1 Overall performance comparison

The performance of Audio (A), Video (V), and Audio-Visual (AV) localization methods is evaluated by calculating the track loss rate (TLR) and the Average Euclidean Distance between estimated positions and ground-truth values (AED), as presented in Tables 1 and 2.

The following part delves into the discussion regarding Tables 1 and 2.

- The AV algorithm results in an average of 3.36 lost frames for both speakers in both sequences, which is a reduction from the average TLR of 35.58 and 14.19 in the individual audio and visual modes, respectively. Additionally, the average AED in all modes is 4.09, showing a decrease from 8.91 in the audio method and 4.61 in the visual method.
- When two speakers speak simultaneously in the audio method, the TLR increases by more than five compared with when only one speaker speaks. The average TLR for both sequences when both speakers speak at the same time is 40.50, whereas it is 34.66 when only one speaker speaks. Additionally, the AED increased from 7.90 in the single-speaker mode to 9.89 in the two-speaker mode. It is worth noting that the number of lost frames for one speaker is consistently higher than that for the other speaker. For instance, in the period from 201 to 300, the first speaker experiences only nine lost frames, whereas the second speaker has 60 lost frames.
- During head rotation or when background light increases, alterations in TLR become evident. In Seq1, the initial speaker rotated his head for 4 s, leading to a loss of 36 frames. The subsequent speaker rotated his head even faster than the first speaker for 3 s, resulting in a loss of 24 frames. Conversely, in Seq2, the second speaker was exposed to an intense background light while moving for 5 s, causing a loss of 33 frames in the audio algorithm.
- During the time intervals when both speakers speak simultaneously and one speaker is positioned behind the other, the TLR and AED of both the audio and visual methods experience an increase. This is due to the fact that in the video method, the speaker who is positioned behind is not visible, resulting in all frames being lost during this period. Conversely, there were no lost frames for the front speaker in this scenario. Additionally, because both speakers are in close proximity to each other, a higher error in the estimation for the detected frames is obtained, and a portion of the audio frames is lost for both speakers. For instance, in Seq1, between the 54th and 58th seconds, the second speaker is completely obscured by the first speaker. In the visual method, 32 missing frames were noted for the second speaker, whereas the first speaker did not have any missing frames. In contrast, the audio method showed 16 missing frames for the first speaker and 15 missing frames for the second speaker during this period.
- At certain intervals, the AED of the visual method may exceed that of the audio method; however it exhibits a lower TLR. For instance, within Seq1, between 100 and 150, the TLR for the first speaker was 11.76 in the audio algorithm, with no lost frames in the visual algorithm. Moreover, the AED for the audio method was 9.20 across 44 frames, whereas it was 9.53 across 51 frames for the visual algorithm. Similarly, within Seq1, between 451 and 500, the TLR for the second speaker was

37.25 in the audio method, compared to 15.68 in the visual algorithm. The AED for the audio method was 5.77 across 32 frames, while it was 7.56 across 43 frames for the visual algorithm.

Table 1: Track loss rate (%) of Audio (A), Video (V) and Audio-Visual (AV) approaches in 3D localization of two speakers labeled by S1 and S2.

			Duration (Frame No.)					
			100-150	151-200	201-300	400-450	451-500	501-600
Seq1	**A**	S1	11.76	–	9	50.98	–	41
		S2	–	29.41	60	–	37.25	48
	V	S1	0	–	36	5.88	–	2
		S2	–	0	24	–	15.68	34
	AV	S1	0	–	0	0	–	0
		S2	–	0	0	–	0	16
Seq2	**A**	S1	21	–	40	23	–	38
		S2	–	49.01	39	–	54.9	49
	V	S1	7.8	–	18	0	–	0
		S2	–	9.8	35	–	5.88	33
	AV	S1	0	–	2	0	–	0
		S2	–	1.96	14	–	5.88	14

Table 2: $\text{AED}^{(x,y)}$ values attained by Audio (A), Video (V) and Audio-Visual (AV) approaches in 3D localization of two speakers labeled by S1 and S2.

			Duration (Frame No.)					
			100-150	151-200	201-300	400-450	451-500	501-600
Seq1	**A**	S1	9.2	–	11.1	8.86	–	9.01
		S2	–	7.38	10.2	–	5.77	10.58
	V	S1	9.53	–	9.05	8.58	–	3.14
		S2	–	3.26	2.97	–	7.56	4.59
	AV	S1	2.34	–	4.87	7.2	–	3.9
		S2	–	3.1	4.4	–	5.17	5.43
Seq2	**A**	S1	8.29	–	10.8	8.5	–	9.4
		S2	–	6.67	11.43	–	8.72	6.67
	V	S1	2.13	–	2.38	3.72	–	2.82
		S2	–	2.27	2.00	–	4.54	5.01
	AV	S1	3.37	–	3.59	2.75	–	2.48
		S2	–	3.06	4.28	–	4.46	5.19