A. Moulson, M. Walley, Y. Grewe, R. Oldfield, B. Shirley, and U. Scuda,
"Practical Implementation of Automated Next Generation
Audio Production for Live Sports,"
J. Audio Eng. Soc., vol. XX, no. X, pp. 1–9 (2024 X).
https://doi.org/10.17743/jaes.2022.0151.

Practical Implementation of Automated Next Generation Audio Production for Live Sports*

AIMÉE MOULSON,[1] *AES Member*, **MAX WALLEY,**[2]
(aimee.moulson@iis.fraunhofer.de) (max@salsasound.com)
YANNIK GREWE,[1] **ROB OLDFIELD,**[2] **BEN SHIRLEY,**[2,3] **AND ULLI SCUDA,**[1] *AES Member*
(yannik.grewe@iis.fraunhofer.de) (rob@salsasound.com) (ben@salsasound.com) (ulli.scuda@iis.fraunhofer.de)

[1]*Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany*
[2]*Salsa Sound Ltd., Manchester, UK*
[3]*Acoustic Research Center, University of Salford, Salford, UK*

Producing a high-quality audio mix for a live sports production is a demanding task for mixing engineers. The management of many microphone signals and monitoring of various broadcast feeds mean engineers are often stretched, overseeing many tasks simultaneously. With the advancements in Next Generation Audio codecs providing many appealing features, such as interactivity and personalization to end users, consideration is needed as not to create further work for production staff. Therefore, the authors propose a novel approach to live sports production by combining an object-based audio workflow with the efficiency benefits of automated mixing. This paper describes how a fully object-based workflow can be built from point of capture to audience playback with minimal changes for the production staff. This was achieved by integrating Next Generation Audio authoring from the point of production, streamlining the workflow, and thus removing the need for additional authoring process later in the chain. As an exemplar, the authors applied this approach to a Premier League football match in a proof-of-concept trial.

0 INTRODUCTION

In the fast-paced environment of live sports, having an audio workflow that is reliable and efficient is crucial. With the prevalence of immersive audio in high-profile sports events such as the Premier League in the United Kingdom [1–3], it is important to design a workflow that embraces these new technologies while also minimizing additional work for audio engineers. Traditionally, audio has been broadcast in a channel-based approach whereby a fixed number of channels are transmitted to the end user with a defined playback format e.g., 5.1 [4].

In contrast, object-based audio (OBA) aims to preserve individual audio elements as objects accompanied by object-specific metadata along the entire chain, providing greater flexibility in playback and interactivity to the end user [5, 6]. The nature of sports events means they are well suited for an object-based production, with the combination of many different auditory elements—such as on-field, crowd, and commentary components—all able to be interacted with as audio objects to enhance the audience's personalized experience of the game.

However, transitioning to a fully object-based workflow has long been a challenge for broadcasters, with reliance on traditional channel-based systems often preferred. In the demanding environment of live broadcasting, creating further work for production staff with additional complex workflows is not desired. Therefore, creating a solution to alleviate some of these pressures could greatly aid in the adoption of object-based workflows.

The increased proliferation of Artificial Intelligence (AI) has brought about transformative changes across numerous domains. AI for audio [7] has been used for many years for applications including noise reduction, speech recognition, speech synthesis [8, 9], automated music mastering [10], and, more recently, generative music [11]. With the technology's increased prevalence in recent years, its application in audio for live sports productions [12] using audio classifiers could help to improve some of the challenges outlined above.

In this paper, the authors discuss the use of artificial neural networks for automated mixing of live sports broadcast for improved efficiency of OBA productions. The presented solution builds on audio recognition work [13], intelligently

*To whom correspondence should be addressed, email: aimee.moulson@iis.fraunhofer.de. Last updated: February 20, 2024

mixing live microphone inputs in accordance with the detection of salient audio components in the feeds to compose a mix that best captures the sounds of the event.

1 LIVE SPORTS MIXING

The goal of a live sports broadcast is to accurately convey the emotion and suspense of a game, a crucial element of this is through the audio mix. When discussing live sporting events, it is first important to highlight the vast differences between sporting genres. Factors such as location (stadium, indoor/outdoor), sport type (team/individual), and additional sport specific factors (motor, equestrian, sailing, etc.) all play a major role in how the audio engineer will approach a mix [14]. However, principally, there are three main elements to be managed:

1. Diegetic sounds (field-of-play mix)
2. Crowd/Atmosphere
3. Speech/Commentary

Diegetic sounds are typically sounds in a game such as whistle blows, ball kicks, or racquet hits that help audiences follow important cues in a game. In the case of stadium sports such as football, these diegetic sounds are captured using several shotgun microphones positioned around the field of play. To create a stable and coherent mix, the engineer needs to balance all field of play microphones to best capture the diegetic sounds [15]. This is a highly skilled task requiring the mixing engineer to "chase" the ball around the pitch by selecting the closest microphone(s) to the audio event and requiring a high level of concentration and time [16].

Crowd/Atmosphere sounds provide a sense of immersion and "being there" for audiences. These frequently act as the bedrock of the mix providing a level of cohesion between the diegetic sounds and commentary. To capture crowd or atmospheric sounds, mix engineers commonly position multiple microphones or a microphone array at the event with the aim to capture the homogeneous, comprehensive crowd/atmospheric sound. Recent advancements in adoption of 3D audio for sports mean that larger multichannel productions such as 5.1+4H are becoming commonplace [17].

Speech/Commentary provides audiences with a running narrative of the sports event. Subtle vocal cues such as the speed of the commentators' voice or how loud they are speaking help to control the pace and emotion of a game. Ensuring the commentator is audible is of high priority for a mix engineer. Attention is required to defend against speech intelligibility issues, a problem that has long been a cause for audience complaints in broadcasting [18, 19].

By dissecting the key elements of any sports mix, it is possible to see the vast number of different tasks a mix engineer must execute to create a high-quality mix. Furthermore, often these micro tasks happen simultaneously, requiring the engineer to make quick decisions to manage the overall mix. This can lead to an unwanted amount of focus on tasks such as managing the field of play microphones, compared to other necessary tasks.

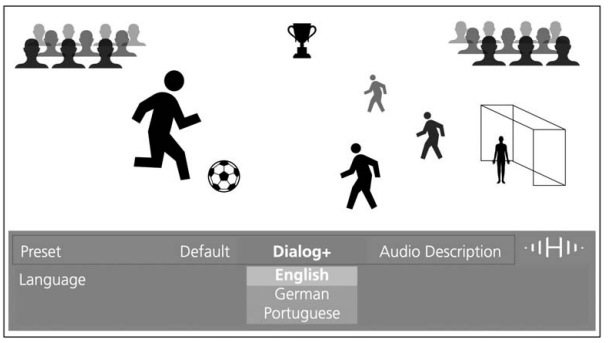

Fig. 1. Schematic of a TV program with user interface for different available presets and audio objects to choose from.

2 NEXT GENERATION AUDIO

In recent years, with the growing popularity of OBA, new codec systems have emerged to support this approach. Next Generation Audio (NGA) codecs, such as DTS:X [20], AC-4 [21], and MPEG-H [22], allow for the transmission of one or a combination of channel, OBA, and scene-based audio. NGA systems provide accompanying metadata from the point of creation, containing information such as content type, object position, and movement [23]. This metadata, defined in NGA codecs, enable enhanced interactivity, personalization, and immersive audio features for audiences.

Due to the transmission of OBA, audio objects, such as dialogue, can be adjusted relative to the background ambience, helping improve accessibility for those who are hard of hearing and provide greater clarity of the narrative for audiences [24, 25]. The use of object switching is also possible with NGA systems, allowing audiences to switch between available language options (see Fig. 1). To enable personalization, pre-defined presets allow audiences to select from a range of options such as "Commentary Off" or "Audio Description" [26]. Furthermore, the transmission of descriptive metadata allows for the separation between channel layout the content is produced in and playback/reproduction channel layout, which can be described as "universal delivery." This is due to the audio objects being rendered at the point of playback. Having a decoupled playback allows engineers to produce one metadata-enriched mix, with the knowledge that the end user's playback device will select the optimal reproduction layout (e.g., 5.1, stereo, binaural, etc.) [27].

Previous trials have demonstrated the benefits of NGA systems such as the cross-collaborative trial conducted by the European Broadcasting Union for the European Athletics Championships in 2018. The trial demonstrated immersive 3D Audio with multiple language options and audio description [28]. Since then, multiple trials have been undertaken [29–31] to refine workflows and improve understanding of the technology. In 2020, the Brazilian Terrestrial System Forum undertook a technical evaluation to consider which NGA system would be implemented into the new TV

3.0 specification. The evaluation saw MPEG-H Audio selected as sole mandatory codec and is expected to be on air in 2025 [32–35]. As the NGA transitions from field trials to implementation, attention is needed on currently available practical solutions to ensure the success and longevity of the technology.

3 AI SYSTEMS

For the system presented in this paper, the authors have implemented an AI/machine learning approach to automated mixing that performs real-time audio event detection on incoming microphone feeds to classify the audio to automate mixing and segment audio objects. The following section provides the background to audio AI and the constraints of real-time implementation.

AI systems involve two key processes: training a model on labeled datasets and applying this trained model to unseen data. The training phase demands sizable, varied datasets for accurate model creation. Some more generic audio AI tasks such as speech recognition can leverage premade datasets [36]; specialized tasks such as specific audio event classification, however, often require custom datasets to be made specifically for the task. Example cases of the target sounds need to be extracted with accompanying labels stating the class/type of sound. The tagging of real-world data in this case is labor-intensive but crucial for accurate model training.

Once the dataset has been acquired, digital signal processing (DSP) is applied to the samples to optimize the data input for the pattern recognition task in question. The choice of DSP depends on the audio content and AI objectives. It is common to utilize representations like spectrograms and Mel Frequency Cepstral Coefficients, rather than temporal descriptors, because they provide an efficient data-rich representation of the incoming audio.

When training the model, the training loop involves processing the incoming data, making predictions, scoring via a loss function, and adjusting the model accordingly. Several metrics are used to monitor the learning process, aiding in optimization, but this can be a fiddly process, and it is important to monitor statistics to achieve the optimal performance. These metrics are vital in training since they allow different versions of the training process to be compared, allowing the designer to optimize the process to achieve the best-performing model. Different versions can be created through changing the dataset, the representation of the data, and the model architecture itself. Due to the complexity of the training process, an AI development framework, such as TensorFlow [37] or PyTorch [38], is often used.

Complications are encountered when running audio AI in a real-time environment because exceeding the time limit imposed by the audio buffer duration in the processing can cause glitches/errors in the output audio. Therefore, the speed of the model is critical, using faster languages, such as C or C++ can help to improve the performance and efficiency of the system. If the TensorFlow framework is being used, the model can be run using the C API but not trained. Therefore, a system must be designed where training is performed in Python and run in C. This can make replication of the DSP data representation very difficult across the two languages.

It is common practice to make real-time audio processing code "real-time safe." This consists of removing code that will not complete within a deterministic amount of time, such as functions including dynamic memory allocation and system calls. Often, it is impossible to be certain that AI code is real-time safe since the implementation is determined by a large framework that was not designed to work within these limitations.

Since an audio glitch is unacceptable in any professional audio production, a fallback option must be implemented in case the AI does not complete its task in time, especially if performing the AI on multiple live audio channels. For instance, it would be preferable to mark a section as not containing the target sound if the processing is taking too long. This potentially comes at the cost of detection accuracy but preserves an error-free audio output, which in this case is more important.

4 AUTOMATED MIXING

For live audio mixing, the AI-based audio event detection holds significant promise for alleviating burdens for engineers across multiple formats [16] through the effective management of many microphone signals with assistance in mixing. If done properly, AI can be a very useful tool for the automatic creation of audio stems from the incoming signals. For example, managing all the discrete field-of-play microphones at a live sports event to produce a single stem typically requires significant experience of the sound engineer and necessitates significant attention.

Automating these processes using AI allows engineers to redirect their attention toward other vital aspects of the mix, such as monitoring and scene authoring. This is especially important in live environments, particularly with the introduction of NGA, which necessitates the separate creation and monitoring of *audio objects* and *beds*. Consequently, existing teams already burdened with responsibilities may face additional workloads when managing channel-based workflows alongside these emerging demands.

To utilize AI for audio mixing tasks, it is imperative that the machine understands mix semantics of the given genre, not only for the training of the AI, as discussed in SEC. 3, but also the mix decisions it informs. Often, the diegetic sounds are the hardest to capture in an object-based framework because they must be captured, segmented, and kept as separate elements throughout the signal/broadcast chain to allow personalization of and interaction with the sound scene.

A system is presented that applies machine learning to the task of automating the detection and separation of key diegetic sounds for football using Salsa Sound's AI-based software platform, MIXaiR. The premise of this approach is to train a deep neural network to recognize particular sounds of interest in the given context (ball impacts and whistle-blows in this context) and to either separate them out as objects and/or automate the microphone faders to create a single "field-of-play" object.

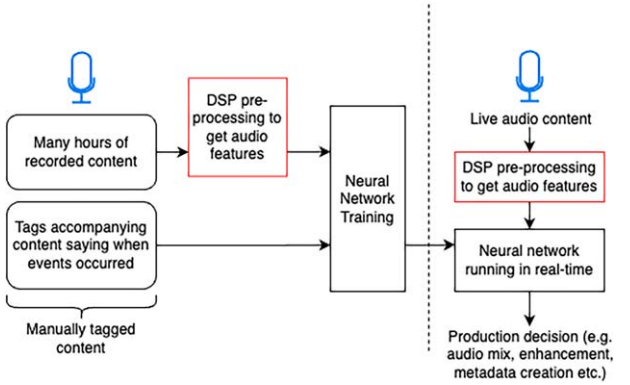

Fig. 2. Machine learning approach for detection and separation of diegetic sounds in the scene.

For this example, the neural network was trained on a large database of ball impact sounds and referee whistle sounds taken from hundreds of broadcast microphones from a variety of different football contexts (large and small crowd/stadium, etc.). Mel spectrograms and multi-band amplitude envelopes provide the input to the neural network with a filter applied for the removal of sounds similar to ball impacts, such as drums. The signal flow diagram is shown in Fig. 2.

This base framework can be adapted to many different genres, on the condition that the neural network has been trained with the target sounds for the given context, e.g., ball bounces and squeaky shoes in basketball, the sounds of a baseball bat connecting with the ball, etc. The mix semantics may change across different sports, but the principle of using an artificial neural network to recognize the narratively important sounds in real time is the same.

For crowd sources, the software is able to manipulate any major spatial audio format and up/down-mix it to the target format of the channel-based output or target bed format for NGA workflows. The output is automatically processed to have the desired dynamic range, level, and characteristics for the given genre and can then be used as a mix stem in later processes.

For the commentary/speech aspects of the mix, automatic dynamic range control runs on the input microphone, and the level is managed intelligently based on voice activity detection. Optionally, noise reduction can be applied to the incoming signal to ensure that the signal-to-noise ratio is not only high but also consistent such that there is no change in the spatial field with the fluctuation in background noise, which can cause the spatial imaging to collapse and expand dependent on the level of crowd noise in the mic feed. It is also possible to manage the mixing of multiple commentary/speech feeds into a single mix or they can be kept separate if required for later points in the signal chain.

Importantly, the output of this stage of the process is a set of audio objects and beds that can either be used to compose channel-based mixes or in NGA scenes as described below.

5 NEXT GENERATION AUDIO INTEGRATION

To combine the benefits of automated mixing and the practicality of having NGA authoring at the point of production, the MPEG-H system was selected for implementation. This section outlines how MPEG-H was integrated into the automated-based mixing software and what considerations were needed to ensure a logical and efficient workflow. The signal path and processing blocks in the automation software discussed can be seen in Fig. 3.

Audio inputs from the chosen sources, such as microphone inputs or interfaces, constitute the first stage in the software's signal chain. These inputs can be routed to different subgroups/stems. The user has the option to select which type of subgroup, pitch, commentary, or crowd, to use based on its intended use. Within each subgroup type is an audio processing block for automated mixing of specific elements of a mix. Each subgroup has a set of processing and control elements specific to that subgroup type, as outlined in SEC. 4.

Subgroups can then be routed to output groups, which output the resultant automated audio mixes. When creating output groups, the user specifies the output group type (such as channel-based or MPEG-H). Any necessary format conversions, such as upmixing or downmixing, are performed automatically at this stage. The user can then route each output group to output channels of their audio hardware. The automation software allows for the addition of signal processing plugins for any of these processing stages and for monitoring through designated channels on their device.

A clear interface is vital in maintaining simplicity, and this is achieved by several rules that are aimed to create a sense of familiarity with the engineer while masking some of the complexities of the tasks they are trying to complete. The left panel view of the interface remains constant, displaying an overview of the signal path with specific controls and information relating to any selected item shown on the right. This exposes the engineer to only the specific information they need at any time. Audio signal flow through the software is always represented through an ordered top-to-bottom display to aid natural understanding of the software signal flow. Audio monitoring is essential to an engineer for ensuring mix quality and diagnosing problems; therefore, any signal stage can be monitored through a monitoring button that is always visible on the left pane of the interface. This allows quick switches between monitored components as well as information as to what is currently being monitored.

MPEG-H integration into the automation software was designed in such a way as to follow the design rules and aims mentioned above. Consistency combined with familiarity in the software lowers the bar of entry to an engineer wanting to create an NGA workflow. A further benefit is that an engineer can work with MPEG-H simultaneously with other desired output formats and mixes. A separate production environment is not necessary to create channel-based mixes at the same time as an NGA mix.

MPEG-H integration simply adds a dedicated output group type designed specifically for MPEG-H. Subgroup/stem inputs to this outgroup become the components of an MPEG-H scene (objects or channels) [3]. These components can be dragged and dropped into user-creatable switch groups and presets to provide viewer personaliza-

Fig. 3. MIXaiR interface with signal flow top to bottom on left panel. Right panel shows edit view of any highlighted left-panel object, in this case, the MPEG-H output group with authoring controls.

tion. If a component is incompatible with one of these destinations, a warning is shown, and it is not added.

Components can be set metadata and interactivity constraints through the software interface. These constraints determine what parameters an end user can control (e.g. gain or position) for an object and can be toggled on and off by the version of that component in any preset. In keeping with other signal stages, MPEG-H presets can be monitored individually through the software's monitoring system for quick quality control of each user option. The output of the authored MPEG-H output group is MPEG-H Production Format (MPF). An MPF consists of 16 audio channels where the 16th channel is a control track containing serial metadata. The control track is a timecode-like audio signal and thus, it can be carried, edited, or recorded like any other audio track. The MPF is easy to integrate into existing serial digital interface (SDI), multichannel audio digital interface (MADI), and Audio over IP (AoIP)–based signal flows that are used for broadcast [39].

Latency management is also a key concern in live broadcasts, especially during events where additional audio feeds are added at different points in the chain. To accommodate this, the software provides the option to adjust the latency of incoming audio signals to compensate and allow alignment of audio to each other. The resulting MPF signal can be frame-synchronized to a video reference using standard broadcast equipment while also maintaining the authored metadata.

6 PROOF OF CONCEPT

Trials were undertaken with the automated-based mixing software to demonstrate what is possible with today's NGA-compatible equipment. The goal was to provide an effective and efficient workflow to ease NGA adoption in the industry while retaining familiarity of mix practices for engineers. The following sections detail how this was undertaken.

6.1 Microphone Plan and Signal Chain

The initial test case for the trial was a Premier League football match. A baseline microphone setup was employed to provide real-time, object-based, immersive mixes. As would be used in a traditional broadcast model, shotgun microphones were placed around the field to capture the field-of-play sounds (see Fig. 4). This was supported by a suspended Ambisonics microphone that was used to create the immersive bed for the broadcast mix. To best utilize the personalization features of NGA, an additional Ambisonics microphone was also employed, positioned nearer to the fans on the right side of the stadium. Both Ambisonics crowd microphones were SoundField ST450 outputting B-Format signals, which were further decoded in the automated mixing software to the target format. The commentary feeds used standard lip microphones.

For this trial, all of the microphone feeds were recorded such that the content could be trialed/demoed offline at a

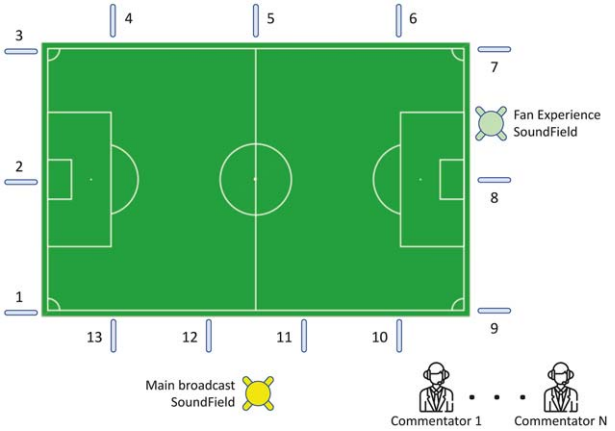

Fig. 4. Microphone set up at the football proof of concept.

Table 1. The chosen MPEG-H presets for the trial along with the individual objects used to create each preset.

Default Preset	Fan Crowd Preset	No Commentary Preset
Broadcast Bed	Broadcast Bed	Broadcast Bed
Pitch FX	Pitch FX	Pitch FX
Commentary Switch Group: -English -German	Commentary Switch Group: -English -German	
	Alternative crowd	

later date with the inputs fed into the automated mixing software as live through a Digital Audio Workstation.

6.2 Production Plan

For this trial, a total of four presets were made available to end users to demonstrate the range of possible mix options. These preset options can be seen in Table 1.

A unique element of the production in this trial was the ability to have an additional crowd presentation for audiences due to the second Ambisonics microphone positioned in the stadium. An additional personalized preset was proposed with the second crowd microphone whereby audiences were then able to listen to an alternative crowd in the stadium providing a "home" and "away" crowd choice. In addition, this production utilized the use of switch groups, allowing end users to select between English and German commentary.

Due to the automated mixing software's integration of MPEG-H, the mixing engineer has the benefit of curating the personalization and interactivity settings for the end user at the point of production. The MPEG-H monitoring in the software provides engineers the ability to quickly monitor how these defined settings sound to the end user in the mix by soloing individual presets. This ensures mixing engineers still have total control over what the audience hears.

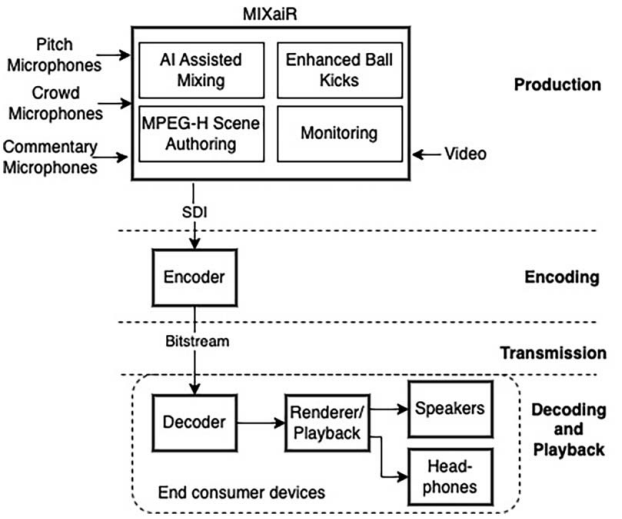

Fig. 5. An overview of an object-based live chain using the proposed workflow.

6.3 Live Chain Example

A live trial of this chain was demonstrated at the International Broadcasting Convention 2022 whereby an end-to-end demonstration of the MPEG-H stream could be viewed in real time [40]. For the purposes of demonstration, the individual microphone feeds recorded at the football match were played via a Digital Audio Workstation and routed to an instance of automated mixing software from a laptop. The output stream was then sent via SDI to a Spin Digital encoder for delivery. Viewers were able to interact with the stream in real time on a playback device (see Fig. 5).

7 FUTURE WORK

The solution presented in this paper focuses on existing SDI-based workflows that support many current broadcast contexts. However, the broadcast landscape is changing, with standard practice moving increasingly toward IP-based workflows and cloud-based production. The proposed SDI-based workflow and corresponding metadata authoring (discussed in SECS. 5 and 6) is fully compatible with future IP-based workflows using transport protocols such as SMPTE 2110 [41, 42]. Audio and metadata would be transmitted in a container over IP using a serialized representation of the Audio Definition Model (ADM) also known as S-ADM [43, 44]. To ensure full interoperability with established NGA content production and distribution systems, MPEG-H uses an ADM Profile that defines constraints on the ADM and Serial ADM streams. For MPEG-H Audio, this is defined in ISO/IEC 23008-3 [45].

To further enhance the system presented in this paper, future work will be undertaken to include the ability to capture and edit the metadata/control track directly from the

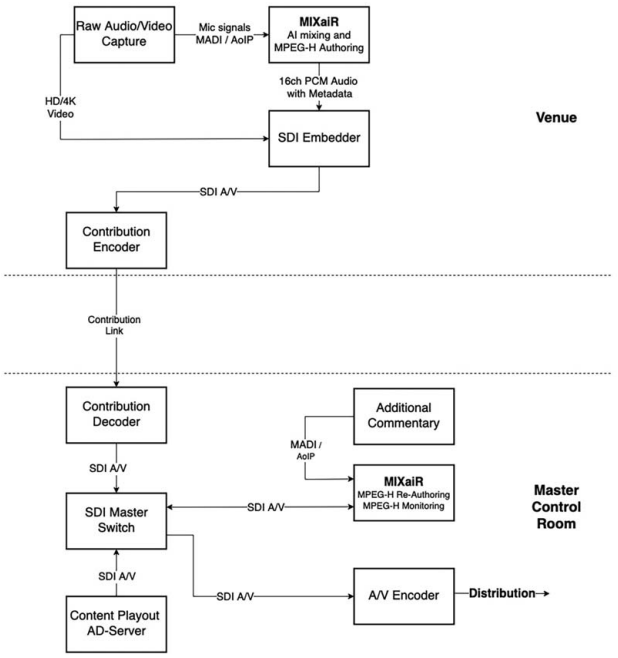

Fig. 6. A proposed workflow for a live production with contribution chain to Master Control Room where additional objects are being added. MADI, multichannel audio digital interface; PCM, pulse-code modulation.

software. This would allow for authoring and re-authoring to happen at different points within the production chain, whereby different audio components may be available (e.g., additional commentaries, audio description, etc.).

Fig. 6 demonstrates how a second instance of the MIXaiR system could be used in the Master Control Room. This would allow an initial production to be made on site with the live microphone feeds and an MPEG-H scene to be authored that could then be edited at subsequent points in the broadcast chain. In addition, the authors plan to also incorporate an encoding and packaging module into the software. This would reduce the need for complex/expensive hardware infrastructure and facilitate a software-only approach, which will be easy to deploy across on-premises and cloud-based environments. Furthermore, Secure Reliable Transport (SRT) and Network Device Interface (NDI) connectivity already built in to the MIXaiR software allows for easy network connection and distribution of content across multiple instances, preserving the control track if carrying PCM.

Lastly, the authors have presented this solution in the context of a football production, but the same software and underlying principles can be applied to any content where there are well-produced stems that can be included as objects and beds within the MPEG-H scene. This can be achieved through the automated mixing engine, which is equally applicable for other sports such as ice hockey and basketball. Aside from live sports, the solution also adds significant value to entertainment and other live broadcast events, giving viewers the ability to interact with and customize their audio presentation.

8 CONCLUSION

The proposed approach to live sports productions demonstrated how the use of automated mixing can be used to reduce pressures on mix engineers by automating certain tasks, such as managing the field of play microphones. A proof-of-concept trial showed how an OBA workflow could be achieved in a live sports broadcast with today's NGA-ready equipment. The result was a live mixed football match streamed with various interactivity and personalization options including the choice of a typical "Broadcast Crowd" or "Away Fan Crowd." These presets provided users with alternative interactivity and personalization options that could not have been achieved using traditional channel-based workflows.

9 ACKNOWLEDGMENT

This journal paper was revised based on an AES paper presented at the International Conference on Spatial and Immersive Audio in Huddersfield [46]. The authors would like to thank Manchester City Football Club and Spin Digital for their respective cooperation.

10 REFERENCES

[1] D. Daley, "Tech Focus: Immersive Sound, Part 1 — The Technology Is Increasingly Integrated Into Broadcast Sports," *Sports Video Group* (2019 Dec.). https://www.sportsvideo.org/2019/12/05/tech-focus-immersive-sound-part-1-the-technology-is-increasingly-integrated-into-broadcast-sports/.

[2] H. Stenzel and U. Scuda, "Producing Interactive and Immersive Sound for MPEG-H: A Field Test for Sports Broadcasting," presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), paper 9211.

[3] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H Audio—The New Standard for Universal Spatial/3D Audio Coding," *J. Audio Eng. Soc.*, vol. 62, no. 12, pp. 821–830 (2014 Dec.).

[4] ITU, "Multichannel Stereophonic Sound System With and Without Accompanying Picture," *Recommendation ITU-R BS.775-4* (2022 Dec.).

[5] R. Bleidt, A. Borsum, H. Fuchs, and S. Weiss, "Object-Based Audio: Opportunities for Improved Listening Experience and Increased Listener Involvement," *SMPTE Motion Imaging J.*, vol. 124, no. 5, pp. 19–20 (2014 Oct.). https://doi.org/10.5594/M001546.

[6] R. Bleidt, D. Sen, A. Niedermeier, et al., "Development of the MPEG-H TV Audio System for ATSC 3.0," *IEEE Trans. Broadcast.*, vol. 63, no. 1, pp. 202–236 (2017 Mar.).

[7] S. Latif, H. Cuayáhuitl, F. Pervez, et al., "A Survey on Deep Reinforcement Learning for Audio-Based Applications," *Artif. Intell. Rev.*, vol. 56, pp. 2193–2240 (2023 Mar.). https://doi.org/10.1007/s10462-022-10224-2.

[8] G. Korvel, P. Treigys, G. Tamulevicus, J. Bernataviciene, and B. Kostek, "Analysis of 2D Feature Spaces for Deep Learning-Based Speech Recognition," *J. Audio*

Eng. Soc., vol. 66, no. 12, pp. 1072–1081 (2018 Dec.). https://doi.org/10.17743/jaes.2018.0066.

[9] A. Triantafyllopoulos, B. W. Schuller, G. İymen, et al., "An Overview of Affective Speech Synthesis and Conversion in the Deep Learning Era," *Proc. IEEE*, vol. 111, no. 10, pp. 1355–1381 (2023 Oct.). https://doi.org/10.1109/JPROC.2023.3250266.

[10] D. Moffat and M. B. Sandler, "Approaches in Intelligent Music Production," *Arts*, vol. 8, no. 4, paper 125 (2019 Sep.). https://doi.org/10.3390/arts8040125.

[11] J.-P. Briot, G. Hadjeres, and F.-D. Pachet, "Deep Learning Techniques for Music Generation – A Survey," *arXiv preprint arXiv:1709.01620v4* (2017 Sep.). http://arxiv.org/abs/1709.01620.

[12] R. Oldfield, B. Shirley, and J. Spille, "Object-Based Audio for Interactive Football Broadcast," *Multimed. Tools Appl.*, vol. 74, pp. 2717–2741 (2015 Apr.). https://doi.org/10.1007/s11042-013-1472-2.

[13] S. Hershey, S. Chaudhuri, D. P. W. Ellis, et al., "CNN Architectures for Large-Scale Audio Classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135 (New Orleans, LA) (2017 Mar.). https://doi.org/10.1109/ICASSP.2017.7952132.

[14] U. Scuda, H. Stenzel, and D. Baxter, "Using Audio Objects and Spatial Audio in Sports Broadcasting," presented at the *57th Convention of the Audio Engineering Society* (2015 Mar.), paper 5-2.

[15] D. Baxter, *Immersive Sound Production: A Practical Guide* (Focal Press, Waltham, MA, 2022).

[16] R. Oldfield, B. G. Shirley, and D. Satongar, "Application of Object-Based Audio for Automated Mixing of Live Football Broadcast," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), paper 9454.

[17] N. Duarte, "Immersive Audio in Large-Scale Sports Production," IBC 2022 Tech Paper (2022 Oct.). https://www.ibc.org/technical-papers/ibc2022-tech-papers-immersive-audio-in-large-scale-sports-production/9232. article.

[18] M. Armstrong, "From Clean Audio To Object Based Broadcasting," BBC White Paper WHP 324 (2016 Oct.).

[19] C. Mathers, "A Study of Sound Balances for the Hard of Hearing," Tech. Rep. 1991-03 (1991 Jan.).

[20] ETSI, "DTS-UHD Audio Format; Delivery of Channels, Objects and Ambisonic Sound Fields," *ETSI TS 103 491* (2017 Apr.).

[21] ETSI, "Digital Audio Compression (AC-4) Standard; Part 2: Immersive and Personalized Audio," *ETSI TS 103 190-2* (2018 Feb.).

[22] ISO/IEC, "Information Technology - High Efficiency Coding and Media Delivery in Heterogeneous Environments - Part 3: 3D Audio," *ISO/IEC Standard 23008-3* (2019 Feb.).

[23] EBU, "Recommended Strategy for Adoption of Next-Generation Audio (NGA)," *Recommendation EBU-R.151* (2018 Dec.).

[24] L. Ward, "Casualty Accessible and Enhanced (A and E) Audio: Trialling Object-Based Accessible TV Audio," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), eBrief 563.

[25] L. A. Ward and B. G. Shirley, "Personalization in Object-Based Audio for Accessibility: A Review of Advancements for Hearing Impaired Listeners," *J. Audio Eng. Soc.*, vol. 67, no. 7/8, pp. 584–597 (2019 Aug.). https://doi.org/10.17743/jaes.2019.0021.

[26] J. Paterson and H. Lee (Eds.), *3D Audio* (Routledge, New York, NY, 2021).

[27] Y. Grewe, P. Eibl, C. Simon, et al., "MPEG-H Audio Production Workflows for a Next Generation Audio Experience in Broadcast, Streaming and Music," presented at the *151st Convention of the Audio Engineering Society* (2021 Oct.), eBrief 649.

[28] F. de Jong, D. Driesnack, A. Mason, et al., "European Athletics Championship: Lessons from a Live, HDR, HFR, UHD, and Next-Generation Audio Sports Event," *SMPTE Motion Imaging J.*, vol. 128, no. 3, pp. 1–10 (2019 Apr.). https://doi.org/10.5594/JMI.2019.2895582.

[29] Fraunhofer-IIS, "Football Fans Around the World Experience the Worldcup in Immersive and Personalized MPEG-H Audio," https://www.iis.fraunhofer.de/en/pr/2022/20221208_mpeg-h_audio_worldcup.html (2022 Dec.).

[30] M. Firth and D. Marston, "Live Next Generation Audio Trial at Eurovision," https://www.bbc.co.uk/rd/blog/2023-06-eurovision-next-generation-audio (2023 Jun.).

[31] A. Turnwald, C. Simon, and U. Scuda, "Eurovision Song Contest 2018 - Immersive and Interactive," presented at the *30th Tonmeistertagung* (Düsseldorf, Germany) (2018 Nov.).

[32] ABNT, "Digital Terrestrial Television. Video Coding and Multiplexing, Part 2: Audio Coding," NBR 15602-2:2020 (2020 May).

[33] SBTVD, "Testing and Evaluation Report: TV 3.0 Project - Audio Coding," Brazilian Ministry of Communications (2021 Dec.).

[34] A. Murtaza, S. Meltzer, Y. Grewe, et al., "MPEG-H Audio System for SBTVD TV 3.0 Call for Proposals," *SET Int. J. Broadcast Eng.*, vol. 7, pp. 30–46 (2021 Dec.). https://doi.org/10.18580/setijbe.2021.3.

[35] C. Cosme, "Next Generation TV in Brazil during the Football World Cup 2022," https://www.ibc.org/ibc2023-tech-papers-next-generation-tv-in-brazil-during-the-football-world-cup-2022/10240. article (accessed).

[36] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," *arXiv preprint arXiv:1804.03209* (2018 Apr.).

[37] M. Abadi, A. Agarwal, P. Barham, et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," White Paper (2015 Nov.). http://download.tensorflow.org/paper/whitepaper2015.pdf.

[38] A. Paszke, S. Gross, F. Massa, et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *arXiv preprint arXiv:1912.01703* (2019 Dec.).

[39] P. Eibl, Y. Grewe, D. Rieger, and U. Scuda, "Production Tools for the MPEG-H Audio System," presented at the *151st Audio Engineering Society Convention* (2021 Oct.), eBrief 651.

[40] H. McLean, "Salsa Sound Hits World First at IBC 2022 With True Object-Based Live Audio Automation With Fraunhofer IIS," *Sports Video Group* (2022 Sep.). https://www.svgeurope.org/blog/headlines/salsa-sound-hits-world-first-at-ibc-2022-with-true-object-based-audio-alongside-fraunhofer/.

[41] SMPTE, "Professional Media Over Managed IP Networks: SMPTE ST 291-1 Ancillary Data," *SMPTE Standard ST 2110-40:2018* (2018 Apr.).

[42] A. Hilber, A. Weiss, and M. Brockmann, "Dynamic Metadata in an SMPTE ST 2110 Environment. A Proof of Concept," EBU Tech. Rev. (2021 Dec.).

[43] ITU, "Audio Definition Model," *Recommendation ITU-R BS.2076-2* (2019 Oct.).

[44] ITU, "A Serial Representation of the Audio Definition Model," *Recommendation ITU-R BS.2125-1* (2022 May).

[45] ISO/IEC, "Information Technology — High Efficiency Coding and Media Delivery in Heterogeneous Environments — Part 3: 3D Audio," *ISO/IEC Standard 23008-3:2022* (2022 Aug.).

[46] A. Moulson, M. Walley, Y. Grewe, et al., "Object-Based Workflows in Live Sports Broadcasting Using AI-Based Mixing," in *Proceedings of the AES International Conference on Spatial and Immersive Audio* (2023 Aug.), paper 31.

THE AUTHORS


Aimée Moulson


Max Walley


Yannik Grewe


Rob Oldfield


Ben Shirley


Ulli Scuda

Aimée Moulson graduated from the University of Huddersfield with a degree in Music Technology and Audio Systems. After working as an engineer in the broadcast industry, Aimée received her M.Sc. in Broadcast Engineering from Birmingham City University. Since 2021, she has been working at Fraunhofer IIS as a Sound Engineer, producing immersive and object-based content for the MPEG-H ecosystem. She specializes in immersive audio for sports productions and systems integration for Next Generation Audio workflows.

•

Max Walley is the lead software developer and a researcher at Salsa Sound. He graduated with first class honors in Broadcast Audio and Music Technology from the University of the West of England. His work is primarily focused on development of large-scale audio applications and libraries based around AI and DSP for use in the broadcast sector. He also has research interests in network audio, hardware emulation, and broadcast metadata technologies.

•

Yannik Grewe received his M.Eng. degree in "audiovisual media–sound" with a thesis on perception and reproduction of floor level sound. He joined Fraunhofer IIS in 2013 and serves today as senior manager for media technologies and business development, focusing on the MPEG-H 3D Audio ecosystem. His research in 3D-Audio production and reproduction technologies has resulted in several publications and conference contributions. He is involved in producing immersive, object-based music applications and immersive and interactive audio for major live events, such as the European Athletics Championships, Eurovision Song Contest, Rio de Janeiro Carnival, Rock in Rio, or Football World Cup Qatar 2022.

•

Rob Oldfield has worked in the audio/broadcast industry for over 17 years and gained his Ph.D. from the University of Salford in object-based, immersive audio in 2013. He is passionate about innovation, sports, and broadcast. Rob is co-founder of Salsa Sound, which has developed a patented AI engine that automates and enhances audio mixing for live sports broadcast. Salsa's technology improves mix quality, consistency, and production efficiency.

•

Ben Shirley is Co-Founder and Director of Salsa Sound and Associate Professor in Audio Technology at the Acoustics Research Centre, University of Salford, UK. He received his Ph.D. from the University of Salford in 2013 on methods for improving TV sound for people with hearing impairments. Dr. Shirley's research activity at Salford includes personalizable broadcast audio and audio accessibility solutions for people with sensory impairments.

•

Ulli Scuda obtained his degree as a sound engineer and his Ph.D. at the Film University Potsdam-Babelsberg. His practical experience in various playback formats, from binaural to wave field synthesis, eventually led him to Fraunhofer IIS, where he currently leads the SoundLab group in the field of audio and media technologies. There, he primarily supports developments in the areas of Next Generation Audio, immersive audio, and object-based audio.