Immersive Live Concert: A Multi-Sensory Experience Based on Real-Time Lyrics Detection from Spatial Audio Data

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

Gianluca Fadda
gianluca.fadda@unica.it
Università degli Studi di Cagliari
Cagliari, Italy

Vlad Popescu
vlad.popescu@unitbv.ro
Universitatea Transilvania din Brașov
Brașov, Romania

Maurizio Murroni
maurizio.murroni@unica.it
Università degli Studi di Cagliari
Cagliari, Italy

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

Abstract

This paper presents a multi-sensory solution aimed at enhancing the immersive experience of remote audiences for a live blues concerts. Using the XRBLUES pilot of the HEAT project as a case study, the approach synchronizes olfactory and visual stimuli for remote viewers. The solution focuses on audio-based detection to identify key moments in the concert's lyrics that trigger specific scents, such as ocean air and wind effects. These sensory cues are synchronized in real-time with the live concert, delivered via scent dispensers for remote users wearing XR headsets. Real-time communication is facilitated through an MQTT-based system to ensure minimal latency. The paper evaluates the effectiveness of this audio-based cue detection approach by evaluating Automatic Speech Recognition (ASR) models, using the Vosk Python library. This work demonstrates how scent-enhanced XR technologies can be integrated into live music events, offering a more engaging and immersive experience for remote concertgoers.

Keywords

Extended Reality, Multi-Sensory Experiences, Audio Detection

How to cite this paper:
Anderson Augusto Simiscuka, Gianluca Fadda, Vlad Popescu, Maurizio Murroni, and Gabriel-Miro Muntean. 2025. Immersive Live Concert: A Multi-Sensory Experience Based on Real-Time Lyrics Detection from Spatial Audio Data. In *Proceedings of ACM IMX Workshops, June 3 - 6, 2025.* SBC, Porto Alegre/RS, Brazil, 5 pages. https://doi.org/10.5753/imxw.2025.1139

Figure 1: Live blues concert streamed in real-time

1 Introduction

Music concerts are primarily an in-person auditory experience. While live streams of concerts are common, they fail to replicate the same experience of attending a concert. Advances in Extended Reality (XR) and communications technologies provide new opportunities to replicate live concert experiences for remote audiences, even in real-time performance settings [7, 13]. Recent innovations in multi-sensory systems, particularly olfactory interfaces, enable the integration of scents into virtual environments, creating more immersive and engaging experiences [9, 12].

This work focuses on a live blues concert where two key parts of the lyrics trigger ocean air scent and wind effects for both in-person and remote participants. The concert is held at a blues bar (see Fig. 1) with participants watching the live performance and others experiencing it remotely through XR headsets, in a Unity-based 3D environment (see Fig. 2). The concert is transformed into an immersive 3D environment, allowing remote users to navigate within it, moving alongside the musicians. As they explore, the sound dynamically adapts, providing a personalized auditory experience based on their position. In the future, this approach will be expanded to incorporate volumetric holograms of the musicians, significantly increasing the amount of data transmitted. The ocean air scent and wind effects are triggered by specific audio cues in the lyrics, creating a more immersive experience for both in-person and remote audiences. Remote viewers experience the concert through XR headsets, with synchronized scents and effects delivered based on audio-based cues.

By combining audio-based cue detection, network communication, and scent dispensers, it is possible to deliver synchronized olfactory stimuli and wind effects to remote audiences watching the concert through the headsets. The solution identifies specific audio cues within the concert's lyrics that trigger the ocean air scent and

https://doi.org/10.5753/imxw.2025.1139

Figure 2: Blues concert presented as an immersive 3D environment created in Unity

wind effects. These cues are processed in real-time to ensure the corresponding scents and effects are released at the appropriate moments for remote users.

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

- Audio-based cue detection for real-time scent and wind effect triggering.
- MQTT-based communication for low-latency interaction between the concert stage and remote users.
- Evaluation of audio cue detection using two Vosk Python ASR models.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 provides an overview of the HEAT project, within which the blues concert is included. Section 4 details the system architecture and methods used to detect audio cues for triggering scents and wind effects. Section 5 evaluates the accuracy of the audio cue detection. Finally, Section 6 concludes the paper and discusses potential directions for future research.

2 Related Works

Recent advances in multi-sensory technologies have opened up new opportunities for enhancing human-computer interaction, particularly in immersive environments such as virtual reality (VR). One area that has attracted attention is the integration of olfactory feedback, which has shown to improve immersion and user engagement. For instance, Yan et al. demonstrated how VR olfactory interfaces could significantly enhance realism, leading to increased user satisfaction [14]. Similarly, Hirata and Suzuki's work on multisensory systems for remote interaction underscores the importance of synchronizing sensory stimuli in real-time for a more cohesive experience [2].

Extended Reality (XR) technologies are increasingly being utilized to enhance both in-person and remote experiences. In particular, live performances have benefitted from the addition of XR elements, such as immersive virtual environments and the use of augmented reality (AR) to integrate multi-sensory effects into stage productions [4]. These innovations have extended beyond the stage to offer new ways for audiences to interact with performances, particularly those involving remote participation.

A significant challenge in these applications, however, lies in the integration of olfactory feedback. The ability to synchronize scent with other sensory inputs, such as visuals and sound, is challenging. Network limitations and the complexity of real-time delivery can negatively affect the overall experience, especially in VR setups. Nonetheless, WebXR and similar technologies enable support to 360° streaming, motion tracking, and olfactory integration [5, 8, 10] in web browsers, making access to multi-sensory experiences in virtual environments more accessible.

Despite these advancements, olfactory feedback remains relatively underutilized in many VR applications, even with studies indicating that over 85% of mulsemedia VR applications see improvements in engagement and user satisfaction when sensory modalities are integrated [1, 6].

Challenges such as network constraints and the need for efficient data compression can negatively impact the Quality of Experience (QoE) for remote users [3, 11, 16]. To address these issues, lightweight communication protocols such as MQTT have become essential for real-time systems, particularly those relying on the Internet of Things (IoT). The low-latency nature of MQTT ensures that sensory feedback, including olfactory cues, can be delivered with minimal delay, which is needed for maintaining synchronization between remote users and the ongoing performance [15].

Incorporating these multi-sensory elements into XR applications, particularly for live performances, can significantly improve the immersion and engagement of remote audiences.

3 The HEAT Project and the XRBLUES Pilot

The work presented in this paper is part of the XRBLUES pilot within the HEAT project. The pilot aims to create an immersive concert experience for both remote and in-person users, allowing them to interact and share moments through holograms and multisensory media.

Funded by the EU Horizon Europe program, the HEAT project explores innovative scenarios and pilots that enable users to be virtually teleported into hyper-realistic, navigable 3D environments. This allows them to fully immerse themselves in the atmosphere of the experience and share it with others, regardless of physical location, including those co-located in the same space. HEAT focuses on integrating cutting-edge immersive technologies such as point cloud and holographic imaging, multi-sensory media, and social XR, all within a multi-user communication system that supports real-time interaction.

At the heart of HEAT is a modular and flexible communication pipeline capable of supporting a wide range of media types—from traditional audio and video to multi-sensory and holographic content. This system ensures efficient encoding, processing, storage, real-time streaming, and rendering. The HEAT project is designing, deploying, and evaluating various content delivery solutions in real-world settings, including a blended learning environment, a contemporary theatre performance, a blues concert (as demonstrated in this paper), and an opera production.

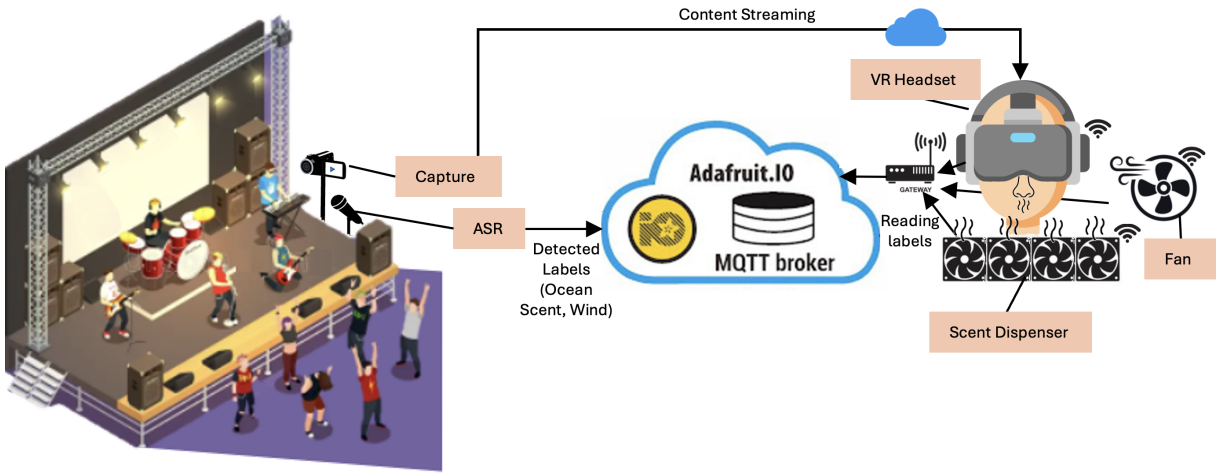

Figure 3: System architecture for the multi-sensory concert experience.

Figure 4: Olfaction dispenser

4 System Architecture

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

4.1 Multi-Sensory Devices

The scent dispensers, positioned at the remote users' locations, are shown in Fig. 4. Supplied by Inhalio, these dispensers feature slots for custom-made scent cartridges containing Ocean Air scent beads designed specifically for the pilot's requirements. Behind each slot, fans blow air across the beads to release the corresponding scents. The dispensers are equipped with Wi-Fi connectivity, allowing them to receive commands through the olfaction API. This API controls the fans, enabling them to be turned on and off, and adjusts the scent intensity based on HTTP requests, ensuring a dynamic and responsive scent delivery system.

A fan is employed to create the wind effect and is connected to a Shelly smart plug. The fan is activated through the Shelly API, which operates via HTTP requests, allowing for precise control of the wind effect in response to the MQTT broker commands.

4.2 Audio-Based Scene Detection

The Vosk Python library, as seen in Figs. 5 and 6, is used for automatic speech recognition (ASR) in key scenes involving ocean air and wind effects. Audio captured from the singer's microphone through the audio mixer is transmitted to a PC, where the Vosk model processes the captured audio.

The Vosk library computes confidence in ASR by assessing the likelihood that a given transcription is accurate, based on both the acoustic model and the language model. The acoustic model evaluates the audio features to identify phonemes, while the language model predicts the most probable word sequences based on context. Vosk combines these models to generate a hypothesis of the recognized speech and computes a confidence score, which reflects how certain the system is about the accuracy of the transcription. This score is derived from the probability of the recognized words and their alignment with the audio features, with higher probabilities indicating greater certainty. The final confidence score provides an estimate of how likely it is that the transcription is correct.

The application utilizes the English language models vosk-model-en-us-0.22 (2.87GB) and vosk-model-small-en-us-0.15 (70.9MB), with the larger model offering higher accuracy in recognizing key phrases from the song "Californian Winds" by Fred Sunwalk, such as "take that road, take to northwest" and "I got nothing to lose alright." These phrases trigger the corresponding ocean air scent and wind effect for remote users.

Once the relevant words are detected, the Python application sends a message containing the ocean air or wind effect label to the Adafruit MQTT broker. The remote dispensers are activated by another Python script that reads from the broker and sends a GET request to the dispensers.

```
Word: the, Start: 625.19, End: 634.1, Confidence: 100.00%
Word: take, Start: 634.85, End: 635.12, Confidence: 100.00%
Word: that, Start: 635.12, End: 635.3, Confidence: 100.00%
Word: road, Start: 635.3, End: 635.72, Confidence: 100.00%
Word: take, Start: 635.96, End: 636.23, Confidence: 100.00%
Word: to, Start: 636.23, End: 636.41, Confidence: 78.76%
Word: northwest, Start: 636.41, End: 637.13, Confidence: 68.39%
Sentence Confidence: 92.45%
Recognized Text: the take that road take to northwest
```

(a)

```
Word: the, Start: 677.51, End: 682.28, Confidence: 100.00%
Word: i, Start: 682.76, End: 682.91, Confidence: 89.41%
Word: got, Start: 682.916357, End: 683.21, Confidence: 100.00%
Word: nothing, Start: 683.24, End: 683.6, Confidence: 100.00%
Word: to, Start: 683.6, End: 683.78, Confidence: 100.00%
Word: lose, Start: 683.78, End: 684.11, Confidence: 100.00%
Word: all, Start: 684.11, End: 684.32, Confidence: 57.28%
Word: right, Start: 684.32, End: 684.77, Confidence: 57.28%
Sentence Confidence: 87.99%
Recognized Text: the i got nothing to lose all right
```

(b)

```
Word: got, Start: 288.57, End: 288.84, Confidence: 86.72%
Word: californian, Start: 288.84, End: 289.59, Confidence: 85.80%
Word: wins, Start: 289.62, End: 290.07, Confidence: 77.06%
Word: that's, Start: 290.13, End: 290.37, Confidence: 100.00%
Word: right, Start: 290.37, End: 290.79, Confidence: 100.00%
Sentence Confidence: 89.92%
Recognized Text: got californian wins that's right

Word: got, Start: 295.29, End: 295.53, Confidence: 100.00%
Word: california, Start: 295.53, End: 296.16, Confidence: 79.57%
Word: and, Start: 296.16, End: 296.25, Confidence: 79.57%
Word: whence, Start: 296.28, End: 296.64, Confidence: 24.95%
Word: that's, Start: 296.648027, End: 296.94, Confidence: 95.34%
Word: right, Start: 296.94, End: 297.21, Confidence: 100.00%
Sentence Confidence: 79.91%
Recognized Text: got california and whence that's right
```

(c)

Figure 5: Real-time Automatic Speech Recognition with Vosk (vosk-model-small-en-us-0.15)

```
Word: take, Start: 77.46, End: 77.82, Confidence: 100.00%
Word: that, Start: 77.82, End: 78.06, Confidence: 100.00%
Word: road, Start: 78.06, End: 78.48, Confidence: 100.00%
Word: take, Start: 78.6, End: 78.87, Confidence: 100.00%
Word: to, Start: 78.87, End: 79.02, Confidence: 62.83%
Word: northwest, Start: 79.02, End: 79.83, Confidence: 100.00%
Sentence Confidence: 93.81%
Recognized Text: take that road take to northwest
```

(a)

```
Word: i, Start: 15.12, End: 15.27, Confidence: 100.00%
Word: got, Start: 15.27, End: 15.51, Confidence: 100.00%
Word: nothing, Start: 15.51, End: 15.9, Confidence: 100.00%
Word: to, Start: 15.9, End: 16.02, Confidence: 100.00%
Word: lose, Start: 16.02, End: 16.32, Confidence: 100.00%
Word: all, Start: 16.32, End: 16.53, Confidence: 100.00%
Word: right, Start: 16.53, End: 16.95, Confidence: 100.00%
Sentence Confidence: 100.00%
Recognized Text: i got nothing to lose all right
```

(b)

```
Word: got, Start: 883.88, End: 884.27, Confidence: 100.00%
Word: californian, Start: 884.27, End: 885.17, Confidence: 100.00%
Word: wins, Start: 885.23, End: 885.89, Confidence: 68.81%
Word: that's, Start: 885.98, End: 886.34, Confidence: 100.00%
Word: right, Start: 886.34, End: 886.76, Confidence: 100.00%
Sentence Confidence: 93.76%
Recognized Text: got californian wins that's right

Word: got, Start: 895.34, End: 895.61, Confidence: 100.00%
Word: californian, Start: 895.64, End: 896.45, Confidence: 100.00%
Word: winds, Start: 896.51, End: 897.14, Confidence: 50.84%
Word: that's, Start: 897.26, End: 897.56, Confidence: 100.00%
Word: right, Start: 897.56, End: 897.92, Confidence: 100.00%
Sentence Confidence: 90.17%
Recognized Text: got californian winds that's right
```

(c)

Figure 6: Real-time Automatic Speech Recognition with Vosk (vosk-model-en-us-0.22)

4.3 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.

5 Testing and Results

The Python library Vosk, used for speech recognition, assigns a confidence score to each recognized word in real-time. This score is then averaged across sentences to determine the reliability of detected phrases. Through empirical testing, it was found that sentence confidence values above 70% were effective for reliably triggering corresponding scent events. For example, when the first sentence of the song “Californian Winds" by Fred Sunwalk (“Take that road, take to northwest") is detected, it triggers the ocean air scent, aligning with the context of the blues song.

The wind effect, is triggered with the line “I got nothing to lose alright", right before the line “got californian winds."

Despite using two different models, vosk-model-en-us-0.22 (full-size, 2.87GB) and vosk-model-small-en-us-0.15 (smaller model, 70.9 MB), both were able to achieve satisfactory accuracy, though with some differences in performance. The larger model showed consistently higher confidence scores across the entire sentence, while the smaller model yielded lower confidence, as seen in Figs. 5 and 6. This indicates that, although the smaller model has reduced accuracy, it still functions well for real-time event triggering in this multi-sensory setup.

Fig. 6 shows the sentence recognition results for the full-size model, achieving higher average sentence confidence (above 90%, up to 100%), which corresponds to better scent and wind effect triggering accuracy. In contrast, the small model, as illustrated in

Fig. 5, returned lower sentence confidence values (ranging from 79% to 92%) and make some mistakes for some words, but still performed well for the core phrases that activate the scents and wind effects.

6 Conclusion

This paper demonstrated the feasibility of integrating ASR with multisensory effects for a live blues concerts. Using “Californian Winds" by Fred Sunwalk as a case study, we showcased how olfactory and visual stimuli can be synchronized to enhance the immersive experience for remote audiences.

The integration of ASR for scent delivery proved effective in triggering ocean air scents and wind effects based on specific song lyrics. The solution was able to synchronize the scent and wind release with key moments of the performance.

For future work, additional scents and effects will be tested across other parts of the concert to further assess the versatility of the system.

Acknowledgments

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). We express our gratitude to Fred Sunwalk for performing in the concert and allowing us to test the song “Californian Winds" as part of this study.

References

[1] Kalliopi Apostolou and Fotis Liarokapis. 2022. A Systematic Review: The Role of Multisensory Feedback in Virtual Reality. In *2022 IEEE 2nd International Conference on Intelligent Reality (ICIR)*. IEEE, 39–42.

[2] Y. Hirata and K. Suzuki. 2019. Multisensory Feedback Systems for Enhanced Remote Interaction. *IEEE Transactions on Haptics* 12, 3 (2019), 291–302.

[3] Zhiqian Jiang, Xu Zhang, Yiling Xu, Zhan Ma, Jun Sun, and Yunfei Zhang. 2021. Reinforcement Learning Based Rate Adaptation for 360-Degree Video Streaming. *IEEE Transactions on Broadcasting* 67, 2 (2021), 409–423. https://doi.org/10.1109/TBC.2020.3028286

[4] Dan Lisowski, Kevin Ponto, Shuxing Fan, Caleb Probst, and Bryce Sprecher. 2023. Augmented Reality into Live Theatrical Performance. In *Springer Handbook of Augmented Reality*, Andrew Yeh Ching Nee and Soh Khim Ong (Eds.). Springer, 433–450. https://doi.org/10.1007/978-3-030-67822-7_18

[5] B. Maclntyre and T. F. Smith. 2018. Thoughts on the Future of WebXR and the Immersive Web. In *Proc. IEEE Int. Symposium on Mixed and Augmented Reality Adjunct (ISMAR)*. 338–342.

[6] M. Melo, G. Gonçalves, P. Monteiro, H. Coelho, J. Vasconcelos-Raposo, and M. Bessa. 2020. Do Multisensory Stimuli Benefit the Virtual Reality Experience? A Systematic Review. *IEEE Trans. Vis. Comput. Graphics* (2020), 1–20.

[7] Krzysztof Pietroszek, Manuel Rebol, and Becky Lake. 2022. Dill Pickle: Interactive Theatre Play in Virtual Reality. In *Proc. ACM Symposium on Virtual Reality Software and Technology (VRST)* (Tsukuba, Japan). Article 51, 2 pages. https://doi.org/10.1145/3562939.3565678

[8] T. Plantefol, A. A. Simiscuka, A. Yaqoob, and G.-M. Muntean. 2025. CNN-based 360° Scene Recognition for Automatic Generation of Omnidirectional Scent Effects. *IEEE Transactions on Multimedia* (2025).

[9] Anderson Augusto Simiscuka, Dhairyasheel Avinash Ghadge, and Gabriel-Miro Muntean. 2023. OmniScent: An Omnidirectional Olfaction-Enhanced Virtual Reality 360° Video Delivery Solution for Increasing Viewer Quality of Experience. *IEEE Transactions on Broadcasting* 69, 4 (2023), 941–950. https://doi.org/10.1109/TBC.2023.3277215

[10] Anderson Augusto Simiscuka, Mohammed Amine Togou, Rohit Verma, Mikel Zorrilla, Noel E. O'Connor, and Gabriel-Miro Muntean. 2022. An Evaluation of 360° Video and Audio Quality in an Artistic-Oriented Platform. In *Proc. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. 1–5. https://doi.org/10.1109/BMSB55706.2022.9828745

[11] Anderson Augusto Simiscuka, Mohammed Amine Togou, Mikel Zorrilla, and Gabriel-Miro Muntean. 2024. 360-ADAPT: An Open-RAN-Based Adaptive Scheme for Quality Enhancement of Opera 360° Content Distribution. *IEEE Transactions on Green Communications and Networking* (2024), 1–14. https://doi.org/10.1109/TGCN.2024.3418948

[12] Irina Tal, Longhao Zou, Alexandra Covaci, Eva Ibarrola, Marilena Bratu, Gheorghita Ghinea, and Gabriel-Miro Muntean. 2019. Mulsemedia in Telecommunication and Networking Education: A Novel Teaching Approach that Improves the Learning Process. *IEEE Communications Magazine* 57, 11 (2019), 60–66. https://doi.org/10.1109/MCOM.001.1900241

[13] Rohit Verma, Anderson Augusto Simiscuka, Mohammed Amine Togou, Mikel Zorrilla, and Gabriel-Miro Muntean. 2025. A Live Adaptive Streaming Solution for Enhancing Quality of Experience in Co-Created Opera. *IEEE Transactions on Broadcasting* (2025), 1–12. https://doi.org/10.1109/TBC.2025.3541875

[14] F. Yan, X. Zhao, and L. Li. 2020. Olfactory Interfaces for Immersive VR: Design and Evaluation. *IEEE Transactions on Human-Machine Systems* 50, 6 (2020), 492–501.

[15] Q. Zhao, Y. Liu, and H. Wang. 2017. MQTT-Based Communication for IoT Applications. *IEEE Internet of Things Journal* 4, 3 (2017), 832–839.

[16] Abid Yaqoob and Gabriel-Miro Muntean. 2021. A Combined Field-of-View Prediction-Assisted Viewport Adaptive Delivery Scheme for 360° Videos. *IEEE Transactions on Broadcasting* 67, 3 (2021), 746–760. https://doi.org/10.1109/TBC.2021.3105022