1 Literature review

A comparison of different notable methods and underlying scenarios for the localization of audio events using audio-visual information is presented in Table 1.

Table 1: Comparison of different scenarios in localization of audio events using audio-visual information.

	Ref./ Year	Noise	Rev.	Simultaneous active Sources	Type/ Number	Moving	Activity Detection	Visual method	Audio method/ features	Record. devices	Localization Params
1	[1]/2008	×	×	✓	Speech/3 Sitting persons	×	✓	VJ face detector	Interaural time difference (ITD)	a pair of microphones and a pair of stereoscopic cameras	3D coordinate (X, Y, Z)
2	[2]/2013	✓	✓	×	Human Speech/3	×	✓	Face encoder	TDOA	two cameras and four microphones into the head of NAO	Azimuth and elevation
3	[3]/2021	×	×	✓	Musical instruments	×	×	VGG19	BiAudioEncoder and ConvLSTM	3Dio binaural microphones, GoPro to record video	pixel-level localization/ 2D
4	[4]/2021	✓	✓	×	Speech/1	×	×	The AdaBoost approach to train the face detection cascade classifier relying on Haar features	TDOA	linear microphone array in Kinect	Azimuth
5	[5]/2022	✓	✓	×	Speech/4-6	✓	✓	Audio-visual Deep network	360 degree voice map based on CNN	AR glasses with a microphone array and RGB camera	2D bounding boxes
6	[6]/2023	✓	✓	×	Speech/4-6	✓	✓	Image Patch Embedding	GCC-PHAT	6 microphones and a camera	Azimuth
7	[7]/2023	✓	✓	×	Speech/2	✓	✓	SeetaFaceEngine2 face detector	GCC-PHAT[8] and SALSA-LITE[9]	16-element microphone array and 11 cameras	2D bounding boxes
8	[10]/2023	✓	✓	✓	Speech/2 or 3	✓	✓	VJ detector: estimates the face and mouth location [11]	probabilistic version of SRP-PHAT	co-located audiovisual sensor	2D bounding boxes
9	Ours	✓	✓	✓	Speech/2	✓	✓	Pose estimation by MediaPipe + Lip activity detection	SALSA-Lite[9]	ReSpeaker USB Mic Array, single webcam	(x, y, z)

References

[1] Khalidov, V., Forbes, F., Hansard, M., Arnaud, E., Horaud, R.: Audio-visual clustering for 3D speaker localization. In: International Workshop on Machine Learning for Multimodal Interaction, pp. 86–97. Springer, ??? (2008)

[2] Cech, J., Mittal, R., Deleforge, A., Sanchez-Riera, J., Alameda-Pineda, X., Horaud, R.: Active-speaker detection and localization with microphones and cameras embedded into a robotic head. In: 13th IEEE-RAS Int. Conf. on Humanoid Robots, pp. 203–210 (2013)

[3] Wu, X., Wu, Z., Ju, L., Wang, S.: Binaural audio-visual localization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2961–2968 (2021)

[4] Zhu, Y.-X., Jin, H.-R.: Speaker localization based on audio-visual bimodal fusion. Journal of Advanced Computational Intelligence and Intelligent Informatics **25**(03), 375–382 (2021)

[5] Jiang, H., Murdock, C., Ithapu, V.: Egocentric deep multi-channel audio-visual active speaker localization. In: CVPR, pp. 10544–10552 (2022)

[6] Zhao, J., Xu, Y., Qian, X., Wang, W.: Audio visual speaker localization from egocentric views. arXiv. org (2023)

[7] Berghi, D., Jackson, P.: Leveraging visual supervision for array-based active speaker detection and localization. IEEE/ACM Trans. Audio Speech Lang. Process. **32**, 1–12 (2023)

[8] Cao, Y., Kong, Q., Iqbal, T., An, F., Plumbley, M.: Polyphonic sound event detection and localization using a two-stage strategy. In: Proceedings of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE 2019), pp. 30–34. New York University, ??? (2019). https://doi.org/10.33682/4jhy-bj81

[9] Nguyen, T.N.T., Jones, D.L., Watcharasupat, K.N., Phan, H., Gan, W.-S.: SALSA-lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays. In: ICASSP, pp. 716–720 (2022)

[10] Sanabria-Macias, F., Marron-Romera, M., Macias-Guarasa, J.: Audiovisual tracking of multiple speakers in smart spaces. Sensors **23**(15) (2023)

[11] Viola, P., Jones, M.: Robust real-time face detection. International Journal of Computer Vision **57**, 137–154 (2004)