Multimodal Speaker Separation in Audio-Visual Scene using Spherical Microphone Array

Jayesh Upadhyay, Priyadarshini Dwivedi, Gyanajyoti Routray*, and Rajesh M. Hegde
Indian Institute of Technology Kanpur, India, *SRM University AP, Amaravati, India.
Email: {jayeshs24, priyadw, rhegde}@iitk.ac.in, *gyanajyoti.r@srmap.edu.in

***Abstract*—Separating overlapping speakers in meeting-room environments remains challenging due to reverberation, source ambiguity, and dynamic speaker interactions. In existing systems, audio-only approaches degrade under adverse acoustic conditions, while audio-visual methods predominantly rely on explicit geometric calibration or depth estimation. This paper proposes a geometry-aware audio-visual speaker separation framework that addresses these limitations by integrating 360-degree visual perception with spherical microphone array (SMA) processing. A co-located 360-degree camera and SMA provide natural audio-visual alignment without explicit calibration. The visual front-end employs the YOLOv8l-pose on equirectangular video to extract speaker directions via nose-keypoint centroids with a confidence-gated bounding-box fallback. These directions steer a null-steering beamformer in 3 dimensional space, followed by SepFormer-based neural speech separation. Visual direction-extraction on the STARSS23 dataset achieves a mean angular error of 2.218° with 96.62% of detections within 5°. The full pipeline achieves an average SDR of 18.27 dB and PESQ of 3.093 in the multiple speaker case.**

I. Introduction

In realistic multi-speaker scenarios such as conference rooms, overlapping speech, reverberation, and dynamic speaker interactions significantly degrade unimodal system performance. Humans naturally integrate auditory and visual cues to resolve such ambiguities, motivating joint audio-visual frameworks for robust speaker separation. Spherical microphone arrays (SMAs) provide a principled framework for capturing 3D sound fields with uniform spatial resolution, making them suitable for direction-of-arrival (DOA) estimation and spatial filtering [1], [2]. Concurrently, 360-degree cameras enable full-scene visual perception. Integrating these modalities presents a promising direction for meeting-room audio reconstruction. Classical subspace-based methods such as MUSIC [3] in the SH domain and intensity-vector-based approaches [4] provide high-resolution localization under ideal conditions but degrade in reverberant multi-source environments [5], [6]. Learning-based approaches using CNNs/CRNNs on Ambisonics signals [7], [8] and modal coherence features [9] improved robustness, though generalization under severe acoustic degradation remains limited. In audio-visual localization, early datasets such as AV16.3 [10] established foundations for multimodal tracking, while recent methods extract speaker metadata from video [11] or combine panoramic video with microphone arrays [12]. YOLO-based detectors with geometric reasoning [13] and equirectangular projections [14] offer direct pixel-to-angle mapping for DOA estimation. Audio-visual DOA estimation has also been explored with permutation-free losses [15] and beamformed acoustic maps [16]. In neural speech separation, transformer-based architectures such as SepFormer [17] achieve state-of-the-art performance but lack explicit spatial grounding. However, reverberation degrades DOA estimation, audio-only methods fail in adverse conditions, visual pipelines face occlusions and distortions, and existing audio-visual systems often rely on explicit calibration or depth estimation.

This work makes the following contributions: (i) a calibration-free audio-visual speaker separation framework exploiting the natural geometric alignment of a co-located 360° camera and spherical microphone array; (ii) a visual DOA estimation pipeline using YOLOv8l-pose with a confidence-gated keypoint centroid, (iii) a hybrid audio pipeline combining visually-steered null-steering beamforming with SepFormer-based neural separation, linked by cross-correlation candidate selection, and (iv) a robust multi-modal speaker separation in presence of noise and reverberation.

II. System Model

The model is developed by considering N active speakers (overlapping in both time and frequency) in a room scenario with noise and reverberation environment. An SMA of radius r comprising M omnidirectional flush mounted microphones, is placed at the center of the room at position $\mathbf{c}_{\text{array}} \in \mathbb{R}^3$. A 360-degree camera is co-located with the array such that the optical center of the camera coincides with $\mathbf{c}_{\text{array}}$ to establish a shared coordinate frame between the acoustic and visual modalities, eliminating the need for explicit extrinsic calibration. The camera captures an equirectangular projection (ERP) frame $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, where the horizontal axis spans the full 360° azimuthal range and the vertical axis spans 180° of elevation. Let $s_i[n]$, $i = 1, \ldots, N$, denote the clean speech signal produced by the i-th speaker. Each source propagates through the room and arrives at the array after undergoing reflections and diffraction characterised by the room impulse response (RIR) $h_{m,i}[n]$ between source i and microphone m. The signal captured at the m-th microphone is

$$x_m[n] = \sum_{i=1}^{N} (h_{m,i} * s_i)[n] + v_m[n], \tag{1}$$

where $*$ denotes linear convolution and $v_m[n]$ represents additive ambient noise at microphone m. For M microphones,

$$\mathbf{x}[n] = \big[x_1[n],\, x_2[n],\, \ldots,\, x_M[n]\big]^T \in \mathbb{R}^M. \tag{2}$$

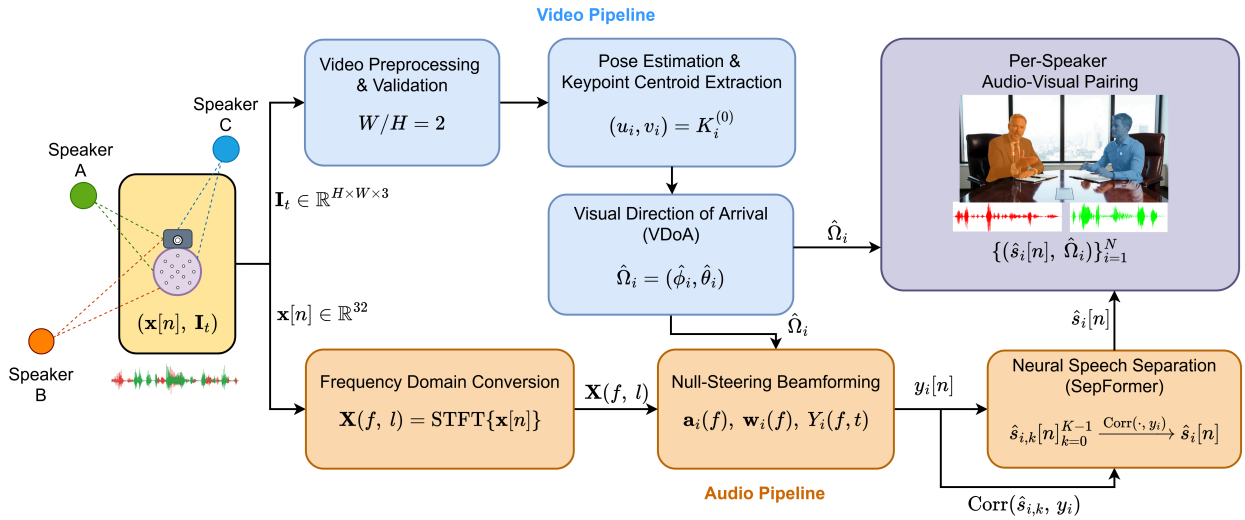

Fig. 1: Proposed multimodal speech separation framework.

Applying the short-time Fourier transform (STFT) to each channel produces the frequency-domain representation

$$\mathbf{X}(f,\ l) = \mathrm{STFT}\{\mathbf{x}[n]\} \in \mathbb{C}^M, \tag{3}$$

where f is the frequency bin index and l the time-frame index. In the STFT domain, the narrowband approximation allows the convolutive mixture in (1) to be expressed as an instantaneous mixture at each time-frequency bin

$$\mathbf{X}(f,\ l) \approx \sum_{i=1}^{N} \mathbf{a}_i(f)\, S_i(f,\ l) + \mathbf{V}(f,\ l), \tag{4}$$

where $S_i(f,\ l)$ is the STFT of source i, $\mathbf{V}(f,\ l) \in \mathbb{C}^M$ is the noise vector, and $\mathbf{a}_i(f) \in \mathbb{C}^M$ is the acoustic steering vector associated with the i-th source at frequency f. The steering vector encodes the inter-microphone propagation delays:

$$\mathbf{a}_i(f) = \left[\, e^{-j2\pi f \tau_{1,i}},\ e^{-j2\pi f \tau_{2,i}},\ \ldots,\ e^{-j2\pi f \tau_{M,i}} \,\right]^T, \tag{5}$$

where $\tau_{m,i} = (d_{m,i} - d_{1,i})/c$ is the relative propagation delay, $d_{m,i} = \|\mathbf{q}_m - \mathbf{p}_i\|$ is the Euclidean distance between the m-th microphone and i-th source $\mathbf{p}_i \in \mathbb{R}^3$, and c is the speed of sound. Now, the objective is to recover an estimate $\hat{s}_i[n]$ of each individual source signal $s_i[n]$, and to associate each recovered signal with its corresponding speaker direction $\hat{\Omega}_i = (\hat{\varphi}_i,\ \hat{\theta}_i)$, where $\hat{\varphi}_i \in [-180°, 180°]$ is azimuth and $\hat{\theta}_i \in [-90°, 90°]$ is elevation, is extracted from the visual modality, resulting in the output set

$$\left\{(\hat{s}_i[n],\ \hat{\Omega}_i)\right\}_{i=1}^{N}. \tag{6}$$

III. Proposed Methodology

The proposed technique consists of three stages: a video pipeline to detect the speakers with their angular direction; an audio extraction pipeline that uses these directions to steer a beamformer; and neural speaker separation SepFormer [17]. The proposed technique is illustrated in Fig. 1.

A. Video pipeline: speaker detection and direction extraction

For each person $i \in \{1, \ldots, N\}$, the model produces $J = 17$ COCO skeletal keypoints [18]

$$\mathbf{K}_i = \left\{(\kappa_{i,j}^x,\ \kappa_{i,j}^y,\ c_{i,j})\right\}_{j=0}^{J-1}, \tag{7}$$

where $(\kappa_{i,j}^x, \kappa_{i,j}^y)$ are pixel coordinates and $c_{i,j} \in [0, 1]$ is the confidence score. Keypoint $j = 0$ (nose) serves as the primary centroid target. The face-level centroid (u_i, v_i) is extracted using a confidence-gated strategy with threshold $\tau_k = 0.2$:

$$(u_i,\ v_i) = \begin{cases} (\kappa_{i,0}^x,\ \kappa_{i,0}^y), & \text{if } c_{i,0} \geq \tau_k, \\ \left(\frac{x_1^{(i)} + x_2^{(i)}}{2},\ y_1^{(i)} + \beta\, h_i^{\mathrm{box}}\right), & \text{otherwise,} \end{cases} \tag{8}$$

where $(x_1^{(i)}, y_1^{(i)}, x_2^{(i)}, y_2^{(i)})$ are bounding-box coordinates, $h_i^{\mathrm{box}} = y_2^{(i)} - y_1^{(i)}$, and $\beta = 0.15$ is the fallback face-level fraction. The primary path exploits the pose model's ability to infer head position from surrounding skeletal geometry even when facial features are obliterated. The fallback activates only when the speaker is turned completely away from the camera. The extracted centroids are mapped to speaker's angular directions using the equirectangular projection formula [14]:

$$\hat{\varphi}_i = \frac{u_i}{W} \times 360° - 180°; \quad \hat{\theta}_i = 90° - \frac{v_i}{H} \times 180°, \tag{9}$$

B. Audio extraction pipeline

The visually estimated directions $\hat{\Omega}_i$ are converted to 3D source positions to construct the steering vectors in (5). Let $\mathbf{f} = [\cos\psi,\ \sin\psi,\ 0]^T$ be the camera forward vector, $\mathbf{r} = [\sin\psi,\ -\cos\psi,\ 0]^T$ the right vector, and $\mathbf{u} = [0,\ 0,\ 1]^T$ the up vector. The unit direction toward speaker i is

$$\hat{\mathbf{d}}_i = \cos\hat{\theta}_i \sin\hat{\varphi}_i\, \mathbf{r} + \cos\hat{\theta}_i \cos\hat{\varphi}_i\, \mathbf{f} + \sin\hat{\theta}_i\, \mathbf{u}, \tag{10}$$

and the assumed position at distance ρ_i is $\mathbf{p}_i = \mathbf{c}_{\mathrm{array}} + \rho_i\, \hat{\mathbf{d}}_i$.

1) Null-steering beamformer: For target speaker i, the constraint matrix combining the target and all interfering steering vectors is:

$$A_i(f) = \big[\mathbf{a}_i(f) \;\; \mathbf{a}_{j_1}(f) \;\; \cdots \;\; \mathbf{a}_{j_{N-1}}(f)\big] \in \mathbb{C}^{M\times N}. \quad (11)$$

To ensure numerical stability and invertibility, the regularized Gram matrix is used [19], [20]

$$R_{i,\mathrm{reg}}(f) = A_i^H(f)A_i(f) + \alpha\frac{\mathrm{tr}(A_i^H A_i)}{N}\, I, \quad (12)$$

where α is the regularisation coefficient. The constraint vector $\mathbf{g} = [1,\, 0,\, \ldots,\, 0]^T \in \mathbb{R}^N$ enforces unit gain for the target and nulls for all interferers. The beamforming weights are [21]

$$\mathbf{w}_i(f) = A_i(f)\, R_{i,\mathrm{reg}}^{-1}(f)\, \mathbf{g} \in \mathbb{C}^M, \quad (13)$$

and the beamformed output is

$$Y_i(f,\, l) = \mathbf{w}_i^H(f)\, \mathbf{X}(f,\, l). \quad (14)$$

The time-domain signal $y_i[n]$ is recovered by inverse STFT.

C. Speaker separation pipeline

The beamformed signal $y_i[n]$ is processed by pretrained SepFormer [17], [22]. SepFormer is run on each of N beamformed inputs, yielding $N \times K$ candidates $\hat{s}_{i,k}[n]$:

$$\{\hat{s}_{i,0}[n],\; \hat{s}_{i,1}[n],\; \ldots,\; \hat{s}_{i,K-1}[n]\} = \mathrm{SepFormer}(y_i[n]). \quad (15)$$

Each candidate is scored against the corresponding beamformed signal $y_i[n]$ using the peak of the normalised cross-correlation:

$$\mathrm{Corr}(\hat{s},\, y) = \max_{\tau} \big|\, \textstyle\sum_n\, \tilde{s}[n+\tau]\; \tilde{y}[n]\, \big|, \quad (16)$$

where $\tilde{s}[n] = \hat{s}[n]/\|\hat{s}\|$ and $\tilde{y}[n] = y[n]/\|y\|$ are the ℓ_2-normalised signals, and the correlation is computed efficiently via the FFT. Because the beamformed signal already encodes the spatial signature of the target direction, this metric serves as a reference-free similarity measure. For each target speaker i, the best candidate is selected as

$$\hat{s}_i[n] = \underset{k\in\{0,\ldots,K-1\}}{\arg\max}\; \mathrm{Corr}\big(\hat{s}_{i,k}[n],\, y_i[n]\big). \quad (17)$$

This candidate selection uses only the beamformed signal $y_i[n]$ as a spatial anchor, exploiting the fact that null-steering beamformer already encodes target speaker's directional signature. No clean or dry reference signals are required, making the pipeline fully self-contained at inference time. The final output is set $\{(\hat{s}_i[n],\, \hat{\Omega}_i)\}_{i=1}^N$ defined in (6), associating each separated stream with its video-derived direction.

IV. Performance Evaluation

To validate the proposed geometry-aware audio-visual framework, the visual and acoustic pipelines are evaluated using real-world 360° video data and controlled multichannel acoustic simulations. The spatial-neural system is benchmarked against a classical spatial-spectral baseline across various acoustic conditions to quantify improvements in separation quality and intelligibility. A demonstration of the complete pipeline can be found at *https://home.iitk.ac.in/ rhegde/spcomdemo/*.

TABLE I: SMIR simulation parameters.

Parameter	**Value**
Room dimensions	$6.5 \times 4.8 \times 3.0$ m
Sampling rate	8 kHz
Reverberation time	$T_{60} = 0.35$ s
Reflection order	3
Array radius	0.042 m
Number of microphones	32
Array height	0.792 m
Number of speakers	3
Assumed speaker distance (ρ_i)	1.5 m
Sphere type	Rigid

A. Datasets

In this work, STARSS23 data corpus is used for performance evaluation [23]. The facial anonymization (blurring) in the STARSS23 dataset is mitigated by employing YOLOv8l-pose [24] instead of standard bounding-box detector. The P6 variant processes 1920×960 equirectangular frames at native resolution. The audio recordings are simulated considering the room dimension of $6.5 \times 4.8 \times 3.0$ m, using the Eigenmike EM32 [25] ($r = 0.042$ m, $M = 32$) placed at the center at height 0.792 m, co-located with the camera. Room impulse responses are generated using the SMIR generator [26] with $T_{60} = 0.35$ s and reflection order 3. The audio source positions corresponds to the speaker positions (N=3) derived from the video-extracted angles. The final recordings are obtained by convolving clean speech signals from SparseLibriMix [27] at 8 kHz with the RIRs and adding Gaussian noise. The parameters are summarized in Table I. The STFT uses a 512-point Hann window with hop size 128 (75% overlap). Further, for the three-speaker case, the `sepformer-libri3mix` model produces K candidate output channels per inference pass.

B. Evaluation metrics

The visual pipeline is evaluated by comparing detected directions against ground-truth annotations using Hungarian assignment [28]. The mean absolute azimuth error (MAAE) ($\overline{|\Delta\varphi|}$), mean absolute elevation error (MAEE) ($\overline{|\Delta\theta|}$), mean angular error (MAnE) ($\overline{\Delta\sigma}$) computed via the haversine formula [29], and angular precision ($\leq 5°$) are considered as metrics to evaluate the performance. The audio pipeline is evaluated using Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Source-to-Artifact Ratio (SAR) from BSS Eval, Perceptual Evaluation of Speech Quality (PESQ) (narrowband, 8 kHz), and Short-Time Objective Intelligibility (STOI) [30]–[32]. All BSS Eval metrics use permutation- invariant assignment and are averaged across speakers. Permutation-invariant evaluation is applied solely for metric computation against the ground-truth references; the pipeline's own candidate assignment uses only the blind correlation-based matching described in Section III-C.

C. Analysis of visual direction estimation

The speaker directions are obtained and compared with ground truth and mentioned in Table II which summaries the visual front-end performance. The system achieves MAnE

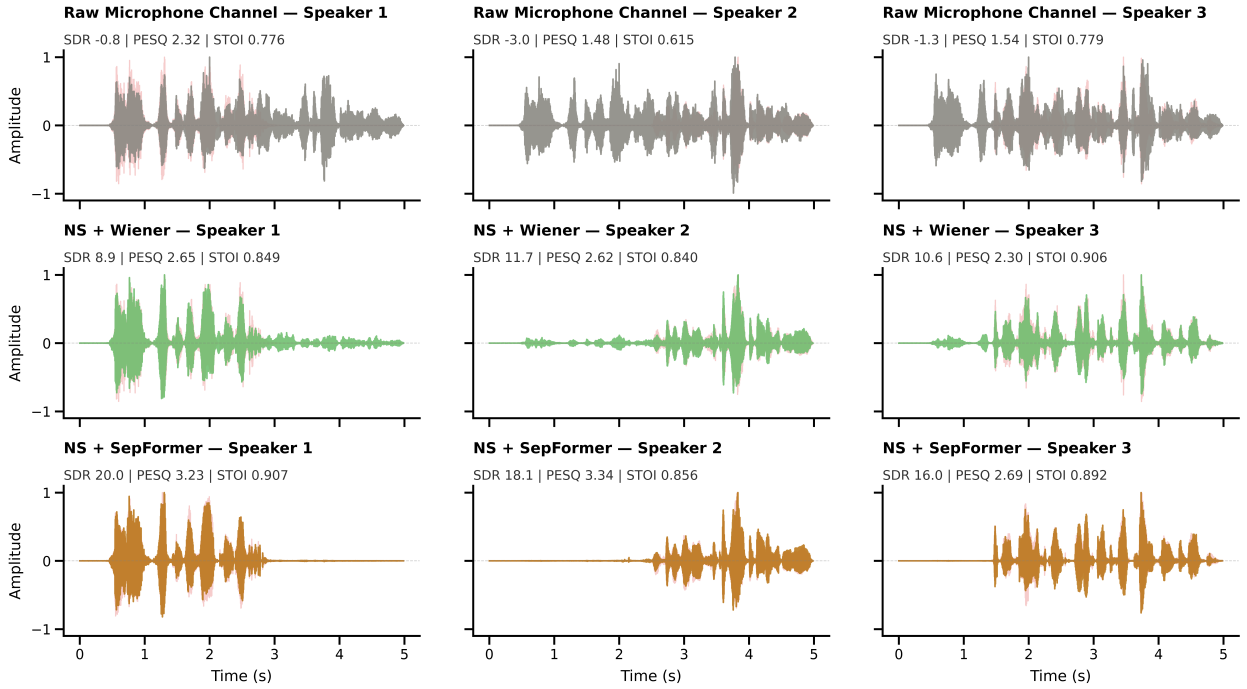

Fig. 2: Waveform comparison across the proposed audio pipeline. The top row shows the raw microphone output against the reverberant reference for Speakers 1 and 2, the middle row shows the null-steering and Wiener (baseline) stages, and the bottom row shows the final NS+SepFormer output.

TABLE II: Spatial Localization Accuracy of the YOLO-Pose Pipeline on Anonymized 360° Equirectangular Video.

Metric Category	Evaluation Parameter	Value
Spatial Accuracy	MAAE ($\overline{\lvert\Delta\varphi\rvert}$)	2.998°
	MAEE ($\overline{\lvert\Delta\theta\rvert}$)	1.112°
	MAnE ($\overline{\Delta\sigma}$)	2.218°
Success Rate	Angular Precision ($\leq 5°$)	96.62%
Model Reliability	Mean Detection Confidence	0.812

TABLE III: Audio pipeline metrics averaged across all speakers. All stages evaluated against the reverberant single-speaker reference. NS+Wiener is a classical baseline; NS+SepFormer is the proposed system.

Stage	SDR (dB)	SIR (dB)	SAR (dB)	PESQ	STOI
32-ch Raw Microphone	-1.93	-1.82	19.34	1.619	0.775
Null-Steering (NS)	4.47	5.45	13.15	2.120	0.823
NS + Wiener (baseline)	9.82	13.76	12.70	2.529	0.862
NS + SepFormer	**18.27**	**23.41**	**20.49**	**3.093**	**0.900**

of 2.218° and angular precision of 96.62% ($\leq$ 5°), with particularly strong vertical stability (1.112° elevation error). These results confirm that the pose-based method successfully bypasses the limitations of facial blurring to provide a robust spatial foundation for downstream beamforming. Since these results are obtained on heavily anonymized video where faces are entirely blurred, pose-based pipeline is expected to achieve equal or better accuracy on unblurred footage, where the nose keypoint can be detected directly with higher confidence.

D. Audio separation performance

The Table III shows the performance of various methods using the evaluation metrics mentioned in Section IV B. It can be observed from the results that the proposed NS+SepFormer outperforms all the other baseline approaches. Further, the waveform comparison in Fig. 2 visually confirms the progressive improvement. In terms of computational cost, the YOLOv8l-pose visual front-end runs in approximately 30 ms per frame on a single GPU, while the SepFormer inference dominates the overall latency. In future extension of the work, real-time operation would require model compression or streaming-compatible architectures.

E. Robustness analysis under reverberation and noise

The performance is also shown for various SNR and reverberation (RT_{60}) values for the robustness analysis. Table IV shows that the system remains highly intelligible across reverberation times, with STOI above 0.85 even at $RT_{60} = 1.00$ s. Table V shows noise performance, STOI stays above 0.79 but drops at lower SNR levels. A noise-aware training strategy or explicit denoising front-end would further enhance the performance in acoustically harsh environments.

V. Conclusion

This work presents a calibration-free audio-visual speaker separation pipeline that combines 360-degree visual direction

TABLE IV: NS+SepFormer across reverberation times (T_{60}).

T_{60} (s)	SDR	SIR	SAR	PESQ	STOI
0.20	18.43	25.46	19.66	3.409	0.944
0.35	12.53	20.86	13.51	2.879	0.901
0.50	9.27	17.96	12.28	2.714	0.858
0.70	9.15	18.09	10.89	2.756	0.856
1.00	8.70	17.75	10.39	2.593	0.854

TABLE V: NS+SepFormer across input SNR levels.

SNR (dB)	SDR	SIR	SAR	PESQ	STOI
5	4.14	15.11	6.77	1.637	0.793
10	7.37	17.51	9.47	2.302	0.865
15	10.02	18.57	11.66	2.585	0.890
20	11.75	19.33	13.06	2.766	0.902
25	12.39	19.86	13.56	2.859	0.903
30	12.57	20.25	13.63	2.887	0.905

extraction, null-steering beamforming, and SepFormer-based neural separation. The visual front-end achieves a mean angular error of 2.218° on the STARSS23 dataset, and the full pipeline achieves an average SDR of 18.27 dB and PESQ of 3.093 in the multiple speaker case, outperforming the classical Wiener baseline on all metrics. Notably, these visual results are obtained on fully anonymized (face-blurred) video, the performance on unblurred footage is expected to be at least as good, underscoring the practical robustness of the approach. The future work will incorporate dynamic beamforming for non-stationary speakers using frame-by-frame detection with temporal association and adaptive steering-vector updates. The fixed assumed speaker distance ρ_i could be replaced by monocular depth estimation. Moreover, noise-aware training of the SepFormer model could be done in extension to this work for further improvements. Also, in the extension of this work we are planning to evaluate the pipeline on real recorded data and assess the impact of visual DOA estimation error on beamformer null depth and separation quality.

ACKNOWLEDGEMENT

This work is supported by the ANRF under project number ANRF/ARG/2025/007629/ENS.

REFERENCES

[1] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2002, pp. II–1781–II–1784.

[2] G. Routray, P. Dwivedi, and R. M. Hegde, "Binaural reproduction of hoa signal using sparse multiple measurement vector projections," in *2021 National Conference on Communications (NCC)*, 2021, pp. 1–6.

[3] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[4] P. Dwivedi, G. Routray, and R. M. Hegde, "Learning based method for near field acoustic range estimation in spherical harmonics domain using intensity vectors," *Pattern Recognition Letters*, vol. 165, pp. 17–24, 2023.

[5] D. Desai and N. Mehendale, "A review on sound source localization systems: D. desai, n. mehendale," *Archives of Computational Methods in Engineering*, vol. 29, no. 7, pp. 4631–4642, 2022.

[6] P. Dwivedi et al., "Spherical harmonics domain-based approach for source localization in presence of directional interference," *JASA Express Letters*, vol. 2, no. 11, p. 114802, 11 2022.

[7] P.-A. Grumiaux et al., "Improved feature extraction for crnn-based multiple sound source localization," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 231–235.

[8] P. Dwivedi, G. Routray, and R. M. Hegde, "Octant spherical harmonics features for source localization using artificial intelligence based on unified learning framework," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 8, pp. 3845–3857, 2024.

[9] P. Dwivedi et al., "Improving source tracking accuracy through learning-based estimation methods in sh domain: A comparative study," *IEEE Trans. on Artificial Intelligence*, vol. 5, no. 8, pp. 3974–3984, 2024.

[10] G. Lathoud et al., "Av16.3: An audio-visual corpus for speaker localization and tracking," in *Machine Learning for Multimodal Interaction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 182–195.

[11] D. Berghi and P. J. B. Jackson, "Audio-visual talker localization in video for spatial sound reproduction," in *Proceedings of DAFx*, 2024.

[12] H. Jiang, C. Murdock, and V. K. Ithapu, "Egocentric deep multi-channel audio-visual active speaker localization," in *Proceedings of CVPR*, 2022.

[13] X. L. et al., "Visual-based spatial audio generation system for multi-speaker environments," 2025.

[14] R. S. et al., "Audio–visual object removal in 360-degree videos," *The Visual Computer*, vol. 36, pp. 2117–2128, 2020.

[15] Q. W. et al., "Deep learning based audio-visual multi-speaker doa estimation using permutation-free loss function," in *Proceedings of ISCSLP*, 2022.

[16] B. J. R. et al., "Beamformed 360-degree sound maps: U-net-driven acoustic source segmentation and localization," 2025.

[17] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proceedings of ICASSP*, 2021, pp. 21–25.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conf. on computer vision*. Springer, 2014, pp. 740–755.

[19] B. D. Carlson, "Covariance matrix estimation errors and diagonal loading in adaptive arrays," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 24, no. 4, pp. 397–401, 1988.

[20] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[21] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[22] M. R. et al., "Speechbrain: A general-purpose speech toolkit," 2021.

[23] K. S. et al., "STARSS23: A Dataset of Spatial Recordings of Real Scenes with Spatial and Temporal Annotations of Sound Events," in *Proceedings of NeurIPS Datasets and Benchmarks*, 2023.

[24] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023.

[25] mh acoustics, "Em32 eigenmike microphone array release notes," 2013.

[26] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: algorithm and applications," *Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1462–1472, 2012.

[27] S. Cornell, "SparseLibriMix: An open-source dataset for generalizable speech separation with variable overlap ratio," https://github.com/popcornell/SparseLibriMix, 2020.

[28] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[29] R. W. Sinnott, "Virtues of the haversine," *Sky and Telescope*, vol. 68, no. 2, p. 158, 1984.

[30] E. V. et al., "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[31] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq): A new method for speech quality assessment of telephone networks and codecs," in *Proceedings of ICASSP*, 2001, pp. 749–752.

[32] C. H. T. et al., "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.