VoiceDirect: mmWave and Audio Signal Fusion for User Localization and Speaking Direction Estimation

Mahathir Monjur
University of North Carolina at Chapel Hill
Chapel Hill, USA
mahathir@cs.unc.edu

Shahriar Nirjon
University of North Carolina at Chapel Hill
Chapel Hill, USA
nirjon@cs.unc.edu

ABSTRACT

Estimating the facing direction of a speaker holds immense significance, from improving user experiences in smart homes through seamless interaction with IoT devices to delivering targeted responses by enabling spatial awareness. This capability is key in advancing natural communication technologies. Recent studies have introduced audio-based techniques that estimate which IoT device a speaker is addressing, utilizing the non-uniform radiation pattern of speech signals. However, these methods typically require a distributed system with multiple microphone arrays to perform effectively and pose scalability challenges. In this paper, we introduce VoiceDirect, a pioneering system capable of accurately determining a user's speaking direction from any location within a room. VoiceDirect employs a standalone smart hub, outfitted with a mmWave radar and a co-located microphone array. The benefits of utilizing the complementary nature of acoustic and mmWave signals are manifold. Firstly, the audio signal aids the mmWave radar in precisely estimating the location of a speaker among multiple humans present in a room. Secondly, this precise location of the speaker aids in normalizing and preprocessing both audio and mmWave signal. Finally, the point cloud of the speaker extracted from mmWave provides an estimation of the speaker's upper body pose, which in turn helps in estimating the head orientation from the speech radiation pattern, even in the presence of noise and multipath effects. Thus, VoiceDirect integrates information from both the acoustic radiation pattern and mmWave-generated body pose, inferring the speaking direction with exceptional accuracy. Our dataset, collected from extensive real-world experiments involving over $6,000$ voice commands by 8 speakers in 6 distinct environments, shows that VoiceDirect achieves a median error as low as 19°, significantly outperforming existing systems.

CCS CONCEPTS

• **Computer systems organization → Embedded systems**.

KEYWORDS

Speaking direction, device arbitration, mmWave and audio sensing, deep learning

1 INTRODUCTION

Voice communication significantly enhances the user experience with smart devices by enabling natural, hands-free interaction for tasks, information retrieval, and control. It streamlines multitasking and task execution efficiently in our daily lives. Personalization is fostered through voice recognition and emotional connections, reducing cognitive load by eliminating the need for specific commands. Context-awareness maintains meaningful conversations and simplifies complex tasks, promoting safety through distraction-free engagement, and bridging language barriers with translation features. Voice assistants continually improve accuracy and responsiveness through user interactions, ultimately revolutionizing interaction with smart devices by offering an intuitive, personalized, and efficient interface that caters to diverse needs and contexts.

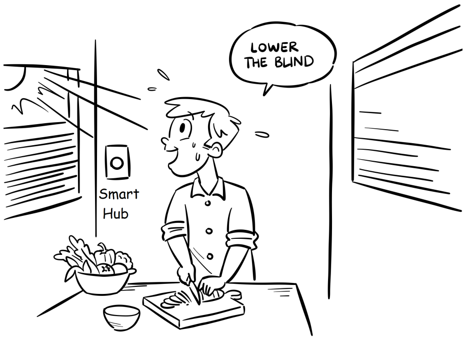


Figure 1: We envision a future where we will interact with smart devices by facing or looking at them and speaking to their direction. A single smart hub device having on-board audio and mmWave sensing capability will facilitate the device arbitration by estimating the user's direction of speaking.

Unfortunately, voice communication with smart devices poses a practical challenge in scenarios where multiple smart devices are present in our surroundings, requiring us to recall each device's name or identifier for interaction. For instance, envision a living room furnished with two smart blinds on its windows. In this scenario, a user cannot simply issue a command such as, "Alexa/Okay Google, lower the blind," to control each blind individually. Instead, they must assign distinct names to the blinds and issue a command like, "Alexa/Okay Google, lower the blind [identifier or name]." This approach swiftly becomes unscalable as the number of smart devices increases. Given that many of us struggle with remembering people's names, the prospect of recalling names or identifiers for dozens of IoT devices in every indoor space where we reside or work quickly transforms into a nightmare.

A more instinctive approach to interacting with smart devices, one that minimizes cognitive load, involves facing or looking at the device and speaking directly to it. This mirrors social situations where eye contact is established to signify our intended conversation partner in a group setting. Recent literature [6, 27, 31] has explored this approach with some degree of success, albeit frequently accompanied by impractical assumptions, unrealistic experimental settings, and/or large estimation errors. Fundamentally, these methods leverage the non-uniform radiation pattern of high-frequency components in speech to infer a user's speaking direction. Due to acoustic multipaths and inaccurate distance estimation capability using audio signals, these methods perform poorly when estimating the speaking direction, e.g., [31] reports an average estimation error of 57^0, and [6] fails to distinguish two sources that are less than 45^0 apart.

The work by Romit et al. [27] surpasses prior solutions through the introduction of a distributed microphone array-based approach. This approach includes two to six 4-channel or 6-channel microphone arrays designed to estimate line-of-sight power, subsequently determining the speaking direction with an average error of approximately 8°-37°. However, their solution involves equipping even the most basic smart devices, such as light bulbs, faucets, and locks, with multi-channel microphone arrays, which is not only cost-prohibitive but also intrusive. Furthermore, the system solves the speaking direction estimation problem in a more constraint setting where the user is looking directly at one of the devices that is also receiving the speech signal, and hence, the whole solution boils down to a N-class classification problem where N is the number of smart devices/ microphone arrays present in the room. As a result, developing a system comprising a standalone smart hub capable of accurately estimating a user's speaking direction from any location within a room remains an open problem.

To solve this problem, we introduce VoiceDirect, which employs a multi-modal fusion approach for accurate estimation of a user's speaking direction using a combination of a microphone array and a co-located mmWave radar on a smart hub device. The use of the mmWave radar in VoiceDirect is motivated by an in-depth empirical experiment where we systematically study the limitations of single-point audio-only speaking direction estimation techniques:

- First, we observe that estimating speaking directions is extremely challenging without knowledge of the user's location relative to the smart hub. Variations in speech intensity and multi-paths contribute to this complexity.
- Second, we demonstrate that if we know the user's precise location, line-of-sight signals from the speaker can be extracted and normalized to account for multi-paths and relative distance to obtain a better estimation of the speaking angle. However, there is significant room for improvement in estimation accuracy.
- Finally, we demonstrate that the impreciseness of single-point, audio-only location estimation results in significantly poor results in speaking direction estimation.

Based on these observations, we conclude that an additional sensing modality, such as an mmWave radar co-located with the microphone array, is necessary. A mmWave radar not only provides fast and accurate user localization but also helps improve the overall accuracy of the user's head and upper body pose estimation while being as minimally privacy-invasive as possible compared to cameras.

The fusion of acoustic and mmWave signals in VoiceDirect is inspired by how humans perceive an intended communication request using audio-visual cues. Typically, when someone hears their name being called, they use the sound to roughly locate where the other person (the caller) is situated. They then orient their head, face, and upper body in that approximate direction to accurately locate the caller using gaze. Finally, they focus their senses to concentrate and suppress noise from other directions. Likewise, in VoiceDirect, the smart hub follows a three-phase process to estimate the user's speaking direction:

- First, a preamble of the audio signals is used to roughly estimate the speaker's location. The mmWave radar, which by itself cannot detect who is speaking, utilizes the approximate location to perform beamforming, obtaining a high-precision relative location (distance and direction) of the speaker.
- Second, both the audio and mmWave radar signal streams are filtered to suppress signals from non-line-of-sight directions, and the signals are normalized to compensate for the effect of distance.
- Third, spatial, temporal, and frequency-domain neural features are extracted from mmWave and acoustic streams, and a cross-modal attention network is employed to accurately estimate the user's speaking direction.

Here, the speaking direction refers to the speaker's head orientation, which can have a slight offset from the speaker's upper body orientation. While we primarily rely on the normalized and filtered speech radiation pattern for precise speaking direction estimation, integrating the point cloud from mmWave with the audio signal in the final phase helps to reduce estimation errors caused by high levels of noise and multipath effects. Furthermore, the integration of mmWave also allows to do self-supervised calibration in unseen environment which is greatly beneficial in real-world scenarios.

We have implemented VoiceDirect using a commercially-available off-the-shelf 76GHz-81GHz mmWave radar (TI AWR1843) and a 4-channel microphone array (ReSpeaker V2) to ensure the reproducibility of our results and the widespread adoption of the system. The integration of the TI AWR1843 (retail price $299) into a smart hub increases the cost of the system, but it significantly enhances the sensing capabilities of the system. These enhancements are also beneficial for other radar-enabled smart home services, such as improved audio signal processing and better understanding of user and environmental context. Furthermore, the extra computational capabilities of the AWR1843 evaluation board contributes mostly to the high retail price since the antenna-on-package (AWR 1843AOP) costs only $26, which is expected to decrease even further when produced at scale. For computation, we use a HP Linux computer with 8 GB memory. After the full signal is captured, the processing time of the system including data pre-processing for both modality and DNN inference is 1.2 seconds for each utterance.

To assess the performance of VoiceDirect, we collect an audio-mmWave multimodal dataset through real-world experiments involving over $6,000$ voice commands issued by 8 users across 6 different rooms. As baselines for comparison, we employ single-device, 2-device, and 3-device microphone array-only systems for user localization and speaking direction estimation, following the algorithm outlined in [27]. Our findings indicate that VoiceDirect exhibits a median speaking direction estimation error of 19° across all users and indoor environments. This marks an improvement of 43°, 33°, and 16° over the respective single-device, 2-device, and 3-device audio-only baselines. We also conduct a real-time evaluation of VoiceDirect by depicting scenarios pertaining to device arbitration, which demonstrates that VoiceDirect successfully identifies the smart device in 80% of instances across all scenarios.

2 PRELIMINARY STUDY

We conduct an experiment to understand the limitation of single-point, audio-only solutions to a user's speaking direction estimation problem.

2.1 System Model Geometry

The system model consists of three entities – a smart hub, one or more human speakers, and multiple IoT devices. A minimal scenario is depicted in Figure 2 where a human speaker issues a voice command to an IoT device while looking at it. The IoT device does not have speech recognition capability. The smart hub listens to user's commands and controls the IoT device accordingly.

The geometry of the scenario is described using three parameters (r, θ, ϕ). The speaker's location with respect to the smart hub is described by the distance r and angle θ. The facing or speaking direction is denoted by the angle ϕ. Both the smart hub and the IoT device are stationary. The human speaker is free to issue voice commands from any location.

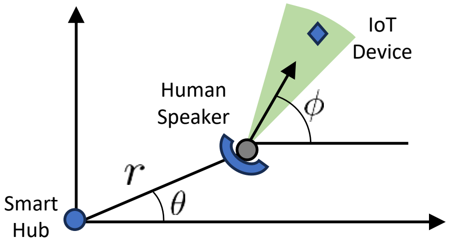


Figure 2: The geometry of the system model.

Although the model above is a 2D planer representation of a 3D real-world scenario, this is sufficient for most use cases since the user's location estimate is refined by using an average human height and we assume that no two IoT devices are placed at the same facing angle ϕ.

2.2 Testbed

We set up a testbed for multi-channel audio data collection in a 30×20 square feet indoor living room that contains typical furniture such as couches, tables, a computer desk, chairs, and cabinets. The interior walls are made of standard drywall panels, while the floor is wood, with a portion of the study area covered by a large carpet. A four-channel, far-field microphone array (ReSpeaker [4]) is placed on a 3-feet-high table surface, approximately 2 feet away from one wall. This setup provides an experimental area covering $0 < r < 16$ feet and $0 < \theta < 180$ degrees, ensuring adequate room coverage for accurate data collection.

We identify and mark 16 fixed locations in the experiment area. Four volunteers (one person at a time) stand at each of these fixed locations and issue 7 voice commands while facing 7 different objects in the room. This process ensures comprehensive data coverage across multiple facing angles. In total, we collect over 600 utterances, spanning a wide range of facing angles, from $0 < \phi < 360$ degrees, providing a robust dataset for our analysis.

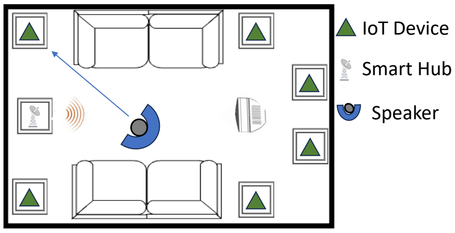


Figure 3: Testbed data collection setup.

2.3 Speaking Direction Estimation

A fundamental property of speech signals is that their radiation pattern is not omnidirectional for all frequencies. While the lower frequency components of human speech are omnidirectional, the higher frequency components are emitted with maximum energy in the direction the speaker is facing. In Figure 4, we illustrate two speech radiation patterns for a speaker facing two different directions, where different colors represent different frequencies. By leveraging this asymmetry, we can design an algorithm that learns to recognize the radiation pattern of the received signals at the smart hub to infer the user's speaking direction. This approach allows us to capture the natural directional bias of speech and utilize it for speaking direction estimation.

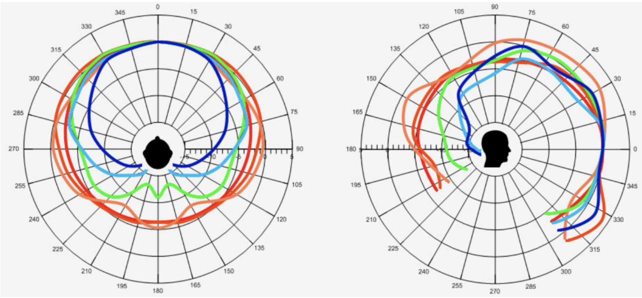
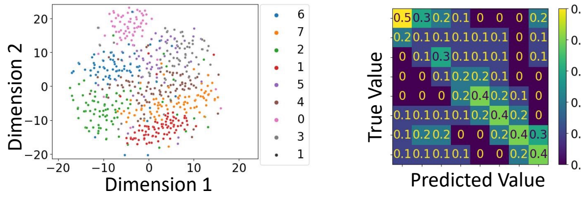


Figure 4: Radiation pattern of human speech [27].

As our initial approach, we implement a deep convolutional neural network (CNN) that takes a 3–5 second utterance as input and infers the speaker's direction as output. The CNN architecture is composed of 6 convolutional layers for feature extraction, followed by 3 fully connected layers that map the extracted features to the final output. We employ a contrastive loss function, which encourages the model to position two utterances with the same speaking angle close to each other in the embedding space, while utterances with different speaking angles are pushed farther apart. This setup enhances the network's ability to distinguish subtle variations in speech radiation patterns associated with different speaking directions. While the CNN performs regression to predict a continuous set of angles, we discretize the output into 8 distinct angle classes post-DNN execution to simplify the classification task. For evaluation, we divide the collected dataset, using 80% for training and the remaining 20% for testing, enabling us to assess the accuracy and generalization capability of the classifier across unseen data.

Figure 5(a) shows the t-SNE plot of feature embedding in two dimensions. We observe that the feature embedding for different



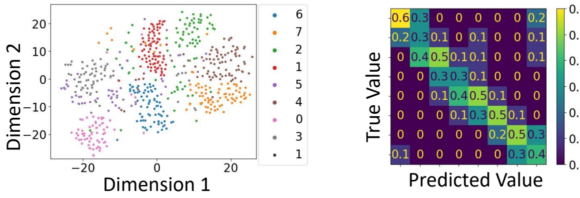
(a) tSNE Plot **(b) Confusion Matrix**

Figure 5: Location-unaware speaking direction estimator's ability to distinguish speaking angle ϕ is poor.

speaking angles are not very separable. This happens because even though the speaker may face different directions while speaking, due to their location and intensity of speaking, utterances from different speaking directions are mapped to neighboring points on the embedding space. For example, a speaker speaking towards the smart hub from far and a speaker speaking away from the smart hub from near are often confusing to the classifier as the two utterances have similar intensity and the subtle difference among the spectral components across different frequencies become difficult to distinguish. Figure 5(b) shows the corresponding confusion matrix. Overall, a location-unaware, single-point, audio-only approach results in an average speaking direction estimation error of 62°, confirming similar recent studies [31], which is not sufficient for IoT device arbitration.

2.4 Speaking Direction Estimation with Known User Location

Since we have ground truth user locations in our dataset, we conducted an additional experiment where the speaker's location information, represented as (r, θ)—the distance and angular position—is used as prior knowledge alongside the utterances to improve the estimation of the speaking direction ϕ. By incorporating this spatial information, we enhance the model's ability to contextualize the audio data. To reduce the effect of distance on the received signals, we normalize them with respect to r. Additionally, to minimize the impact of multipath interference and non-line-of-sight (NLOS) signals, which can degrade the signal quality, we apply a delay-sum beamforming algorithm [20], which focuses on extracting line-of-sight (LOS) signals from the speaker. After these preprocessing steps, the normalized and beamformed signals, along with the speaker's location data, are fed into a six-layer CNN with 16 convolutional filters that are designed to capture both spatial and frequency-domain features. The network is followed by three fully connected layers, which process the extracted features for the final speaking direction estimation. Similar to the earlier experiment, we utilize a contrastive loss function to ensure that utterances with similar speaking directions are closely aligned in the embedding space, while those from different directions are spaced farther apart. We use 80% of the dataset for training and reserve 20% for testing, allowing us to assess the improvement in speaking direction estimation when the speaker's location is incorporated as prior knowledge.



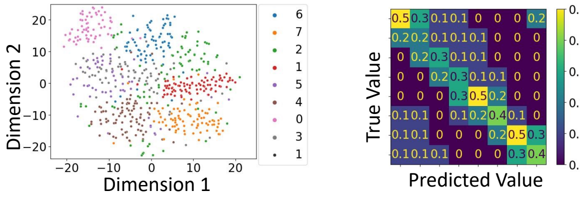
(a) tSNE Plot **(b) Confusion Matrix**

Figure 6: Location-aware speaking direction estimator's ability to distinguish speaking angle ϕ is better.

Figure 6(a) shows the t-SNE plot of feature embedding in two dimensions. We observe that the feature embedding for different speaking angles are more separable than the location-unaware setting. Figure 6(b) shows the corresponding confusion matrix. Overall, a location-aware, single-point, audio-only approach results in an average speaking direction estimation error of 27°, which is better than location-unaware algorithms, but there are still room for improvement.

2.5 Speaking Direction Estimation with Estimated User Location

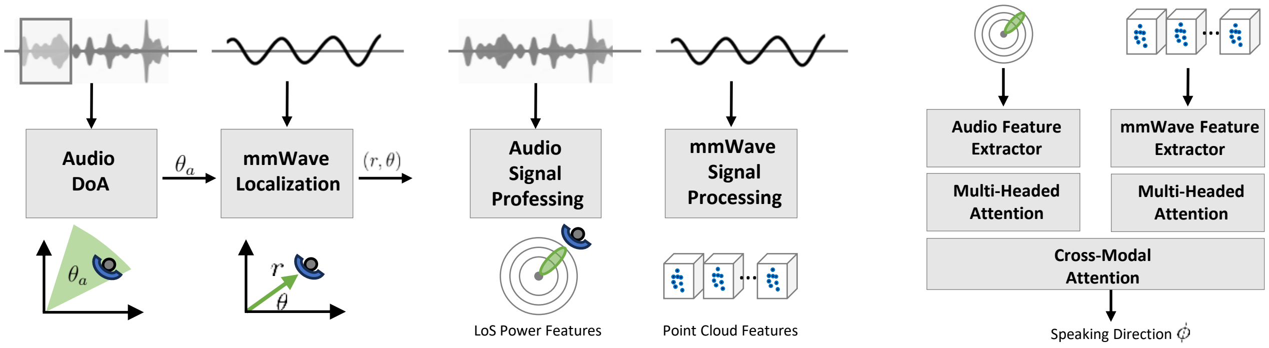
Knowing the user's location (r, θ) improves the accuracy of user's speaking direction ϕ estimation. Estimating the location itself, however, is still a challenge. Specifically, audio-based localization solutions are susceptible to environmental factors such as noise and reverberation. State-of-the-art solutions that use a single microphone array report meter-level resolution in distance estimation. Such large error margins is practically useless in speaking direction estimation given that even the perfect knowledge of the user's location cannot entirely solve the problem. Yet, for completeness of this study, we repeat the location-aware speaking direction estimation experiment with the modification that the user's location (r, θ) is estimated from the audio signals.



(a) tSNE Plot **(b) Confusion Matrix**

Figure 7: Speaking direction estimator's ability to distinguish speaking angle ϕ degrades rapidly when we use estimated location from audio.

Figure 7 shows that when we estimate the user's location using the four-channel audio and use that information to estimate speaking direction, the separability of the classes drastically reduces. This happens due to the location estimation error which is 1.1m on average for distance and 19 degrees for angle of arrival. While this error can be reduced by using multiple microphone arrays, such solutions would require multiple smart hubs in the room, which goes against our assumption that each room has a single smart hub.



(a) Speaker Localization

(b) Signal Filtering and Processing

(c) Speaking Direction Estimation

Figure 8: Overview of VoiceDirect.

3 OVERVIEW OF VOICEDIRECT

VoiceDirect combines the complementary nature of acoustic and mmWave signals to accurately infer a user's speaking direction in indoor environments which is otherwise difficult to achieve by using either of the individual signal modalities alone. It is a single-device, software-only solution that uses commercially available off-the-shelf microphone arrays and mmWave radar modules, and neither requires multiple coordinated sensors nor any modifications to off-the-shelf mmWave radars and microphone arrays. VoiceDirect is agnostic to the indoor environment and human voice, and its fusion network is generalizable beyond the device arbitration problem and can inspire future radar and audio fusion systems for detecting complex, long-term patterns in the indoor environment.

While designing VoiceDirect, we considered several different approaches. One option was to train an end-to-end deep neural network (DNN) to estimate the facing direction directly from raw mmWave and audio signals. However, this approach would require extensive training data for effective generalization due to the sparsity and limited speaker information present in mmWave signals. Therefore, preprocessing steps are crucial to extract meaningful patterns before training. Additionally, single-modality approaches, like mmWave-based mesh generation algorithms ([28, 29]), can estimate body pose and orientation but are incapable of determining head orientation. To overcome these limitations, we designed VoiceDirect in multiple stages, leveraging the complementary strengths of both audio and mmWave signals. Figure 8 illustrates the three phases of VoiceDirect's end-to-end operation: speaker localization, signal filtering and processing, and speaking direction estimation. Each of these phases is briefly described in this section.

3.1 Phase 1 – Speaker Localization

Since single-point, audio-only indoor localization solutions are inaccurate, VoiceDirect employs a two-step process to estimate the speaker's location.

Audio-Based DoA Estimation: VoiceDirect uses a preamble of the audio signals (e.g., the wake-word "Alexa" or "OK Google") captured by the microphone array to estimate the direction and arrival (DoA) of audio, which provides an approximate angle θ of the user's location.

mmWave-Based Localization: VoiceDirect processes the raw mmWave signals through a series of signal processing steps to obtain a set of point clouds, one of which contains the speaker. The approximate angle of the user's location is used to select the point cloud that corresponds to the speaker and the centroid of the point cloud is used to compute the exact location (r, θ) of the speaker.

3.2 Phase 2 – Signal Filtering and Processing

Audio and mmWave signals are conditioned to compensate for multi-paths and relative distance, and to prepare them for the next phase.

Audio Signal Processing: The speaker's location information is used to filter out unwanted sounds by beamforming [20] at the microphone array towards the line-of-sight direction between the user and the smart hub. The power of the line-of-sight signal is extracted and normalized by the square of the user's distance to compensate for the relative distance between the user and the smart hub.

mmWave Signal Processing: The mmWave-generated point cloud corresponding to the speaker is tracked and updated throughout the duration of the speech command. A time-sequence of point clouds where each point is represented by its 3D coordinates, velocity, and intensity of reflected signals is prepared for the next phases.

3.3 Phase 3 – Speaking Direction Estimation

A multi-modal sensor fusion network is used to estimate the user's speaking direction, which has a two-part operation:

mmWave and Acoustic Feature Extraction: mmWave point clouds go through a hierarchical feature extraction network that extracts features from individual points, spatial features from the cluster of points on a point cloud, and temporal features from a sequence of point clouds. Acoustic features are extracted by a deep convolutional network.

Multi-Modal Fusion Network: mmWave and acoustic features are fed to a network which consists of a CNN-based feature encoder, followed by a multi-modal attention-based fusion module, followed by a classification module to estimate the user's speaking direction.

4 SPEAKER LOCALIZATION

The goal of this phase is to estimate the speaker's location using acoustic and mmWave signals. It begins with the wake-word, where acoustic analysis provides a rough angular position, which is then refined by the mmWave-based algorithm for precise localization.

4.1 Audio-Based DoA Estimation

Numerous solutions have been proposed in the literature to estimate the direction of arrival (DoA) of acoustic signals that include time difference of arrival (TDoA)-based methods [12], energy-based strategies [22], and more recently, approaches rooted in deep neural networks (DNNs) [5, 10, 25]. Because of their superior performance, VoiceDirect employs a deep learning-based technique.

DoA Estimation Basics: The principle behind acoustic DoA estimation is that depending on the angular position of the source, consecutive microphones on the array receive the same signal but with different delays. Since the position of the microphones are fixed, these delays directly map to the DoA. In practical acoustic environments where noise and reverberation is common, the mapping is often done by data-driven learning approaches.

DoA Estimation in VoiceDirect: VoiceDirect employs a method that involves the computation of the *Generalized Cross-Correlation Phase Transform* (GCC-PHAT [13]) across pairs of microphone channels. It removes the effect of magnitude from each signal so that a speech does not get ignored in the presence of large noise of reflection. We use the GCC-PHAT of the first 0.5s of the audio signal, starting from the wakeword, as input to a convolutional neural network to estimate the DoA. We use a standard 4 convolution layers followed by 2 GRU layers for DoA estimation.

4.2 mmWave-Based Localization

The initial DoA estimation provides VoiceDirect with a reference search area, which the mmWave radar module uses to precisely locate the user. This is crucial for distinguishing a stationary speaker among a group of moving or stationary individuals. The steps for accurately localizing the speaker are as follows:

Step 1 – VoiceDirect initiates a sequence of pre-processing operations on the raw mmWave I-Q signals. First, 1-D FFT is performed to extract the range bins from the complex signals. Second, a Doppler-FFT is performed along the chirp dimension. Third, a cell averaging *Constant False Alarm Rate* (CFAR) [21] technique across both the range and Doppler dimensions is applied to compute a threshold, which is applied on the FFT outputs to identify points with high Signal-to-Noise Ratios (SNR). A notably low threshold is used to specifically mitigate the influence of stationary clutter and noise.

Step 2 – VoiceDirect conducts an azimuthal scan in the range of $[-90°, 90°]$ with 1° resolution beamforming. Employing Minimum Variance Distortionless Response (MVDR) [24] beamforming, we calculate the Angle of Arrival (AoA) for objects in each 100ms signal (one frame), creating an AoA spectrum. Elevation data is derived from signals at elevation antennas. We then use a threshold to identify the strongest reflections across range and azimuth angles.

Step 3 – After calculating the azimuth and elevation angles for the strongest reflections, we gather these points for each frame during the voice command. VoiceDirect then uses the DB-SCAN [9] density-based clustering algorithm to group dense data points and discard low-density clusters. We compute the centroids of these clusters and apply a threshold to exclude those with too few points, typically representing static objects like walls and furniture. The final step involves selecting the cluster whose centroid aligns with the audio-based DoA estimate within a specified tolerance range.

5 SIGNAL FILTERING AND PROCESSING

The goal of this phase is to effectively filter the signals within both modalities to preserve only those signals that are directly relevant to the speaker's activity.

5.1 Audio Signal Processing

Step 1 – Delay and Sum Beamforming: Using the user's angular position θ, we perform acoustic beamforming to minimize noise and multi-path reflections. This is achieved through the delay and sum beamforming technique [20], which calculates the time delay Δ_i for each microphone based on θ and microphone array geometry by using the following equation:

$$\mathrm{x} = \sum_{\mathrm{i}} \mathrm{x_i}(\mathrm{t} - \Delta_{\mathrm{i}}) \tag{1}$$

This method aligns and enhances coherent speech signals while suppressing incoherent noise and multi-paths across the array by summing the time-adjusted signals.

Step 2 – Line-of-Sight Power Estimation: From the line-of-sight acoustic signals, we compute the line-of-sight power by performing FFT and taking the square of the amplitude. This is further multiplied by the square of the user's distance to normalize the effect of distance on signal power: $\mathrm{P}_{\mathrm{norm}}^{\mathrm{LoS}} = \mathrm{A}^2 \times \mathrm{r}^2$. Here, A is the amplitude of the signal and r is the user's distance. The normalization operation ensures that the line-of-sight power is scaled in accordance with the user's distance, which helps estimating the true radiation pattern of the speech.

Step 3 – Feature Extraction: To form the audio channel input to the neural network that is used in the next phase of VoiceDirect, along with the line-of-sight power, we include the received signal power from all microphones in order to compensate for any inaccuracies in line-of-sight power estimation that we may incur in a practical system. Thus we generate a feature vector having $T \times F \times (M+1) \times 2$ dimensions, where M represents the number of microphones, T and F represent the time and frequency bins, and the factor of 2 is due to the inclusion of both phase and amplitude.

5.2 mmWave Signal Processing

To create the mmWave channel input for the neural network in the next phase of VoiceDirect, we take the 3D coordinates, velocity, and signal power of the reflected points at each frame. The neural network, detailed in the following section, processes these time-sequenced point clouds to discern patterns and relationships. This helps it learn representations of the speaker's body orientation and pose, aiming to accurately estimate the user's speaking direction.

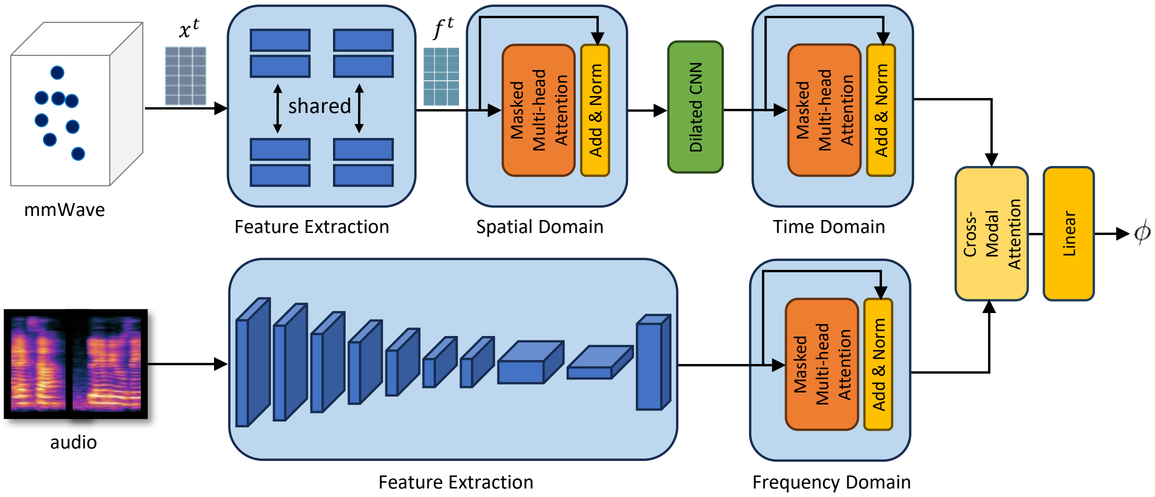


Figure 9: Deep Neural Network Overview

6 SPEAKING DIRECTION ESTIMATION

The goal of this phase is to apply a multi-modal deep neural network (DNN) on the processed audio and mmWave signals from the previous phase to estimate the speaking direction. Figure 9 shows the architecture of the DNN which is organized into three key components: (a) a mmWave radar feature extractor; (b) an audio feature extractor; and (c) a multi-modal attention module.

6.1 mmWave Radar Feature Extractor

In VoiceDirect, each mmWave radar frame (captured at 10 fps) is processed at a time to compute a point cloud. To make the frame size consistent, a predetermined number of 20 3D points is used to represent each point cloud. In instances where the number of points in a particular frame is lower than 20, padding with zero is employed for the remaining slots. A sequence of 5 consecutive point clouds or frames is used to extract the mmWave features. Thus, the input vector has the size of $B \times T \times N \times 5$, where B represents the batch size, T indicates the number of frames, and N signifies the quantity of 3D points within each frame.

Point Feature Extraction: This module processes the feature vectors associated with each 3D point. The feature vector contains the x, y, and z coordinates of individual 3D point, alongside their Doppler velocity and reflected signal power.

Each point within a frame undergoes processing via two separate multi-layer perceptrons (MLP). Notably, these MLPs possess shared weights, signifying that the same MLP layer is applied across all 3D points. If we designate the input vectors as x_i^t, where $i \in [1, N]$ denotes the point index within each frame, and $t \in [1, T]$ indicates the frame index, the outcome of the shared MLPs is an output feature vector, f, as represented by the equation:

$$\mathrm{f}_\mathrm{i}^\mathrm{t} = \mathrm{MLP}(\mathrm{x}_\mathrm{i}^\mathrm{t}; \omega) \qquad (2)$$

Here, the symbol ω represents the weight parameters associated with the MLP layers.

Spatial-Domain Attention Module: In order to estimate the body orientation and shape information of the speaker from the point cloud, the internal structure and the relationship among the points on a point cloud are learnt. To aggregate the point feature vectors at each frame, we employ a multi-head attention [26] module.

The attention mechanism is integral to the functioning of the model as they increase the network's ability to focus on relevant aspects of the point cloud and establish meaningful spatial relationship among various points. In VoiceDirect, we implement a multi-headed attention [7] module featuring 32 attention heads. This mechanism comprises three dense layers that compute the key, query, and value components from the feature vectors extracted from the point cloud. During the computation of attention scores, we apply a mask that identifies the 3D points present in a frame and distinguishes them from the zero-padded ones. The output from all the attention heads are concatenated and linearly transformed to form the final output.

Time-Domain Attention Module: Once the spatial-domain features capturing the inherent local structure within the point cloud is extracted, we apply a dilated convolutional neural network to convolve across both the time- and the point cloud feature dimensions. Another instance of a masked multi-headed self attention module is used to aggregate information across all frames. This radar feature vector is then fused with the audio feature vector for speaking direction estimation.

6.2 Audio Feature Extractor

CNN Operation: The filtered and processed acoustic signals produced in the previous phases that contains acoustic signals relevant to only the speaker is passed through a CNN consisting of seven convolutional layers and subsequent max pooling operations.

Multi-Headed Attention: The feature vector after convolution is fed into a multi-headed attention module to capture the intricate relationships within the acoustic signals. Specifically, the attention scores are computed in the frequency domain since the most informative feature to extract from speech (for speaking direction estimation) is the relationship among intensities at different frequency levels.

6.3 Multi-Modal Fusion Network

After extracting feature vectors from both mmWave radar and audio modalities, VoiceDirect integrates them using a cross-modal attention module. This begins with linear layers computing keys (K) and queries (Q) for each modality. Cross-modal attention scores are then computed for the radar, using the radar-derived key and audio's query and value. The process is mirrored for the audio modality, inverting the key, query, and values. The outcomes are concatenated and fed to another linear layer for estimating the user's speaking direction. This multimodal fusion step combines insights from mmWave point cloud, which primarily offers information about the user's upper body pose, with audio signals that capture additional head orientation relative to the upper body.

6.4 Self-Supervised Calibration

mmWave radar signals are highly susceptible to noise and environmental interference, while speech radiation patterns can vary significantly between individuals, introducing further challenges for accurate speaker localization and direction estimation. To address these challenges, we propose a self-supervised calibration procedure that lasts for approximately two minutes before deploying VoiceDirect in a new user's environment. During this calibration phase, the user is instructed to move in random directions while speaking, allowing the system to capture both mmWave radar and audio data in sync. The mmWave radar operates with a frame periodicity of 100ms, and we utilize 5 consecutive frames (covering 500ms) to track the speaker's movement. By averaging the speaker's position across these frames, we achieve a precise location estimate, which is critical for normalizing the corresponding speech signal. This location normalization helps in compensating for distance-related attenuation in audio signals. In addition to tracking the user's location, we also monitor the movement direction to infer the speaker's facing direction. While slight discrepancies between the head and body orientation are expected, the user is asked to face directly forward during calibration, minimizing such variations and ensuring accurate initial fine-tuning of the model.

Once the calibration data is captured, we segment it into smaller, 500ms trainable units, assuming minimal displacement of the user during each unit. To further enhance the model's learning, we employ a sliding window approach with a stride of 250ms to extract overlapping segments from the full two-minute calibration session. This method significantly increases the amount of training data available for fine-tuning the deep neural network (DNN). By training on these overlapping windows, the model learns to adapt dynamically to the speaker's unique speech radiation pattern, compensating for individual differences and environmental noise. A major advantage of this approach is that it requires no manual labeling of ground truth data, as the system automatically aligns the mmWave radar information with the audio signals, using the radar as a reliable reference for speaker orientation.

Through this self-supervised calibration process, VoiceDirect is able to personalize the model for each user and environment, significantly enhancing its accuracy and robustness. The system can perform reliably even in previously unseen or acoustically challenging environments, as the calibration allows it to generalize effectively across a wide variety of users and spaces. This method ensures that VoiceDirect maintains high precision in real-world scenarios, providing consistent and accurate speaker localization and direction estimation.

7 SYSTEM IMPLEMENTATION

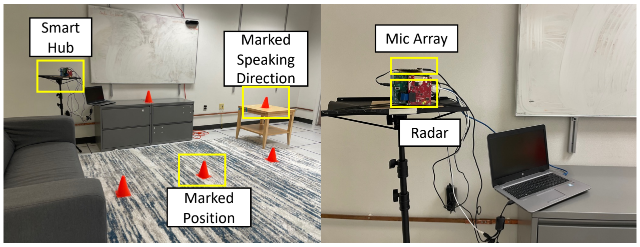


Figure 10: Experimental Setup.

7.1 Hardware Setup

We implement VoiceDirect using a low-cost microphone array [4], an off-the-shelf AWR1843 radar [1] equipped with a data collection board DCA1000EVM [2], and a laptop, as shown in Figure 10.

mmWave Radar: In VoiceDirect, we use the AWR1843 radar, a commercially available portable ($8.3cm \times 6.4cm$, $30g$) mmWave device that is capable of real-time data capture. The compact form factor and lightweight design make it ideal for embedding into smart home systems. The mmWave radar has 3 transmitting antennas and 4 receiving antennas, which allows it to generate highly accurate spatial and velocity information. The frequency of the RF starts at 77 GHz and increases after each chirp with a frequency slope of $60Hz$, enabling precise distance and velocity measurements. The detailed configuration of our FMCW radar is shown in Table 1. This configuration enables our radar to achieve a range resolution of $4.3cm$ and a maximum range of $9.02m$, making it well-suited for indoor applications where high-resolution motion tracking is required.

Microphone Array: VoiceDirect uses the commercially available ReSpeaker Mic Array ($70mm \times 70mm \times 13.3mm$) that has 4 high-performance digital microphones capable of far-field voice capture. It uses 5V power supply from micro USB and can easily be connected to a laptop to capture and stream audio data at a maximum of $16kHz$ sampling rate.

7.2 Software Setup

We connect and control the mmWave radar with a laptop running Linux operating system utilizing the mmWave-SDK from TI [3]. The mmmWave-SDK enables VoiceDirect to change the chirp configuration and use a software trigger to emit chirps in real-time. We

Parameter	Value	Parameter	Value
No. of frames	30	Frame periodicity	100
No. of chirps	64	No. of ADC samples	256
ADC start time	3	Idle time	3
Frequency slope	60	Ramp end time	64

Table 1: Configuration of the mmWave radar.

develop a simple UDP-based program to collect packets from the device and parse them into data frames. The program also triggers the microphone array at the same time to start capturing and streaming the audio data. The microphone array has a voice detection mode which continuously listens for the presence of voice, and whenever speech is detected, the whole system is triggered to start capturing data from both radar and audio modalities.

7.3 Neural Network Setup

For training and inference in our system, we use a 500ms signal length, equivalent to 5 mmWave radar frames and $8,000$ audio samples, as input to the DNN. Although the speech signal, including the wakeword and voice command, usually lasts $1-3$ seconds, we repeatedly run the DNN on the same command with a 250ms stride until the command ends. During inference, we average the most similar $\lfloor \frac{N}{2} + 1 \rfloor$ speaking direction estimations, where N is the number of DNN executions on the signal. This approach is chosen because large movements of user can disrupt speaker localization and signal processing. The 500ms window minimizes these effects, assuming minimal indoor movement impact within this duration.

7.4 Micro-benchmark

To evaluate the performance of VoiceDirect in real-time operation, we conducted micro-benchmark tests on a typical consumer-grade laptop. All inference and processing tasks were performed on an HP Linux laptop equipped with 4 GB of RAM and an Intel Core i5 processor. The micro-benchmark focused on measuring the time required for each step in the inference pipeline, including audio Direction of Arrival (DoA) estimation, radar and audio signal preprocessing, and the final multimodal deep neural network (DNN) used for speaking direction estimation.

The total time required for each iteration of the inference pipeline is 1.2 seconds. Table 2 provides a breakdown of the time taken by each component of the pipeline.

Step	Time (seconds)
Audio DoA Estimation	0.2
Preprocessing (Radar and Audio)	0.7
Multimodal DNN Inference	0.5
Total	**1.2**

Table 2: Time taken for each step in the inference pipeline.

The multimodal DNN used in this process consists of 5M trainable parameters. The model was trained using the Adam optimizer with an initial learning rate of 1×10^{-3}. A StepLR learning rate scheduler was employed, with a step size of 100 epochs and a scheduler gamma of 0.9, allowing the learning rate to decrease every 100 epochs. The model was trained over 1000 epochs, using both audio and mmWave radar data collected from our experimental dataset. Training was conducted on a machine with higher computational resources, but the micro-benchmark for inference was performed on the HP Linux laptop to simulate real-world performance in resource-constrained environments.

8 EVALUATION

8.1 Experimental Setup

Data Collection: In our experiments, 8 participants were asked to stand at various locations in an indoor environment and give 6 different voice commands, such as "Alexa, play music" and "Ok Google, turn on," while facing different directions. These directions were pre-determined by placing dummy IoT devices at various locations around the environment. Between 5 and 10 such IoT devices were marked, where the participants directed their gaze while uttering the voice commands. All participants spoke at their normal speech speed and volume, in the presence of background noise such as HVAC systems, air conditioning, and outside chatter. We randomly placed the hub at different locations within the environment, ensuring that the user remained within the mmWave radar's field of view. The recommended position for placing VoiceDirect is in a corner of the room at a 45° angle for optimal coverage.

We collect data in 6 different environments including residential living rooms, lab spaces, and large conference rooms in a commercial building. In total, we collect $6,000$ pairs of 3-second samples (i.e., the mmWave and audio signal), on which, we apply a 80%-20% split for creating the training and evaluation datasets.

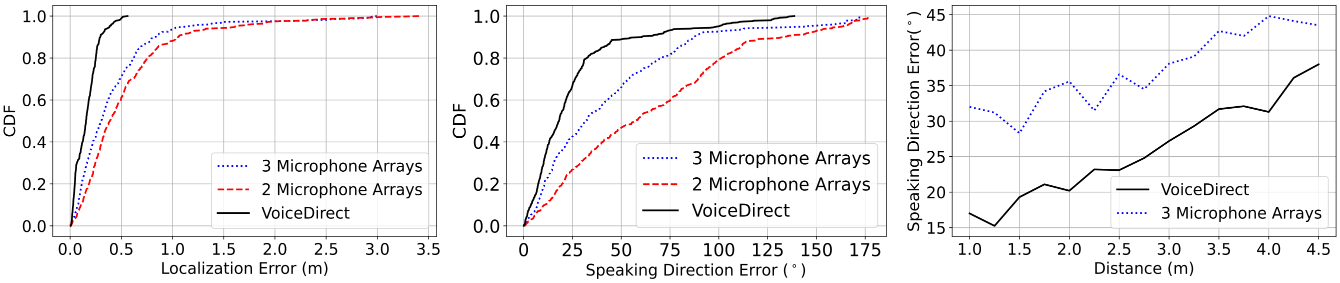
Metrics and Baseline: We evaluate our system for both user localization and speaking direction estimation. We use the Euclidean distance as the metric for user localization and the mean absolute error as the metric for speaking direction. In all of our experiments, we report the median error as well as the CDF plot to demonstrate the performance of the system.

We compare the performance of VoiceDirect against two multi-device, audio-only solutions as the baselines. Specifically, we use a 2-device and a 3-device audio-only distributed system where the devices use triangulation to estimate the user location and share information with each other to estimate the radiation pattern of the speech signal, and consequently the speaking direction, by following the algorithm described by the authors in [27].

8.2 Overall Performance

In Figure 11 we compare the performance of VoiceDirect against a 2-device and a 3-device audio-only solutions for both user localization and speaking direction estimation. Figure 11a shows the CDF plot of localization errors, where we can see that for audio-only solutions, the 90%tile localization error is more than 80cm and the maximum error is 3.4m. In comparison, VoiceDirect has a maximum localization error of 60cm, which is a significant improvement over audio-only algorithms.

Figure 11b shows the CDF plot for estimating speaking direction. The median speaking direction error for VoiceDirect, 2-device, and



(a) CDF Plot for Localization

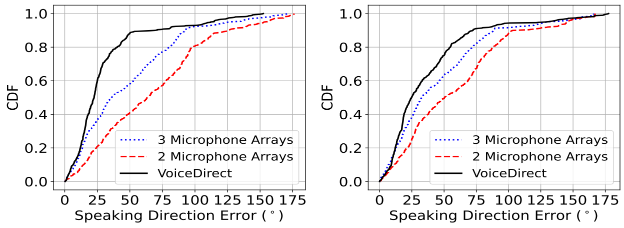
(b) CDF Plot for Speaking Direction

(c) User Distance vs Speaking Direction Error

Figure 11: Comparison of VoiceDirect with audio-only multi-device algorithms.

3-device solutions are 19°, 52°, and 35° respectively. Furthermore, in 70% of test cases, VoiceDirect achieves a speaking direction estimation error of less than 25°. This level of accuracy is sufficient for correct device arbitration among several IoT devices uniformly distributed around the room. Finally, less than 10% of the test data exhibits a large speaking direction error, leading to a long tail in the CDF plot. The primary contributors to these instances are incorrect localization due to noisy and complex multi-path environment and rapid movement of the speaker. We also show how VoiceDirect is affected by the user-to-device distance in Figure 11c.

8.3 Cross Utterance Performance



(a)

(b)

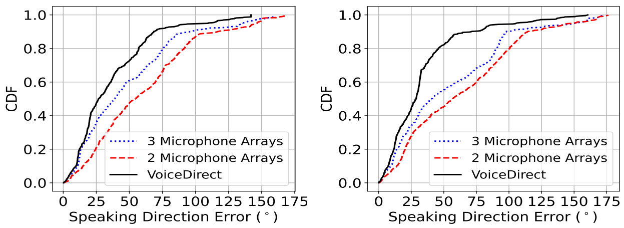
Figure 12: Speaking direction estimation error for (a) unseen utterances and (b) unseen environments.

To estimate the robustness of VoiceDirect to variations in different voice commands, we reserve two specific voice commands solely for evaluation and exclude them from the training process to ensure they remain unseen by the model. From Figure 12a, we observe that VoiceDirect is significantly more robust to unseen voice commands than audio-only solutions, as it does not rely solely on the audio modality for estimation. Instead, the fusion of radar and audio signals allows VoiceDirect to better generalize across different commands. The median speaking direction error for VoiceDirect, 2-device, and 3-device solutions are 23°, 60°, and 34° respectively, clearly demonstrating the advantage of multi-modal fusion. This proves that by integrating radar and audio data, VoiceDirect achieves greater resilience in estimating speaking direction, even for voice commands it has never encountered during training, making it highly adaptable in real-world applications.

8.4 Cross Environment Performance

To find out how VoiceDirect performs in a new environment, we split our dataset and hold data from two environments (out of 6) for evaluation only. Figure 12b shows that VoiceDirect attains a median speaking direction error of 25°. Although the difference between VoiceDirect and 3-microphone solution gets slightly lower in unseen environment due to radar's dependency on environment multipath, VoiceDirect still outperforms audio-only solutions.

8.5 Cross Speaker Performance



(a)

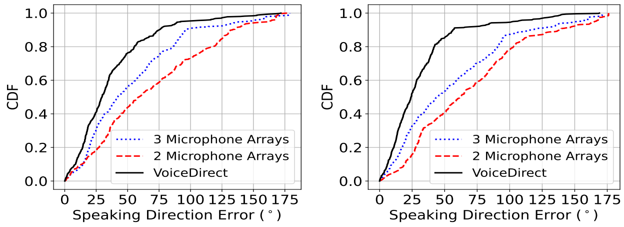
(b)

Figure 13: Speaking direction estimation error for (a) unseen speaker and (b) moving speaker.

For this experiment, we leave two speakers out from our dataset for evaluation only. Figure 13a shows the CDF plot of speaking direction estimation errors for unseen speakers. We see that the median speaking direction error for VoiceDirect, 2-device and 3-device solutions are 29°, 52°, and 38°, respectively. Overall, VoiceDirect outperforms audio-only algorithms even when the speaker is completely unseen.

8.6 Effect of Mobility

To assess the impact of speaker mobility on performance, we conduct sessions where the speaker moves or engages in activities while issuing voice commands. As depicted in Figure 13b, VoiceDirect greatly outperforms audio-only algorithms in such instance due to radar's superior performance in capturing human motion.



(a)

(b)

Figure 14: Speaking direction estimation error in the presence of (a) multiple speakers and (b) environment noise.

8.7 Effect of Multiple Speakers

In VoiceDirect, the radar's ability to filter out the speaker's point cloud depends on the initial DoA estimation from audio, since higher DoA error makes the location refinement using radar more error-prone. For this experiment, we arrange a data collection session where one speaker gives voice commands and 1/2 persons talk in the background. The audio-based DoA algorithm estimates the DoA of the separated voice command only. Figure 14a shows the CDF plot for speaking direction estimation errors in the presence of multiple speakers. The median error for VoiceDirect decreases from 19° to 29°, however it still outperforms audio-only algorithms. Even during high DoA estimation error, VoiceDirect can localize the speaker better by using radar, unless the estimated DoA completely shifts towards the other speaker. This makes VoiceDirect more capable of handling noisy environment. We also show the result of speaking direction estimation in the presence of environment noises such as HVAC in Figure 14b.

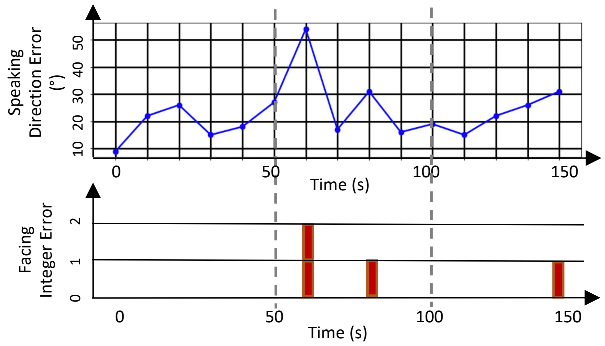


Figure 15: Real-time speaking direction estimation of a single user.

8.8 Real-time Evaluation

Experimental Setup: In addition to evaluating VoiceDirect on a test set, we design a real-time deployment scenario focusing on device arbitration. We invite three volunteers into a room for evaluation, each participating in three different sessions. The room contains five IoT devices. Before each session, the system is fine-tuned using self-supervised learning as mentioned in 6.4. The sessions last 150 seconds, consist of 15 voice commands and are divided into three sub-sessions. In the first sub-session, the user gives voice commands to five different IoT devices while standing in the middle of the room. Next, the user naturally moves within the room while issuing commands to the same devices. Finally, the user starts giving commands near the smart hub and gradually moves further away. After signal processing and DNN execution, the smart hub sends the inference results to a server.

Experimental Results: Figure 15 displays the speaking direction estimation results for a single user session. The first row shows the speaking direction error in degree, while the second row shows the facing integer error as mentioned in [27]. The facing integer error quantifies the number of devices between the one the user is actually addressing and the one identified by VoiceDirect's inference. From the figure, we can see that the speaking direction error surpasses 40° in 1 instance, leading to a facing integer error of 2. Overall, VoiceDirect attains a 80% accuracy and a reduced median error of 22°.

9 RELATED WORK

Sound Source Localization: The literature on sound source localization is rich with diverse methodologies designed to accurately determine the spatial origin of acoustic signals. Proposed solutions include using time difference of arrival(TDoA) based approaches [12], energy based approaches [22] and recently DNN based approaches [5, 10, 25] achieving exceptional accuracy.

Speaking Direction Estimation: Existing DNN-based speaking direction estimation works mostly use single mic array [6], [31]. This limits the performance significantly as in [6] the authors reported 65.4% accuracy for eight-class classification task, and in [31] the authors reported an average speaking direction error of 57°. In [27] the authors proposed signal processing based solution using multiple microphone arrays to estimate the speaking direction. However, it requires every smart device to include a microphone array. [32] reports 23-degree median error with two microphone arrays, but the evaluation is done only on speaker sitting on a chair in a controlled lab environment.

Human Pose Estimation using mmWave Radar: The literature of using mmWave radar in indoor environment mainly focuses on estimating human poses and mesh [23, 28–30], healthcare [33], security [11] etc. Authors in [30] extracted point cloud from range-Doppler heatmap and feed them to a PointNet like architecture for estimating human pose. [14] uses two AWR1843 radar for more accurate estimation of human body.In [23] authors proposed a human activity recognition system using mmWave radar. However, these systems focus on body pose only and can not detect head orientation due to the limitation of mmWave signal resolution.

Audio and mmWave Fusion: Fusion of radar and audio signal has mostly been proposed in audio acquisition systems where radar aids audio modality in capturing and enhancing speech signal in the presence of extreme noise [8, 15, 19]. However to our knowledge, VoiceDirect is the first system that fuses radar and audio signal for indoor localization and device arbitration problem.

10 CONCLUSION AND FUTURE DIRECTION

In this paper, we present VoiceDirect, a novel end-to-end fusion system that combines mmWave radar and audio signals for speaker localization and speaking direction estimation. By leveraging the complementary properties of mmWave radar, which excels in spatial resolution and motion detection, and acoustic signals, which capture voice features, VoiceDirect provides a more accurate and robust estimation of both the location and speaking direction of individuals giving voice commands. The system is designed to perform effectively across a wide range of indoor environments, including rooms with varying levels of clutter and acoustic conditions. Our experimental results demonstrate that VoiceDirect significantly outperforms prior works in both localization accuracy and speaking direction estimation, due to its multimodal approach that mitigates issues such as noise and multipath effects typically encountered by audio-only or radar-only systems.

Looking ahead, future research on VoiceDirect could focus on addressing several key challenges to improve its robustness, scalability, and versatility. One promising direction is extending the system to work in larger, multi-room environments, which would require further development in device coordination and communication to manage overlapping detection zones and ensure seamless tracking across connected spaces. Additionally, expanding the dataset to incorporate a more diverse population, including users with varying accents and age groups, would enhance the system's ability to generalize and perform reliably across different demographics. Future work could also explore implementing such systems in resource-constrained, energy-harvesting environments, following works like [16–18], to improve sustainability and operational efficiency. Finally, integrating advanced features such as gesture recognition, emotion detection, or user intent inference could broaden the system's applications beyond basic speaking direction estimation, enabling its use in fields such as human-computer interaction, immersive virtual reality experiences, and smart environments.

ACKNOWLEDGMENTS

This work was supported, in part, by grants NSF CAREER 2047461 and NIH 1R01LM013329-01.

REFERENCES

[1] Awr1843 single-chip 76-ghz to 81-ghz industrial radar sensor evaluation module. https://tinyurl.com/4wawbrh3.

[2] Dca1000evm real-time data-capture adapter for radar sensing evaluation module. https://tinyurl.com/4y2amc8c.

[3] Mmwave-sdk. https://tinyurl.com/yf9jmdx8.

[4] Respeaker mic array v2.0. https://tinyurl.com/bdhup89e.

[5] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2018.

[6] K. Ahuja, A. Kong, M. Goel, and C. Harrison. *Direction-of-Voice (DoV) Estimation for Intuitive Speech Interaction with Smart Devices Ecosystems*, page 1121–1131. Association for Computing Machinery, New York, NY, USA, 2020.

[7] J.-B. Cordonnier, A. Loukas, and M. Jaggi. Multi-head attention: Collaborate instead of concatenate, 2021.

[8] Y. Dong and Y.-D. Yao. Secure mmwave-radar-based speaker verification for iot smart home. *IEEE Internet of Things Journal*, 8(5):3500–3511, 2020.

[9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[10] E. L. Ferguson, S. B. Williams, and C. T. Jin. Sound source localization in a multipath environment using convolutional neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2386–2390. IEEE, 2018.

[11] B. Gonzalez-Valdes, Y. Alvarez, S. Mantzavinos, C. M. Rappaport, F. Las-Heras, and J. A. Martinez-Lorenzo. Improving security screening: A comparison of multistatic radar configurations for human body imaging. *IEEE Antennas and Propagation Magazine*, 58(4):35–47, 2016.

[12] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati. Real-time passive source localization: A practical linear-correction least-squares approach. *IEEE transactions on Speech and Audio Processing*, 9(8):943–956, 2001.

[13] B. Kwon, Y. Park, and Y.-s. Park. Analysis of the gcc-phat technique for multiple sources. In *ICCAS 2010*, pages 2070–2073. IEEE, 2010.

[14] S.-P. Lee, N. P. Kini, W.-H. Peng, C.-W. Ma, and J.-N. Hwang. Hupr: A benchmark for human pose estimation using millimeter wave radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5715–5724, 2023.

[15] T. Liu, M. Gao, F. Lin, C. Wang, Z. Ba, J. Han, W. Xu, and K. Ren. Wavoice: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, SenSys '21, page 97–110, New York, NY, USA, 2021. Association for Computing Machinery.

[16] Y. Luo and S. Nirjon. Smarton: Just-in-time active event detection on energy harvesting systems. In *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 35–44, Los Alamitos, CA, USA, jul 2021. IEEE Computer Society.

[17] M. Monjur, Y. Luo, Z. Wang, and S. Nirjon. Soundsieve: Seconds-long audio event recognition on intermittently-powered systems. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, MobiSys '23, page 28–41, New York, NY, USA, 2023. Association for Computing Machinery.

[18] M. Monjur and S. Nirjon. An empirical analysis of perforated audio classification. In *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*, IASA '22, page 25–30, New York, NY, USA, 2022. Association for Computing Machinery.

[19] M. Z. Ozturk, C. Wu, B. Wang, M. Wu, and K. R. Liu. Radio ses: mmwave-based audioradio speech enhancement and separation system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1333–1347, 2023.

[20] I. Rakotoarisoa, J. Fischer, V. Valeau, D. Marx, C. Prax, and L.-E. Brizzi. Time-domain delay-and-sum beamforming for time-reversal detection of intermittent acoustic sources in flows. *The Journal of the Acoustical Society of America*, 136(5):2675–2686, 11 2014.

[21] H. Rohling. Radar cfar thresholding in clutter and multiple target situations. *IEEE Transactions on Aerospace and Electronic Systems*, AES-19:608–621, 1983.

[22] X. Sheng and Y.-H. Hu. Energy based acoustic source localization. In *Information Processing in Sensor Networks*, pages 285–300. Springer, 2003.

[23] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava. Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar. In *Proceedings of the 3rd ACM Workshop on Millimeter-Wave Networks and Sensing Systems*, mmNets'19, page 51–56, New York, NY, USA, 2019. Association for Computing Machinery.

[24] M. Souden, J. Benesty, and S. Affes. A study of the lcmv and mvdr noise reduction filters. *IEEE Transactions on Signal Processing*, 58(9):4925–4935, 2010.

[25] R. Takeda and K. Komatani. Sound source localization based on deep neural networks with directional activate function exploiting phase information. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 405–409. IEEE, 2016.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.

[27] Y.-L. Wei, R. Li, A. Mehrotra, R. R. Choudhury, and N. Lane. Inferring facing direction from voice signals, 2021.

[28] Q. Xie, Q. Deng, T. Y. Cheng, P. Zhao, A. Patel, N. Trigoni, and A. Markham. mmpoint: Dense human point cloud generation from mmwave. In *BMVC*, pages 194–196, 2023.

[29] H. Xue, Q. Cao, Y. Ju, H. Hu, H. Wang, A. Zhang, and L. Su. M4esh: mmwave-based 3d human mesh construction for multiple subjects. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, SenSys '22, page 391–406, New York, NY, USA, 2023. Association for Computing Machinery.

[30] H. Xue, Y. Ju, C. Miao, Y. Wang, S. Wang, A. Zhang, and L. Su. mmmesh: Towards 3d real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, pages 269–282, 2021.

[31] J. J. Yang, G. Banerjee, V. Gupta, M. S. Lam, and J. A. Landay. *Soundr: Head Position and Orientation Prediction Using a Microphone Array*, page 1–12. Association for Computing Machinery, New York, NY, USA, 2020.

[32] Q. Yang and Y. Zheng. Model-based head orientation estimation for smart devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–24, 2021.

[33] Q. Zhai, X. Han, Y. Han, J. Yi, S. Wang, and T. Liu. A contactless on-bed radar system for human respiration monitoring. *IEEE Transactions on Instrumentation and Measurement*, 71:1–10, 2022.