Visual-based spatial audio generation system for multi-speaker environments*

Xiaojing Liu[1], Ogulcan Gurelli[1], Yan Wang[2], and Joshua Reiss[1]

***Abstract*— In multimedia applications such as films and video games, spatial audio techniques are widely employed to enhance user experiences by simulating 3D sound: transforming mono audio into binaural formats. However, this process is often complex and labor-intensive for sound designers, requiring precise synchronization of audio with the spatial positions of visual components. To address these challenges, we propose a visual-based spatial audio generation system - an automated system that integrates face detection YOLOv8 for object detection, monocular depth estimation, and spatial audio techniques. Notably, the system operates without requiring additional binaural dataset training. The proposed system is evaluated against existing Spatial Audio generation system using objective metrics. Experimental results demonstrate that our method significantly improves spatial consistency between audio and video, enhances speech quality, and performs robustly in multi-speaker scenarios. By streamlining the audio-visual alignment process, the proposed system enables sound engineers to achieve high-quality results efficiently, making it a valuable tool for professionals in multimedia production.**

***Index Terms*—Spatial audio, multi-model, audio-visual system**

I. Introduction

In multimedia production, including films, advertisements, teleconferencing and video games, achieving seamless alignment between character dialogue and corresponding visual elements is essential for creating immersive experiences. As humans rely on multi-modal cues to interpret and engage with real-world events [1], the demand for high-quality audio-visual experiences continues to grow. With the rise of three dimensional (3D) audio, virtual reality (VR), and augmented reality (AR), the importance of accurate spatial audio alignment has become increasingly evident [2]. Visual-based audio spatialization has become a prominent area of research due to its broad applications in AR [3], VR [4], social video sharing [5] [6] and audio-visual video understanding [7]. Effective audio-visual spatialization enhances realism, enabling audiences to feel as though they are present within the environment.

Currently, most post-production professionals and audio engineers manually adjust spatial audio parameters on digital platforms, relying on visual cues to determine the positions of sound sources. This process is highly labor-intensive and requires significant time and effort.

*This work was supported by Queen Mary University of London and the China Scholarship Council (CSC).

[1]School of Electronic Engineering and Computer Science, Center for Digital Music (C4DM), Queen Mary University of London, United Kingdom. xiaojing.liu@qmul.ac.uk, ec21698@qmul.ac.uk, joshua.reiss@qmul.ac.uk

[2]Department of Electrical Engineering, Xidian University, China. wangyan97@stu.xidian.edu.cn

Some researchers have already explored generating spatial audio based on video input. Lin et al. [7] proposed a model for generating binaural audio from visual frames and monaural audio inputs, demonstrating its effectiveness through comparisons with other models on the FAIR-Play dataset (binaural audio clips recorded in a controlled music room) and the MUSIC-Stereo dataset (a diverse collection of audio-visual clips from musical performances). Ruohan Gao and Kristen Grauman [8] introduced Mono2 Binaural, a deep network that takes a mixed monaural audio and its accompanying visual frame as input, using a ResNet-18 network to extract visual features and U-NET to extract audio features.

However, these systems [7]–[9] rely heavily on large binaural datasets, which pose significant challenges, such as requiring specialized equipment, controlled recording environments, precise audio-visual synchronization, and labor-intensive annotation processes. Additionally, they face issues with overfitting when handling multiple audio tracks, further complicating the training and optimization process. They also struggle to achieve precise spatial positioning with more than two audio sources, frequently leading to compromised sound quality or reduced immersion, ultimately impacting the user experience. This lack of fidelity not only affects the user experience but also poses challenges in professional multimedia production environments, where rapid adaptation and precision are essential.

Pseudo Binaural [10] has been introduced as an innovative model to generate visually coherent binaural audios from multiple sound sources without requiring recorded binaural data. However, for these system the coordinates for Azimuth and Elevation must be manually pre-defined.

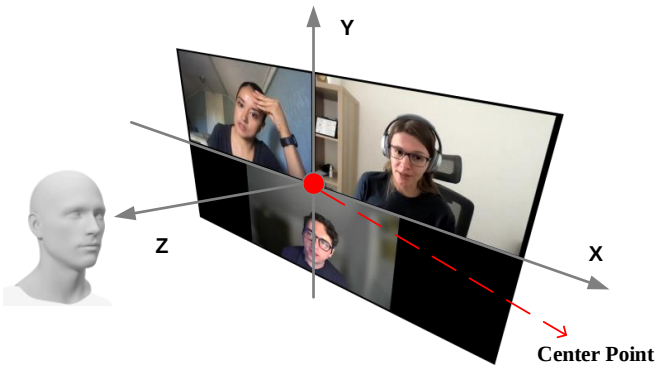


Fig. 1: Spatialisation of audio representing **X**, **Y** and **Z** coordinates.

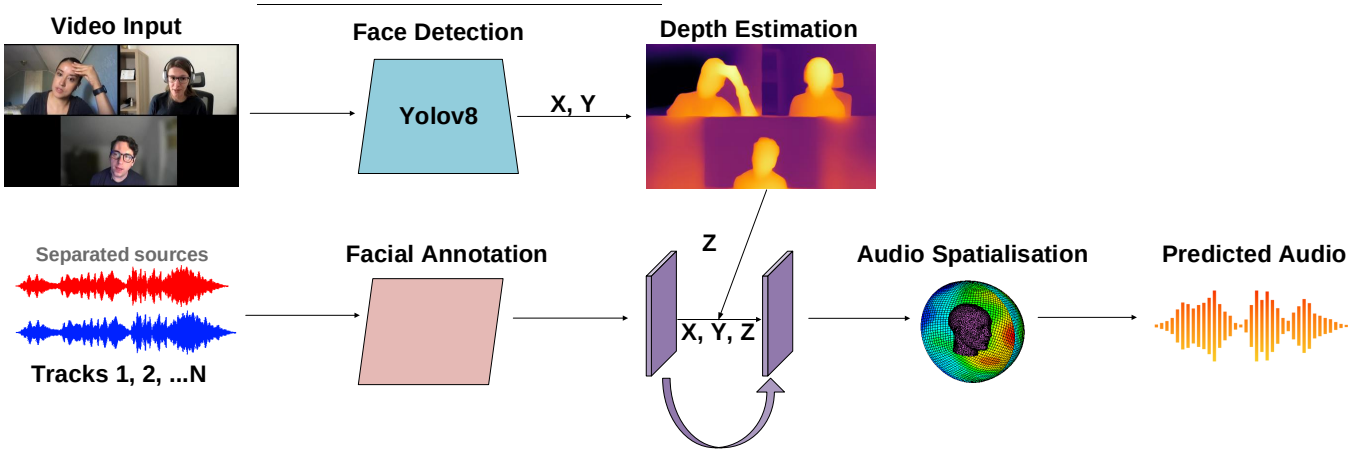


Fig. 2: Flowgraph of the system showcasing the main processing pipeline.

To address these challenges, we propose a visual-based spatial audio generation system designed to support multi-speaker scenarios while eliminating the dependency on large-scale binaural datasets. In this approach (see Fig. 1), the center of each video frame is assumed to be the origin of a Cartesian coordinate system. Using this framework, the facial positions of individual speakers are accurately calculated, providing spatial cues for generating realistic spatial audio.

Our system integrates object detection using YOLOv8 [11] with the WIDER FACE Dataset [12] and pre-trained Depth Estimation model Depth Anything [13], to enhance the precision of spatial audio alignment by extracting visual cues with high accuracy, thus facilitating the automatic adjustment of audio sources based on spatial positioning in real-time.

To assess the effectiveness of our system objectively, we compare its performance against existing audio-visual spatial audio generators using established metrics: Perceptual Evaluation of Speech Quality (PESQ), Short-Time Fournier Transform (STFT) Distance, Envelope (ENV) Distance, and Mean Opinion Score (MOS). These metrics enable a comprehensive evaluation of our system's ability to maintain audio quality and synchronization under varying conditions, demonstrating its applicability in high-fidelity audio post-production environments.

Our contributions can be summarized as follows:

Elimination of Large Binaural Dataset Dependency: Unlike existing systems that heavily rely on extensive binaural datasets, which require specialized recording setups and labor-intensive annotations, our method generates spatial audio without the need for large scale binaural datasets.

Multi-Speaker Scenario Support: In contrast to traditional methods that struggle with precise spatial positioning beyond two audio sources, our system effectively supports multiple speakers, maintaining spatial accuracy and immersion.

The using in real-world: The proposed system can streamline sound design workflows, it makes a contribution to audio/music producer in the Post-production audio mixing, or improves user experience in teleconferencing application (e.g. Zoom).

Audio quality and Immersive Promotion: Our method outperforms existing works in audio quality, particularly in speech scenarios, as shown in Table III and Table IV.

II. Methodology

Our model[1] consists of two main components: visual processing and audio processing. For visual processing, we compare current YOLO models and employ YOLOv8 in combination with the Depth Estimation model Depth Anything (see Sec. II-A). Details of audio source preparation and facial annotation are provided in Sec. II-B and II-C. In the audio processing pipeline, we implement two spatialization methods: HRTF convolution and a 3D algorithmic approach, as detailed in Sec. II-D. The overall workflow of the system is illustrated in Fig. 2.

A. Visual

Object detection facilitates the localization of spatial positions, particularly azimuthal information, of characters or objects in real-time, acting as anchors for their corresponding audio sources. By generating precise bounding boxes around detected objects, the object detection model ensures consistent spatial alignment, even in dynamic scenes where sources may move or overlap.

YOLO (You Only Look Once) is a widely used object detection model renowned for its speed and accuracy. First introduced by Joseph Redmon et al [14] in 2016, YOLO has undergone numerous advancements, with the latest iteration, YOLOv10, incorporating state-of-the-art techniques for enhanced performance and efficiency. This makes it well-suited for applications requiring real-time tracking and localization in complex environments. To evaluate advancements in YOLO's performance and its suitability for real-time applications, we conducted comparative testing across different YOLO versions using the WIDER FACE dataset. This dataset comprises 32,203 images and 393,703 labeled faces, with significant variations in scale, pose, and occlusion. It is divided into training (40%), validation (10%), and testing

[1] https://youtu.be/lWbx58GmP-o

(50%) sets, ensuring a balanced distribution. The results are summarized in Table I and Table II.

The experiments are conducted using PyTorch 1.7.0 on an NVIDIA RTX 3090 GPU with CUDA 11.0. The system is configured with a 15-core Intel Xeon Platinum CPU and 80 GB RAM, providing sufficient computational resources for training and evaluation.

TABLE I: Performance comparison of YOLO models (P: Precision, R: Recall, mAP50: mean average precision at IoU threshold 0.5, mAP50-95: mean average precision at IoU thresholds ranging from 0.5 to 0.95).

Module	P	R	mAP50	mAP50-95
YOLOv5-n	0.838	0.598	0.726	0.342
YOLOv10-n	0.829	0.554	0.633	0.316
YOLOv8-n	0.845	0.588	0.669	0.366
YOLOv-Face2	**0.896**	**0.666**	**0.735**	**0.397**
YOLOv5-s	0.872	0.655	0.696	0.347

TABLE II: Speed comparison of YOLO models (measured in milliseconds, ms).

Model	Pre-process	Inference	NMS
YOLOv5-n	0.2	6.1	0.8
YOLOv10-n	0.1	1.2	**0.0**
YOLOv8-n	**0.1**	**0.6**	0.5
YOLOv-Face2	16.3	1.0	17.4
YOLOv5-s	16.3	1.0	17.4

Through Table I and Table II, bold numbers denote highest performance. YOLOv8-n [11] demonstrates a well-balanced trade-off between accuracy and speed. It achieves a mean average precision of 0.669 (mAP50) and 0.366 (mAP50-95), outperforming YOLOv10-n in both precision (P) and recall (R) metrics while maintaining competitive performance against YOLOv-Face2. Although YOLOv-Face2 achieves slightly higher accuracy, its increased inference time makes it less ideal for real-time applications. YOLOv8-n, with its processing and inference times of just 0.1 ms and 0.6 ms per image, significantly surpasses models like YOLOv5-s and YOLOv-Face2 in speed. This balance of accuracy and efficiency makes YOLOv8-n particularly suitable for scenarios requiring rapid and precise visual tracking to enhance spatial audio generation.

In the results of YOLOv8-n, we obtain the center coordinates $(x_{\text{center}}, y_{\text{center}})$ of the detected bounding box (as shown in 3), which are normalized by default. To adapt these coordinates for audio processing, we first normalize $(x_{\text{center}}, y_{\text{center}})$ them into the range $[-1, 1]$.

$$\begin{aligned} \mathbf{X} &= 2x_{\text{center}} - 1, \\ \mathbf{Y} &= 1 - 2y_{\text{center}} \end{aligned} \tag{1}$$

In three-dimensional space and spatial audio processing, the Cartesian Coordinate System, defined by **X** (left and right),**Y** (up and down), and **Z** (front and back) coordinates.

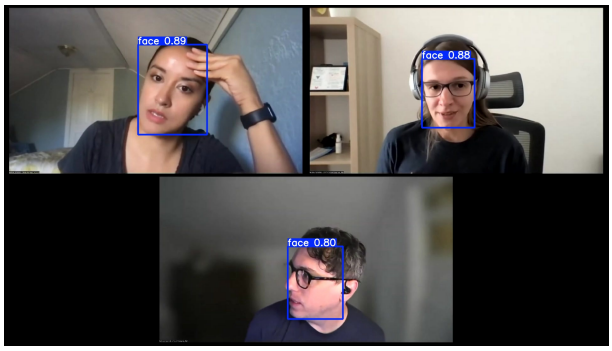
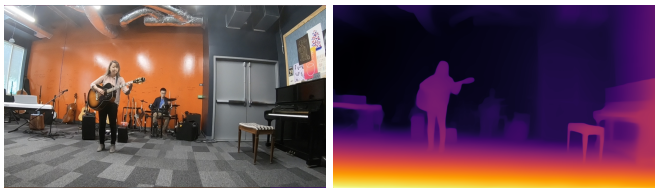


Fig. 3: YOLOv8-n Output: Object Detection with Bounding Box Predictions

Object detection models, such as YOLOv8-n, provide accurate **X** and **Y** coordinates in a 2D plane, enabling real-time detection and localization of individual speakers within an image frame. To extend this to 3D space, depth information (**Z**-coordinate) must be incorporated, offering a more realistic representation of the environment.

For this purpose, we employ depth estimation techniques to compute the **Z**-coordinate, representing the distance between the camera and each detected facial individual. Following the methodology outlined in [13], we utilize the pre-trained ViT-S model from the Depth Anything framework for robust monocular depth estimation. This model excels in handling diverse environments and arbitrary images, delivering precise depth measurements. We chose the Depth Anything ViT-S model for its demonstrated speed efficiency and accuracy in depth estimation tasks, which are crucial for accurately estimating the distance between the sound source and the camera (audience). An example of the results obtained using this model is illustrated in Fig. 4.



(a) Raw Image (b) Depth Anything

Fig. 4: Depth estimation model visualization.

In (2), g_{min} and g_{max} represents the range of gray scale value in the depth map, ranging from 0 to 255 [13]. The O represents the gray scale value calculated from the depth estimation at the location of the image corresponding to each facial detection in YOLOv8-n. d_{min} and d_{max} are the minimum and maximum self-defined values of the actual distance **Z** in the scene (in meters), set to range from 0.1 to 5.

$$\mathbf{Z} = d_{\text{max}} - (\text{O} - \text{g}_{\text{min}}) \cdot \frac{d_{\text{max}} - d_{\text{min}}}{\text{g}_{\text{max}} - \text{g}_{\text{min}}} \tag{2}$$

By combining YOLOv8-n's **X** and **Y** coordinate with

the output **Z** calculated from depth estimation model, our approach ensures a higher level of realism and accuracy in tracking and mapping the individual speakers' position in 3D space.

B. Audio Source Preparation

To prepare the audio input, we employ different source separation models depending on the scenario. For multiple speaker scenarios, we utilize Conv-TasNet [15], a convolutional time-domain audio separation network known for its efficiency and strong performance in speech separation tasks. For music scenarios, we adopt Demucs [16], which enables the separation of musical tracks such as vocals, drums, bass, and other instruments. All models used are pre-trained without additional fine-tuning. In cases involving more than two concurrent speakers, where the source separation quality becomes insufficient, we directly use pre-separated audio tracks provided by the dataset (Sec. III-A).

C. Facial Annotation

After source separation, we assign speaker identities to each separated audio track using annotations. This identity labeling enables alignment with visual information to support subsequent spatialization. It is important to note that our work primarily focuses on visual and spatial audio synchronization, rather than speaker diarization or audio-speaker identification association. Future studies may explore automatic speaker identification and alignment to enhance practical deployment.

D. Audio

To achieve accurate spatialization while maintaining high audio quality, we adopt two methods for audio processing. In the first method, we employ the SADIE II Database[2] (Subject: KEMAR) [17], which offers precise Head-Related Transfer Function (HRTF) measurements specifically designed for virtual environments. The spatial coordinates (**X**, **Y**, **Z**) from the visual cues are converted into azimuth and elevation angles. These angles are then used to select the corresponding HRTF files, which are convolved with the individual audio tracks.

In the second method, we implement a 3D audio positioning algorithm that enhances the spatial perception of stereo signals by simulating sound propagation in different directions. The algorithm comprises three main steps.

Left-Right Positioning: The algorithm adjusts the signal strength of the left and right channels to control the perceived horizontal position of the sound, effectively creating a stereo panning effect. The signal $\vec{S}$ represents the input audio as a 1-D mono track, which is split into two channels (left and right) by applying the following (3), where **X** is a horizontal positioning factor which is calculated from (1).

$$\begin{aligned} \text{Left channel} = x^L = \vec{S} \cdot \frac{1-\mathbf{X}}{2}, \\ \text{Right channel} = x^R = \vec{S} \cdot \frac{1+\mathbf{X}}{2} \end{aligned} \tag{3}$$

[2]https://www.york.ac.uk/sadie-project/database.html

Up-Down Positioning: The elevation is adjusted through frequency filtering, primarily by enhancing or attenuating high frequencies to simulate changes in the sound source's vertical angle.

In (4), $\mathcal{F}(f)$ represents the frequency adjustment factor, which modifies the energy distribution across frequencies. The **Y** is the output from the equation (1). The variable f denotes the frequency in Hz. $S(f)$ represents the original frequency spectrum of the stereo signal which is the output from (3), while $S'(f)$ is the adjusted spectrum after applying the frequency adjustment factor.

$$\begin{aligned} \mathcal{F}(f) &= 1 + \mathbf{Y} \cdot \left(\frac{f}{1000}\right)^{1.5}, \\ S'(f) &= S(f) \cdot \mathcal{F}(f) \end{aligned} \tag{4}$$

Front-Back Positioning: The algorithm simulates the proximity of the sound source by reducing the volume and adding reverberation. For distant sound sources, the volume is lower, and the reverberation effect is more pronounced.
The output signal $s_{\text{output}}(t)$ is computed as:

$$s_{\text{output}}(t) = \frac{\text{signal}(t)}{\mathbf{Z}} + \alpha \cdot \frac{\text{signal}(t - \Delta t_{\text{samples}})}{\mathbf{Z}} \tag{5}$$

where

$$\Delta t_{\text{samples}} = \frac{\mathbf{Z} \cdot f_s}{v} \tag{6}$$

and f_s is the sampling rate, $v = 343\,\text{m/s}$ is the speed of sound, and $\alpha = 0.3$ is the reverberation intensity factor. t is the time variable representing the sampling point of the audio signal. $signal(t)$: The input audio signal at time t. $signal(t - \Delta t_{\text{samples}})$: A delayed version of the signal.

III. Experiment

A. Ground Truth

To evaluate the effectiveness of the system, we utilized two datasets during the testing phase for performance comparison.

Speech: The audio stimulus was collected from the LibriSpeech dataset [18] or extracted from multiple-speakers video in YouTube. We designed three scenarios with different speakers from 2 tracks, 3 tracks and 5 tracks. Each audio clip is approximately 10 seconds long. Each scenario is edited, with its corresponding soundtracks redesigned and synchronized with the visuals. To align the audio with the visual spatial information and provide a reference for evaluation, we manually adjusted the audio panning using the digital audio platform REAPER. This process ensures that the audio accurately corresponds to the spatial positions of visual elements.

Music: We utilized the FAIR-Play dataset [8], which contains 1,871 binaural audio and video clips of musical performances totaling 5.2 hours, for comparative testing in our experiments.

TABLE III: Objective evaluation results for speech metrics. The numbers in bold denote the best performance.

Method	MOSNet			Nb_PESQ			PESQ			STOI		
Speaker Number	**2**	**3**	**5**	**2**	**3**	**5**	**2**	**3**	**5**	**2**	**3**	**5**
Mono2Binaural	2.814	3.447	3.022	3.706	3.634	3.078	3.423	2.215	2.089	0.976	0.912	0.898
PseudoBinaural	**2.954**	2.854	3.017	3.719	0.320	3.071	3.422	1.025	2.068	0.978	0.155	0.898
Our system (3D)	2.902	3.182	2.796	**4.019**	3.136	2.957	**4.198**	2.835	1.789	**0.981**	0.913	0.868
Our system (HRTF)	2.797	**3.475**	**3.040**	3.856	**3.718**	**3.379**	3.960	**2.275**	**2.558**	0.979	**0.921**	**0.912**

B. Test process

We compare our proposed system with the following baselines: Mono2 Binaural [8] and Pseudo Binaural [10]. During the evaluation, each test file was normalized to a loudness level of -23 LUFS, resampled to 16 kHz, and processed using STFT with a Hann window of 25 ms, a hop length of 10 ms, and an FFT size of 512. The audio files were stereo-channel with a bit depth of 16 bits.

C. Evaluation Metrics

To rigorously assess the performance of our proposed system, we conducted a comprehensive evaluation using multiple metrics that capture both perceptual and objective aspects of audio quality. Each system's output was compared with the ground truth reference audio (III-A), and the results were subsequently analyzed across systems. Specifically, we compared the audio predicted by our system (system-generated) against reference audio samples (collected from the dataset) to determine the fidelity and accuracy of our audio processing algorithm. The test methods include:

STFT Distance: The Euclidean distance between the complex spectrograms of the reference signal $\mathbf{x}$ and the predicted signal $\tilde{\mathbf{x}}$ for the left and right channels:

$$\mathcal{D}_{\text{STFT}} = \|\mathbf{X}^L - \tilde{\mathbf{X}}^L\|_2 + \|\mathbf{X}^R - \tilde{\mathbf{X}}^R\|_2 \tag{7}$$

where $\mathbf{X}^L$, $\mathbf{X}^R$, $\tilde{\mathbf{X}}^L$, and $\tilde{\mathbf{X}}^R$ denotes the complex-valued spectrograms of $\mathbf{x}^L$, $\mathbf{x}^R$, $\tilde{\mathbf{x}}^L$, and $\tilde{\mathbf{x}}^R$, respectively.

Envelope (ENV) Distance: Quantifies the Euclidean distance between the envelopes of ground-truth and predicted signals [19]. The envelope of the signal $x(t)$ is represented as $E[x(t)]$.

$$\begin{aligned}\mathcal{D}_{\text{ENV}} = &\|E[x^L(t)] - E[\tilde{x}^L(t)]\|_2 \\ &+\|E[x^R(t)] - E[\tilde{x}^R(t)]\|_2\end{aligned} \tag{8}$$

PESQ (Perceptual Evaluation of Speech Quality): Assesses perceptual audio quality, widely used in telecommunication applications [20]. **Nb** stands for Narrowband, referring to a frequency range of 300 Hz to 3400 Hz.

STOI (Short-Time Objective Intelligibility): Evaluates the intelligibility of speech signals [21].

MOSNet: Predicts the Mean Opinion Score (MOS) to estimate perceived audio quality [22].

IV. Result and discussion

Table III compares the performance of different systems across various speech metrics (MOSNet, NbPESQ, PESQ, and STOI) for 2, 3, and 5 audio tracks. Our proposed system includes two variants: HRTF based and 3D algorithmic approach. While the 3D approach achieves the highest scores in simpler scenarios (e.g., 2Track, with PESQ = 4.198 and NbPESQ = 4.019), the HRTF-based approach demonstrates greater robustness in more complex scenarios. Notably, for the 5Track setup, the HRTF-based approach outperforms the 3D approach in MOSNet (3.040 vs. 2.796) and STOI (0.912 vs. 0.868), highlighting its ability to maintain perceptual quality and intelligibility under challenging multi-speaker conditions.

Table IV compares the STFT and ENV distance metrics for different systems in the Speech and Fair-Play datasets, highlighting the performance distinctions. The baseline methods, Mono2 Binaural and Pseudo Binaural, achieve consistent results with relatively low STFT and ENV distances, but struggle to capture spatial cues effectively, as indicated by higher STFT values. Pseudo Binaural slightly outperforms Mono2 Binaural in ENV distance on the Fair-Play dataset but exhibits similar limitations on the Speech dataset. In contrast, the proposed system using HRTF achieves competitive ENV distances but the highest STFT values. This is due to the HRTF convolution introducing phase delay, which results in significant phase angle differences and consequently a large STFT distance. Meanwhile, the 3D approach demonstrates superior performance across all metrics and datasets, achieving the lowest STFT and ENV distances.

Furthermore, our system generally achieves lower STFT and ENV distances on the Fair-Play dataset compared to the Speech dataset. This discrepancy can be attributed to the

TABLE IV: Comparison of STFT and ENV distance metrics for different systems on Speech and Fair-Play datasets. Values in bold represent the best performance.

	Speech		Fair-Play	
	STFT	ENV	STFT	ENV
Mono2 Binaural	0.304	0.128	0.101	0.049
Pseudo Binaural	0.330	0.112	**0.093**	**0.048**
Our system (3D)	**0.220**	**0.094**	0.151	0.063
Our system (HRTF)	2.051	0.127	0.841	0.084

Fair-Play dataset being instrument-based rather than speaker-based, whereas our system relies on object detection in human faces. Additionally, the spatial positioning mismatch between instruments and human faces in the visual component introduces inaccuracies in spatial audio localization, further impacting the performance on the Fair-Play dataset.

These results confirm the robustness and effectiveness of our proposed system, particularly the 3D approach, in achieving perceptually accurate and spatially consistent audio synthesis across diverse datasets, underscoring its potential for real-world applications in spatial audio processing.

V. Limitations and Future Work

After the source separation stage, it is necessary to annotate the speaker identity for each separated audio track. This identity labeling facilitates alignment with visual information to support subsequent spatialization. Specifically, the labeling step addresses the problem of speaker diarization with visual cues—i.e., determining which audio stream corresponds to which visible speaker.

Speaker diarization is itself a complex task that warrants dedicated investigation [23]. While prior studies have established important foundations in this area [24] [25], the focus of this work is not on speaker diarization, but rather on leveraging visual cues to guide spatialization and enhance audio quality.

The diarization challenge will be addressed in future work. We aim to explore the integration of facial features (e.g., lip motion), spectral differences in audio, and synchronized visual-audio onset detection, combined with a novel unsupervised neural network. The goal is to achieve robust multi-speaker spatial audio generation without requiring manual annotations.

Furthermore, due to page limitations, we do not include subjective listening tests in this paper; however, they will be incorporated in future work to further deepen the analysis of the objective experimental results.

VI. Conclusion

In this paper, we proposed a spatial audio generation system based on visual cues. The system aimed at simplifying the transformation of mono audio into binaural audio in multi-speaker scenarios without relying on binaural dataset. The system integrates object detection, depth estimation, and audio spatialization. Extensive experimental evaluations demonstrate that our system outperforms existing spatial audio generation systems across various metrics. The results highlight significant improvements in spatial consistency between audio and visual components, enhanced speech quality, and robust performance in complex multi-speaker environments.

References

[1] J. Ni, H. Tang, S. T. Haque, Y. Yan, and A. H. Ngu, "A survey on multimodal wearable sensor-based human action recognition," *arXiv preprint arXiv:2404.15349*, 2024.

[2] F. Tao and C. Busso, "Aligning audiovisual features for audiovisual speech recognition," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 2018, pp. 1–6.

[3] H. Kim, L. Remaggi, and A. H. Philip J.B. Jackson, "Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360° images," *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2019.

[4] D. Li, T. R. Langlois, and C. Zheng, "Scene-aware audio for 360° videos," *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[5] C. Reynolds, M. Reed, and P. Hughes, "Analysis of a distributed processing model for spatialized audio conferences," in *2008 IEEE International Conference on Multimedia and Expo (ICME)*, 2008, pp. 461–464.

[6] S. Mehrotra, W. ge Chen, Z. Zhang, and P. A. Chou, "Realistic audio in immersive video conferencing," in *2011 IEEE International Conference on Multimedia and Expo (ICME)*, 2011, pp. 1–4.

[7] L. Wang, R. Sanchez-Matilla, and A. Cavallaro, "Audio-visual sensing from a quadcopter: dataset and baselines for source localization and sound enhancement," *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[8] R. Gao and K. Grauman, "2.5d visual sound," *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2019.

[9] Y.-B. Lin and Y.-C. F. Wang, "Exploiting audio-visual consistency with partial supervision for spatial audio generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2056–2063.

[10] X. Xu, H. Zhou, Z. Liu, B. Dai, X. Wang, and D. Lin, "Visually informed binaural audio generation without binaural audios," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2021.

[11] M. Sohan, T. Sai Ram, R. Reddy, and C. Venkata, "A review on yolov8 and its advancements," in *International Conference on Data Intelligence and Cognitive Informatics*. Springer, 2024, pp. 529–545.

[12] Y. Shuo, L. Ping, L. C. Change, and X. Tang, "Wider face: A face detection benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2024.

[14] J. Redmon, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016.

[15] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[16] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Demucs: Deep extractor for music sources with extra unlabeled data remixed," *arXiv preprint arXiv:1909.01174*, 2019.

[17] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A perceptual evaluation of individual and non-individual hrtfs: A case study of the sadie ii database," *Appl. Sci.*, vol. 8, no. 11, 2018.

[18] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia *et al.*, "Libritts: A corpus derived from librispeech for text-to-speech," *Interspeech*, 2019.

[19] P. Morgado, N. Nvasconcelos, T. Langlois, and O. Wang, "Self-supervised generation of spatial audio for 360 video," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[20] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends *et al.*, "Peaq-the itu standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc.*, vol. 48, no. 1/2, pp. 3–29, 2000.

[21] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[22] C. Lo, T.-Y. Hsiao, H. Kawai, and J.-H. Chou, "Mosnet: Deep learning based objective assessment for voice conversion," *Interspeech*, 2019.

[23] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[24] K. Li, F. Xie, H. Chen, K. Yuan, and X. Hu, "An audio-visual speech separation model inspired by cortico-thalamo-cortical circuits," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[25] D. Dimitriadis and P. Fousek, "Developing on-line speaker diarization system." in *Interspeech*, 2017, pp. 2739–2743.