(511A.1) 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.

(511A.2)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.

(5111A.3) 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).

(511B.1) Known Position: 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? 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.

(511C.1) Neural Network: 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)?

Response: Overall the network was trained on 8 speakers, and 2 speakers were left out of the training set for cross speaker evaluation. The neural network shown in figure 10 outputs the facing direction only.

(511E.1) Integration of mmWave: 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.

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.

(511E.2) Audio radar fusion: 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.

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.

(511E.3) Multi-subject evaluation and latency: 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.

Response: In Section 8.7, we report our result on multiple people moving in the environment, however assuming one speaker at a time. We will add a new evaluation where there are other speakers in the background. The total processing time of our system is 1.2 second- reported in the introduction section.