MobiSys 2024 Paper #511 Reviews and Comments
==========
Paper #511 SoundGaze: Acoustic and mmWave Signal Fusion for Enhanced Speaking Direction Estimation

Review #511A
==========
Overall merit
—----------
2. Weak reject

Reviewer expertise
—----------
3. Knowledgeable

Paper summary
—----------
Summary: This paper proposes SoundGaze, a multimodal speaker orientation estimation system based on a microphone array and an mmWave radar. SoundGaze leverages the complementary nature of audio and mmWave signals to first estimate the user's location and then estimate the user's speaking direction using a data-driven approach. This paper conducts a series of experiments with 8 participants to evaluate SoundGaze's performance and its robustness under the conditions of different utterances, environments, users, levels of mobility, and environment noise.

Strengths
—----------
Estimating a user's speaking direction is an unaddressed problem in the community and the authors propose a novel idea that uses a multimodal design such that the speaker's orientation can be estimated from a standalone device.

Weaknesses
—----------
1. [A1]Motivation. The authors argue that SoundGaze is superior to pure audio-based designs in that (i) SoundGaze only requires one standalone device while pure audio-based systems need to equip every target device with a microphone array and (ii) SoundGaze has better performance. However, research [A] successfully decodes the user's head orientation with only two dedicated microphone arrays with a 23-degree error reported in the paper. The authors should also discuss the difference between SoundGaze and [A] to better show the motivation of this paper. In addition, I don't think SoundGaze has a better performance compared with the baseline with a 3-device setup. This is because the cross-subject performance presented in Fig. 15 shows almost the same performance between SoundGaze and the 3-device baseline. Although the other evaluations (Secs. 8.2-8.4, 8.7) show SoundGaze is better, this is reasonable because

SoundGaze is data-driven while the baseline [39] is not and I assume that the experiments except the one in Sec. 8.5 are not cross-subject which means the training set in these experiments contains samples from the same subject. To better show the superiority of SoundGaze to other methods, I suggest the authors use leave-one-subject-out (LOSO) for evaluation.

Response: While [Yang et al.] report a 23-degree median error with a two-microphone array in a controlled setup with the speaker seated, our system, tested in various environments including living rooms and office spaces with both static and mobile speakers and different levels of environmental noise, consistently outperforms setups with two or three audio-only devices, even in cross-speaker evaluations. We also evaluated our system with completely unseen users, however we argue that one of the key advantages of integrating mmWave radar with a single microphone array is the capability for unsupervised fine-tuning. This allows the system to learn and adapt to the speech radiation patterns of previously unseen users, achieving superior performance in any environment compared to existing techniques.

2. [A2]Contribution. The contribution of this paper is not clear and the authors should better summarize what challenges this paper solves and what contributions this paper makes to the community. Specifically, the techniques used in the paper look standard and lack novelty. It seems that the performance gain is the direct consequence of the introduction of the new modality - mmWave - while the devices themselves, the signal processing algorithms, and the models that make up the system are all standard methods. It is necessary for the authors to explicitly show the technical novelty in the paper since MobiSys values technical contributions.

   Response: We designed a careful fusion of audio and mmWave signals to localize a speaker and estimate speaking direction using a single smart hub. Our innovative DNN-based multimodal fusion network is the first of its kind, specifically tailored for integrating audio and mmWave data for speaking direction estimation. Furthermore, we introduced a novel method for self-supervised fine-tuning of these DNNs, leveraging radar data as the ground truth, allowing for practical deployment in real-world scenarios with remarkable accuracy.

3. [A3]Missing details. I find the paper misses a few important details. (i) What is the field-of-view (FoV) of the mmWave radar? If the radar itself has a narrow FoV, it may not be able to cover the entire room. (ii) Fig. 4 needs more explanations, such as, what are different colors stand for and what are the differences between the two subplots. (iii) The DoA estimation module seems to be DNN-based. What is the DNN architecture? (iv) Fig. 9 is not referenced. Also, what do "speaker" and "human" stand for respectively? I assume they refer to the same thing - the speaking user. (v) In Sec. 6.1 the mmWave point cloud is processed to have a fixed number of points - 20. How to process the point cloud into 20 points if it originally has more than 20 points?

Response: The FoV of mmWave is 120°, which is enough to effectively cover the entire room if placed in a corner. The different colors in Figure 4 are for different frequencies and the two plots are showing the speech radiation pattern for two different speaking directions. We use a standard 4 convolution layers followed by 2 GRU layers for DoA estimation. In figure 9, we show a scenario with 2 people present, where one of them is giving a voice command (speaker). We select the 20 points based on elevation (upper body).

4. [A4]Typos. (i) The error is 19 degrees in the abstract but 21 degrees in the last paragraph of the introduction. (ii) The name of the proposed system is "SoundGaze" in the text but "SoundStand" in the figures in Sec. 8.

   Response: All the typos are fixed.

[A] Qiang Yang and Yuanqing Zheng. 2021. Model-based Head Orientation Estimation for Smart Devices. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5, 3, Article 136 (Sept 2021), 24 pages. https://doi.org/10.1145/3478089

Review #511B
==========
Overall merit
—---------
2. Weak reject

Reviewer expertise
—---------
2. Some familiarity

Paper summary
—---------
This paper introduces a multi-modal system, SoundGaze, combining a microphone array and mmWave radar on a smart hub to estimate a user's speaking direction in the room.

Strengths
—---------
1. The paper demonstrates results outperforming existing systems.

2. The design of a multi-modality model integrating mmWave Radar and audio signals. The synergy between these modalities enhances speaker localization accuracy.

3. Real world implementations and experiment evaluations.

Weaknesses
—----------

1. The paper draws on existing work in mmWave localization, pose estimation, and sound direction estimation. The novelty of the idea is not explicitly clarified. While it seems to combine user location and sound direction estimation for improved accuracy, the paper could benefit from a more explicit explanation of the unique contributions of SoundGaze. Also from the results in Section 2.4, it should be very accurate when the position (r, alpha) is known. Why is the estimation error so large compared to the results of SoundGaze in section 8.2 (median error of 45° vs. 21°)? Further clarification is needed on the discrepancies.

   Response: Section 2.4 shows preliminary results using audio only. Section 8.2 shows results for audio-radar fusion.

Comments for authors
—----------

By comparing the CDF results in Fig. 14/15 and Fig. 12, the performance of 2&3 microphone arrays improve in unseen environments, which is counter-intuitive. The authors should provide clarifications here. In particular, Fig. 15 shows the the performance of of SoundGaze is similar to that of with three microphone arrays for unseen speakers, and the improvement of SoundGaze (by fusing the information from mmWave radar and microphone array) is insignificant compared to fusing the information from multiple microphone arrays.

Sections 4.1 and 4.2, discussing existing algorithms, could be moved to literature review.

Please change the name in the figures (SoundStance) to match the name in the context (SoundGaze).

Review #511C
==========

Overall merit
—----------

3. Weak accept

Reviewer expertise
—----------

2. Some familiarity

Paper summary
—----------
The paper proposes SoundGaze, a system that uses an instrumented smart hub (with a mmWave radar and a microphone array) to localize the speaker and subsequently estimate the speaking direction. The authors did a real-world implementation of the proposed system with varying IoT devices and voice commands. The authors evaluated their system under various conditions, such as different environments, mobility of the speaker, environment noise etc., and showed that their system outperforms other multi-device setups/algorithms (proposed in one of the papers they have cited).

Strengths
—----------
1. The problem is timely and the proposed solution of single device-based speaker orientation estimation sounds promising

2. The experimental setup and the conditions that the authors have considered while evaluating the system are impressive

Weaknesses
—----------
1. the paper uses mmWave radar for speaker localization -- this has been extensively studied and well-researched in the literature. Similarly, the use of microphone arrays to estimate the angle of arrival of sound has also been well-adopted in the past. This questions the technical novelty of the paper. several key factors are not explained or justified in the paper. For example, a taxonomy of a list of devices and voice commands experimented, an explicit explanation of the novelty of their contributions, etc., are missing in the current version of the paper.

   Response: Please Refer to A2[Contribution].

Comments for authors
—----------
While the paper is fairly well-written, it's still missing several key factors. For example, a list of voice commands used during the experiments is not stated in the paper. When new commands are introduced, it's still unclear whether the good performance is because of the lexical similarity between the new and old commands. Further, as mentioned previously, the key novelty of their approach is not detailed in the paper. At first glance, it looks like the novelty is the use of a single but relatively costly piece of hardware, however, later in the paper, it's observed that a 3 microphone array setup performs as well as SoundGaze, especially with the unseen speaker. In the new environment, SoundGaze marginally outperforms the 3 microphone array setup. The reasoning/key takeaways behind the results presented are not addressed in the paper.

When mobility is introduced, how well/bad the multi-device setup is performed? Fig 16, only shows SoundGaze, not for the multiple microphone arrays.

Was the neural network trained on the audio signals from a single user or a group of users who participated in the experiments? Given that when a new user is introduced, the error in detecting the orientation of the speaker goes high, it seems like the model was trained using a single user's data. Further, instead of directly using the raw audio samples, would it help to derive MFCC features and use them as input? For a given audio sample and the corresponding mmWave frames, what is the classification output from the DNNs? Does it output a facing angle, or a combination of (r, theta, and phi)?

there are discrepancies in the paper, specially when addressing the name of the system. The text describes the system as SoundGaze, while the Figures 12-17 mentions the name SoundStance -- please correct.

Review #511D

==========

Overall merit

—----------

2. Weak reject

Reviewer expertise

—----------

3. Knowledgeable

Paper summary

—----------

The paper introduces SoundGaze, a novel multimodality sensing system designed to determine a user's speaking direction from any location within a room. This is accomplished through a standalone smart hub equipped with a mmWave radar and a co-located microphone array. The system initially utilizes audio signals to provide a rough estimate of the user's location. Subsequently, mmWave beamforming is employed to attain a precise relative location between the smart hub and the user. After filtering out the impact of non-line-of-sight (NLOS) signals, SoundGaze extracts features from both sources and utilizes a cross-modal attention network to estimate the user's speaking direction. Through extensive real-world experiments, the authors demonstrate that SoundGaze achieves a median error of 19 degrees or lower.

Strengths

—----------

1. The concept of creating a lightweight sensing system that accurately determines users' speaking direction using only mmWave and audio signals is intriguing. Its application, where users can interact with smart devices by simply looking at them and speaking in

their direction, is compelling. This has the potential to eliminate the need for additional hardware attachments on those devices.

2. The paper is well-presented and well-organized. The authors provide sound reasoning and a solid preliminary study to advocate for the integration of mmWave radar and a microphone array in SoundGaze.

Weaknesses
—----------

1. The random selection method with an 80%-20% split for training and testing datasets to assess the proposed DNN model may have considerably influenced the outcomes. For a more robust evaluation, the k-fold cross-validation should be implemented instead.

   Response: For overall evaluation, we now conduct 5-fold cross-validation and report average results.

2. The Discussion section should be included to delve deeper into the current weaknesses and explore potential solutions envisioned by the authors for future improvements in the system.

   Response: In newly added Section 6.4 we discuss some of our improvement of the system, specially utilizing mmWave radar for unsupervised fine-tuning.

Comments for authors
—----------

. The authors should include a figure depicting the placement of the 9 IoT devices distributed around the testing room to bolster their conclusion regarding the system accuracy.

. In Section 8.6, in contrast to the median error of 25 degrees observed during natural movements, the error for fast movements should increase to 35 degrees, NOT reduce as indicated in the paragraph.

. Figure 3 has not been referenced in the main text of the manuscript.

. The system name, SoundGaze, in Figure 12-17 is not correct (i.e., SoundStance).

. There are typos in the manuscript that the authors should carefully revise. For example, in Section 2.2, "... approximately 2 feel away ..." should be "2 feet away".

Review #511E

==========

Overall merit

——————

2. Weak reject

Reviewer expertise

——————————

3. Knowledgeable

Paper summary

——————————

In this paper, the authors propose a system called SoundGaze to enhance the interactions between humans and smart hubs by combining a microphone array with a mmWave radar to localize the subject and detect the user's speaking direction. The framework of SoundGaze contains three steps: firstly, the system localizes the user by utilizing audio-based angle estimation and mmWave radar-based accurate localization; secondly, the point clouds are generated by filtering and processing both audio signals and mmWave signals; Lastly, a multi-modal fusion network is used to estimate the user's speaking direction. The authors also build a real-world testbed using a commercial microphone array and a mmWave radar, and conduct extensive experiments to demonstrate the proposed system outperforms the systems with only 2 or 3 microphone arrays.

Strengths

——————

1. The problem of the user's speaking direction estimation is very interesting and important.

2. The authors conduct preliminary research to investigate the limitation of single-point, audio-only solutions to estimate the user's speaking direction.

3. The authors conduct comprehensive experiments to compare the results of the proposed system and the systems with 2 or 3 microphone arrays.

Weaknesses

——————————

1. The motivation for equipping smart hubs with mmWave radar is not clear.

   Response: In section 1, we discuss the benefit of integrating mmWave with a smart hub. First of all, mmWave allows us to accurately localize a speaker- which is very error-prone using only audio. Secondly, mmWave gives us an estimate of front vs back detection and helps to minimize the search space for speech radiation patterns. Finally, mmWave allows us to employ a self-supervised approach to make the system robust and practical for deployment in any environment. Additionally, smart devices equipped

with mmWave radar such as Amazon Halo rise at $139, are commercially available and feasible in real-world.

2. It is not clear how the head orientation can be accurately estimated by fusing the acoustic and mmWave signals.

   Response: Although the head and body may not align perfectly, significant angular differences are uncommon during natural communication. Consequently, while mmWave point cloud data can constrain estimates of speaking direction based on body orientation, audio signals excel at determining the most probable speaking direction within this limited range.

3. The scenario with multiple subjects in the room is not evaluated.

   Response: In newly added Section 8.7, we report our result on people talking in the background and other types of environment noise.

4. The authors conduct real-time experiments but don't report the latency of the system.

   Response: The total processing time of our system is 1.2 second- reported in the introduction section.

Comments for authors

—----------

The motivation for equipping smart hubs with mmWave radar is not clear. In the paper, the mmWave radar should be integrated with the smart hub for accurate speaking direction estimation. However, the authors do not justify the motivation and feasibility of using such an integrated system in real-world deployment.

It is not clear how the head orientation can be accurately estimated by fusing the acoustic and mmWave signals. In real-world scenarios, when the user talks to the hub, there may be an angular offset between the user's head and body. It is reasonable to utilize the radar signals to estimate the orientation of the human body since the torso of the subject has strong reflections and the locations of limbs provide hinds for the body orientation. However, it is not clear how the head orientation is estimated using the noisy and sparse point cloud of the mmWave signals, especially when there is an offset between the user's head and body. It is suggested that the authors provide a detailed justification for it.

The scenario with multiple subjects in the room is not evaluated. It is common to have multiple people in the same room. The authors are suggested to conduct experiments when there are multiple people in the room, or even multiple people are speaking. In addition, the authors are suggested to clearly state the difference among the rooms in the cross-room experiments.

The authors conduct real-time experiments and report only the speaking direction error and the facing integer error. The performance of the real-timeness of such a voice system is not clear.