Mobile Microphone-Based Sound Source Detection and Localization across Diverse Acoustic Scenes

Soma Sugioka[1] and Yuki Matsuda[1 [0000−0002−3135−4915]]

Okayama University, Okayama, Okayama 700–0082, Japan
soma.sugioka@cocolab.jp, yukimat@okayama-u.ac.jp

Abstract. This study investigates beep detection and eight-direction sound source localization using only the built-in microphones of an iPhone 12 Pro. The system combines Goertzel-based detection of a 440 Hz beep with a residual-dense localization model using left and right Log-Mel spectrograms, inter-channel phase difference, and inter-channel level difference. Direction-labeled simulated audio was generated from measured device-related transfer functions, and a real-recording dataset containing 800 samples from ten acoustic scenes was constructed. Standard real-recording and take-split evaluations achieved accuracies of 97.50% and 98.33%, respectively, whereas leave-one-scene-out evaluation yielded a mean accuracy of 34.25%, indicating limited generalization to unseen scenes. In an on-device test, all 160 reference beeps were detected, but exact localization accuracy and accuracy within 45 degrees under the original 1.5 m condition were 10.00% and 36.25%, respectively. These results demonstrate controlled-condition detection feasibility while identifying unseen-scene generalization and localization under the original 1.5 m condition as the principal remaining challenges.

Keywords: Smartphone sound source localization · Built-in microphones · Beep detection · Device-related transfer function.

1 Introduction

Mobile safety applications for cyclists would benefit from the ability to detect approaching vehicles and estimate their directions, thereby improving awareness of the surrounding traffic environment. Acoustic sensing provides a potential means of obtaining such information, but conventional sound source localization generally relies on dedicated microphone arrays, which are impractical for everyday mobile use. To address this limitation, our long-term goal is to realize acoustic-event detection and direction estimation using only a smartphone mounted on a bicycle. Smartphones integrate multiple built-in microphones with sufficient computing capability for on-device signal processing, making them sensing platforms that require no additional audio hardware. However, their microphones are more closely spaced than those of dedicated arrays, and their placement is constrained by the device enclosure. Moreover, inter-channel time, phase, and level

differences are affected not only by device geometry but also by reverberation, background noise, source characteristics, and the recording path.

As a controlled first step toward this long-term goal, this study focuses on a beep with a known frequency and performs sound detection and direction estimation using only the built-in microphones of an iPhone 12 Pro[1]. By restricting the target signal to a 440 Hz pure tone lasting 2 s, the system can detect the presence of the beep in a frequency-selective manner and perform localization on the detected audio segment. Detection and localization are implemented as separate processes. A Goertzel-based detector triggers localization, and a deep learning model classifies the stereo recording into eight directions at 45-degree intervals.

The study does not target a specific user demographic or seek to replace human auditory perception of the beep; it evaluates the technical feasibility of smartphone-based machine detection and direction estimation.

Simulated audio generated from device-specific directional characteristics can reduce the cost of collecting a large amount of direction-labeled real audio. By analogy with HRTFs, we refer to the directional transfer characteristics introduced by the iPhone enclosure and microphone placement as device-related transfer functions (DRTFs) throughout this paper. We measure the DRTFs of the iPhone 12 Pro and convolve them with the beep signal to generate training audio. Although simulation provides directional labels efficiently, a model trained on simulated audio does not necessarily transfer directly to real recordings. Real environments introduce reverberation, background noise, loudspeaker characteristics, and recording levels that differ from those of the DRTF measurement. We therefore combine simulated audio with real recordings and evaluate generalization to unseen acoustic environments.

To investigate this problem, we construct a multi-scene dataset in which the device, source, and geometry are fixed while the recording environment is varied. The completed dataset contains 800 beep recordings across ten scenes, with ten takes for each of eight directions in every scene. Each scene represents one independent indoor or outdoor acoustic environment, enabling scene-level evaluation in previously unseen environments.

The contributions of this work are threefold. First, we present an end-to-end system that detects a beep and estimates one of eight source directions using only built-in iPhone microphones, without an external microphone array or additional sensors. Second, we construct a localization pipeline that combines iPhone-specific DRTFs with Log-Mel, inter-channel phase difference (IPD), and inter-channel level difference (ILD) features and optimizes a residual-dense model using direction classification and metric learning. Third, we collect recordings in ten independent environments under fixed acquisition conditions and evaluate generalization to unseen environments using leave-one-scene-out cross-validation. The evaluation also reveals a substantial gap between within-scene and unseen-scene performance, clarifying the present limitations of the learned localization model.

[1] `https://support.apple.com/en-us/111875`

2 Related Work

2.1 Smartphone-Based Acoustic Sensing for Mobile Safety

Smartphone-based acoustic sensing includes approaching-vehicle detection without direction estimation [4] and active obstacle localization using self-emitted signals [8]; neither passively localizes an external sound source. In contrast, Küçük *et al.* implemented real-time CNN-based passive speech-source localization on an Android smartphone using two built-in microphones [5]. Whereas their method localizes speech over a 0°–180° range, we target a known beep in eight azimuth classes around an iPhone.

2.2 Deep Learning-Based Detection and Localization with Compact Microphone Arrays

Goli and van de Par proposed a residual-dense model for speech detection and localization using four or six hearing-aid microphones [3]. Their training and evaluation used RAZR simulations and measured HRTFs, including unseen simulated rooms, but did not examine transfer to real recordings [9]. We adapt this design to two built-in iPhone microphones, add metric learning, and use Goertzel-based beep detection followed by eight-direction classification with Log-Mel, IPD, and ILD features. Generalization is evaluated using real recordings separated at the scene level.

2.3 HRTF Modeling and PCA-Based Directional Interpolation

A head-related transfer function (HRTF) describes the acoustic path from a source to a listener's ear and contains direction-dependent spectral and temporal characteristics introduced by the head, torso, and pinnae. Because measuring an individual's HRTFs at many directions requires considerable time and effort, methods have been studied for estimating unmeasured responses from a limited set of measurements.

For a source at azimuth θ, let $P_0(f)$ denote the free-field sound pressure and $P_e(\theta, f)$ the pressure observed at the ear-canal entrance. The frequency-domain HRTF and its corresponding head-related impulse response (HRIR) are expressed as

$$H_e(\theta, f) = \frac{P_e(\theta, f)}{P_0(f)}, \qquad h_e(\theta, n) = \mathcal{F}^{-1}\{H_e(\theta, f)\}\,. \tag{1}$$

This representation separates the source signal from the directional filtering introduced along the propagation path.

Matsui and Ando proposed a PCA-based individualization method that derives orthogonal bases from a set of HRTFs and estimates partially measured

HRTFs as weighted sums of those bases [6]. Wang *et al.* decomposed HRTF magnitude spectra into direction-independent principal components and direction-dependent component weights and interpolated unmeasured directions by modeling the spatial variation of the weights [7]. These approaches separate high-dimensional directional spectra into a small number of shared bases and direction-dependent weights, reducing the number of measurements required for directional interpolation.

The terms device-related transfer function (DRTF) and device-related impulse response (DRIR) are used in this paper by analogy with HRTF and HRIR, respectively. They represent the frequency- and time-domain directional responses introduced by the iPhone enclosure and microphone placement rather than by the human body. We apply the PCA-based representation used in HRTF interpolation to these device responses. Specifically, the left and right magnitude spectra measured at eight azimuths are reduced by PCA, and the component weights are interpolated along the azimuth using periodic cubic splines to generate responses at one-degree intervals. Thus, we adapt the general PCA-based interpolation principle to device responses on the horizontal plane rather than directly adopting a particular HRTF interpolation algorithm.

2.4 Simulation-to-Real Generalization in Acoustic Localization

Evaluation confined to simulated signals does not establish whether the learned directional cues transfer to real recordings. Audio generated from measured transfer functions does not fully reproduce the reverberation, background noise, loudspeaker response, and recording gain encountered in different real environments. Consequently, localization performance obtained within simulation may not be maintained for real recordings or unseen acoustic scenes.

We train the model using simulated audio and real recordings and evaluate generalization using leave-one-scene-out cross-validation. The training and test data are separated at the scene level rather than only at the recording level, preventing information from the evaluation scene from entering the training data. Rather than proposing a general-purpose domain-adaptation method, this study examines whether a controlled smartphone localization system generalizes to unseen real acoustic environments.

In relation to these studies, we investigate an integrated system comprising beep detection using only built-in iPhone microphones, DRTF-based generation of direction-labeled audio, eight-direction localization using a residual-dense model with metric learning, and unseen-scene evaluation across multiple real environments. The central contribution is not to introduce each component in isolation, but to adapt a hearing-aid detection and localization model to a two-microphone smartphone configuration, augment it with metric learning, and evaluate it as an on-device detection and localization pipeline in real acoustic environments.

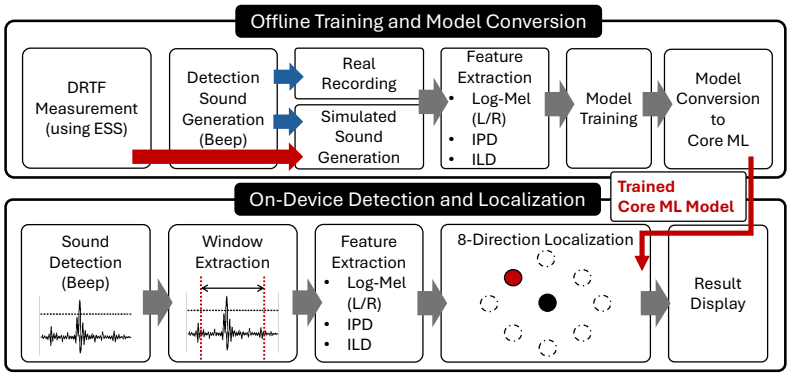


Fig. 1. End-to-end overview of the proposed system. The upper lane shows model development using DRTF measurement, simulated and real recordings, feature extraction, training, and Core ML conversion. The lower lane shows on-device beep detection and eight-direction sound source localization.

3 System Overview

The proposed system detects a known 440 Hz beep and estimates its direction of arrival as one of eight classes at 45-degree intervals. The overall system is divided into two phases: deep-learning model development and on-device detection and localization. The end-to-end workflow is summarized in Fig. 1.

The localization model is trained using DRTF-convolved beeps and real recordings from non-held-out scenes, converted to Core ML, and integrated into the iPhone application [1]. DRTF measurement and simulated-data generation are described in Section 4, the real-recording dataset in Section 5, and localization-model training in Section 6.

The on-device detection and localization phase consists of six processes: stereo recording, beep detection, localization-window extraction, acoustic feature generation, eight-direction classification, and result presentation. The iPhone 12 Pro captures 44.1 kHz stereo audio from its front and bottom microphones. For beep detection, the system averages the two channels into a mono signal and monitors the frequency region around 440 Hz. When the detection criterion is satisfied, the system forms a 2 s stereo localization window by combining the most recent 0.5 s of buffered audio with 1.5 s of audio acquired after detection. The system then computes five-channel features comprising left and right Log-Mel spectrograms, cos(IPD), sin(IPD), and ILD. The trained model outputs a direction class and its estimated probability. A direction estimate is accepted when the maximum class probability is at least 0.40; otherwise, the result is displayed as uncertain. Finally, the iOS application presents the estimated direction on an eight-direction radar-style interface.

3.1 Beep Detection

The target is a 440 Hz pure-tone beep with a duration of 2 s. The input-node tap is configured with a buffer size of 4,096 samples, and the left and right channels of each stereo buffer converted to 44.1 kHz are averaged to form a mono signal. The nominal detection-window length and update interval are therefore both approximately 92.9 ms. For a mono buffer $x[n]$ of length N, the Goertzel algorithm computes the state at frequency f [2].

$$s_f[n] = x[n] + 2\cos(\omega_f)s_f[n-1] - s_f[n-2], \quad \omega_f = \frac{2\pi f}{f_s}, \tag{2}$$

where $s_f[-2] = s_f[-1] = 0$ and f_s is the sampling frequency. The power at f is obtained from the final two states as

$$P(f) = s_f^2[N-1] + s_f^2[N-2] - 2\cos(\omega_f)s_f[N-1]s_f[N-2]. \tag{3}$$

Let $\mathcal{F}_{\mathrm{t}} = \{420, 430, 440, 450, 460\}$ Hz and $\mathcal{F}_{\mathrm{g}} = \{250, 300, 600, 700, 900, 1200\}$ Hz denote the target and guard frequency sets, respectively. The band-power ratio R and RMS level L_{RMS} are

$$R = \frac{\sum_{f\in\mathcal{F}_{\mathrm{t}}} P(f)}{\sum_{f\in\mathcal{F}_{\mathrm{g}}} P(f)}, \qquad L_{\mathrm{RMS}} = 20\log_{10}\left(\sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x^2[n]}\right) \text{ dBFS}. \tag{4}$$

A buffer is counted as a hit when $R \geq 5.0$ and $L_{\mathrm{RMS}} \geq -60$ dBFS. These thresholds were provisionally selected during implementation of the current prototype rather than optimized on the evaluation data, and they are held fixed across all evaluation conditions. The detector declares a beep after three consecutive hits, corresponding to a nominal minimum duration of approximately 278.6 ms. A 2.2 s cooldown prevents repeated detections during the same 2 s beep. The detector does not estimate the source direction; its sole role is to determine the audio segment passed to the localization model.

3.2 Direction Definition

The reference orientation is defined as 0 degrees when the iPhone screen directly faces the loudspeaker. Positive angles increase clockwise when the iPhone is viewed from above, forming eight classes from 0 to 315 degrees in 45-degree increments. This rotation appears counterclockwise when viewed from the loudspeaker. The direction convention and fixed recording geometry are shown in Fig. 2.

4 DRTF Measurement and Simulated Data Generation

To incorporate the direction-dependent acoustic characteristics introduced by the iPhone enclosure and microphone placement, we measure the directional

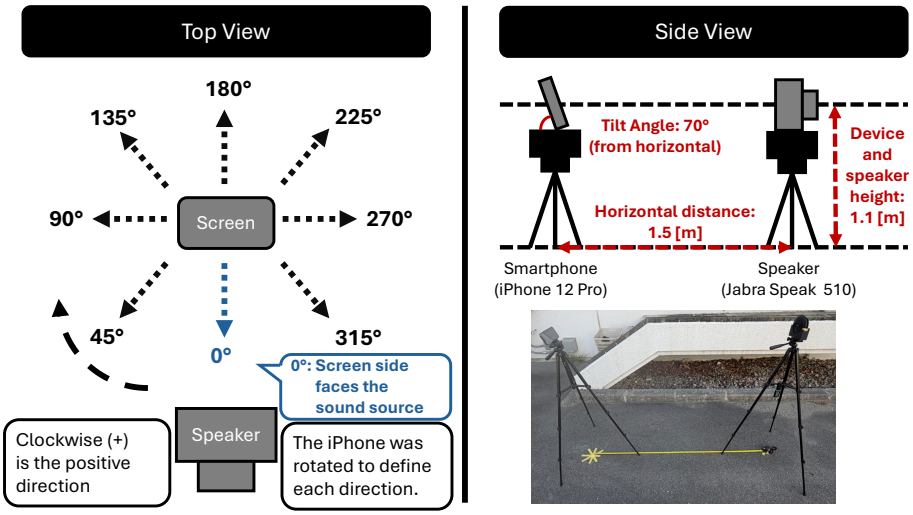


Fig. 2. Definition of the source directions and geometry of the controlled recording setup. The eight azimuth classes increase clockwise from 0° on the screen-facing side of the iPhone. The iPhone and loudspeaker are positioned 1.5 m apart at a height of 1.1 m.

device responses referred to in this paper as DRTFs by analogy with HRTFs. The iPhone 12 Pro and a Jabra Speak 510 loudspeaker[2] are placed at the same height, and stereo responses are acquired for eight directions at 45-degree intervals. For each direction, three takes of a playback signal comprising 1 s of silence, a 10 s ESS from 20 Hz to 20 kHz, and 1 s of silence are recorded at 44.1 kHz.

Each raw recording lasted more than 13 s, and a 12 s segment from 1 s to 13 s after the start of recording was retained for subsequent processing. The inverse filter is generated by reversing the ESS in time and applying an exponentially decaying envelope corresponding to a -6 dB/octave correction. The impulse response of each channel is estimated by FFT convolution of the recording with the inverse filter. The earlier of the left and right peaks is used as a common anchor, and 2,048 samples before and after that anchor are extracted from both channels, yielding a 4,096-sample response. This common framing preserves the inter-channel time difference, after which the three takes for each direction are averaged.

4.1 DRTF Representation and PCA-Based Interpolation

Analogously to the HRTF formulation, let $s[n]$ be the known sweep signal and $y_{c,i}[n]$ the signal recorded at channel $c \in \{L, R\}$ for measured azimuth θ_i. Given the inverse sweep filter $g[n]$, the measured DRIR is estimated by

$$\widetilde{h}_{c,i}[n] = y_{c,i}[n] * g[n], \tag{5}$$

[2] `https://www.jabra.com/supportpages/jabra-speak-510`

where $*$ denotes convolution.

The three responses at each azimuth are averaged, and a direct-path window $w_{c,i}[n]$ is applied to the averaged response $\bar{h}_{c,i}[n]$. After an FFT and one-twelfth-octave smoothing, the DRTF magnitude supplied to PCA is represented in decibels as

$$D_{c,i}[k] = 20\log_{10}\left(\left|\mathcal{S}_{1/12}\left\{\mathcal{F}\left[w_{c,i}[n]\bar{h}_{c,i}[n]\right]\right\}[k]\right| + \epsilon\right), \tag{6}$$

where $\mathcal{S}_{1/12}\{\cdot\}$ denotes one-twelfth-octave smoothing of the complex spectrum and ϵ prevents taking the logarithm of zero.

For each channel, the directional DRTF magnitudes are arranged as $\mathbf{D}_c \in \mathbb{R}^{N_d \times F}$, where $N_d = 8$ is the number of measured azimuths and F is the number of frequency bins. The directional mean and centered matrix are

$$\boldsymbol{\mu}_c = \frac{1}{N_d}\sum_{i=1}^{N_d}\mathbf{D}_{c,i:}, \tag{7}$$

$$\mathbf{X}_c = \mathbf{D}_c - \mathbf{1}\boldsymbol{\mu}_c. \tag{8}$$

The centered matrix is decomposed using singular value decomposition (SVD), and its right singular vectors are used as frequency-domain principal component bases:

$$\mathbf{X}_c = \mathbf{U}_c\boldsymbol{\Sigma}_c\mathbf{V}_c^{\mathsf{T}}, \tag{9}$$

$$a_{c,i,k} = \mathbf{X}_{c,i:}\mathbf{p}_{c,k},\ \mathbf{p}_{c,k} = \mathbf{V}_{c,:,k}. \tag{10}$$

We retain $K = 5$ components independently for the left and right channels.

For each component, the eight measured weights $a_{c,i,k}$ are interpolated with a periodic cubic spline satisfying continuity between 0° and 360°, producing $\widetilde{a}_{c,k}(\theta)$ at an arbitrary azimuth. The DRTF magnitude at an interpolated azimuth is reconstructed as

$$\widehat{\mathbf{D}}_c(\theta) = \boldsymbol{\mu}_c + \sum_{k=1}^{K}\widetilde{a}_{c,k}(\theta)\mathbf{p}_{c,k}^{\mathsf{T}}. \tag{11}$$

In this study, Equation (11) is evaluated at one-degree intervals to obtain 360 directional magnitude responses.

For each averaged DRIR, the direct response is extracted using a 6-ms Hann window beginning 3 ms before the channel peak and is zero-padded to 8,192 points. The resulting spectrum is smoothed at a resolution of one-twelfth octave. Five-component principal component analysis is applied separately to the left and right magnitude spectra, and the component weights are interpolated at one-degree intervals using periodic cubic splines. In parallel, the ITD at each of the eight measured directions is obtained by cross-correlating the averaged left and right DRIRs, and the ITDs are likewise interpolated at one-degree intervals using a periodic cubic spline.

A 4,096-sample minimum-phase DRIR is reconstructed from each interpolated magnitude spectrum using the homomorphic method. The interpolated

ITD is then applied to either the left or right DRIR as a combination of integer delay and frequency-domain fractional delay. The resulting SOFA representation contains 360 directions, two channels, and 4,096 samples per response. Finally, the direction-dependent DRTFs are convolved with the 2 s beep to generate simulated audio. The complete DRTF measurement and simulated-audio generation pipeline is summarized in Fig. 3.

4.2 Simulated Dataset Construction

After convolution, the two channels are jointly normalized with a shared scale factor so that their maximum absolute amplitude is 0.95. A common gain is then applied to both channels at ten levels from 0.1 to 1.0 in increments of 0.1, thereby preserving the inter-channel level difference. Each result is trimmed or zero-padded to 88,200 samples, yielding 2 s, two-channel audio at 44.1 kHz. Among the 360 one-degree azimuths, the 72 azimuths divisible by five are used for validation and the remaining 288 azimuths are used for training. The resulting simulated dataset contains 2,880 training recordings and 720 validation recordings. Each azimuth is assigned to the nearest of the eight direction classes at 45-degree intervals.

An additional set of 24 simulated recordings is retained for diagnostic checks only. Because its azimuths overlap the training azimuth set, it is not treated as an independent test condition for the research questions. The retained PCA representation and interpolation are validated in Fig. 4.

5 Multi-Scene Real-Recording Dataset

5.1 Recording Protocol

Real recordings are captured at 44.1 kHz in stereo using the front and bottom microphones of an iPhone 12 Pro fixed in landscape orientation. The same Jabra Speak 510 loudspeaker used for DRTF measurement is placed 1.5 m from the iPhone, and both devices are positioned at a height of 1.1 m. The source emits a 440 Hz beep for 2 s at a loudspeaker volume held constant throughout the recordings. Eight directions are recorded at 45-degree intervals from 0 to 315 degrees. The recording geometry and direction labels are illustrated in Fig. 2. The fixed acquisition conditions are summarized in Table 1.

5.2 Dataset Composition and Preprocessing

A scene is defined as one independent recording environment. Across scenes, environmental characteristics such as the room or outdoor location, reverberation, and background noise vary, whereas the device, microphone configuration, signal, distance, height, orientation, and loudspeaker volume remain fixed. The completed dataset comprises ten scenes, SC001–SC010, with 10 takes for each of eight directions in every scene, yielding 800 recordings containing the beep.

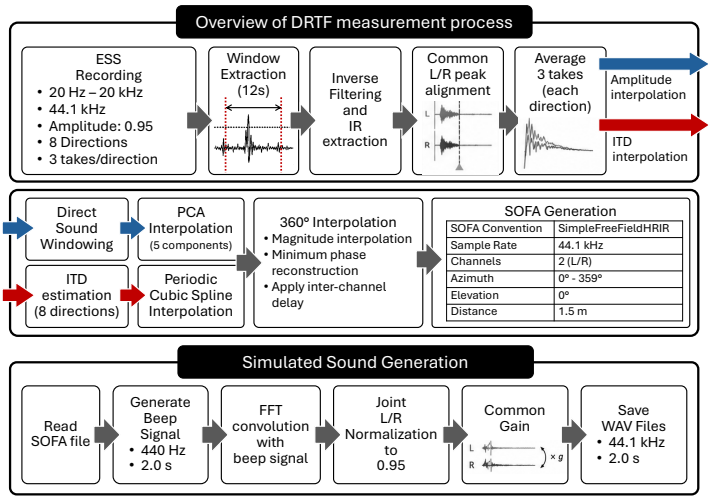


Fig. 3. DRTF measurement and simulated-audio generation pipeline. The upper diagram shows ESS-based response measurement, interpolation, and SOFA generation. The lower diagram shows convolution with the generated beep, joint L/R normalization, common-gain scaling, and WAV output.

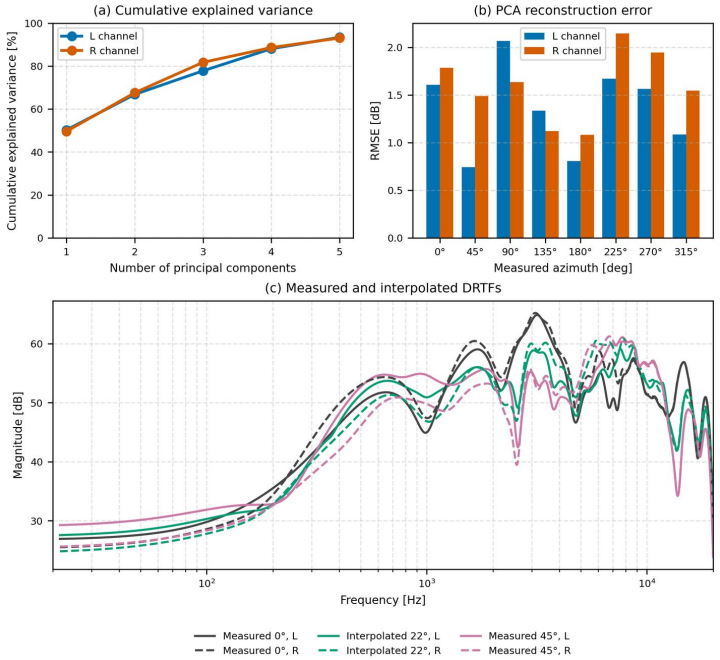


Fig. 4. Analysis of the PCA-based representation and an example of directional interpolation for the left and right DRTFs: (a) cumulative explained variance for the first five components, (b) reconstruction RMSE at the eight measured azimuths using five components, and (c) measured magnitude responses at 0° and 45° together with the response interpolated at 22°.

Table 1. Fixed conditions for the real recordings.

Item	Setting
Smartphone	iPhone 12 Pro
Orientation	Landscape
Microphones	Front and bottom built-in microphones
Sampling	44.1 kHz, stereo
Target signal	440 Hz pure tone, 2 s
Source–device distance	1.5 m
Device and source height	1.1 m
Directions	8 classes at 45-degree intervals
Loudspeaker volume	Fixed across all recordings



Fig. 5. Representative indoor and outdoor recording environments used for dataset construction: (a) and (b) indoor scenes, and (c) and (d) outdoor scenes.

In addition, 10 ambient recordings are collected in each scene, yielding 100 noise recordings used as augmentation material for simulated training audio. Representative indoor and outdoor recording environments are shown in Fig. 5.

For each real recording, the onset of the sustained 440 Hz component is detected from the ratio between the 400–480 Hz band and adjacent reference bands in the channel-wise STFT. A common 2 s segment is then extracted from both channels and normalized to a peak magnitude of 0.95 using a single scale factor for both channels, thereby preserving the ILD. The processed files are verified to contain two channels and 88,200 samples at 44.1 kHz. The indoor/outdoor classification of the ten scenes is listed in Table 2.

Table 2. Indoor/outdoor classification of the real-recording scenes.

Scene ID	Environment	Scene ID	Environment
SC001	Indoor	SC006	Outdoor
SC002	Outdoor	SC007	Outdoor
SC003	Indoor	SC008	Outdoor
SC004	Outdoor	SC009	Outdoor
SC005	Outdoor	SC010	Outdoor

6 Sound Source Localization Model

6.1 Acoustic Features

The model input is a five-channel representation consisting of left and right Log-Mel spectrograms, the cosine and sine of the IPD, and the ILD. The two Log-Mel channels represent timbral and direction-dependent spectral changes observed at each microphone. The IPD represents inter-channel arrival differences, particularly at lower frequencies, and its cosine and sine avoid the discontinuity of wrapped phase. The ILD represents inter-channel amplitude differences, including the shadowing effect of the device enclosure.

The preprocessing uses a sampling rate of 44.1 kHz, an FFT length of 1,024 with a Hann window, a hop length of 512, 64 Mel bands, and 173 time frames, producing an input tensor of size $5 \times 64 \times 173$. A frame is treated as silent when its mean left-channel Log-Mel energy is more than 25 dB below the maximum frame value; the cosine and sine IPD channels are set to zero in those frames. The global mean and standard deviation are calculated jointly from the left and right Log-Mel channels and used to normalize both channels, while the ILD is scaled by the same standard deviation.

6.2 Network and Training Objective

The localization model is a late-fusion residual-dense CNN that processes the five input features in two branches before fusion. The overall architecture of the localization model is shown in Fig. 6. The Log-Mel branch receives the left and right Log-Mel features, whereas the spatial branch receives cos(IPD), sin(IPD), and ILD. In each branch, an initial 3×3 convolution followed by batch normalization and ReLU maps the input to 64 channels. Each branch then uses three residual dense blocks (RDBs), with 2×2 max pooling after each of the first two blocks. Each RDB contains three convolutional layers with a growth rate of 32 and densely concatenates the output of each layer with the inputs of subsequent layers. A 1×1 local-fusion convolution maps the concatenated representation back to 64 channels before residual addition with the block input.

After the third RDB, adaptive average pooling in each branch produces an 8×1 frequency–time representation, which is flattened into 512 dimensions and transformed by a fully connected layer, batch normalization, and ReLU into a

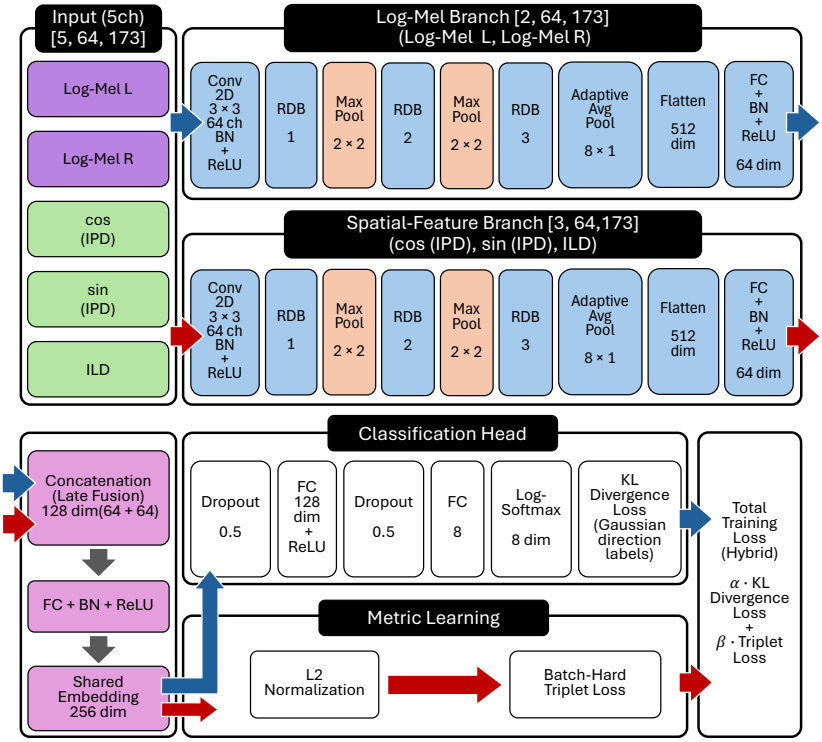


Fig. 6. Architecture of the proposed localization model: dual-branch feature extraction for the Log-Mel and spatial features, and late fusion, shared embedding, classification head, and training objectives.

64-dimensional branch representation. The two branch outputs are concatenated and transformed by another fully connected layer, batch normalization, and ReLU into a 256-dimensional embedding. The classification head applies dropout with a probability of 0.5 before each of two fully connected layers, which map the embedding to 128 dimensions and then to eight direction outputs; ReLU follows the first layer. For metric learning, the 256-dimensional embedding is L2-normalized.

For direction classification, the target is a circular Gaussian distribution centered on the ground-truth direction with a standard deviation of 15°, and KL divergence is computed against the model's log-softmax output. Metric learning uses batch-hard triplet loss. For each anchor, the farthest sample from the same direction class is selected as the hardest positive, and the closest sample from a different direction class is selected as the hardest negative, using a margin of 0.5. The total training loss is the weighted sum of KL divergence and triplet loss, with weights of 1.0 and 0.3, respectively.

6.3 Training Configuration and Data Augmentation

Weak augmentation is applied only to the simulated training recordings; no augmentation is applied to real, validation, or test recordings. With probability 0.5, a smooth time-varying envelope shared by the two channels is generated from five control points whose gains are sampled from 0.85 to 1.0. With probability 0.5, a randomly selected real ambient recording is mixed at an SNR sampled from 45 to 65 dB. After feature extraction, a single time mask covering at most 5% of the frames is applied with probability 0.5 at the same position in all five feature channels. White noise and independent left–right gains are not used. For each leave-one-scene-out fold, ambient recordings from the held-out scene are excluded, leaving 90 noise recordings available for augmentation.

The implementation uses AdamW with a learning rate of 5×10^{-5} and weight decay of 10^{-4}. The batch size is 64 and the maximum number of epochs is 100. The learning rate is halved when the monitored loss does not improve for three epochs, with a minimum learning rate of 10^{-6}. The monitored loss is the equally weighted mean of the KL classification losses on the simulated and real-recording validation sets. Training stops early if this value does not improve for 15 epochs, and the checkpoint with the lowest monitored loss is retained.

7 Evaluation

7.1 Research Questions

The evaluation addresses the following research questions: **RQ1**: Can a model trained with simulated audio and real recordings from non-held-out scenes perform eight-direction localization in unseen real acoustic environments? **RQ2**: What performance does the system achieve for beep detection and direction estimation on an iPhone under the controlled on-device test conditions?

7.2 Evaluation Protocols and Metrics

For RQ1, we perform ten-fold leave-one-scene-out cross-validation over SC001–SC010, holding out 80 recordings from one scene in each fold. The proposed model is trained using 2,880 simulated recordings and 576 real recordings from the other nine scenes and selected using 720 simulated and 144 real validation recordings. For comparison, a parameter-free DRTF template-matching baseline is evaluated on the same 800 held-out recordings. For each channel, the complex amplitude at exactly 440 Hz is obtained by projecting the entire signal onto a complex sinusoid without using an STFT. A Hann window is applied to each 2 s recording before projection, whereas the measured DRTFs are projected without windowing. The ILD and IPD are defined as $\mathrm{ILD} = 20 \log_{10}(|A_L|/|A_R|)$ and $\mathrm{IPD} = \arg(A_L A_R^*)$, respectively. For direction i, the distance is

$$d_i = \left(\frac{\mathrm{ILD}_x - \mathrm{ILD}_i}{\sigma_{\mathrm{ILD}}} \right)^2 + 2\left[1 - \cos\left(\mathrm{IPD}_x - \mathrm{IPD}_i\right)\right], \tag{12}$$

Table 3. Experimental matrix for model evaluation. The 24 simulated recordings in the standard condition are used only for diagnostic evaluation because their azimuths overlap those of the simulated training set.

Condition	Real recordings train/val./eval. (T0xx)	Simulated recordings train/val./eval.	Separation	Role
Standard	560 (01, 02, 04, 06, 07, 09, 10) / 160 (03, 08) / 80 (05)	2,880 / 720 / 24 (diagnostic)	Take-level; SC001–10 shared	Reference evaluation and deployed model
Take-split A	560 (01, 02, 04, 06, 07, 09, 10) / 160 (03, 08) / 80 (05)	2,880 / 720 / –	Take-level; SC001–10 shared	Auxiliary to RQ1
Take-split B	560 (02, 04, 05, 06, 07, 09, 10) / 160 (03, 08) / 80 (01)	2,880 / 720 / –	Take-level; SC001–10 shared	Auxiliary to RQ1
Take-split C	560 (01, 02, 04, 05, 06, 07, 09) / 160 (03, 08) / 80 (10)	2,880 / 720 / –	Take-level; SC001–10 shared	Auxiliary to RQ1
LOSO (10 folds)	576 / 144 from nine non-held-out scenes / 80 from one held-out scene	2,880 / 720 / –	Scene-level; each scene held out once	Primary evaluation for RQ1
On-device	– / – / 160 reference beeps at 0.35 and 1.5 m; 19.935 min	– / – / –	Operational test	Primary evaluation for RQ2

Table 4. Localization performance of the model trained using the standard split.

Test set	N	Accuracy	Macro F1	Mean error	Median error	Within 45°
Combined	104	98.08%	98.07%	1.73°	0°	98.08%
Simulated	24	100.00%	100.00%	0.00°	0°	100.00%
Real recordings	80	97.50%	97.49%	2.25°	0°	97.50%

where σ_{ILD} is the standard deviation of the ILDs of the eight DRTF templates. The direction minimizing d_i is selected. Because the baseline requires neither training nor evaluation-set tuning, no fold-specific fitting is performed.

Localization performance is evaluated using eight-class accuracy, macro F1, mean and median angular error, accuracy within 45 degrees, confusion matrices, and per-scene results. Angular error is calculated as the shortest circular distance, treating 0 and 360 degrees as adjacent. Detection performance is evaluated using precision, recall, F1, false-positive rate, and false detections per unit time. The evaluation conditions and data splits are summarized in Table 3. Take-split evaluates separation within shared scenes, whereas LOSO evaluates unseen-scene generalization. The on-device test uses 1.5 m as the original evaluation condition and 0.35 m as an auxiliary diagnostic condition intended to increase the received level through closer source placement. The 0.35 m condition is not intended to represent a practical bicycle–vehicle separation.

7.3 Results

Standard Training Results The checkpoint selected under the standard training split was evaluated using 24 simulated test recordings and 80 real test recordings from SC001–SC010.

As shown in Table 4, all 24 simulated test recordings were classified correctly. For the real-recording test set, 78 of 80 recordings were classified correctly. The

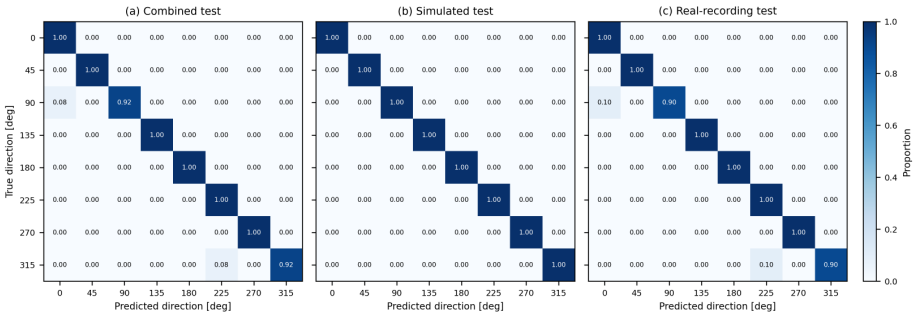


Fig. 7. Row-normalized confusion matrices for the standard training model: (a) combined test set, (b) simulated test set, and (c) real-recording test set.

Table 5. Take-split cross-validation results for SC001–SC010. Mean and SD denote the mean and sample standard deviation over the three folds, respectively.

Fold	Take	N	Epoch	Accuracy	Macro F1	Weighted F1	Macro precision	Macro recall	Mean error	Median error
A	T005	80	51	100.00%	100.00%	100.00%	100.00%	100.00%	0.00°	0.00°
B	T001	80	75	95.00%	94.99%	94.99%	95.20%	95.00%	6.19°	0.00°
C	T010	80	67	100.00%	100.00%	100.00%	100.00%	100.00%	0.00°	0.00°
Mean	–	80	64.33	98.33%	98.33%	98.33%	98.40%	98.33%	2.06°	0.00°
SD	–	0	12.22	2.89%	2.89%	2.89%	2.77%	2.89%	3.57°	0.00°

two errors were a 90° source classified as 0° and a 315° source classified as 225°. The corresponding row-normalized confusion matrices for the combined, simulated, and real-recording test sets are shown in Fig. 7.

Take-Split Cross-Validation Results Table 5 reports the three-fold take-split results. Mean accuracy was 98.33% (SD 2.89%); Folds A and C achieved 100.00%, whereas Fold B achieved 95.00%. Thus, performance remained high under take-level separation within shared scenes.

Leave-One-Scene-Out Results Table 6 reports the scene-wise LOSO results. The proposed model achieved mean accuracy and macro F1 of 34.25% and 31.08%, respectively, compared with 27.00% and 18.17% for the DRTF template baseline. However, it achieved higher exact accuracy in only five of the ten scenes. The baseline also yielded a slightly lower mean angular error (65.19° versus 66.49°) and higher accuracy within 45° (60.88% versus 56.25%). Thus, the advantage of the learned model was limited to exact-class metrics and a subset of scenes (Fig. 8). In Fig. 9, errors were not confined to adjacent classes.

On-Device Operational Results Table 7 reports localization performance over 80 detected events under the original 1.5 m condition and the auxiliary 0.35 m condition. Exact accuracy, accuracy within 45°, and mean angular error

Table 6. Scene-wise localization performance in leave-one-scene-out evaluation. Mean and SD denote the mean and sample standard deviation over the ten held-out scenes, respectively.

Scene	Accuracy	Macro F1	Mean angular error	Median angular error	Within 45°	Within 90°
SC001	0.00%	0.00%	127.69°	135.00°	10.00%	33.75%
SC002	17.50%	12.03%	69.75°	45.00°	57.50%	71.25%
SC003	18.75%	13.27%	88.31°	90.00°	47.50%	56.25%
SC004	12.50%	5.95%	73.13°	90.00°	47.50%	80.00%
SC005	65.00%	57.99%	24.19°	0.00°	88.75%	93.75%
SC006	21.25%	17.10%	68.06°	45.00°	62.50%	75.00%
SC007	15.00%	17.90%	114.75°	135.00°	21.25%	37.50%
SC008	31.25%	26.72%	68.63°	90.00°	45.00%	75.00%
SC009	78.75%	77.43%	19.13°	0.00°	90.00%	92.50%
SC010	82.50%	82.46%	11.25°	0.00°	92.50%	100.00%
Mean	34.25%	31.08%	66.49°	63.00°	56.25%	71.50%
SD	29.76%	30.14%	39.01°	52.82°	28.27%	22.75%

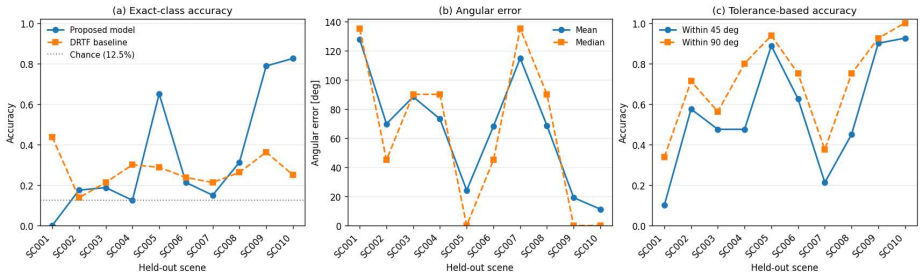


Fig. 8. Scene-wise performance in leave-one-scene-out evaluation: (a) exact-class accuracy of the proposed model and DRTF template baseline, with the dotted line indicating the 12.5% chance level; (b) mean and median angular errors of the proposed model; and (c) its accuracy within 45° and 90°.

Table 7. On-device localization performance under the original 1.5 m condition and the auxiliary 0.35 m condition for true-positive beep detections.

Distance	N	Exact accuracy	Within 45°	Mean angular error
0.35 m	80	51.25%	76.25%	41.63°
1.5 m	80	10.00%	36.25%	76.50°

were 10.00%, 36.25%, and 76.50° at 1.5 m and 51.25%, 76.25%, and 41.63° at 0.35 m, respectively.

7.4 On-Device Implementation

The trained localization model is converted to Core ML and integrated into an experimental iOS application. The application performs recording, beep detection, feature extraction, and direction inference on the device and displays the

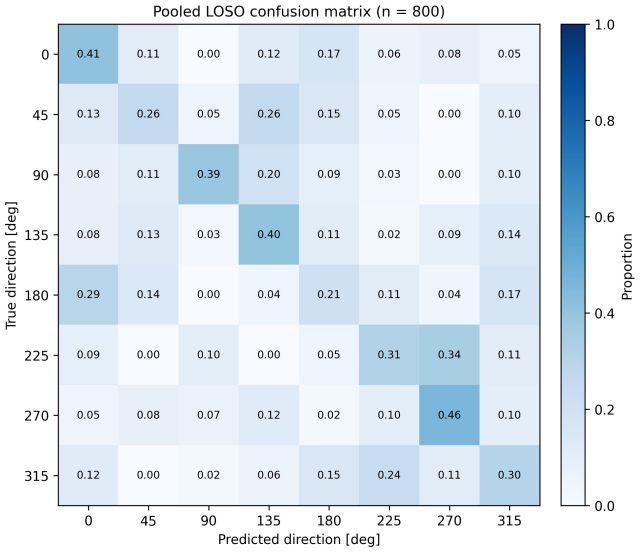


Fig. 9. Row-normalized confusion matrix pooled over the 800 held-out predictions from the ten leave-one-scene-out folds.

Table 8. Beep-detection performance in the on-device operational test. The false-positive rate (FPR) is the number of unmatched detections divided by the number of negative 0.1-s logging frames.

Distance	Ref.	TP	FP	FN	Precision	Recall	F1	FPR (%)	FP/min	Duration (min)
0.35 m	80	80	0	0	100.00%	100.00%	100.00%	0.000%	0.000	9.028
1.5 m	80	80	2	0	97.56%	100.00%	98.77%	0.045%	0.183	10.907
Overall	160	160	2	0	98.77%	100.00%	99.38%	0.026%	0.100	19.935

estimated direction on the screen. This display was used to confirm inference results during the experiment and was not designed or evaluated as a cyclist-facing notification mechanism. The application interface is shown in Fig. 10, and the deployment configuration is summarized in Table 9.

8 Discussion

The standard real-recording and take-split accuracies were 97.50% and 98.33%, whereas mean LOSO accuracy was 34.25%, compared with 27.00% for the DRTF template baseline. However, the proposed model outperformed the baseline in only five of ten scenes and showed no advantage in the angular-tolerance metrics. Overall, the relative performance of the learned model remained scene-dependent. No inferential indoor–outdoor comparison was conducted because only two scenes were indoor and relevant acoustic quantities were not measured.

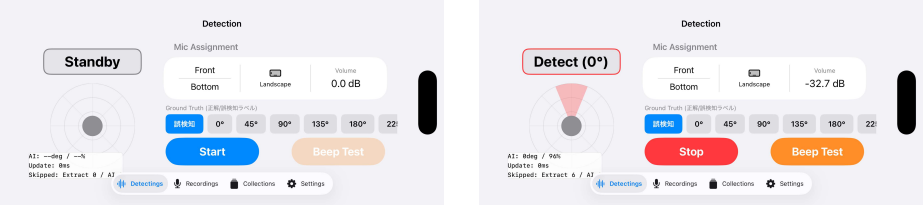


Fig. 10. Experimental iOS application used in the on-device localization test: (a) standby state and (b) a detected beep displayed at 0°. The controls at the bottom were used to record ground-truth labels for evaluation.

Table 9. Experimental environment and deployment configuration for the on-device operational test.

Item	Specification	Item	Specification
Playback device	Jabra Speak 510	Recording/ inference device	iPhone 12 Pro Experimental iOS
Operating system	iOS 26.5.2 (23F84)	Application	Hybrid localization application
Model	model (epoch 59)	Deployment format	Core ML
Model size	Approximately 1.4 MB	Audio duration per inference	2.0 s
Localization confidence threshold	0.40	Loudspeaker–smartphone distance	0.35 m and 1.5 m

Localization performance was higher under the auxiliary 0.35 m condition than under the original 1.5 m condition. However, because sound level and SNR were not measured, this difference cannot be attributed solely to an increase in the received level. The 0.35 m result also does not represent performance at a practical bicycle–vehicle separation. Limited localization performance at 1.5 m and systematic errors such as 45° sources predicted as 135° remain challenges.

This study is limited to one iPhone model and loudspeaker, the evaluated fixed device placement, a stationary 440 Hz tone, and two source distances. Performance under other placement or carrying conditions was not evaluated. The evaluation did not include real traffic noise or bicycle motion, and the sound level, SNR, and reverberation characteristics of the evaluation recordings were not measured. The results therefore do not establish performance with other devices, moving sources, or general sounds, and their causal interpretation remains limited. The detection-aligned localization window used on-device differed from the onset-aligned training inputs, which may have affected localization performance. The separate effects of simulated and real training data were not evaluated.

9 Conclusion

This study implemented on-device beep detection and eight-direction localization using the built-in microphones of an iPhone 12 Pro. Both the standard and take-split accuracies were approximately 98%, but mean LOSO accuracy

was 34.25%. All 160 reference beeps were detected on the device, but exact localization accuracy under the original 1.5 m condition was 10.00%. The results show controlled-condition detection feasibility, but unseen-scene generalization and localization under the original test condition remain challenges. Broader validation requires a larger and more diverse dataset spanning acoustic conditions, distances, devices, loudspeakers, moving sources, and real vehicle sounds.

References

1. Apple Inc.: Core ML. `https://developer.apple.com/documentation/coreml/`, Apple Developer Documentation, accessed July 24, 2026
2. Goertzel, G.: An Algorithm for the Evaluation of Finite Trigonometric Series. The American Mathematical Monthly **65**(1), 34–35 (1958). https://doi.org/10.2307/2310304
3. Goli, P., van de Par, S.: Deep Learning-Based Speech Specific Source Localization by Using Binaural and Monaural Microphone Arrays in Hearing Aids. IEEE/ACM Transactions on Audio, Speech, and Language Processing **31**, 1652–1666 (2023). https://doi.org/10.1109/TASLP.2023.3268734
4. Kawanaka, S., Kashimoto, Y., Firouzian, A., Arakawa, Y., Pulli, P., Yasumoto, K.: Approaching Vehicle Detection Method with Acoustic Analysis Using Smartphone for Elderly Bicycle Driver. In: Proceedings of the 10th International Conference on Mobile Computing and Ubiquitous Networking. pp. 1–6 (2017). https://doi.org/10.23919/ICMU.2017.8330069
5. Küçük, A., Ganguly, A., Hao, Y., Panahi, I.M.S.: Real-Time Convolutional Neural Network-Based Speech Source Localization on Smartphone. IEEE Access **7**, 169969–169978 (2019). https://doi.org/10.1109/ACCESS.2019.2955049
6. Matsui, K., Ando, A.: Estimation of Individualized Head-Related Transfer Function Based on Principal Component Analysis. Acoustical Science and Technology **30**(5), 338–347 (2009). https://doi.org/10.1250/ast.30.338
7. Wang, L., Yin, F., Chen, Z.: Head-Related Transfer Function Interpolation through Multivariate Polynomial Fitting of Principal Component Weights. Acoustical Science and Technology **30**(6), 395–403 (2009). https://doi.org/10.1250/ast.30.395
8. Wang, Z., Tan, S., Zhang, L., Yang, J.: ObstacleWatch: Acoustic-Based Obstacle Collision Detection for Pedestrian Using Smartphone. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**(4), 1–22 (2018). https://doi.org/10.1145/3287072
9. Wendt, T., van de Par, S., Ewert, S.D.: A Computationally-Efficient and Perceptually-Plausible Algorithm for Binaural Room Impulse Response Simulation. Journal of the Audio Engineering Society **62**(11), 748–766 (2014). https://doi.org/10.17743/jaes.2014.0042