OBJECT-BASED AUDIO FOR LIVE SPORTS AUDIO

RG Oldfield Salsa Sound Ltd. Manchester, UK
BG Shirley University of Salford, Salford, UK

1 ABSTRACT

Object-based and so-called Next Generation Audio (NGA) is beginning to gain significant traction for both motion picture and also increasingly in broadcast as content creators and consumers are looking for more immersive and customizable audio experiences. Of particular interest in the broadcast context is how to utilize audio objects in live sports to facilitate these experiences and enable content that can be played back over traditional systems and also for 360 video video/VR contexts. There are however significant challenges for the use of audio object techniques for such a live scenario primarily in terms of the capture of the individual audio objects at the production end and also the best rendering paradigms at the user end.

In this paper we overview NGA theory and the current state-of-the-art in terms of audio object capture techniques and present an investigation into the application and challenges of NGA for live sports production.

2 INTRODUCTION

The emergence of so-called Next Generation Audio (NGA), and object-based audio (OBA), requires a new set of production tools to enable the capture of the audio scene with individual sources captured as separate 'objects' and tagged with metadata. For motion picture, where sources are generally captured separately or created in post-production, the creation of audio objects is achieved predominantly with non-live workflows and is relatively straightforward. Applying NGA principles to live contexts however introduces several logistical difficulties as in a live scene it is most likely not possible to have discrete microphones for each of the sources (which are often transient and unpredictable in nature), making source separation and the generation of metadata describing the scene difficult.

For example, currently, for a standard sports broadcast, the sound engineer will typically mix the signals from the available microphones around the field-of-play dynamically to produce a channel-based mix intended for a target reproduction setup. This approach does not allow the separation of sound sources and provides little or no metadata/information about the captured content. New capture and production methodologies therefore need to be employed to allow the sound scene to be split into its individual components and each one tagged with the necessary metadata (duration, location, event type etc.) for manipulation later on. In a complex scene such as football match this is a non-trivial task.

In this paper, we present a solution for live sports production for NGA broadcast. Our approach utilises audio pattern-matching algorithms to automatically detect and extract audio objects from the scene using an artificial neural network. Using only the microphones that are present in a standard broadcast environment, the Salsa system automatically picks out only content that matches audio signatures of events considered salient in the context (ball-kicks, racquet hits, referee whistle-blows etc.). In the analysis engine Salsa uses time-difference-of-arrival triangulation to provide object location and generates other metadata as described below. Additionally Salsa is able to use the signal statistics of the extracted sounds to apply appropriate processing to result in 'clean' audio objects with greatly reduced crowd noise.

2.1 Next Generation Audio

Traditionally broadcast audio has adhered to a channel-based paradigm where content producers have created audio mixes for reproduction over specific loudspeaker setups such as stereo, 5.1, 7.1 etc. Such a channel-based system thus offers very limited opportunities for interactivity, personalisation and immersion and consequently, in recent years, there has been an appetite for more flexible and immersive audio technologies[1] termed 'Next Generation Audio' or NGA. NGA technologies aim to preserve as much of the audio scene as possible all the way through the production chain such that the end-user can create a personalised mix based on their own reproduction equipment and requirements. For example, using NGA a user may be able to select components of the audio scene that they would like to be emphasised or diminished, or to choose a preferred commentary feed, e.g. in an alternate language, or biased toward a specific team. Other personalisation can include the levels of audio objects such as the commentary and background noise that can be particularly useful in facilitating accessible audio solutions, such as for hearing impaired viewers[2] or switching between different crowd feeds.

NGA can be considered to consist predominantly of 3 key technologies namely, channel-based, scene-based (ambisonics) and object-based audio. NGA aims to capture as much information of the audio scene as possible at the production end and retain it's indivisula components throughout the entire broadcast workflow such that it can be manipulated later and optimised for reproduction on whatever reproduction equipment is available to the viewer.

Channel-based Audio

Channel-based audio is the paradigm currently used for broadcast and defines the audio scene with reference to a specific loudspeaker layout (2.0, 5.1, 9.1 etc.) where each channel defines the feed for the corresponding loudspeaker and therefore offers very limited manipulation at the receiver end. In order to maintain existing production workflows, and in order not to break existing good practice, channel-based audio remains a key component of NGA systems with the components usually referred to as a channel 'bed' or a 'channel object' that can be processed/manipulated independently of other objects.

Ambisonics

Ambisonics dates back to the 1970's with the work of Michael Gerzon[3] and is a means of describing an audio scene in terms of its spherical harmonic components such that the scene can be efficiently captured/encoded, transmitted and then decoded for reproduction. The more recent introduction of higher order ambisonics (HOA) brings increased spatial resolution to the format[4] and is the preferred audio format for several popular 360° video platforms. Interaction within an ambisonics scene is generally limited to rotation and, to some extent, zoom and in some NGA descriptions[5] is considered to be an 'object' (see below) or forms a 'bed' on to which other sources can be superimposed.

Object-based audio

The concept of so-called 'object-based audio'[1] is to separate an audio scene into its individual components (often called 'audio assets' or 'objects'). Although terminology can vary across the industry, the audio components are essentially the discrete sound sources that make up the scene

and are often superimposed on to a 'bed' that could for example be a 9.1 channel or ambisonics mix of the ambient components within the scene. An object-based paradigm can feasibly contain any or all of the aforementioned technologies with 'objects' being a broad term that can include discrete sound sources, HOA and channel-based content. For the purpose of this paper we will focus our attention on an object-based paradigm seeking to define the discrete audio sources/objects within the scene.

2.2 NGA experiences

NGA allows a move towards content that can be tailored to the specific rendering environment, whether that be for a specific reproduction system, for a listener demographic (e.g. hearing impaired) or to allow altering the audio render to adapt to a user or production-driven visual viewpoint. The ability of NGA rendering to be adapted to a specific context is one of the key market drivers and one that can provide end-users with perhaps the most significant improvement in the service offering. Broadly speaking, adaptive content falls in to two main categories (*interaction* and *personalisation*).

2.2.1 Interaction

With the increased proliferation of 360° video content, it is becoming more important than ever to perform spatial audio rendering of a scene to match the immersive video content such that the end-user is more fully able to 'buy in' to the immersive experience that is being presented. NGA technologies allow automated interaction with the content such as the aforementioned 360° video scenario and also allows for the manual interaction such as moving sources around in broadcast gamification or just to suit the desires of the viewer.

2.2.2 Personalisation

NGA technologies provide potential for users to move away from one-size-fits-all content consumption and to tailor their media experiences according to their requirements or needs. Having all of the audio components of a scene available as separate objects means that the user can be allowed to change the relative levels of sources (e.g. the level of the commentator or sound effects), to select alternate commentaries in other languages or from the perspective of their team. NGA can also facilitate accessibility by enabling remixing of the audio content based on hearing impairment, or enabling an audio description audio object for visually impaired and blind viewers.

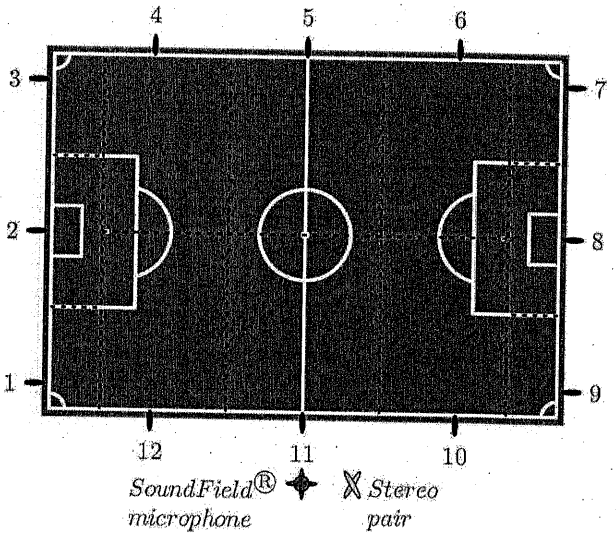

Figure 1 – Example microphone setup for a standard broadcast of a football match where numbers 1-12 represent shotgun microphones

The addition of further crowd microphones also allows for content to replayed to match the experience from different perspectives such that viewers could 'sit' with either the home fans or the away fans. If these crowd 'objects' are captured spatially e.g. using ambisonics microphones these custom crowd perspectives can be rendered spatially accurately over whatever system is available at the reproduction end.

2.3 NGA capture

One of the biggest tasks facing content producers as they seek to embrace NGA technologies is how to capture the scene in such a way that as much content and information about the content is preserved as possible while retaining well established production workflows. In addition to the audio content being captured, it is imperative that substantial metadata describing the content is also captured or extracted from the scene to enable correct rendering at the user end. There are existing solutions for ambisonic capture, e.g. using first order ambisonic (FOA) microphones such as the SoundField or HOA microphones like the Eigenmike, but capture of discrete sources in an object-based format requires new microphone and processing techniques[6]. In the following sections of this paper we describe a novel solution for analysis, processing and extraction of audio objects and metadata in the context of sports broadcast.

3 NGA FOR SPORTS BROADCAST

One of the biggest commercial drivers for NGA technology is for live sports broadcast and in recent years several companies have been running trials to attempt to show how NGA can be integrated in to current workflows. Up to now these trials have been predominantly about converting a standard broadcast mix in to an object-based transport stream to provide an immersive audio experience, putting the viewer 'in the crowd' at the big game, and not about the capture of the individual objects. In the context of live sports it is a non-trivial task to be able to separate out individual sound sources: in capturing a live football game for example, the players are not close-miked and there is generally a high level of background noise from supporters in the stadium.

3.1 Extracting objects

It has been shown that audio objects from a complex scene can be extracted from standard pitch-side microphones at a football match (see Figure 1) using signal processing techniques[7]. The basic operation of the algorithm is to scrutinise each of the pitch-side microphones to see whether they contain any content, which is considered salient to the present context as defined by an acoustic feature template for each desired sound that may be present on the field-of-play. The specifics of the acoustic feature template will be dependent upon the sounds of interest to be extracted in the current context. For example when scrutinising microphone feeds for the sound of a referee's whistle the template would look for harmonic content with a fundamental in the region of frequencies of a standard referee's whistle (~4kHz), using the signal cepstrum, and will extract the sound using appropriate thresholding.

Another example could be the ball-kick where the sound is characterised by broadband transient energy with specific attack and decay durations as per Figure 2. The template can contain any signal feature which can be easily measured in the incoming signal and that describes the significant features of the signal in question.

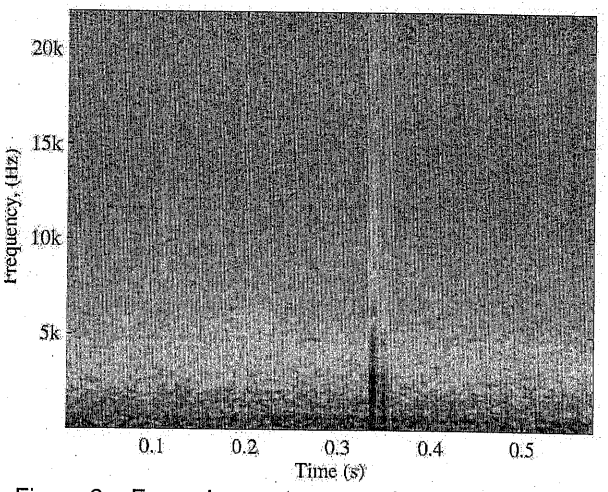

Figure 2 – Example spectrogram of a football ball-kick showing broadband transient energy

3.2 Generating metadata

In an object-based paradigm the creation of metadata is vitally important to give semantic meaning to the content/scene description and to facilitate adaptive rendering. Whilst the metadata can cover almost any aspect of the scene, of primary importance are the location of the audio objects in the scene, identifiers as to object

type and the start and end points of when the audio objects are active.

3.2.1 Extracting source type

As the processing algorithm is analysing the audio ingest for content matching a specific acoustic feature template, it is able to write to the metadata stream the type of source that has been detected and hence what sources are within the scene e.g. referee's whistle, crossbar hit, header, kick, racquet-hit etc. This metadata can be a powerful source of scene analysis to inform the audio production workflow, for example to apply specific processing/enhancement to content of a specific type. So a filter that enhances ball-kick sounds, applies signal compression or other effects that should only be applied to sources of a specific nature.

3.2.2 Extracting temporal duration

Depending upon which object-based codec is to be employed within the NGA framework it can be beneficial to generate metadata indicating the duration of each of the objects within the scene. This is particularly useful if it is desired to reduce the bit-rate of the broadcast stream as an object needs only be encoded when it is active in the scene, giving rise to so-called *short-term audio objects*. These short-term objects can be created or destroyed dynamically within the broadcast, reducing the need to transmit objects (e.g. a microphone channel) for the duration of the broadcast when it only contains salient information for limited durations, thus reducing audio compression redundancy. The algorithm presented here enables accurate time stamps to be determined for each of the audio objects within the scene and these can act as metadata stamps to the audio content.

3.2.3 Extracting object location

One of the most significant challenges and important aspect of the metadata creation is the determination of the location of the audio objects within the scene. This is important for any spatial rendering at the receiver end or in a 360° video setting where sources surround the listener.

In non-real time contexts it may be possible to add tracking devices to a close miked source which could provide a real-time dynamic description of the source location. However for sports broadcast this is seldom possible, so sound sources within the scene must be extracted using the sensors available at the scene.

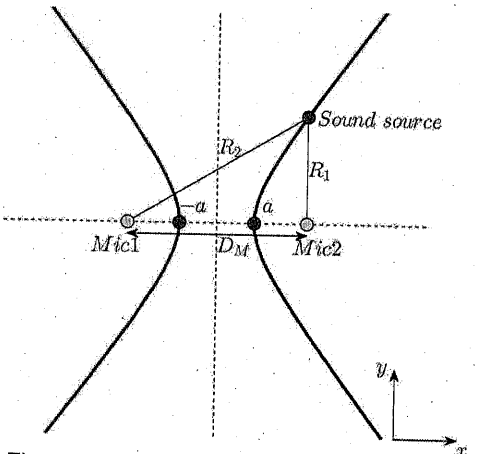

Figure 3 – Source geometry for localisation algorithm

Within the audio object extraction framework presented in this paper, time-difference-of-arrival audio triangulation is utilised to locate the position of the sound source on the field-of-play. Once sound events are identified as important by the use of acoustic feature templates, the algorithm determines if the same event is detected in more than one microphone. The time-difference-of-arrival (TDOA) between microphones can then be used to perform a triangulation to locate the source. If the sound source is detected in two microphones, the TDOA between the signals received at each microphone can be computed using the cross-correlation between the audio feeds received at each microphone. This TDOA allows the calculation of the relative distances from the source to each microphones, as given by equation 1 where the geometry is given by Figure 3.

$$R_1 = R_2 + c\Delta t = R_2 + \Delta R \qquad 1$$

Where Δt is the time-difference-of-arrival between the sound arriving at *Mic 1* and *Mic 2* and *c* is the

speed of sound. There are many source positions that satisfy equation 1, plotting the possible source positions, reveals that they can be located along a pair of hyperbolae between the two microphones describing a set of all points in a plane such that the difference of the distances from two fixed points (foci) is constant. In this case the foci are the microphone positions. It can be shown that for a hyperbola the relative distances between R_1 and R_2 is given as

$$\Delta R = R_2 - R_1 = 2a \qquad 2$$

Where *a* is the absolute value along the x-axis of the vertices of the hyperbola. The equation for the hyperbola is then given as

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \qquad 3$$

Where *b* is derived from the asymptotes of the hyperbola as:

$$b = \sqrt{\frac{D_M{}^2}{4} - a^2} \qquad 4$$

Thus a hyperbola can be plotted taking the positive solution of *x* along which the source will be situated.

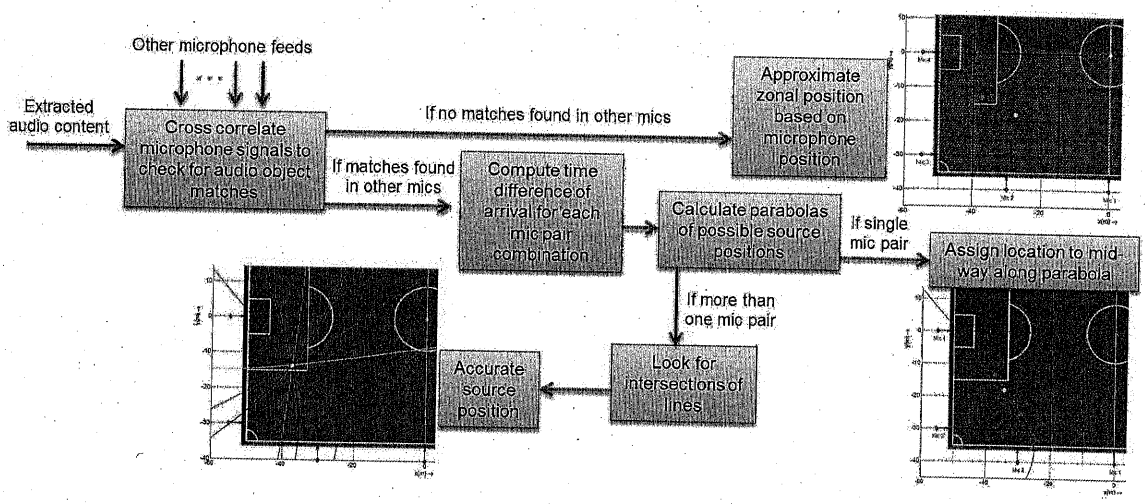

Figure 4 – Localisation schemes dependent on number of microphones detecting audio event. Cyan & yellow dots represent true and estimated source positions respectively

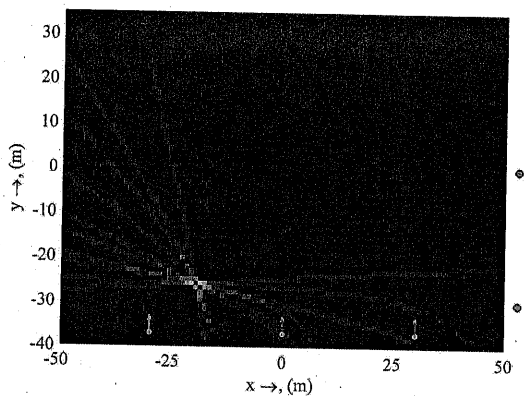

Figure 5 – Efficient source using a triangulation heat map approach

It should be noted that this derivation constitutes a specific geometry where the two microphones are along the x-axis and equally spaced around the origin, thus for microphones not at these positions a coordinate transform is required to plot the correct hyperbola.

There are a number of different localisation scenarios depending upon how many microphones have detected the source as shown in Figure 4. If only one microphone picks up the source, then there is insufficient information to localise it so it is positioned approximately in the centre of the pick up region of the microphone. If two microphones pick up the source it is possible to assume a source location along the mid-point of the hyperbola as defined by equation *3*. The best-case scenario is that the source is detected in three or more microphones such that a complete triangulation can be performed. In order to do this a means of determining the points of intersection of multiple hyperbolae needs to be employed. In a real-time situation this can be a complex task, so here a more pragmatic and computationally efficient methodology is adopted where a grid of potential positions is defined over the field-of-play. This forms a grid of zeroes and each time a hyperbola is plotted over this grid the value of the cell in which the hyperbola crosses is increased by one such that after multiple hyperbolae are plotted there forms a heat map of potential source positions as shown in Figure 5 the maxima of this heat map corresponds to the source location. Increasing the size of the pixels in the grid increases the chance of intersection of the hyperbolae and although decreasing the granularity of the source location, increases the efficiency of the algorithm. Once the location of the sources within the scene have been ascertained, this information is written to the metadata stream and can be used within the viewer's renderer for e.g. source positioning within the sound stage or for correcting for the temporal offset between audio and video events resulting from the discrepancy between the speed of light and sound.

3.3 Capturing the crowd

As mentioned above, any scene needs to have both the audio objects and also the crowd components or 'bed'. In the context of a sports broadcast the true sense of immersion for the audience depends somewhat upon the capturing and rendering of the crowd 'object'. In this context the crowd is captured using a SoundField ambisonics microphone such that the B-format signals can be decoded to fit any target reproduction system.

Of particular interest is the capturing and rendering of the crowd from more than one crowd perspective. This allows for the optional rendering of the audio from the point of view of the home or away fans for example. This can help fans engage more with the content more as they can 'sit' with the fans from the team that they are supporting. From an outside broadcast perspective, it provides an additional overhead in that more than one crowd mix needs to be captured but it allows for regional customisations to be made easily such that a single outside broadcast truck can meet the broadcast needs of multiple audiences without having a rigid one-size-fits-all paradigm.

4 CONCLUSIONS

This paper has described how so-called Next Generation Audio techniques can be applied to live sports broadcast, resulting in audio content that can benefit from personalisation, interactivity and a greater degree of immersion. In order to facilitate the NGA paradigm it is important that new capture techniques are utilised, enabling the separation of the audio sources within the scene and the creation of metadata to describe the sources within the scene. This paper has presented an algorithm that is capable of the automatic extraction of the audio and writing of metadata describing the scene applied to a football context.

Although this paper has focused primarily upon football as an exemplar case, the techniques apply to many sports, where audio templates can be created and the content from the available microphones can be scrutinised for matching content. When matches are found, intelligent production choices can be made to produce audio objects and extract their metadata.

5 REFERENCES

1. Shirley, B., et al., 2014, Platform Independent Audio, in Media Production, Delivery and Interaction for Platform Independent Systems: Format-Agnostic Media, Wiley: UK.
2. Shirley, B. and Oldfield, R., 2015, Clean audio for TV broadcast: An object-based approach for hearing-impaired viewers. J. Audio Eng. Soc., 63(4): p. 245-256.
3. Gerzon, M.A., 1985, Ambisonics in multichannel broadcasting and video. J. Audio Eng. Soc., 33(11): p. 859-871.
4. Bertet, S., et al., 2013, Investigation on localisation accuracy for first & higher order ambisonics reproduced sound sources. Acta Acustica united w. Acustica, 99(4): p.642-657
5. Herre, J., et al., 2015, MPEG-H audio—the new standard for universal spatial/3D audio coding. J. Audio Eng. Soc., 62(12): p. 821-830.
6. Oldfield, R., B. Shirley, & J. Spille., 2014, An object-based audio system for interactive broadcasting. in 137th Conv. Audio Eng. Soc.
7. Oldfield, R.G., B. Shirley, & J. Spille, 2015, Object-Based Audio for Interactive Football Broadcast. Multimedia Tools & Applications, 74(8), p. 2717-2741