FORMAT AGNOSTIC RECORDING AND SPATIAL AUDIO REPRODUCTION OF FOOTBALL BROADCASTS FOR THE TELEVISION

R G Oldfield University of Salford, Salford, UK
B G Shirley University of Salford, Salford, UK
J Spille Technicolor, Hannover, Germany

1 INTRODUCTION

Traditionally the audio for television broadcasts of football is recorded using 12 shotgun microphones positioned around the pitch for the on-pitch sounds with a Soundfield® microphone or stereo pair used for the ambient crowd noise. It is the sound engineer's job to control the level of the crowd noise in the mix; as such they will only raise the fader of a shotgun microphone when the play is near by. We present an algorithm that performs this mixing process automatically. The algorithm extracts only the key on-pitch sounds (ball kicks and whistle blows) and determines their position on the pitch. The extracted 'audio objects' can then be positioned in space for reproduction using any spatial audio system. This corresponds to a paradigm shift for such broadcasts where on-pitch sounds are currently panned to front centre. This work forms part of the EU funded project, FascinatE, which aims at more interactive and immersive broadcasting and which is format agnostic.

2 STANDARD PRACTICE FOR TV BROADCAST OF FOOTBALL

The broadcast of football on the television typically has a dynamic visual scene accompanied by a static audio scene. The dynamic visual scene consists of several cuts made between a variety of cameras with varying pan, tilt and zoom as the director chooses. The stationary audio scene is a mix of the commentary feed, the pitch sounds and an ambient crowd recording. As a standard, the pitch sounds and commentary are panned front central with the crowd noise (recorded in stereo, 5.1 or ambisonics) providing the only spatial aspect. The audio is static because it is unaffected by camera cuts, pans and zooms. The crowd noise is most often recorded using a single Soundfield® ambisonic microphone or a stereo pair suspended high above the crowd so as to capture the ambience rather than individuals in the crowd. Sounds on the field of play are recorded using twelve highly directional shotgun microphones arranged around the pitch as shown in Figure 1. Whilst these microphones are chosen to be as directional as possible, they still pick up much of the crowd noise, either from the rear of the microphone or from the crowd on the opposite side of the pitch. If all of these microphones were left high in the mix the result would be an audio mix that was just a wash of crowd noise. If left low in the mix there would be no on-pitch sounds, which would be unrealistic. Consequently the sound engineer will track the action on the pitch and will only raise the levels of a microphone when the action is near by and there is likely to be some on-pitch sounds to pick up. This process can be laborious for the sound engineer and also means that the only sounds that are picked up are the ones around the main action which may not tell the whole story of the match. If for example one player shouts for his team mate to pass the ball at one end of the pitch they would not be picked up by the microphones as the action would have not yet reached that point on the pitch. Also if there were to be an altercation between two players or another event auxiliary to the play during the game, the corresponding sound would be unlikely to be picked up using this approach. A further problem can be the sound engineer's reactions; if the play switches quickly from one end of the pitch to the other he/she will have to quickly move one set of faders down and the others up, if this is not done in time the audio will not get picked up.

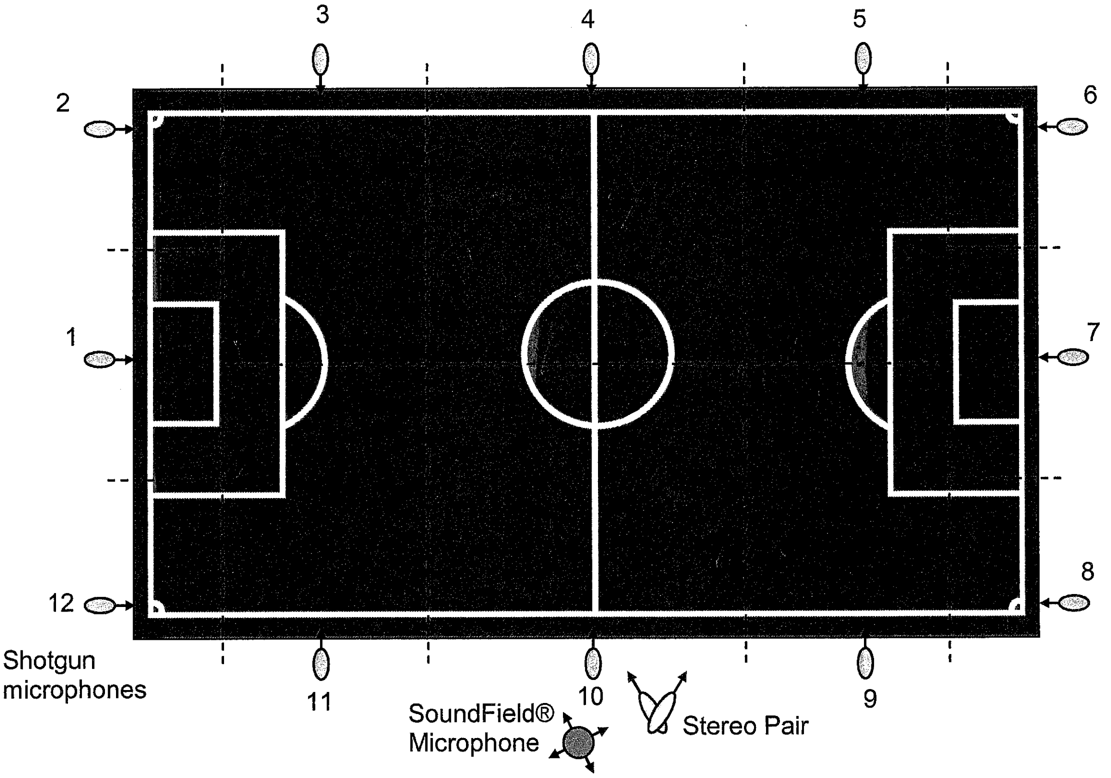

Figure 1. Typical Microphone setup for and English Premier League match (dotted lines show the principal capture zone for each microphone)

Automatic mixing applications have been suggested in previous research (e.g. Cengarle et al[1]). These require that the engineer tracks the position of the on-pitch action using a remote device such as a tablet PC and then the microphone signals are automatically mixed from this data. The authors of this paper have also proposed an automatic mixing algorithm that does not require manual tracking[2]. From the audio feeds alone, the algorithm detects when a significant on-pitch sound has occurred and adds the corresponding microphone feed into the mix accordingly. In this paper we present an algorithm that can not only automatically mix the on-pitch audio but can extract the audio content, position and onset time of the on-pitch sounds for extraction as so called 'audio objects' which can be transmitted in place of the traditional audio mix (which is specific to a given loudspeaker setup). This new approach allows multiple on-pitch sounds at any time and also allows a spatially dynamic mix during reproduction to match the dynamic visual scene enabling more immersive viewing and allowing an interactive audio experience. Moreover the output can be spatially mixed for any audio reproduction format at the user end. This facilitates a format agnostic rendering which is one of the goals of the EU funded FascinatE Project[3] of which this research forms a part.

3 THE FASCINATE PROJECT – A NEW PARADIGM FOR TELEVISION BROADCASTING

The FascinatE Project, funded under EU FP7, is developing an end to end AV broadcast system that will provide an immersive, interactive experience for live events. The system will allow users to create and navigate their own user-defined AV scene based on their preferences, production choices or by free pan and zoom control. FascinatE stands for Format-Agnostic SCript-based INterAcTive Experience and has the objective of allowing a completely customisable viewing experience where the viewer will be able to make choices as to which area of interest he/she would like to view on the pitch and they will have liberty to navigate around the visual scene by zooming, panning etc. For the audio side of the project the sound has to match the visual content and therefore it is important that a realistic and robust audio scene be recorded for quality spatial reproduction whatever the user's navigational decisions[4]. FascinatE attempts to transfer control operations from the production side to an end-user terminal. This enables a better adaptation not only for the audience but also for different end user devices. These can vary from modern displays in connection with large loudspeaker setups down to mobile devices combined with headphones, thus FascinatE provides an interactive experience and is also format agnostic. By making production operations at the user end, the FascinatE scene composer needs to broadcast the necessary components for the scene to be recreated at the user end. In terms of audio this means a paradigm shift from broadcasting a pre-mixed audio stream to broadcasting the elements needed to make a mix that is dependant on the user preferences and production decisions. These elements are sound fields and audio objects. A sound field describes the sound from all directions at one place and can be recorded using a SoundField®, Eigenmike® or any other multi-channel sound field microphone. Audio objects are discrete sound sources that also make up the audio scene and are defined below.

4 AUDIO OBJECTS

An audio object describes a sound source of which the content, position, onset time and duration are known. Audio objects are fundamental to format agnostic reproduction as they can easily be rendered to whatever loudspeaker setup is present at the user end. Currently, broadcasters will mix and broadcast the audio for a specific loudspeaker setup such as stereo 5.1 etc. This method thus introduces a compromise at the user end if the reproduction system does not match the broadcast format. Broadcasting a stream of audio objects with an accompanying sound field would allow a spatial mix at the render end rather than at the production end so any audio system can be accommodated from a simple stereo system right up to higher order ambisonics, wave field synthesis and even reproduction systems not yet in use. For the more complex reproduction systems, broadcasting audio objects rather than loudspeaker signals is of particular importance as the number of loudspeakers used in these systems could potentially run into the hundreds resulting in a huge demand on bandwidth.

Audio objects can be loosely grouped into two main categories: *explicit* and *implicit*. The grouping of audio objects depends on the method and accuracy of capture. A brief description of each type can be found below.

4.1 Explicit Audio Objects

Explicit audio objects are objects that directly represent a sound source and have a clearly defined position within the coordinate system. This could include a sound source that is recorded at close proximity either by microphone or by a line audio signal and is either tracked or is stationary with defined coordinates. Example: instruments close miked in a performance which are static and have little or no crosstalk from other sources. Close miking and position tracking however is not always possible in practice – as in the case of football where individual players can not be close miked. Thus, not all audio objects are explicit.

4.2 Implicit Audio Objects

Implicit audio objects represent sound sources in a more indirect manner, these could include signals that are picked up by more distant microphone techniques or by microphone arrays where the source of sound is distinct from the receiving device or where the audio object may be derived from several recording sources. Example: in the first FascinatE Premier League test shoot the sound of the ball being kicked is derived from one or more shotgun microphones around the pitch. In this instance the ball could not be tracked so areas of the pitch are defined as implicit audio objects that are either active or inactive depending on whether there is relevant sound activity in that region at any given time.

For the capture of a football match most of the audio objects are implicit, as there are limitations on the number and placement of microphones (it is not possible to close mike and track the players), the exception is the commentator who is close miked.

5 EXTRACTING AUDIO OBJECTS

In the case of a football match, as most of the audio objects are implicit they must be derived from the microphone signals available at production. In this section we present an algorithm that ingests the audio from the shotgun microphone feeds, analyses them and assesses whether the content contains a significant on-pitch audio event (OPAE) and if so, creates an audio object. In the context of a football match, there are principally two dominant categories of OPAE corresponding to ball kicks and whistle blows. If the algorithm detects such an event in the audio feed, it determines that action is taking place in the vicinity of that microphone; this information can be combined with information from other microphones to determine the position and so facilitate the extraction of an audio object. For a standard broadcast, this can be used to retrospectively add that microphone signal into the broadcast mix for the window of time in which the audio event occurred (so providing an automatic mixing algorithm). For the FascinatE scenario the audio object with content, location, onset time and duration can be broadcast for object based format agnostic rendering.

5.1 Extracting object content

The methodology for detecting and extracting audio objects differs depending on the type of OPAE that is to be detected. The following, briefly describes how ball kicks and the referee's whistle blows can be extracted from the pitch-side microphones

5.1.1 Ball-Kicks

To successfully detect a ball-kick from the audio data, it is necessary to analyse the key characteristics of the audio generated by such a kick. Figure 2 shows the spectrogram of a typical ball kick as recorded at a football match in the English Premier League.

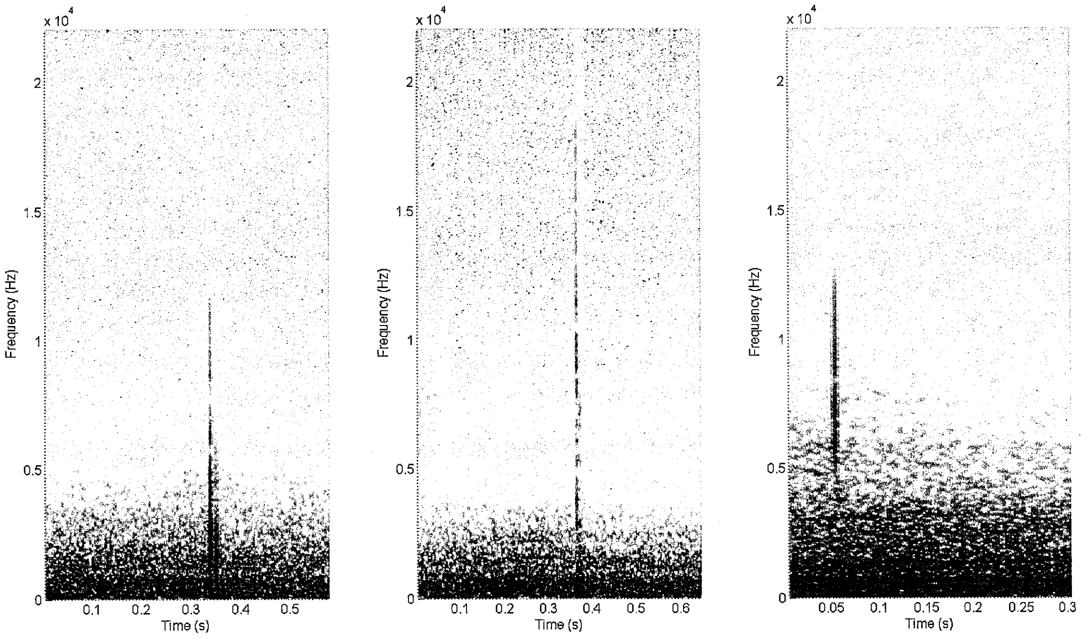

Figure 2: Spectrogram from some typical ball kicks from a live broadcast

As can be seen from Figure 2 the kick of a football is characterised by transient energy in the audio feed particularly at low frequencies. This is in contrast to the wash of crowd noise that contains few transients. This feature can be exploited to extract the audio from the microphone signal. The spectrogram in Figure 2 is calculated from the power spectral density ($PSD(\omega)$) which is a measure of how the energy of a signal $S(t)$ is distributed with frequency and is calculated from the square of the signal's Fourier transform:

$$PSD(\omega)=\left|\frac{1}{2\pi}\int_{-\infty}^{\infty}s(t)e^{-j\omega t}dt\right|^2 \qquad (1)$$

The algorithm presented here calculates the power spectral density of the signal at specific frequencies thus producing a signal envelope at these frequencies. The gradient of this envelope is then calculated at each of the analysis frequencies to determine the significance of any transients that it contains. When this gradient exceeds a given threshold value of the analysis band, it is determined that a significant transient has occurred and thus it can be considered to be a ball-kick. When a ball-kick is detected the original microphone signal is multiplied by an amplitude envelope between the time intervals of the detected ball kick and is classed as an audio object, the time is also embedded in the metadata of the audio object such that it can be synchronised with the video content and other audio events during rendering.

5.1.2 Whistle blows

The technique for extracting the referee's whistle differs from the ball kicks. The spectrogram of a typical blow from a referee's whistle as recorded during a broadcast of a live football match is shown in Figure 3, it is characterised by its harmonic content, having a fundamental of approximately 4 kHz. This feature can be exploited and used to detect and extract the sound of the whistle by using the signal's cepstrum.

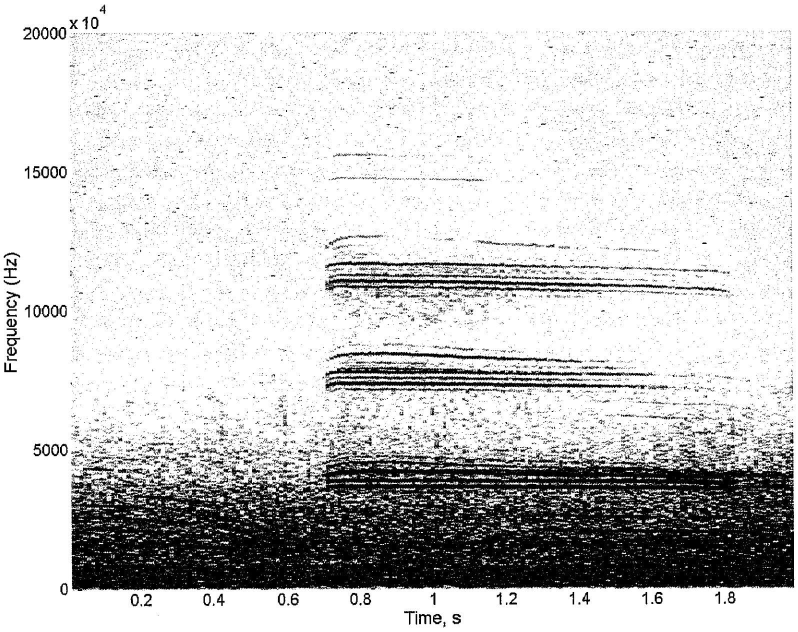

Figure 3: A spectrogram of a typical referee's whistle blow

The cepstrum is often used in speech processing to determine the fundamental frequency of formants and for other scenarios where pitch detection is needed. The cepstrum highlights periodicity in a signal and is therefore ideally suited to detect a whistle blow. The cepstrum, $c[n]$is defined as the Fourier analysis of the logarithmic amplitude spectrum of the input signal:

$$c[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|S(\omega)\right| e^{jnw} d\omega \qquad (2)$$

Where $S(\omega)$ is the frequency spectrum of the signal.

If the input signal contains many harmonics, its spectrum will exhibit, peaks at the harmonic frequencies, whose spacing are determined by the fundamental frequency of the harmonic signal. The cepstrum picks out the periodicity of this spectrum which correspond to the harmonics in the signal. The peak value of the signal's cepstrum for a harmonic input relates to its fundamental frequency. The x axis is the quefrency in temporal units, reciprocally related to frequency.

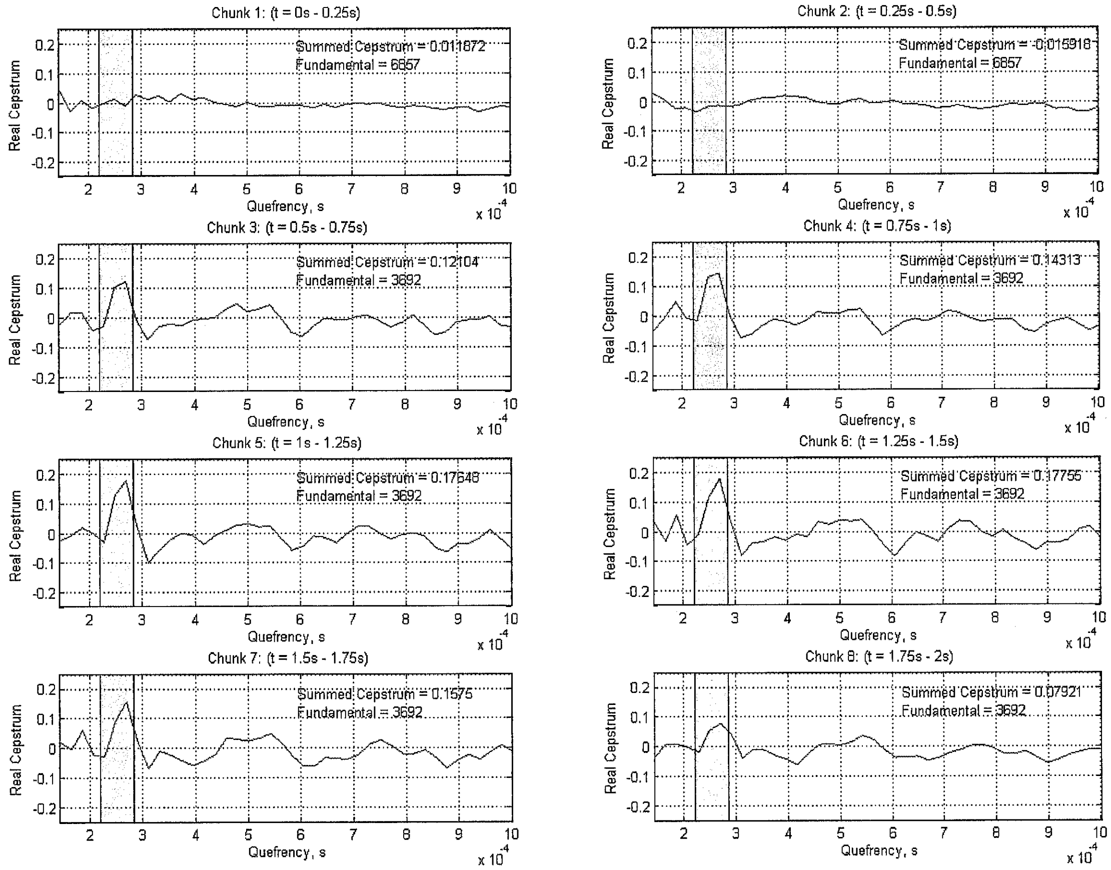

Figure 4: Real cepstrum of a whistle blow in 0.25s chunks

From figure 4 it can be concluded that there is a whistle blow present in the signal between ~0.5 seconds and ~1.75 seconds with a fundamental of 3.7 kHz which concurs with the spectrogram in Figure 3. The red band in the plots corresponds to the analysis frequency range, i.e. the range of frequencies in which the fundamental frequency of the referee's whistle is likely to occur in (3.5 - 4.5 kHz). To determine whether or not a whistle blow has occurred, the values of the cepstrum coefficients in this range are summed and if these exceed a set threshold value, it is determined that a whistle blow has been detected.

The algorithm processes the input signal in chunks and calculates the cepstrum for each 0.25 second segment, if the above criteria are met, a whistle blow has been detected and that chunk of microphone feed is extracted as an audio object. In order for the onset and offset of the microphone signal to not be perceptually obvious a attack and decay ramp is applied with a 1.5s duration, the subsequent increase in level introduced with the addition of the microphone feed is not problematic as it is masked by the crowd noise from the ambient microphones (as is the case for a standard broadcast when the engineer raises the level of shotgun microphones manually).

One potential pitfall of the Audio object extraction algorithm for both the ball-kicks and the whistle-blows is the potential for sounds from the crowd to be detected as an OPAE. This can be eliminated if accurate information of the source location can also be extracted from the microphone signals. If it can be determined whether the source is behind the microphones, it is from the crowd and therefore should not be extracted as an audio object. To resolve this issue further work is being done utilising dual shotgun microphones at each location.

5.2 Extracting object position

Determining the location of an OPAE is critical to the success of the audio object approach, assigning a location to an audio object allows it to be positioned accurately in the reproduced sound field such that a correct spatial impression of the game is experienced, furthermore for the FascinatE project where users have the ability to interact with the broadcast content, the locations of the objects need to be accurately known such that the audio can updated to match the individual's viewpoint.

Determining the position of the OPAE is difficult in the example used here because of limitations in the number of available microphones used for the capture. Sounds can only be localised accurately if they are picked up by more than one microphone at a given time, the greater the number of microphones that pick up the OPAE the more accurate the positioning of the source will be. If only one microphone picks up the OPAE, positioning the source on the pitch in the sound field can be done by simply looking at which microphone has energy at that moment in time and positioning its signal as the audio object in the centre of the zone it covers (as shown in Figure 1). This has the undesirable effect that there can be a substantial mismatch in visual source location and the corresponding audio object which may be noticeable in the FascinatE rendering, especially if the user zooms close into the pitch where the relative difference in position will be greater.

If more than one microphone picks up a signal within a given time window it is possible to determine whether or not the sound source being picked up is the same audio event by looking at coherence between the microphone signals. If this is within previously determined limits it is assumed that the sound source is the same. Time delay estimation[5] techniques can then be used to find the time delay between the sound arriving at each microphone and from this the object's position can be inferred. This time difference can be used to approximately position the source on the pitch although it can only tell the azimuth direction. In terms of depth, the object will have to be positioned somewhere on a line running through the centre of the active zones.

6 CONCLUSIONS

A new approach to the recording and rendering of football for television broadcasts has been presented. The new approach centres around the use of so called audio objects which allow a fully customisable and interactive audio rendering for the user. The algorithm presented in this paper is able to analyse the signals from the microphones positioned around the pitch and extract the significant on-pitch audio events. The algorithm identifies key features in the audio that correspond to ball-kicks and whistle-blows and extracts audio with these features. Furthermore the position of the audio event, the onset time and the duration are extracted and used to generate an audio object.

As part of the EU funded project, FascinatE, these audio objects can be broadcast in place of a pre-mixed audio stream. This enables not only a format agnostic mix and rendering at the rendering terminal but also allows the audio to match the visual scene as the user interacts and makes their own production/viewing decisions.

7 REFERENCES

1. G. Cengarle, T. Mateos, N. Olaiz, and P. Arumí, "A New Technology for the Assisted Mixing of Sport Events: Application to Live Football Broadcasting." *Proc. 128th Conv. Audio Eng. Soc. London, UK (2010).*

2. R. Oldfield and B. Shirley , "Automatic Mixing and Tracking of On-Pitch Football Action for Television Broadcasts", *130th Conv. Audio Eng Soc*, London, UK, May 2011.

3. FascinatE Project. Official FascinatE Website: http://www.fascinate-project.eu, 2011.

4. JM. Batke, J. Spille *et al*, "Spatial Audio Processing for Interactive TV Services", *130th Conv. Audio Eng Soc*, London, UK, May 2011.

5. C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.* ***ASSP-24****, 320–327 (1976).*