Enhancing XR Theatre with Remote Scent Delivery Using Audio Speech Recognition and Convolutional Neural Network-Based Scene Detection

Anderson Augusto Simiscuka
Dublin City University
Dublin, Ireland
andersonaugusto.simiscuka@dcu.ie

Matis Picoreau
Dublin City University
Dublin, Ireland
matis.picoreau2@mail.dcu.ie

Gabriel-Miro Muntean
Dublin City University
Dublin, Ireland
gabriel.muntean@dcu.ie

Gianluca Fadda
University of Cagliari
Cagliari, Italy
gianluca.fadda@unica.it

Maurizio Murroni
University of Cagliari
Cagliari, Italy
maurizio.murroni@unica.it

Mario Montagud
i2CAT Foundation
Rome, Italy
mario.montagud@i2cat.net

Massimo Mancini
Sardegna Teatro
Cagliari, Italy
massimo@sardegnateatro.it

Marco Carli
Università degli Studi Roma TRE
Rome, Italy
marco.carli@uniroma3.it

Federica Battisti
Università degli Studi di Padova
Padua, Italy
federica.battisti@unipd.it

Vlad Popescu
Transilvania University of Brașov
Brașov, Romania
vlad.popescu@unitbv.ro

***Abstract*—This paper presents a multi-sensory solution designed to enhance the immersive experience of remote audiences for theatrical plays. Using the *Macbettu* play as a case study, the approach synchronises olfactory and visual stimuli for remote viewers. The solution considers two approaches for generating scents: audio detection that identify the scenes within the play that require specific scents, based on automatic speech recognition (ASR), and convolutional neural networks (CNN) for scene recognition triggering the scent dispensers. When the key scenes are detected during the live play, via network communications, olfaction dispensers release the corresponding scent for remote users viewing the play on VR headsets. Real-time communication is managed through an MQTT-based approach. The paper also compares user feedback for CNN-based recognition versus the audio-based approach, focussing on scent delay and user perceived Quality of Experience (QoE) of the generated remote scents in multiple distances from a scent source. This work illustrates how scent-enhanced XR technologies can be integrated into virtual theatre experiences to engage remote audiences.**

***Index Terms*—Extended Reality, Olfaction, CNNs, Audio Speech Recognition, Multi-Sensory Experiences, Theatre Technology**

I. Introduction

Theatre is a powerful medium for storytelling, yet it remains a primarily in-person visual and auditory experience. This limitation can hinder the ability to fully immerse remote audiences. Advances in Extended Reality (XR) and communications technologies provide new opportunities to replicate theatrical experiences for remote audiences, even in live performance settings [1], [2]. Recent innovations in multi-sensory systems, particularly olfactory interfaces, enable the integration of scents into virtual environments, creating more immersive and engaging experiences [3], [4].

This work focuses on the play *Macbettu*, a Sardinian theatrical adaptation of Shakespeare's *Macbeth*, to explore multi-sensory audience engagement. The play is produced by Sardegna Teatro in Cagliari, Italy. During one dramatic scene, ashes and smoke are present (see Fig. 1). In another scene, red wine is thrown onstage (see Fig. 2). These scenes with real scent sources create a more immersive theatrical experience. The play is being adapted as a hybrid live theatre performance with remote audiences experiencing it through XR headsets, viewing reconstructed holographic point clouds in real-time.

By combining scene detection, network communication, and scent dispensers, it is possible to deliver synchronised olfactory stimuli to those remote audiences experiencing the play remotely through the headsets. To identify scenes where wine and ash scents are present in the in-person performance, this paper introduces an audio-based cue detection method and compares it with a convolutional neural network (CNN) approach for scene recognition. Both methods generate scene labels and timestamps, which are uploaded to a cloud broker and used to trigger the release of the corresponding scents via olfactory dispensers for remote users, also considering their distances to the scent source in the 3D space. This approach is strongly related to the Internet of Musical Things paradigm, which integrates sensors, wearables, and interactive multi-sensory performance environments [5].

In this paper, we propose and test a solution that introduces:

- Automatic Speech Recognition (ASR) for real-time scene detection.
- CNNs (i.e. MobileNetV3 and ResNet-18) for visual scene recognition.
- MQTT-based communication for low-latency interaction between the theatre stage and remote users.
- Comparative evaluation of audio-based and CNN-based approaches, and the impact of user distance to scent

source.

The remainder of this paper is organised as follows. Section II outlines related work. Section III provides an overview of the play Macbettu. Section IV details the solution architecture and methods for scene recognition based on visual and audio cues. Section V presents the accuracy of scene detection and the results of perceptual tests evaluating viewers' Quality of Experience (QoE) in relation to their distance from the scent source. Finally, Section VI concludes the paper and discusses directions for future research.

II. Related Works

Multi-sensory technologies have been increasingly applied to enhance human-computer interaction. Olfactory displays, integrated into virtual environments, have shown promise in improving immersion. Yan et al. explored VR olfactory interfaces, showing significant gains in user engagement and realism [6]. Similarly, Hirata and Suzuki developed multi-sensory systems for remote interaction, highlighting the importance of synchronising sensory feedback in real time [7].

XR technologies are increasingly used in theatre to create immersive and interactive experiences [1]. These range from head-mounted VR performances to live productions blending AR and multi-sensory effects [8]. Notable examples include *Sleep No More*, *The Under Presents*, and XR-enhanced productions by the Royal National Theatre and Royal Shakespeare Company. Other initiatives include the Extended Reality Theatre (XRT) by Staatstheater Nürnberg and the VR opera *Out of the Ordinary/As an nGnách*, developed under the EU-funded TRACTION project [9].

XR in live performances also enhances accessibility via features like audio description and tactile feedback, and supports multi-sensory elements such as scent or motion to improve immersion [10], [11]. Integrating scent into VR presents challenges, especially regarding bandwidth and synchronisation with other sensory modalities [12]. Technologies such as WebXR support 360° streaming with motion tracking and olfactory integration [13], [14], [15]. In [16] the authors employed music information retrieval methods to extract in real-time audio cues to control a haptic device during a multisensory performance.

Studies report that over 85% of mulsemedia VR applications improve engagement and satisfaction [17], [18], though olfactory feedback remains underutilised. Network constraints and compression can reduce QoE in remote VR setups [19], [20], [21].

The use of CNNs in scene recognition has advanced significantly. While some efforts demonstrated ResNet-18 offers good accuracy for scene detection for scent generation, the authors did not explore real-time or near real-time options [22], [23]. Szegedy et al. introduced GoogLeNet for efficient image recognition, which paved the way for lightweight models like MobileNetV3 [24]. MobileNetV3's efficiency and accuracy make it well-suited for real-time or resource-constrained environments, such as live theatre.

Fig. 1. Character in an environment covered in ash and smoke

Fig. 2. Character pouring wine in Macbettu

Lightweight communication protocols such as MQTT have proven essential for real-time systems based on IoT [25]. The low overhead of MQTT ensures minimal latency, which is critical for synchronising remote sensory feedback in XR applications.

III. Macbettu

Macbettu is a theatrical production by Sardinian director Alessandro Serra. This unique reinterpretation of Shakespeare's Macbeth draws deeply from Sardinian cultural traditions. The production reimagines the mystical and ominous world of Macbeth through the lens of local carnival rites, folklore, and archetypes. Traditional Sardinian music and sound design further enrich the sensory landscape, creating a deeply immersive theatrical experience.

The integration of scent technology into the XR adaptation of Macbettu demonstrates the potential of multi-sensory media innovations. Two key scenes trigger scents for remote users, a scene with the scent of ash (as seen in Fig. 1) and a scene where a character pours wine, shown in Fig. 2.

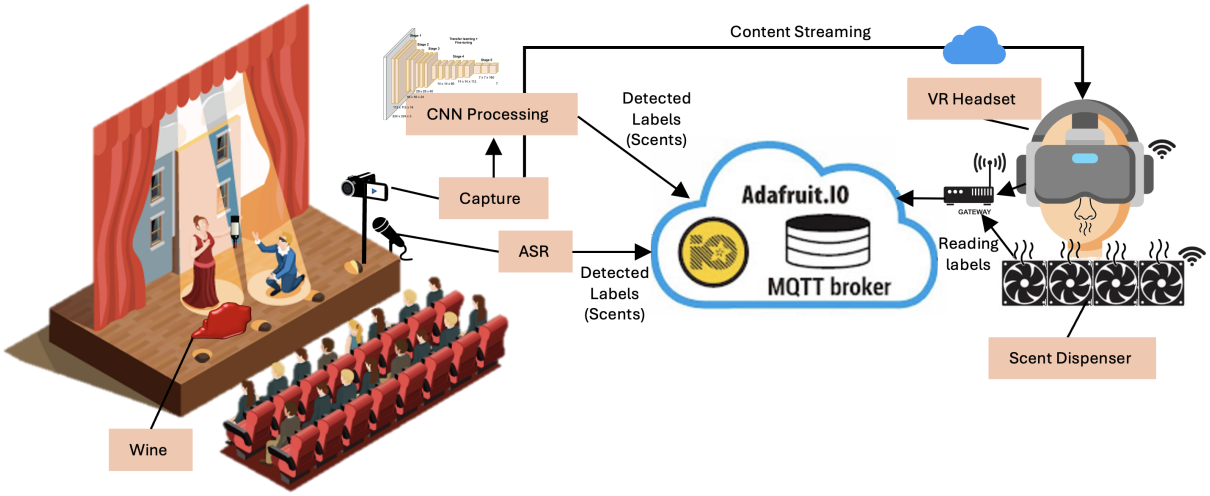

Fig. 3. System architecture for the multi-sensory theatre experience.

Fig. 4. Olfaction dispenser

```
LOG (VoskAPI:ComputeDerivedVars():ivector-extractor.cc:204) Done.
LOG (VoskAPI:ReadDataFiles():model.cc:282) Loading HCL and G from vosk-model-small-it-0.22/graph/HCLr.fst vosk-model-small-it-0.22
LOG (VoskAPI:ReadDataFiles():model.cc:303) Loading winfo vosk-model-small-it-0.22/graph/phones/word_boundary.int
Start speaking...
Word: beve, Start: 0.96, End: 1.39643, Confidence: 66.87%
Word: figlio, Start: 1.39643, End: 1.86, Confidence: 100.00%
Word: mio, Start: 1.86, End: 2.19, Confidence: 100.00%
Word: bevi, Start: 2.19, End: 2.73, Confidence: 46.98%
Sentence Confidence: 78.46%
Recognized Text: beve figlio mio bevi

Word: beh, Start: 4.5, End: 4.8, Confidence: 30.98%
Word: figlio, Start: 4.91702, End: 5.4, Confidence: 100.00%
Word: mio, Start: 5.4, End: 5.73, Confidence: 100.00%
Word: bevi, Start: 5.73, End: 6.24, Confidence: 67.08%
Sentence Confidence: 74.52%
Recognized Text: beh figlio mio bevi

Word: beve, Start: 8.31, End: 8.76, Confidence: 78.79%
Word: figlia, Start: 8.76, End: 9.18, Confidence: 77.68%
Word: mia, Start: 9.18, End: 9.48, Confidence: 77.68%
Word: bevi, Start: 9.487511, End: 9.99, Confidence: 89.45%
Sentence Confidence: 80.86%
Recognized Text: beve figlia mia bevi

Word: base, Start: 11.97, End: 12.417485, Confidence: 83.14%
Word: figlio, Start: 12.417485, End: 12.87, Confidence: 100.00%
Word: mio, Start: 12.87, End: 13.17, Confidence: 53.33%
Word: bevi, Start: 13.17, End: 13.68, Confidence: 77.08%
Sentence Confidence: 78.39%
Recognized Text: base figlio mio bevi
```

Fig. 5. Real-time Automatic Speech Recognition with Vosk (vosk-model-small-it-0.22)

IV. Solution Architecture

The proposed solution consists of the following components: scent dispensers, audio detection, scene recognition and a communications with an MQTT broker, as illustrated in Fig. 3.

```
LOG (VoskAPI:ReadDataFiles():model.cc:279) Loading HCLG from vosk-model-it-0.22/graph/HCLG.fst
LOG (VoskAPI:ReadDataFiles():model.cc:294) Loading words from vosk-model-it-0.22/graph/words.txt
LOG (VoskAPI:ReadDataFiles():model.cc:303) Loading winfo vosk-model-it-0.22/graph/phones/word_boundary.int
LOG (VoskAPI:ReadDataFiles():model.cc:310) Loading subtract G.fst model from vosk-model-it-0.22/rescore/G.fst
LOG (VoskAPI:ReadDataFiles():model.cc:312) Loading CARPA model from vosk-model-it-0.22/rescore/G.carpa
Start speaking...
Word: beve, Start: 2.13, End: 2.579916, Confidence: 48.97%
Word: figlio, Start: 2.58, End: 2.965928, Confidence: 100.00%
Word: mio, Start: 2.97, End: 3.238302, Confidence: 70.58%
Word: bevi, Start: 3.242733, End: 3.66, Confidence: 92.38%
Sentence Confidence: 77.98%
Recognized Text: beve figlio mio bevi

Word: bevi, Start: 6.54, End: 6.87, Confidence: 36.28%
Word: figlio, Start: 6.898787, End: 7.29, Confidence: 99.88%
Word: mio, Start: 7.29, End: 7.552158, Confidence: 89.37%
Word: bevi, Start: 7.56, End: 7.95, Confidence: 87.47%
Sentence Confidence: 78.25%
Recognized Text: bevi figlio mio bevi

Word: bev, Start: 10.71, End: 11.097768, Confidence: 30.39%
Word: figlio, Start: 11.098347, End: 11.477792, Confidence: 54.21%
Word: mia, Start: 11.49, End: 11.759368, Confidence: 26.82%
Word: baby, Start: 11.76, End: 12.12, Confidence: 23.98%
Sentence Confidence: 33.83%
Recognized Text: bev figlio mia baby

Word: bevi, Start: 15.3, End: 15.66, Confidence: 62.22%
Word: figlio, Start: 15.69, End: 16.079941, Confidence: 98.22%
Word: mio, Start: 16.08, End: 16.31863, Confidence: 98.85%
Word: bevi, Start: 16.32, End: 16.71, Confidence: 60.51%
Sentence Confidence: 79.95%
Recognized Text: bevi figlio mio bevi

Word: bevve, Start: 20.79, End: 21.24, Confidence: 58.72%
Word: figlio, Start: 21.247242, End: 21.69, Confidence: 100.00%
Word: mio, Start: 21.69, End: 22.079323, Confidence: 99.59%
Word: beve, Start: 22.08, End: 22.56, Confidence: 33.58%
Sentence Confidence: 78.97%
Recognized Text: bevve figlio mio beve
```

Fig. 6. Real-time Automatic Speech Recognition with Vosk (vosk-model-it-0.22)

A. Scent Dispensers

The scent dispensers located at the remote users' locations are shown in Fig. 4. Provided by Inhalio, the dispenser feature scent cartridge slots that hold custom-made cartridges with scent beads tailored to the play's needs, with fans behind the slots generating scents by blowing wind into the beads. They are equipped with Wi-Fi interfaces, enabling them to receive requests via the olfaction API, which controls the fans, turning them on and off and adjusting the intensity of the scent flow based on HTTP requests.

B. Audio-Based Scene Detection

The Vosk Python library, as seen in Figs. 5 and 6, is used for speech recognition in key scenes involving red wine and ash scents. A microphone is placed near the stage and connected to the PC, where the Vosk model processes the captured audio. The application uses the Italian language models vosk-model-it-0.22 (2.05GB) and vosk-model-small-it-0.22 (91.3MB). The Vosk model processes the speech in real-time, capturing specific words.

Once the relevant words are detected, the Python application sends a message containing the wine label to the Adafruit

TABLE I
TOP-3 PREDICTIONS WITH CONFIDENCE SCORES FOR MOBILENETV3 AND RESNET18

Scene	Top-1 Pred.	Top-2 Pred.	Top-3 Pred.
MobileNetV3 – Trained on Object			
Ash	Unknown (43.96%)	Wine (35.96%)	Ash (20.08%)
Wine	Wine (39.47%)	Unknown (37.27%)	Ash (23.26%)
ResNet18 – Trained on Object			
Ash	Ash (58.87%)	Wine (22.39%)	Unknown (18.74%)
Wine	Wine (56.08%)	Ash (35.52%)	Unknown (8.40%)
MobileNetV3 – Trained on Full Frame			
Ash	Ash (59.99%)	Unknown (25.74%)	Wine (14.27%)
Wine	Wine (41.00%)	Unknown (31.29%)	Ash (27.71%)
ResNet18 – Trained on Full Frame			
Ash	Ash (92.11%)	Wine (4.99%)	Unknown (2.90%)
Wine	Wine (89.22%)	Ash (6.87%)	Unknown (3.91%)

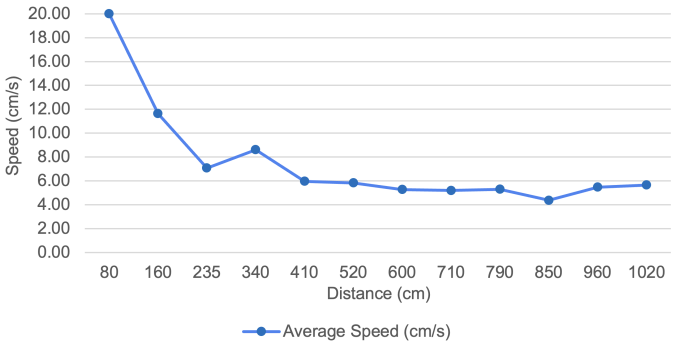

Fig. 7. Average speed of scent dispersion to each chair distance

MQTT broker. The remote dispensers are activated by another Python script that reads from the broker and sends a GET request to the dispenser. The triggering of the dispenser is done with an added delay, which is based on the VR viewer's position in relation to the stage. In the VR scenario, viewers can sit closer or farther from the stage, and the specific delay for these distances will be discussed in Section V. This delay is based on the time it takes for scents to be perceived by viewers at varying distances in a real theatre setting.

C. CNN-Based Scene Recognition

ResNet18 and MobileNetV3 CNN models are compared in the task of recognising visual elements in key scenes involving red wine and ash. These models run locally on the connected PC and process video frames in real-time to detect specific visual cues associated with each scent. When the scenes are identified, the application sends the corresponding label to the Adafruit MQTT broker, triggering the appropriate olfaction device. The activation includes a delay based on the viewer's position in the VR space, simulating the natural dispersion of scents in a theatrical setting.

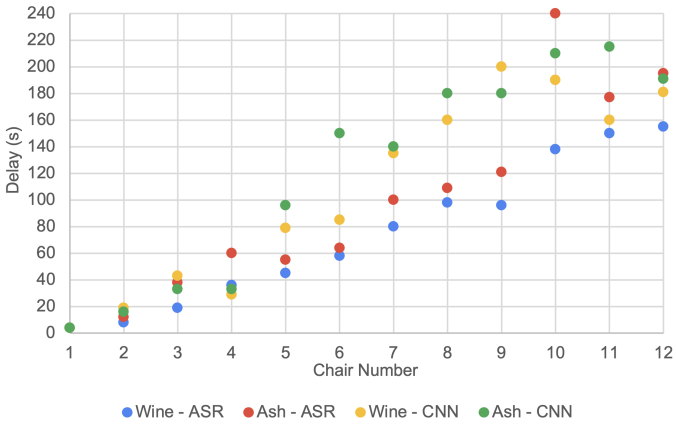

Fig. 8. Time taken for users to detect scent at each chair, categorised by scent and trigger type

D. Communication Framework

The Adafruit MQTT broker facilitates real-time communication between the theatre stage and remote users. This low-latency protocol ensures that sensory cues are delivered with delays below 100ms, maintaining synchronisation between events and sensory feedback.

V. TESTING AND RESULTS

A. Audio-Based Approach

The Python library Vosk employed in speech recognition provides a confidence score for each recognised word in real-time, which we used to calculate an average per sentence. This average confidence value was used to assess the reliability of detected phrases. Through empirical testing, we observed that average word confidence values above 70% were suitable for reliably triggering the corresponding scent events. Despite the significant size difference, the smaller Vosk model also performed well in terms of accuracy in the proposed solution.

Both the full-size and small models consistently returned sentence averages above this threshold when the key words — figlio, mio, and bevi (in English, son, my, drink) — were spoken clearly, validating the feasibility of the approach.

B. CNN-Based Approach

The performance of the CNN models varies notably depending on whether they were trained on isolated objects (props) or the entire scene frame, selected from the actual Macbettu play. A train–test split ratio of 80% to 20% was employed, with files organised into tables according to the labels ash, wine, and unknown. As seen on Table I, when trained on objects alone, such as the wine jar and smoke, both MobileNetV3 and ResNet18 showed moderate classification accuracy. For example, MobileNetV3 trained on objects misclassified the "Ash" scene as "Unknown". ResNet18's object-based model achieved better top-1 accuracy, with confidence below 60%.

In contrast, models trained on full-frame images demonstrated significantly improved performance. MobileNetV3, while still less accurate than ResNet18, correctly predicted

TABLE II
SCENT DELAY, DIFFUSION SPEED, AND ADDITIONAL DELAY FOR CNN IN RELATION TO ASR

Chair No.	Distance to Dispenser	Delay (in seconds)				Average Delay (s)	Average Speed (cm/s)	Add. Delay Delay CNN Wine	Add. Delay Delay CNN Ash
		Wine-ASR	Ash-ASR	Wine-CNN	Ash-CNN				
1	80 cm	4	4	4	4	4.00	20.00	0s	0s
2	160 cm	8	12	19	16	13.75	11.64	11s	4s
3	235 cm	19	38	43	33	33.25	7.07	24s	-5s
4	340 cm	36	60	29	33	39.50	8.61	-7s	-27s
5	410 cm	45	55	79	96	68.75	5.96	34s	41s
6	520 cm	58	64	85	150	89.25	5.83	27s	86s
7	600 cm	80	100	135	140	113.75	5.27	55s	40s
8	710 cm	98	109	160	180	136.75	5.19	62s	71s
9	790 cm	96	121	200	180	149.25	5.29	104s	59s
10	850 cm	138	240	190	210	194.50	4.37	52s	-30s
11	960 cm	150	177	160	215	175.50	5.47	10s	38s
12	1020 cm	155	195	181	191	180.50	5.65	26s	-4s
Avg.								33.17s	22.75s

both scenes with higher confidence than its object-based version. Most notably, ResNet18 trained on full frames performed the best, achieving over 89% top-1 accuracy for both "Ash" and "Wine" scenes. MobileNetV3, however, requires approximately 50% less inference time compared to ResNet18.

C. Scent Delays and User Perceptual Tests

The experimental data collected from twelve participants across twelve seating positions reveals a clear correlation between the distance from the scent dispenser and the delay in scent perception. For the experiment, ethical approval was obtained from the Dublin City University Research Ethics Committee, application number DCUREC2024/175. 12 individuals took part in the tests, including 2 female and 10 male participants, aged between 21 and 59, from different ethnic groups and different backgrounds and professions. Each participant experienced two different seating positions—one closer and one farther from the dispenser.

The goal of the experiment was to measure the real-world diffusion delay of scents triggered by two approaches used in the theatre setting: ASR and CNN. Scents were released over four rounds in the following order: Wine (ASR), Ash (ASR), Wine (CNN), and Ash (CNN). The delays were measured in combination with the physical distance between the dispenser and each viewer. In the XR implementation of the experience, users will also be able to choose their seating position relative to the stage. Based on these findings, the application can simulate scent dispersion delays accordingly, ensuring a realistic and spatially accurate olfactory experience.

As shown in Table II and Fig. 7, the delay increases with distance in both audio-based (ASR) and CNN-based triggers, with average delays ranging from 4 seconds (at 80 cm) to over 180 seconds (at 1020 cm). Correspondingly, the average diffusion speed of the scent decreases from approximately 20 cm/s to 5.2 cm/s as the distance increases, illustrating the non-linear nature of scent dispersion in a real theatre environment.

CNN-based (i.e. using MobileNetV3) triggers generally resulted in slightly longer delays than ASR, with an average additional delay of 33.17 seconds for wine scene and 22.75 seconds for ash scene (see Table II). This difference is largely due to the image processing and recognition time required by CNN models. However, as seen in Fig 8, the overall trend of scent arrival time across seats remains consistent between the two modalities, enabling a uniform mapping of dispersion timing for replication in XR.

TABLE III
PARTICIPANTS' PERCEPTION OF DELAY BASED ON PROXIMITY TO THE DISPENSER

Chair No.	Wine ASR	Ash ASR	Wine CNN	Ash CNN	Avg. Score
1	Str. Agree	Str. Agree	Agree	Agree	3.50
2	Agree	Agree	Agree	Str. Agree	3.25
3	Agree	Disagree	Str. Agree	Str. Agree	3.00
4	Disagree	Disagree	Neutral	Neutral	1.50
5	Neutral	Disagree	Agree	Neutral	2.00
6	Neutral	Agree	Disagree	Disagree	1.75
7	Agree	Agree	Disagree	Disagree	2.00
8	Neutral	Agree	Neutral	Agree	2.50
9	Str. Agree	Str. Agree	Neutral	Neutral	3.00
10	Agree	Neutral	Str. Disag.	Str. Disag.	1.25
11	Agree	Agree	Neutral	Neutral	2.50
12	Neutral	Disagree	Neutral	Neutral	1.75
Avg.	2.67	2.42	2.08	2.17	**2.33**

To validate the realism of the delay experience, participants were asked to rate their perception of scent timing relative to their position (Table III). Overall, feedback indicated a positive correlation between proximity and satisfaction, with viewers closer to the stage expressing stronger agreement that the timing felt appropriate, as seen on Fig. 9. The number of responses for each Likert scale item, categorised by scent and trigger type is available in Fig. 10 Average perception scores, converted from the Likert scale (ranging from strongly disagree = 0 to strongly agree = 4), indicate that both ASR and CNN-based approaches were well-received, as seen in Fig. 11 with the ASR-wine pairing rated highest (2.67) and the CNN-wine pairing slightly lower (2.08), reflecting the impact of longer recognition time.

VI. CONCLUSION

This paper demonstrated the feasibility of integrating machine learning and artificial scent detection into a multisen-

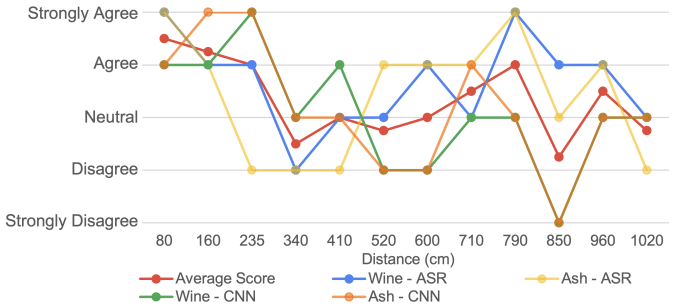

Fig. 9. Likert scale scores per chair distance and average scores, categorised by scent and trigger type

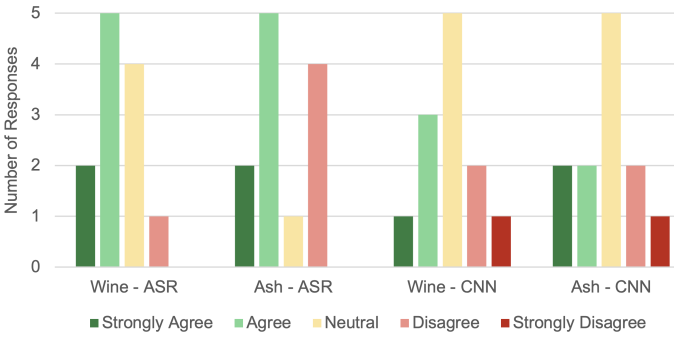

Fig. 10. Number of responses for each Likert scale item, categorised by scent and trigger type

sory theatre experience. Using the play Macbettu as a case study, we showcased how olfactory and visual stimuli can be synchronised to create an immersive experience for remote audiences.

The integration of ASR- and CNN-based triggers for scent delivery proved effective in generating spatially aware and temporally synchronised olfactory effects. Participant feedback confirmed that the timing and accuracy of scent release were generally well-received.

While traditional theatrical environments frequently use established cueing mechanisms such as TimeCode, DMX lighting scenes, or sound effect triggers to synchronise events, these systems often depend on access to stage networks and predefined control tracks, which may not be available or practical in remote or decentralised XR theatre scenarios. In contrast, ASR and CNN-based approaches offer greater autonomy and adaptability by responding directly to the live performance, allowing the solution to remain functional even in venues lacking DMX integration or centralised time-based control. Moreover, different plays may benefit from different cueing strategies: productions with darker lighting conditions may favour ASR-based detection, while those with limited dialogue but strong visual markers may find CNN-based scene recognition more effective.

Nonetheless, we acknowledge that these machine learning-based methods introduce limitations. ASR performance depends on the clarity and frequency of speech, which may vary in productions with sparse dialogue or high levels of

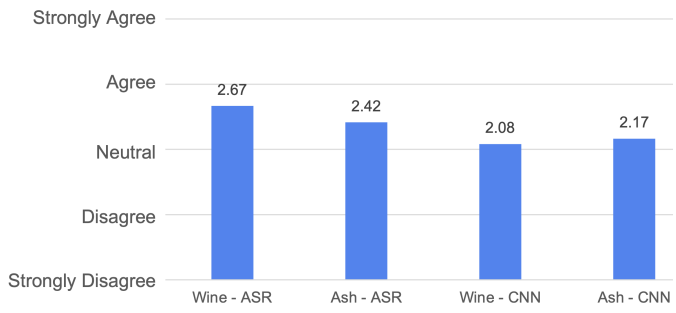

Fig. 11. Average score for each scent and trigger format according to user-perceived delay relative to proximity

repetition. Similarly, CNN-based detection requires adequate lighting and distinctive visual features, which may not always be present in all theatrical contexts. Consequently, while the proposed system performed well in *Macbettu*, further work is required to assess its robustness across diverse theatrical styles and lighting conditions, and to explore potential hybrid solutions that integrate simpler cueing methods where feasible.

Other avenues for future work include testing additional scents across a broader range of scenes to further evaluate the scalability of the solution. Moreover, the feasibility of incorporating an artificial scent detection mechanism, such as an electronic nose (e-nose), will be explored to enable automated scent feedback directly from the stage.

Acknowledgment

This work was supported by Research Ireland via the Research Centres grant 12/RC/2289_P2 (INSIGHT), and by the European Union (EU) Horizon Europe grant 101135637 (HEAT Project).

References

[1] K. Pietroszek, M. Rebol, and B. Lake, "Dill Pickle: Interactive Theatre Play in Virtual Reality," in *Proc. ACM Symposium on Virtual Reality Software and Technology (VRST)*, 2022.

[2] R. Verma, A. A. Simiscuka, M. A. Togou, M. Zorrilla, and G.-M. Muntean, "A Live Adaptive Streaming Solution for Enhancing Quality of Experience in Co-Created Opera," *IEEE Transactions on Broadcasting*, pp. 1–12, 2025.

[3] A. A. Simiscuka, D. A. Ghadge, and G.-M. Muntean, "OmniScent: An Omnidirectional Olfaction-Enhanced Virtual Reality 360° Video Delivery Solution for Increasing Viewer Quality of Experience," *IEEE Transactions on Broadcasting*, vol. 69, no. 4, pp. 941–950, 2023.

[4] I. Tal, L. Zou, A. Covaci, E. Ibarrola, M. Bratu, G. Ghinea, and G.-M. Muntean, "Mulsemedia in Telecommunication and Networking Education: A Novel Teaching Approach that Improves the Learning Process," *IEEE Commun. Mag.*, vol. 57, no. 11, pp. 60–66, 2019.

[5] L. Turchet, C. Fischione, G. Essl, D. Keller, and M. Barthet, "Internet of musical things: Vision and challenges," *IEEE Access*, vol. 6, pp. 61 994–62 017, 2018.

[6] F. Yan, X. Zhao, and L. Li, "Olfactory interfaces for immersive vr: Design and evaluation," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 6, pp. 492–501, 2020.

[7] Y. Hirata and K. Suzuki, "Multisensory feedback systems for enhanced remote interaction," *IEEE Transactions on Haptics*, vol. 12, no. 3, pp. 291–302, 2019.

[8] D. Lisowski, K. Ponto, S. Fan, C. Probst, and B. Sprecher, "Augmented Reality into Live Theatrical Performance," in *Springer Handbook of Augmented Reality*, A. Y. C. Nee and S. K. Ong, Eds. Springer, 2023, pp. 433–450.

[9] "EU Horizon 2020 Traction Project – Opera co-creation for social transformation."

[10] D. Fox and I. G. Thornton, "The IEEE Global Initiative on Ethics of Extended Reality (XR) Report–Extended Reality (XR) Ethics and Diversity, Inclusion, and Accessibility," White Paper, 2022.
[11] G. Minopoulos and K. E. Psannis, "Opportunities and Challenges of Tangible XR Applications for 5G Networks and Beyond," *IEEE Consumer Electronics Magazine*, vol. 12, no. 6, pp. 9–19, 2023.
[12] A. Gallace, M. K. Ngo, J. Sulaitis, and C. Spence, "Multisensory Presence in Virtual Reality: Possibilities and Limitations," in *Multiple Sensorial Media Advances and Applications: New Developments in MulSeMedia*. IGI Global, 2011, ch. 1, pp. 1–38.
[13] B. Maclntyre and T. F. Smith, "Thoughts on the future of webxr and the immersive web," in *Proc. IEEE Int. Symposium on Mixed and Augmented Reality Adjunct (ISMAR)*, 2018, pp. 338–342.
[14] A. A. Simiscuka, M. A. Togou, R. Verma, M. Zorrilla, N. E. O'Connor, and G.-M. Muntean, "An Evaluation of 360° Video and Audio Quality in an Artistic-Oriented Platform," in *Proc. IEEE Int. Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2022, pp. 1–5.
[15] T. Plantefol, A. A. Simiscuka, A. Yaqoob, and G.-M. Muntean, "CNN-based 360° Scene Recognition for Automatic Generation of Omnidirectional Scent Effects," *IEEE Transactions on Multimedia*, 2025.
[16] L. Turchet, T. West, and M. M. Wanderley, "Touching the audience: musical haptic wearables for augmented and participatory live music performances," *Personal and Ubiquitous Computing*, vol. 25, no. 4, pp. 749–769, 2021.
[17] K. Apostolou and F. Liarokapis, "A systematic review: The role of multisensory feedback in virtual reality," in *2022 IEEE 2nd International Conference on Intelligent Reality (ICIR)*. IEEE, 2022, pp. 39–42.
[18] M. Melo, G. Gonçalves, P. Monteiro, H. Coelho, J. Vasconcelos-Raposo, and M. Bessa, "Do Multisensory Stimuli Benefit the Virtual Reality Experience? A Systematic Review," *IEEE Trans. Vis. Comput. Graphics*, pp. 1–20, 2020.
[19] A. A. Simiscuka, M. A. Togou, M. Zorrilla, and G.-M. Muntean, "360-ADAPT: An Open-RAN-Based Adaptive Scheme for Quality Enhancement of Opera 360° Content Distribution," *IEEE Transactions on Green Communications and Networking*, pp. 1–14, 2024.
[20] A. Yaqoob and G.-M. Muntean, "A Combined Field-of-View Prediction-Assisted Viewport Adaptive Delivery Scheme for 360° Videos," *IEEE Transactions on Broadcasting*, vol. 67, no. 3, pp. 746–760, 2021.
[21] Z. Jiang, X. Zhang, Y. Xu, Z. Ma, J. Sun, and Y. Zhang, "Reinforcement learning based rate adaptation for 360-degree video streaming," *IEEE Transactions on Broadcasting*, vol. 67, no. 2, pp. 409–423, 2021.
[22] P. Szabó, A. A. Simiscuka, S. Masneri, M. Zorrilla, and G.-M. Muntean, "A CNN-Based Framework for Enhancing 360° VR Experiences With Multisensorial Effects," *IEEE Transactions on Multimedia*, vol. 25, pp. 3245–3258, 2023.
[23] J. P. Sexton, A. A. Simiscuka, K. Mcguinness, and G.-M. Muntean, "Automatic CNN-Based Enhancement of 360° Video Experience With Multisensorial Effects," *IEEE Access*, vol. 9, pp. 133 156–133 169, 2021.
[24] A. H. et al., "Searching for mobilenetv3," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324.
[25] Q. Zhao, Y. Liu, and H. Wang, "Mqtt-based communication for iot applications," *IEEE Internet of Things Journal*, vol. 4, no. 3, pp. 832–839, 2017.