(511A.1) Motivation: While [Yang et al.] reports 23-degree median error with two microphone arrays, the evaluation is done in controlled setup while the speaker is sitting on a chair. In our system, we experimented in various environments including living room, and office rooms, both static and mobile speakers, and different environment noise. In each case, our system with a single hub device performs better than 2 or 3-device audio only setup, even in cross-speaker evaluation.

(511A.2) Contribution: We design a careful fusion of audio and mmWave signal to localize a speaker and estimate speaking direction using a single smart hub. Our innovative DNN-based multimodal fusion network represents the first of its kind tailored specifically for integrating audio and mmWave data for speaking direction estimation. Furthermore, we've introduced a novel method for self-supervised fine-tuning of these DNNs, leveraging typical behaviors observed during natural communication between individuals for the first time.

(5111A.3) Missing Details: 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: Section 2.4 shows preliminary results using audio only. Section 8.2 shows results for audio-radar fusion.

(511C.1) Neural Network: 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: 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: 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: 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.