Wang, X. Xiao, N. Kanda, T. Yoshioka, and J. Wu, "Target speaker voice activity detection with transformers and its integration with end-to-end neural diarization," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] M. Neri, A. Politis, D. Krause, M. Carli, and T. Virtanen, "Speaker distance estimation in enclosures from single-channel audio," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[14] Y.-L. Wei, B. Islam, R. RAJAGOPALAN, and romit choudhury, "Self-supervised speech enhancement using multi-modal data," 2023. [Online]. Available: https://openreview.net/forum?id=OqPD_6kukm

[15] P. Hoang, J. M. De Haan, Z.-H. Tan, and J. Jensen, "Multichannel speech enhancement with own voice-based interfering speech suppression for hearing assistive devices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 706–720, 2022.

[16] K. Kazama, T. Nakashima, and N. Ono, "Measurement of relative transfer function for own voice in head-mounted microphone array," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2024, pp. 1–5.

[17] J. Pohlhausen, I. Holube, and J. Bitzer, "Near-ear sound pressure level distribution in everyday life considering the user's own voice and privacy," *Acta Acustica*, vol. 6, p. 40, 2022.

[18] P. Pertilä, E. Fagerlund, A. Huttunen, and V. Myllylä, "Online own voice detection for a multi-channel multi-sensor in-ear device," *IEEE Sensors Journal*, vol. 21, no. 24, pp. 27 686–27 697, 2021.

[19] I. López-Espejo, Z.-H. Tan, and J. Jensen, "Keyword spotting for hearing assistive devices robust to external speakers," *arXiv preprint arXiv:1906.09417*, 2019.

[20] ——, "Improved external speaker-robust keyword spotting for hearing assistive devices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1233–1247, 2020.

[21] R. Blandin and M. Brandner, *Influence of the vocal tract on voice directivity*. Universitätsbibliothek der RWTH Aachen, 2019.

[22] M. Brandner, R. Blandin, M. Frank, and A. Sontacchi, "A pilot study on the influence of mouth configuration and torso on singing voice directivity," *The Journal of the Acoustical Society of America*, vol. 148, no. 3, pp. 1169–1180, 2020.

[23] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[24] R. M. Aarts and A. J. Janssen, "Comparing sound radiation from a loudspeaker with that from a flexible spherical cap on a rigid sphere," *Journal of the Audio Engineering Society*, vol. 59, no. 4, pp. 201–212, 2011.

[25] L. Beranek and T. Mellow, *Acoustics: sound fields, transducers and vibration*. Academic Press, 2019.

[26] "Spherical Wave Scattering by a Rigid Sphere," http://ansol.us/Products/Coustyx/Validation/MultiDomain/Scattering/SphericalWave/HardSphere/Downloads/dataset_description.pdf, accessed: 2024-02-04.

[27] R. M. Aarts and A. J. Janssen, "Sound radiation from a resilient spherical cap on a rigid sphere," *The Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2262–2273, 2010.

[28] L. L. Beranek and T. Mellow, *Acoustics: sound fields and transducers*. Academic Press, 2012.

[29] H. Ziegelwanger, W. Kreuzer, and P. Majdak, "Mesh2hrtf: Open-source software package for the numerical calculation of head-related transfer functions," in *22nd international congress on sound and vibration*, 2015.

[30] W. Kreuzer, K. Pollack, P. Majdak, and F. Brinkmann, "Mesh2hrtf/numcalc: An open-source project to calculate hrtfs and wave scattering in 3d," in *Proceedings of the Euroregio BNAM Joint Acoustics Conference*, 2022, pp. 443–452.

[31] Z.-S. Chen, H. Waubke, and W. Kreuzer, "A formulation of the fast multipole boundary element method (fmbem) for acoustic radiation and scattering from three-dimensional structures," *Journal of Computational Acoustics*, vol. 16, no. 02, pp. 303–320, 2008.

[32] L. Gaul, M. Kögl, and M. Wagner, *Boundary element methods for engineers and scientists: an introductory course with advanced topics*. Springer Science & Business Media, 2013.

[33] B. Foundation, "Blender: 3d creation suite," accessed: [01 February 2023]. [Online]. Available: https://www.blender.org

[34] MeshInspector Team, "Meshinspector," 2025, accessed: Jan 02, 2025. [Online]. Available: https://meshinspector.com/

[35] H. Kawano and S. Shimizu, "An effective transformer-based contextual model and temporal gate pooling for speaker identification," *arXiv preprint arXiv:2308.11241*, 2023.

[36] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[37] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[38] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[39] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[40] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.
[41] A. Barcovschi, R. Jain, and P. Corcoran, "A comparative analysis between conformer-transducer, whisper, and wav2vec2 for improving the child speech recognition," in *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 2023, pp. 42–47.
[42] J. M. Kates, *Digital Hearing Aids*. Plural Publishing, 2008.
[43] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European conference on computer vision*. Springer, 2016, pp. 443–450.
[44] H. Yan, J. Zhang, C. Fan, Y. Zhou, and P. Liu, "Lisennet: Lightweight sub-band and dual-path modeling for real-time speech enhancement," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
[45] E. Liu, A. Li, C. Fan, C. Zheng, J. Yi, R. Fu, X. Li, J. Zhou, and Z. Lv, "Sse-net: Toward low-power-consumption spiking neural network for monaural speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 2077–2089, 2026.

Mathuranathan Mayuravaani is a Ph.D. candidate at the School of Engineering and Computer Science, Victoria University of Wellington, New Zealand. She received her B.Sc. Hons (Computer Science) degree from the University of Jaffna, Sri Lanka. Her research interests include speech signal processing and machine learning.

W. Bastiaan Kleijn is Professor at Victoria University of Wellington, New Zealand (since 2010) and a research scientist at Google (since 2011). He was a professor at TU Delft (2011-2021) and KTH Stockholm (1996-2014) and prior to that a Member of Technical Staff in the Research Division of AT&T Bell Laboratories. He holds a PhD in Soil Science and an MSc in Physics from the University of California, Riverside, an MSEE from Stanford University, and a PhD in Electrical Engineering from TU Delft. He is Fellow of the IEEE (1999), the Royal Society of New Zealand (2021), and Engineering New Zealand (2024).

Andrew Lensen is a Senior Lecturer and Programme Director of Artificial Intelligence at Victoria University of Wellington, New Zealand. He holds a B.Sc.(Hons) and a Ph.D. degree (2019) from Victoria University of Wellington. His current research interests include explainable AI, interdisciplinary applications of AI, and the societal impacts of AI use in New Zealand. He is a Senior Member of the IEEE (2025).

Charlotte Sørensen is a Research Scientist at GN Hearing A/S, Denmark. She received the B.S. and M.Sc. degrees in Biomedical Engineering and Informatics, and the Ph.D. degree, from Aalborg University, Denmark. Her research interests include speech signal processing with emphasis on speech intelligibility prediction, measurement of speech intelligibility, and speech enhancement for hearing aid applications.