Tragic Talkers: A Shakespearean Sound- and Light-Field Dataset for Audio-Visual Machine Learning Research

Davide Berghi*
CVSSP, University of Surrey
Guildford, U.K.
d.berghi@surrey.ac.uk

Marco Volino
CVSSP, University of Surrey
Guildford, U.K.
m.volino@surrey.ac.uk

Philip J. B. Jackson
CVSSP, University of Surrey
Guildford, U.K.
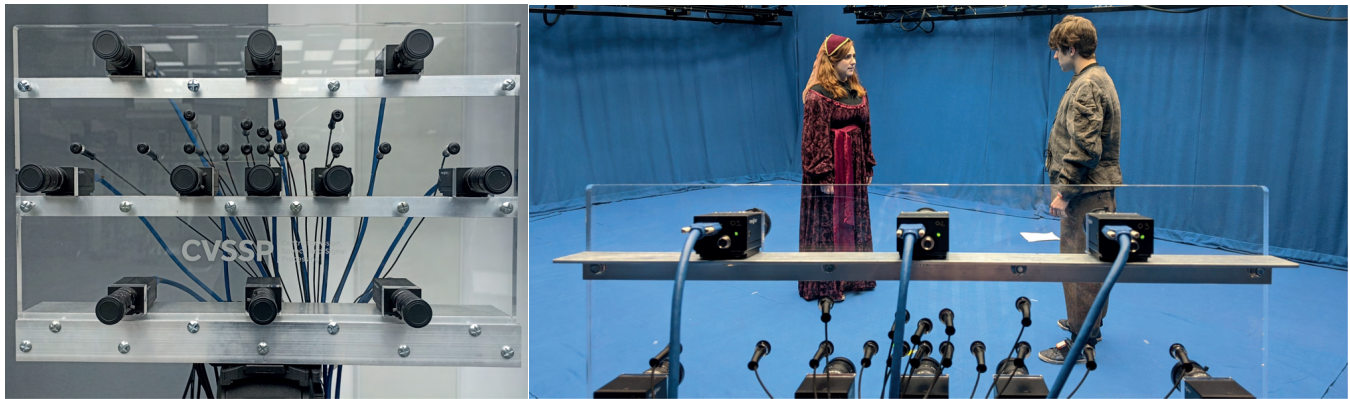p.jackson@surrey.ac.uk

Figure 1: On the left, a photo of one of the two AVA Rigs used to capture the Tragic Talkers dataset. On the right, a photo taken from behind one of the rigs during the dataset capture session.

ABSTRACT

3D audio-visual production aims to deliver immersive and interactive experiences to the consumer. Yet, faithfully reproducing real-world 3D scenes remains a challenging task. This is partly due to the lack of available datasets enabling audio-visual research in this direction. In most of the existing multi-view datasets, the accompanying audio is neglected. Similarly, datasets for spatial audio research primarily offer unimodal content, and when visual data is included, the quality is far from meeting the standard production needs. We present "Tragic Talkers", an audio-visual dataset consisting of excerpts from the "Romeo and Juliet" drama captured with microphone arrays and multiple co-located cameras for light-field video. Tragic Talkers provides ideal content for object-based media (OBM) production. It is designed to cover various conventional talking scenarios, such as monologues, two-people conversations, and interactions with considerable movement and occlusion, yielding 30 sequences captured from a total of 22 different points of view and two 16-element microphone arrays. Additionally, we provide voice activity labels, 2D face bounding boxes for each camera view, 2D pose detection keypoints, 3D tracking data of the mouth of the actors, and dialogue transcriptions. We believe the community will benefit from this dataset as it can assist multidisciplinary research. Possible uses of the dataset are discussed.

CCS CONCEPTS

• **Computing methodologies** → *Virtual reality*; **Scene understanding**; *Vision for robotics*; 3D imaging; **Tracking**; • **Hardware** → *Emerging tools and methodologies*.

KEYWORDS

dataset, multi-view, microphone array, active speaker detection, object-based media

1 INTRODUCTION

In 3D immersive production, there is little room for spatial misalignment between the audio and the visual displays. To provide realistic immersive experiences, accurate positioning of the audio sources with respect to the visual events that trigger them is therefore essential [Berghi et al. 2020; Stenzel and Jackson 2018]. With synthetic audio-visual assets, this is typically achieved through a manual process performed by the producer. The objects are located at the desired positions in the virtual environment enabling six degree-of-freedom (6-DoF) interaction of the end-user. However, to achieve the same result and faithfully reproduce real-world scenes, an object-based audio-visual representation is required [Coleman et al. 2018; Pike et al. 2016]. That is, object events must be extracted from the audio and visual streams and partnered with metadata describing specific attributes of the objects themselves, such as their content or positions in space. An approach to enable real-world object-based media (OBM) production for 6-DoF interaction is by extracting the visual foreground objects and three-dimensionally reconstructing them. This is possible through a multi-view data capture [Hartley and Zisserman 2004]. Meanwhile, the acoustic

Table 1: Summary of existing audio-visual datasets captured with a microphone array or binaural microphones for speech-related research. 'Col.': co-located platform; '# m': number of microphones in the microphone array; 'f_s': audio sampling frequency in kHz; '# c': number of cameras; 'Res.': video resolution; 'Cal.': camera calibration data; 'BG': chroma-uniform background; 'Act.': actors employed; 'int.': intrinsics-only; 'VA': voice activity annotations; 'pos.': position; 'transc.': transcripts

Dataset	Col.	# m	f_s	# c	Res.	Cal.	BG	Act.	Labels
AV16.3 [Lathoud et al. 2004]	✗	8(×2)	16	3	360×288	✓	✗	✗	VA, 3D mouth and head pos.[a]
CAVA [Arnaud et al. 2008]	✓	2	44.1	2+1	1024×768	✓	✗	✗	Corner points
AVDIAR [Gebru et al. 2018]	✓	6	48	2	1920×1200	✓	✗	✗	VA, 2D bbox
SSLR[b] [He et al. 2018]	✓	4	48	1	640×480	✓	✗	✗	VA, 3D tracking data
CAV3D [Qian et al. 2019]	✓	8	96	1	1024×768	✓	✗	✗	VA, 3D mouth pos.
EasyCom [Donley et al. 2021]	✓	6	48	1	1920×1080	int.	✗	✓	VA, head pose, 2D face+head bbox, transc.
Tragic Talkers	✓	16(×2)[c]	48	11(×2)	2448×2048	✓	✓	✓	VA, 2D face bbox, 3D mouth pos.[a], 2D pose keypoints[a], transc.

[a] Only a subset.
[b] SSLR is mostly an audio-only dataset. We refer to the subset named 'Human' that includes video recordings of human speakers.
[c] In addition to two 16-element microphone arrays, audio is recorded with two lavalier microphones and a 4-channel FOA mic too.

sound-field of the audio events must be captured and tracked over time. This is possible through multi-channel audio, for example, recorded with a microphone array [Brandstein and Ward 2001].

Producing such audio-visual content can be extremely costly as it requires expensive equipment and studio time. To the best of our knowledge, there are no available datasets with these characteristics. Existing datasets with multi-view video content usually neglect the audio component [Ma et al. 2020; Pumarola et al. 2019; Trumble et al. 2017]. On the other hand, datasets with multi-channel audio lack video content [Politis et al. 2021] or, when provided, the quality is rather low [He et al. 2018; Lathoud et al. 2004]. Therefore, we propose Tragic Talkers: a sound- and light-fields dataset captured with the aid of two co-located audio-visual sensing devices. Co-locating audio and visual sensors is particularly useful to maintain the same fixed coordinate frame across the two modalities. The dataset consists of drama scenes, therefore, it is suitable for OBM production applications or OBM research. The actors were also provided with dedicated close microphones to record high-quality voice signals useful for reference and association when beamforming algorithms are applied. Furthermore, we offer a range of curated labels to enable research in multiple domains, e.g., active speaker detection, speaker diarization, or body pose detection. A preliminary version of the Tragic Talkers dataset was used in our work [Berghi et al. 2021]. However, the previous version was smaller in size (20 sequences instead of 30) and was only labeled for 2D bounding boxes and voice activity.

The rest of this paper is organized as follows: In chapter 2 existing datasets for 3D multimedia content are presented and their limitations discussed; In chapter 3 the recording set-up employed to capture Tragic Talkers is described; In chapter 4 we motivate the design of the dataset and present the provided labels; In chapter

5 few key uses for the dataset are presented; Finally, in chapter 6 we conclude the paper. Tragic Talkers is available for download at https://cvssp.org/data/TragicTalkers/. The website provides ready-to-use PyTorch python scripts for data loading and audio-visual feature extraction.

2 BACKGROUND

Datasets for 3D audio-visual content production and research are important for a number of reasons. First of all, they represent a useful benchmark for evaluating new algorithms. Secondly, producing new multi-view and multi-channel audio-visual content can be extremely costly and time-consuming. Additionally, consistently-captured data can be employed in machine learning algorithms training. Existing datasets that offer 3D multimedia content are typically designed to serve a specific technical aspect. 3DPeople [Pumarola et al. 2019], CAPE [Ma et al. 2020], SIZER [Tiwari et al. 2020], and Deep Fashion3D [Heming et al. 2020] are examples of real and synthetic datasets for 3D clothing modeling. AVAMVG [López-Fernández et al. 2014] is a multi-view dataset for 3D gait recognition. TotalCapture [Trumble et al. 2017] provides multi-view video data for 3D pose estimation. However, none of the aforementioned datasets provide audio data. Volumetric datasets that also recorded the audio component are HUMAN4D [Chatzitofis et al. 2020], MHAD [Ofli et al. 2013], CMUPanoptic [Joo et al. 2015], and NAVVS [Stenzel et al. 2021], even though only MHAD and NAVVS released the audio recordings. Even in these rare exceptions, the audio sources are recorded with dedicated microphones and not with the intention of capturing the directional cues of the acoustic sound-field of the performance, as it would be possible with microphone arrays or binaural microphones. To the best of our
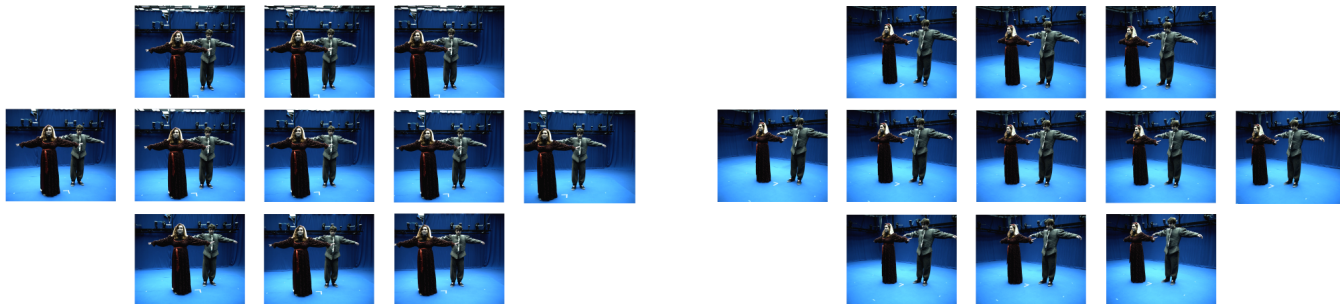
Figure 2: Same frame with actors in T-pose captured from the 11 cameras of AVA Rig 1 (right), and the 11 cameras of AVA Rig 2 (left).

knowledge, Tragic Talkers is the first dataset that simultaneously provides sound- and light-field video content.

In the literature, there exists a variety of datasets captured with microphone arrays for sound source localization and tracking. Some of them are summarized in [Roman and Bello 2021]. In these cases, however, they lack visual content. A good amount of multi-channel audio datasets that also include the visual modality are produced by the Inria group[1] for audio-visual tracking research, including the AVASM dataset [Deleforge et al. 2015], AVDIAR [Gebru et al. 2018], RAVEL [Alameda-Pineda et al. 2013], and CAVA [Arnaud et al. 2008]. Qian *et al.* [Qian et al. 2019] propose a summary of these and other datasets. However, in most cases, the resolution of the video stream is very low. Additionally, they are primarily recorded with the intent of testing speaker tracking algorithms. Therefore, the actors are asked to perform tasks such as counting, talking on top of each other, and in some cases improvising a conversation. In contrast, Tragic Talkers offers high-quality video content captured in a studio with the help of two student actors. The actors play famous lines from the tragedy of Romeo and Juliet and their costumes are chosen to resemble the clothing of the Italian Renaissance period. All these features representing the quality of performance and integrating stagecraft skill, method, and material, are typically referred to by the film industry with the term *production value*.

Similar to [Qian et al. 2019], in Table 1 we propose a summary of audio-visual datasets captured with microphone arrays for speech-related research. We omit some of the obsolete ones and include a few that have been released recently. In the table, a column has been reserved for the employment of actors. Excluding Tragic Talkers, EasyCom [Donley et al. 2021] is the only dataset that proposes an acting performance as it presents groups of people talking at a shammed restaurant table ordering food from an actor restaurant server that collects the preferences and describes the dishes. AV16.3 [Lathoud et al. 2004] is the only dataset of the list that was not captured from a co-located platform but with three cameras located in different positions of the room and their resolution is fairly low. In the other datasets, the visual component is either captured with a single camera [Donley et al. 2021; He et al. 2018; Qian et al. 2019] or with a single stereo pair [Arnaud et al. 2008; Gebru et al. 2018]. Tragic Talkers is the only dataset that is captured with two separate sets of 11 high resolution cameras for light-field video production.

Additionally, Tragic Talkers is captured against a chroma-consistent background facilitating silhouette extraction. Excluding EasyCom [Donley et al. 2021], all the datasets are quite small and therefore unsuitable for learning-based research. Tragic Talkers can be either seen as a 30-sequences dataset captured from 22 different points of view or, considering the two rigs separately, as a 60-sequences dataset captured with a single rig. Viewed this way, it offers 3.8 hours of total annotated video data with audio that is semantically, temporally and spatially coherent. Therefore, we believe it is of sufficient scale as to enable a range of machine-learning research.

3 CAPTURING SET-UP

Tragic Talkers was captured with the aid of two twin Audio-Visual Array (AVA) Rigs. Each AVA Rig is a custom device consisting of a 16-element microphone array and 11 cameras fixed on a flat perspex sheet supported by a tripod. Therefore, the AVA Rig represents a co-located multi-sensing platform. That is, each sensor is in a fixed relative position with respect to the other sensors of the array and the same coordinate frame is shared across the modalities. The array has size 0.6m×0.4m and is compact and lightweight enough to be mounted on a standard production tripod. The 11 cameras (Grasshopper3 USB3[2]) capture synchronized 2448×2048p video frames at 30 fps. They are positioned in three rows and are equally spaced from one another, with the exception of the central camera. The camera array has been designed to support 6-DoF content production for the limited volume of a seated user. Before the capture session, intrinsics and extrinsics of the camera arrays were calibrated using a chart [Hartley and Zisserman 2004; Zhang 1999]. Therefore, the two camera rigs share the same world reference coordinated with the origin being the central camera of AVA Rig 1. Audio is sampled at 48 kHz, 24 bits. The microphone array can be used for sound source localization. It has a horizontal aperture of 450 mm and a vertical aperture of only 40 mm. This results in higher resolution when localizing audio sources along the azimuth direction than elevation, which is consistent with human perception [Strybel and Fujimoto 2000]. The horizontal design of the microphone array is log-spaced to focus on broad frequency coverage from 500 Hz to 8 kHz, to better support the horizontal speech band resolution. After the capture session, audio recordings have been channel-wisely multiplied by gain factors to set their

---

root mean square (RMS) magnitude to a common value and achieve level calibration, as in [Vasudevan et al. 2020].

During the data capture, the two AVA Rigs were placed side by side at a distance of about 2m from one another to provide alternative view points and emulate a plausible camera arrangement in a standard video production set, generating an angle of about 30° to 50° between the rigs and each subject. In addition to the acoustic sound-field captured by the microphone arrays, actors were accoutered with two lavalier microphones to record high-quality voice signals and a first-order ambisonic room microphone (Soundfield DSF1[3]) useful for reverb estimation was placed above the capturing volume. A clapper-board was employed at the beginning and at the end of each take to enable manual alignment of the audio and visual streams. The actors were asked to start and conclude their performance with the T-pose. The entire dataset was captured against a blue background to facilitate silhouette extraction by chroma-keying and 3D reconstruction using camera calibration data. Background images of the empty studio captured with the AVA Rigs are provided.

4 DESIGN OF DATASET

The Tragic Talkers dataset is designed to suit a variety of research areas including, but not limited to, light-field and sound-field processing. The intention to generate high quality content for next-generation 3D audio-visual production led to the decision of employing semi-professional actors with predefined scripts and selected costumes. Therefore, we chose to capture excerpts from Shakespeare's Romeo and Juliet drama, hence the name 'Tragic Talkers', where student actors played the role of the two key characters. In the rest of this paper, we will refer to 'Juliet' as the actress who interpreted Juliet and, similarly, to 'Romeo' as the actor who interpreted Romeo. Tragic Talkers consists of 30 sequences ranging from about 13 to 43 seconds in length (21 seconds on average), yielding over 400K frames (3.8 hours) of video data considering the 22 separate views. In order to produce a diverse set of talking conditions useful to speech-related research, three different types of scenes were recorded: *monologues* with single-character oration performance to provide male-only and female-only speech content; *conversations* with the two actors standing side-by-side exchanging lines from the famous tragedy; and *interactive* scenes where the actors were asked to move within the capture volume while talking, occluding each other and improvising their conversation. The last provides more challenging content for applications such as 2D and 3D audio-visual tracking. Although the dataset does not include scenarios with multiple simultaneous voices, the nature of the recordings allows synthetic creation of audio-visual cocktail parties by combining speech sequences: the audio is mixed, while the respective actors' video extracted from the background using silhouettes is overlaid to create augmented multi-talker sequences.

4.1 Labels

Along with the audio-visual data, we provide an extensive set of labels useful to test a variety of algorithms. Here, we summarize the available labels and explain how they have been generated. First of all, we have manually produced the voice activity (VA) labels

---

Table 2: Speech activity over the dataset's 30 sequences.

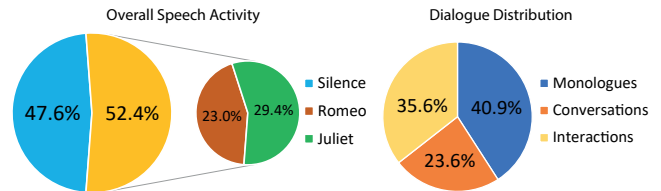| Partition	Duration	Active	Silent	Romeo	Juliet
Development set	519s	271s 52%	248s 48%	116s 43%	155s 57%
Test set	121s	64s 53%	57s 47%	31s 48%	33s 52%
Overall	640s	335s 52%	305s 48%	147s 44%	188s 56%

Figure 3: On the left, a pie chart showing the overall distribution of speech and silence across the dataset and how the speech content is distributed between Romeo and Juliet. On the right, the distribution of monologues, conversations, and interactions.
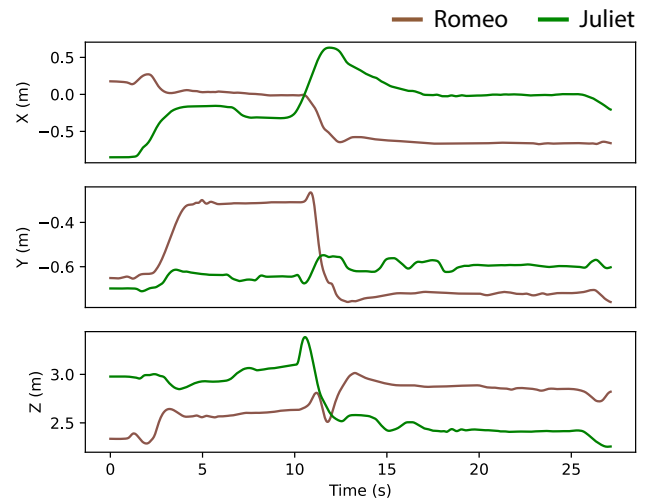

Figure 4: 3D mouth tracking results for Romeo and Juliet in the sequence 'interactive1_t2' of the test set.

with actor ID association. In order to achieve that, we employed the software Adobe Audition[4] to manually find the onset and offset of each speech segment as, alongside the audio tracks, it also provides the visual reference of the spectrograms. Table 2 reports the distribution between speech and silence across the dataset, as well as the distribution of Romeo's speech and Juliet's speech. Secondly, we provide the sequence transcriptions with the respective speaker

---

[3]https://www.soundfield.com/#/products/dsf1

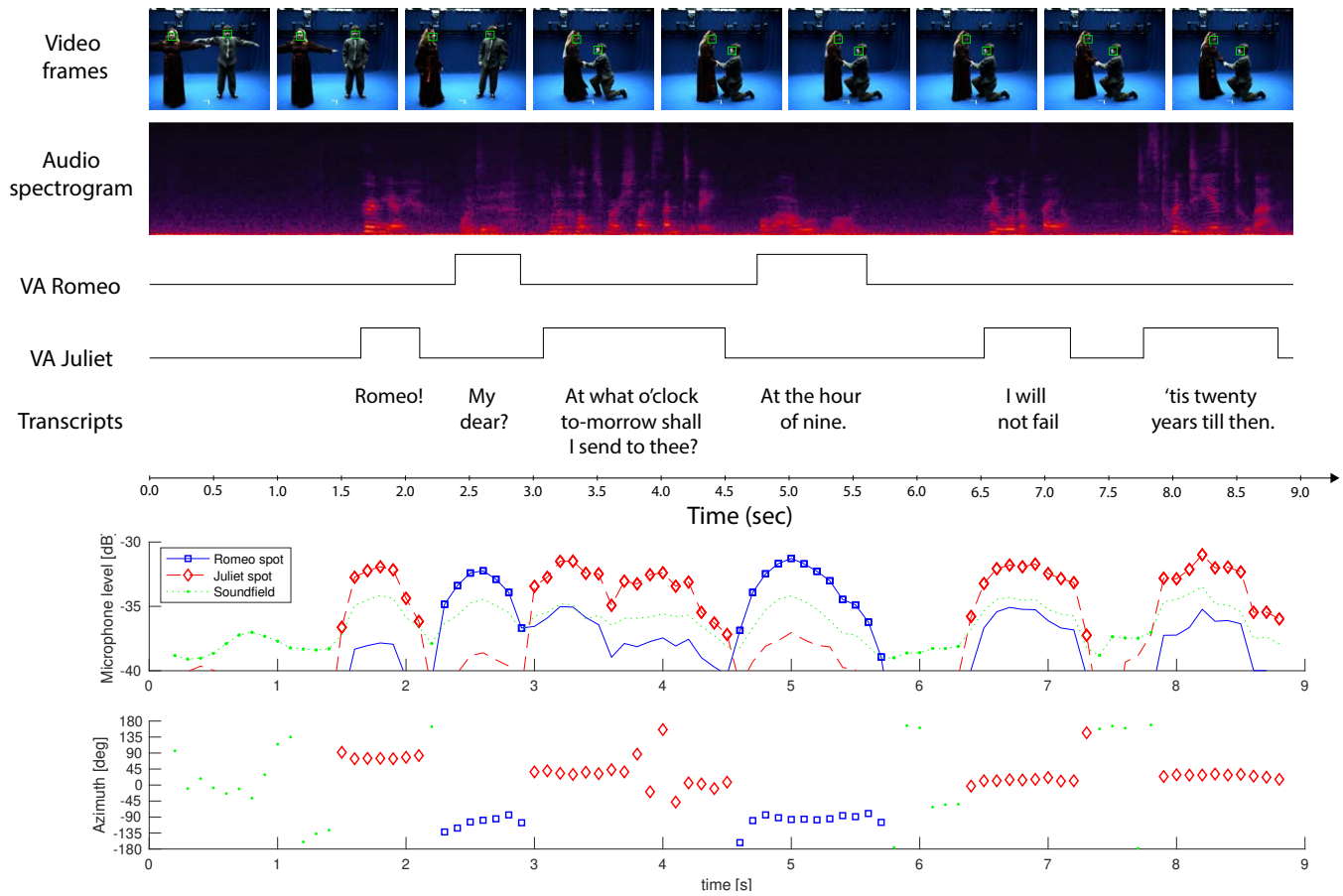[4]https://www.adobe.com/products/audition.html

Figure 5: Audio-visual content of an excerpt from the sequence 'interactive1_t2' depicted by video frames overlaid with detected-face bounding boxes (reported at 1 fps) and audio spectrogram. Alongside are voice activity (VA) labels for Romeo and Juliet, and the speech transcriptions over time. The middle axes below show normalized sound levels at spot microphones worn by Romeo and Juliet, which correspond with the voice activity labels above, and the soundfield microphone mounted in the room. The bottom axes illustrate the active speakers' azimuth angle in relation to the soundfield microphone, exploiting the directivity of the first-order-ambisonic audio signals.

ID. To generate them, we employed Otter.ai[5] Then, we manually adjusted the incorrect transcriptions.

To enable speaker localization and tracking, 2D face bounding boxes for each camera view are provided. Face detection was performed on each video frame with S3FD [Zhang et al. 2017]. When the face detector failed, for instance when the actors are not facing the camera, a bounding box was drawn by a manual educated guess. In this way, the actors are continuously tracked in the 2D domain throughout the entire duration of the sequence. Face tracks were generated from the per-frame detections based on intersection over union (IoU) across adjacent frames. Then, the corners of the bounding boxes have been interpolated over time to provide smooth temporal coherence. In addition to that, we selected 5 sequences from the dataset to constitute a test set. In order to be heterogeneous and representative of the dataset, the test set includes two monologues, one from Romeo and one from Juliet,

one conversation, and two interactive sequences. The test set represents a useful benchmark for learning-based algorithms and it was labeled for 3D mouth position. To generate the 3D mouth labels, we employed OpenPose [Cao et al. 2019] to detect 2D skeletal keypoints of the two actors on the 2D video frames. In order to provide spatial and temporal coherence to accurately track the 3D location of the mouth joint, the 2D detections have been sorted as described in [Malleson et al. 2020]. The sorted 2D detections are combined with confidences to remove the impact of potentially unreliable detections and estimate the 3D locations via camera triangulation. 2D skeleton keypoints detected with OpenPose are provided. Examples of 3D mouth detections over time from one of the interactive sequences are reported in Figure 4.

5 USES OF THE DATASET

Tragic Talkers can be employed for multiple applications. Here we propose some of the key use cases for this dataset. The availability

Figure 6: Examples of possible uses of the dataset. In the top picture, it is proposed an example of audio-visual active speaker tracking application, where the face of Juliet is detected by a face detector while the voice of Romeo is correctly located by the microphone array (yellow asterisk), even though his face is not visible. Below, an example of a partial 3D reconstruction of Romeo and Juliet positioned in a background of our choice.

of 2D and 3D tracking data clearly suggests the employment of this dataset in speaker localization and tracking applications. This can either be addressed with audio-only algorithms, for instance by applying beamforming [Blanco Galindo et al. 2020] or SRP-PHAT [Do et al. 2007], or with audio-visual solutions, as in [Qian et al. 2019]. Tragic Talkers enables object-based spatial audio production through the generation of tracking metadata and automatic association with the high-quality signals from the additional microphones provided. For instance, [Mohd Izhar et al. 2020] proposed to generate automatic 3D spatial metadata with an audio-visual tracker and utilize it to drive a spatial beamformer to steer the listening directivity of the microphone array towards the predicted positions of the audio sources. The beamformed audio signals are

then automatically associated to, and replaced by, the high-quality recording spatialized at the predicted positions.

The large amount of views with which the dataset was captured enables learning-based audio-visual research for 2D active speaker detection with multi-channel audio signals [Berghi et al. 2021; Donley et al. 2021]. In fact, over 3.8 hours of monocular video content can be generated from the 30 sequences provided, all accompanied by multi-channel audio recordings. For instance, in our earlier work [Berghi et al. 2021] where a preliminary version of the Tragic Talkers dataset was employed, we adopted a student-teacher learning pipeline. We proposed to generate pseudo-labels from the video sequences with a pre-trained active speaker detector (teacher) and used them to train a multi-channel audio network (student) to detect and regress the position of the speakers directly in the video frames. The availability of the dialogue transcriptions also allows the speaker diarization task to be performed [Gebru et al. 2018]. The light-field camera array on each of the two AVA Rigs allows 3D dynamic foreground character reconstruction for immersive production. For example, the central camera of the AVA Rigs can be employed as a reference view for an operator during the editing and compositing of the experience. The surrounding cameras can be used for automatic partial 3D reconstruction of Romeo or Juliet. The size of the aperture of the camera array enables 6-DoF movement in the limited volume of a seated user. The partially reconstructed foreground avatars can then be inserted into a virtual environment. An example of 3D reconstruction of Romeo and Juliet inserted in a background location of our choice is proposed in the lower image of Figure 6. Furthermore, the 3D avatars can be accompanied by their respective voices, spatialized as previously described, to produce an object-based audio-visual immersive experience, as suggested in [Schweiger et al. 2022].

The bi-modal multi-view content offered in Tragic Talkers is optimal to evaluate rendering algorithms too. In the context of audio-visual text-to-video synthesis (see, for instance, Synthesia[6]), Tragic Talkers can be employed as an evaluation set to test audio-visual speech rendering algorithms from different viewpoints. This could study different view point interpolation or replace one character with an avatar, and supports formal subjective experiments for perceptual quality evaluation. Another potential use of the data involves the construction of synthetic datasets, for instance, by replacing the background with synthetic locations [Caliskan et al. 2020], or combining the sequences to create different combinations of multi-talker scenarios, which can be useful as part of a data augmentation strategy. In addition to these tasks, we envisage a contribution to fundamental audio-visual machine learning may be explored through the use of these data.

6 CONCLUSION

This paper introduces the Tragic Talkers dataset, the first sound-field and light-field dataset, produced with drama students performing excerpts from Shakespeare's famous tragedy, Romeo and Juliet. Tragic Talkers was captured with the aid of two twin AVA Rigs, each consisting of an 11-element light-field camera array and a 16-element microphone array. High-quality speech data recorded

---

[6]https://www.synthesia.io/

with two lavalier spot microphones as well as a first-order ambisonics room microphone are also provided. The dataset is labeled for voice activity, 2D face bounding boxes, 3D mouth coordinates, and dialogue transcriptions. We discussed potential uses of the dataset, including 2D or 3D speaker tracking, object-based media production, and light-field 3D reconstruction for 6-DoF interaction.

In future work, we aim to use Tragic Talkers to investigate the effect of different temporal, spatial and semantic aspects of audio-visual coherence [Afouras et al. 2022; Arandjelovic and Zisserman 2017].

ACKNOWLEDGMENTS

REFERENCES

Triantafyllos Afouras, Yuki M. Asano, Francois Fagan, Andrea Vedaldi, and Florian Metze. 2022. Self-Supervised Object Detection From Audio-Visual Correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10575–10586.

Xavier Alameda-Pineda, Jordi Sanchez-Riera, Johannes Wienke, Vojtech Franc, Jan Cech, Kaustubh Kulkarni, Antoine Deleforge, and Radu P. Horaud. 2013. RAVEL: An Annotated Corpus for Training Robots with Audiovisual Abilities. *Journal on Multimodal User Interfaces* 7, 1-2 (2013), 79–91. http://hal.inria.fr/hal-00720734/en

R. Arandjelovic and A. Zisserman. 2017. Look, Listen and Learn. In *IEEE/CVF Inter. Conf. on Computer Vision*. 609–617.

Elise Arnaud, Heidi Christensen, Yan-Chen Lu, Jon Barker, Vasil Khalidov, Miles Hansard, Bertrand Holveck, Hervé Mathieu, Ramya Narasimha, Elise Taillant, Florence Forbes, and Radu Horaud. 2008. The CAVA Corpus: Synchronised Stereoscopic and Binaural Datasets with Head Movements. In *Proceedings of the 10th International Conference on Multimodal Interfaces* (Chania, Crete, Greece) *(ICMI '08)*. Association for Computing Machinery, New York, NY, USA, 109–116. https://doi.org/10.1145/1452392.1452414

Davide Berghi, Adrian Hilton, and Philip J. B. Jackson. 2021. Visually Supervised Speaker Detection and Localization via Microphone Array. In *IEEE 23rd Inter. Workshop on Multimedia Signal Processing*.

Davide Berghi, Hanne Stenzel, Marco Volino, Adrian Hilton, and Philip J. B. Jackson. 2020. Audio-Visual Spatial Alignment Requirements of Central and Peripheral Object Events. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 666–667. https://doi.org/10.1109/VRW50115.2020.00184

Miguel Blanco Galindo, Philip Coleman, and Philip J. B. Jackson. 2020. Microphone array geometries for horizontal spatial audio object capture with beamforming. *J. of the Audio Engineering Society* 68, 5 (2020), 324–337. https://doi.org/10.17743/jaes.2020.0025

M. Brandstein and D. Ward. 2001. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer. https://books.google.co.uk/books?id=nND6ObXSNoEC

Akin Caliskan, Armin Mustafa, Evren Imre, and Adrian Hilton. 2020. Multi-view Consistency Loss for Improved Single-Image 3D Reconstruction of Clothed People. In *Asian Conference on Computer Vision*. Springer International Publishing, Cham, 71–88.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Real-time Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

Anargyros Chatzitofis, Leonidas Saroglou, Prodromos Boutis, Petros Drakoulis, Nikolaos Zioulis, Shishir Subramanyam, Bart Kevelham, Caecilia Charbonnier, Pablo Cesar, Dimitrios Zarpalas, Stefanos Kollias, and Petros Daras. 2020. HUMAN4D: A Human-Centric Multimodal Dataset for Motions and Immersive Media. *IEEE Access* 8 (2020), 176241–176262. https://doi.org/10.1109/ACCESS.2020.3026276

Philip Coleman, Andreas Franck, Jon Francombe, Qingju Liu, Teofilo de Campos, Richard J. Hughes, Dylan Menzies, Marcos F. Simón Gálvez, Yan Tang, James Woodcock, Philip J. B. Jackson, Frank Melchior, Chris Pike, Filippo Maria Fazi, Trevor J. Cox, and Adrian Hilton. 2018. An Audio-Visual System for Object-Based Audio: From Recording to Listening. *IEEE Transactions on Multimedia* 20, 8 (2018), 1919–1931. https://doi.org/10.1109/TMM.2018.2794780

Antoine Deleforge, Radu Horaud, Yoav Y. Schechner, and Laurent Girin. 2015. Co-Localization of Audio Sources in Images Using Binaural Features and Locally-Linear Regression. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 23, 4 (apr 2015), 718–731. https://doi.org/10.1109/TASLP.2015.2405475

Hoang Do, Harvey F. Silverman, and Ying Yu. 2007. A Real-Time SRP-PHAT Source Location Implementation using Stochastic Region Contraction(SRC) on a Large-Aperture Microphone Array. In *IEEE Inter. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1. I–121–I–124. https://doi.org/10.1109/ICASSP.2007.366631

Jacob Donley, Vladimir Tourbabin, Jung-Suk Lee, Mark Broyles, Hao Jiang, Jie Shen, Maja Pantic, Vamsi Krishna Ithapu, and Ravish Mehra. 2021. EasyCom: An Augmented Reality Dataset to Support Algorithms for Easy Communication in Noisy Environments. (2021). arXiv:2107.04174 [cs.SD]

I. D. Gebru, S. Ba, X. Li, and R. Horaud. 2018. Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusion. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 40, 5 (2018), 1086–1099.

R. I. Hartley and A. Zisserman. 2004. *Multiple View Geometry in Computer Vision* (second ed.). Cambridge University Press, ISBN: 0521540518.

Weipeng He, Petr Motlicek, and Jean-Marc Odobez. 2018. Deep Neural Networks for Multiple Speaker Detection and Localization. In *IEEE Inter. Conf. on Robotics and Automation*. 74–79.

Zhu Heming, Cao Yu, Jin Hang, Chen Weikai, Du Dong, Wang Zhangye, Cui Shuguang, and Han Xiaoguang. 2020. Deep Fashion3D: A Dataset and Benchmark for 3D Garment Reconstruction from Single Images. In *Computer Vision – ECCV 2020*. Springer International Publishing, 512–530.

Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2015. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 3334–3342. https://doi.org/10.1109/ICCV.2015.381

Guillaume Lathoud, Jean-Marc Odobez, and Daniel Gatica-Perez. 2004. AV16.3: An Audio-Visual Corpus for Speaker Localization and Tracking. In *Proc. of the Machine Learning for Multimodal Interaction Workshop*. 182–185. https://doi.org/10.1007/978-3-540-30568-2_16

David López-Fernández, Francisco J. Madrid-Cuevas, Ángel Carmona-Poyato, Manuel J. Marín-Jiménez, and Rafael Muñoz Salinas. 2014. The AVA Multi-View Dataset for Gait Recognition. In *Activity Monitoring by Multiple Distributed Sensing*, Pier Luigi Mazzeo, Paolo Spagnolo, and Thomas B. Moeslund (Eds.). Springer International Publishing, 26–39. https://doi.org/10.1007/978-3-319-13323-2_3

Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. 2020. Learning to Dress 3D People in Generative Clothing. In *Computer Vision and Pattern Recognition (CVPR)*.

Charles Malleson, John Collomosse, and Adrian Hilton. 2020. Real-Time Multi-Person Motion Capture from Multi-View Video and IMUs. *Int. J. Comput. Vision* 128, 6 (jun 2020), 1594–1611. https://doi.org/10.1007/s11263-019-01270-5

Mohd Azri Mohd Izhar, Marco Volino, Adrian Hilton, and Philip J. B. Jackson. 2020. Tracking Sound Sources for Object-based Spatial Audio in 3D Audio-visual Production. In *Forum Acusticum*. 2051–2058. https://doi.org/10.48465/fa.2020.0884

Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. 2013. Berkeley MHAD: A comprehensive Multimodal Human Action Database. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*. 53–60. https://doi.org/10.1109/WACV.2013.6474999

Chris Pike, Richard Taylor, Tom Parnell, and Frank Melchior. 2016. Object-based 3D Audio Production for Virtual Reality using the Audio Definition Model. *Journal of the Audio Engineering Society* (september 2016).

Archontis Politis, Sharath Adavanne, Daniel Krause, Antoine Deleforge, Prerak Srivastava, and Tuomas Virtanen. 2021. A Dataset of Dynamic Reverberant Sound Scenes with Directional Interferers for Sound Event Localization and Detection. In *Proc. of the 6th Detection and Classification of Acoustic Scenes and Events Workshop*. 125–129.

Albert Pumarola, Jordi Sanchez, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2019. 3DPeople: Modeling the Geometry of Dressed Humans. In *International Conference on Computer Vision (ICCV)*.

X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro. 2019. Multi-Speaker Tracking From an Audio–Visual Sensing Device. *IEEE Trans. on Multimedia* 21, 10 (2019), 2576–2588.

Iran R Roman and Juan Pablo Bello. 2021. Micarraylib: Software for Reproducible Aggregation, Standardization, and Signal Processing of Microphone Array Datasets. In *Proc. of the 6th Detection and Classification of Acoustic Scenes and Events Workshop*.

Florian Schweiger, Chris Pike, Tom Nixon, Matt Firth, Bruce Weir, Paul Golds, Marco Volino, Craig Cieciura, Mohd Izhar, Nick Graham-Rack, Philip J. B. Jackson, and Alex Ang. 2022. Tools for 6-DoF Immersive Audio-visual Content Capture and Production. In *IBC*. https://www.bbc.co.uk/rd/publications/tools-six-depth-of-field-immersive-audiovisual-content-capture-production

Hanne Stenzel, Davide Berghi, Marco Volino, and Philip J.B. Jackson. 2021. Naturalistic audio-visual volumetric sequences dataset of sounding actions for six degree-of-freedom interaction. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 637–638. https://doi.org/10.1109/VRW52623.2021.00201

Hanne Stenzel and Philip J. B. Jackson. 2018. Perceptual thresholds of audio-visual spatial coherence for a variety of audio-visual objects. In *Conf. on Audio for Virtual and Augmented Reality*.

T. Strybel and K. Fujimoto. 2000. Minimum audible angles in the horizontal and vertical planes: Effects of stimulus onset asynchrony and burst duration. *The J. of the Acoustical Society of America* 108 6 (2000), 3092–5.

Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. 2020. SIZER: A Dataset and Model for Parsing 3D Clothing and Learning Size Sensitive 3D Clothing. In *European Conference on Computer Vision (ECCV)*. Springer.

Matt Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. 2017. Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors. In *2017 British Machine Vision Conference (BMVC)*.

Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. 2020. Semantic Object Prediction and Spatial Sound Super-Resolution with Binaural Sounds. In *The European Conf. on Computer Vision*. 638–655. https://doi.org/10.1007/978-3-030-58548-8_37

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and S. Li. 2017. S3FD: Single Shot Scale-Invariant Face Detector. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 192–201.

Zhengyou Zhang. 1999. Flexible camera calibration by viewing a plane from unknown orientations. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Vol. 1. 666–673 vol.1. https://doi.org/10.1109/ICCV.1999.791289