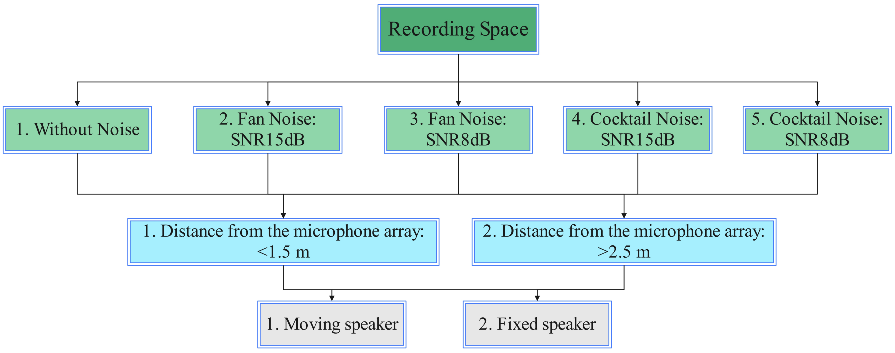
Recording Space
1. Without Noise
2. Fan Noise: SNR15dB
3. Fan Noise: SNR8dB
4. Cocktail Noise: SNR15dB
5. Cocktail Noise: SNR8dB
1. Distance from the microphone array: <1.5 m
2. Distance from the microphone array: >2.5 m
1. Moving speaker
2. Fixed speaker

Fig. 1: Recording conditions and the assigned number labels in single-speaker mode.

1 Assessment of audio localization

We used both the TAU-NIGENS Spatial Sound Events 2021 and 2020 datasets [1, 2] to train the neural network introduced in https://github.com/thomeou/SALSA-Lite. Thereafter, we tested the trained network with each of the test files in the TAU-NIGENS datasets and evaluated its performance on our customized recorded audio data. To evaluate the performance of the trained network on our customized audio recordings, we used the ReSpeaker USB Mic Array https://wiki.seeedstudio.com/ReSpeaker-USB-Mic-Array/ to collect audio data. This particular device comes with four channels specifically designed for recording sound. Several audio files were recorded with three individuals in the scene, where at most two speakers were talking at the same time. Regarding movement, all scenarios including one, two, or three moving speakers have been considered in the recordings.

Four distinct rooms with varying specifications, whose dimensions, types and reverberation times have been fully described in [3], were used for recording. Fan noise and cocktail party noise from https://noises.online were utilized as background noise at two SNR levels, approximately 8 dB and 15 dB. All the considered conditions for a single speaker and two speakers are shown in Figures 1 and 2, respectively. Within the recorded audio files, the initial 10 s consisted of either silence (in clean condition) or noise (if background noise was present), followed by the speaker/speakers engaging in dialogue as per the predetermined scenario from 10 to 30 seconds. Subsequently, there is either silence or background noise from 30 to 40 seconds, leading to the voice of the speaker/speakers from 40 to 60 seconds. During the initial 20-second period, the speaker/speakers remain fixed, while in the subsequent 20-second duration, they are moving.

Let us consider the objective is determining the locations of speakers in one-second intervals. If the SALSA-Lite method can identify a minimum of five 100-ms frames out of 10 in each second, it is classified as a "detection"; otherwise, it is labeled as a "missed detection".

In conclusion, our experimental results revealed although the SALSA-lite method was designed to detect and localize polyphonic sound sources while considering factors such as low reverberation, noise, and moving sound sources, it faces notable challenges

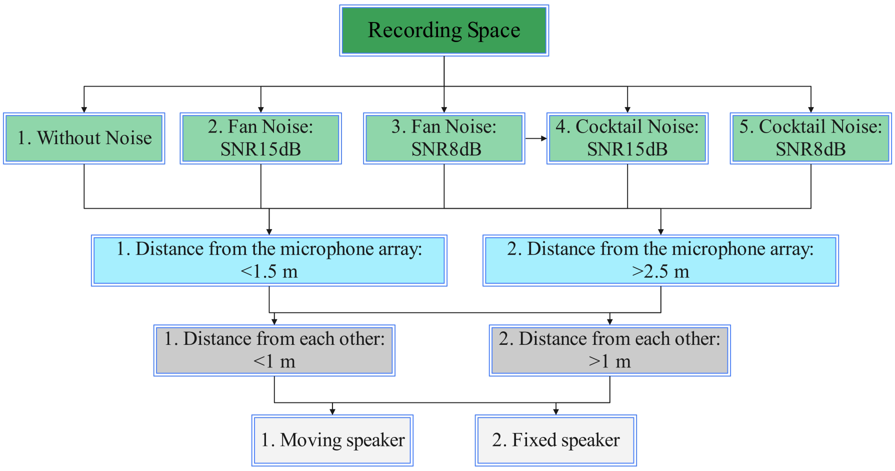
Recording Space
1. Without Noise
2. Fan Noise: SNR15dB
3. Fan Noise: SNR8dB
4. Cocktail Noise: SNR15dB
5. Cocktail Noise: SNR8dB
1. Distance from the microphone array: <1.5 m
2. Distance from the microphone array: >2.5 m
1. Distance from each other: <1 m
2. Distance from each other: >1 m
1. Moving speaker
2. Fixed speaker

Fig. 2: Recording conditions and the assigned number labels in two-speaker mode.

when implemented in practical settings [3]. The ability of the SALSA-lite method to accurately localize human speakers depends on factors such as the distance between speakers, the number of speakers, signal to noise ratio, speakers' voice tones and their movements. As reported in [3], the average missed detection rate in the single-speaker mode is approximately 6.61, while in the two-speaker mode, the missed detection rate is around 14.25.

References

[1] Politis, A., Adavanne, S., Virtanen, T.: TAU-NIGENS Spatial Sound Events 2020. https://doi.org/10.5281/zenodo.4064792 . https://doi.org/10.5281/zenodo.4064792

[2] Politis, A., Adavanne, S., Virtanen, T.: TAU-NIGENS Spatial Sound Events 2021. https://doi.org/10.5281/zenodo.4844825

[3] Esfehani, P., Faraji, N., Almasganj, F.: An experimental study on SALSA-Lite method for localization of multiple human speakers. In: 29th Int. Computer Conf. (2025)