VAMBU SOUND: A MIXED TECHNIQUE 4-D REPRODUCTION SYSTEM WITH A HEIGHTENED FRONTAL LOCALISATION AREA

MARTIN J. MORRELL[1], CHRIS BAUME[2], JOSHUA D. REISS[1]

[1] Centre for Digital Music, Queen Mary University of London, London, UK
[2] BBC Research and Development, London, UK

A system, Vambu Sound, was developed for BBC R&D to create a spatial audio production environment. The specification of the system is to provide good localisation around a main television screen and diffuse sound from around the listener. The developed system uses Vector Base Amplitude Panning for six loudspeakers in front of the listener and Ambisonics for eight loudspeakers in the corners of a cube configuration. The system is made four-dimensional by the incorporation of a dedicated haptic feedback channel within the audio format. The system design and implementation are presented and responses from a demonstration are evaluated.

INTRODUCTION

Through a collaborative project between BBC Research & Development and Queen Mary, University of London, the author was tasked to create a spatial audio system to work alongside an immersive video technology known as 'Surround Video' [1]. The requirements of the system were to match the localisation attributes of surround video; that is highly localisable in a frontal region around a main television screen and a diffuse atmospheric sound from everywhere else around the user.

Previous surround and 3-D reproduction systems were not viable for meeting our needs. Surround technologies such as 5.1 have more speakers at the front which would give a higher degree of directionality, but this is considered a flaw of the system and as such 7.1 adds extra rear speakers to make the reproduction of sounds more uniform. Ambisonics only systems require that speakers are arranged in a regular arrangement around the listener, that is equally positioned in 3-D space. A Vector Base Amplitude Panning only system leads to the lack of diffuseness and a sense of envelopment for the non-frontal section. This method also suffers from sounds being pulled to a speaker location where the sound source is perceived to be tonaly different when at a point that it is reproduced from a single speaker.

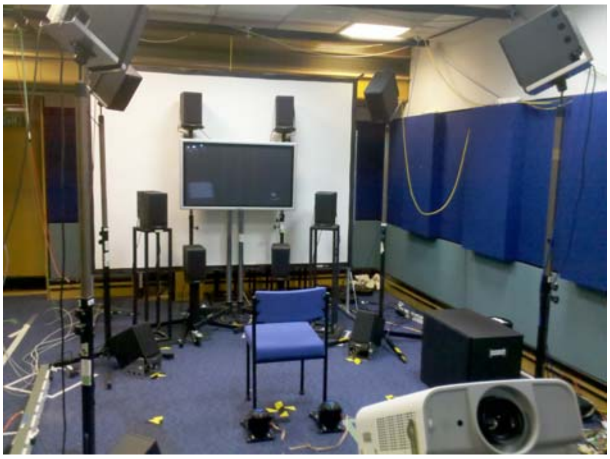

Figure 1 – The system set up in BBC Television Centre, featuring 6 front speakers, 8 speakers in a cube, subwoofer, ButtKickers and surround video system.

The implemented system uses Vector Base Amplitude Panning (VBAP) [2-3] reproduced through 6 loudspeakers around the main television area and first order Ambisonics [4-6] throughout the rest of the listening area using a cube arrangement of 8 speakers (see Figure 1). VBAP was chosen for the highly directional sound sources as it uses between 1 and 3 loudspeakers to place a single source depending on its location within a 3D triplet of speakers. Conversely Ambisonics uses all loudspeakers at once to place a sound source, including an out of phase component. The resultant polar pattern produced by 1st order Ambisonics is quite wide and so is difficult to exactly localise a sound source. For this reason 1st order Ambisonics is well suited to the reproduction of diffuse and ambient sounds whilst the power panning of VBAP is best suited to place prominent sound sources.

The system can be experienced over headphones, using binaural rendering of the speakers, and also features dedicated low-frequency effects (LFE) and vibration effects channels. The vibration is reproduced using the ButtKicker [7] system.

By mixing the names of the technologies used, the final system was entitled ‘Vambu Sound’.

SYSTEM OVERVIEW

Figure 2 shows a simplified flow diagram of the application’s architecture. Nuendo 4 is used as the Digital Audio Workstation (DAW), the processing is done in Max/MSP 5, and the two are linked using the audio connection interface Jack. Control data is sent using MIDI. The tool chain was designed and used on the Mac OSX platform.

Up to 28 mono audio tracks are supported by the system, in addition to an Ambisonics B-format input. There are 16 mono outputs – 14 for the speakers, one for a subwoofer and one for the vibration effects.

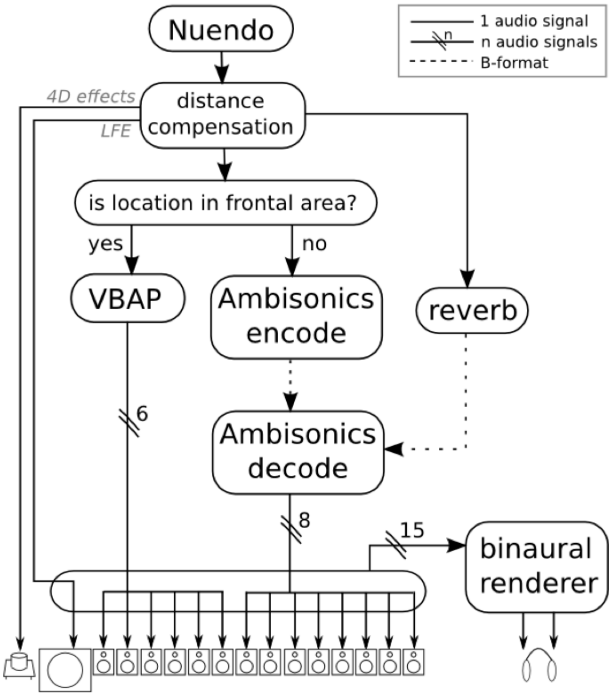

Figure 2 – A simplified block diagram of the internal connections between modules in the Vambu Sound application

Each audio track is encoded independently, and is firstly processed by applying a gain and delay based on the distance of the audio source. After this step, each track has its coordinates transformed by any sound scene rotations in the Cartesian domain. After the new spherical angles are obtained, inside-panning is performed. After this the audio is determined to be in the VBAP frontal section or else in the Ambisonics surrounding section. VBAP audio is sent directly to the speaker feed management. Audio sent to the Ambisonics reproduction is encoded to a first order B-format signal. Reverberation inserts are taken from the sound source prior to the scene transformations. Reverb is encoded to first order-format and outputted based on the level, reverb type and reverb bank. LFE and ButtKicker signals are also taken directly after the audio delay at a user defined level.

The reverberation signals are processed by either the Wiggins VST [10] or convolved with B-format impulse responses [8-9]. Sound scene rotations are then applied. The rotations are applied post-reverberation for the reverb content so that the sound source reverb characteristic remains the same because it is the main characters point-of-view that is changing not the audio location. Reverb signals are then transformed by the Anderson ‘press’ transform [11] to give a sense of movement within a user specified direction.

In the LFE module the dedicated LFE sends are summed and amplified to the master LFE level. In the same module the omnidirectional Ambisonics signal and audio source signal (used when a source is in VBAP area) are bass managed to create a subwoofer signal. The subwoofer signal uses a phase-matched second-order low pass filter [12] at a 120Hz cut off frequency, which is also used by the LFE signal. Prior to the crossover the subwoofer level is altered by a master subwoofer level control. The outputs go to the speaker sends module.

Within the ButtKicker module the send signals for ButtKicker from each audio track are summed. The signal is then amplified according to a master ButtKicker level control. The output is sent to the speaker sends module.

The Ambisonics signals from each audio track and the reverb are sent to a global Ambisonics decoder. The audio signals are first filtered to remove bass content that is handled by the subwoofer using the high-pass part of the phase-matched filter [12]. As is common with Ambisonic decoders, a second-order crossover is used to send the signals to separate low and high frequency decoders. The low decoder is optimised to maximise rV values and the high frequency to maximise rE values [12-13]. Both decoders are configured for a cube arrangement of loudspeakers, where the decoder is increased by $\sqrt{2}$ to perceptually match the level of the VBAP reproduced section.

The user has the option to send the speaker outputs to a binaural renderer. This uses a set of impulse responses [14] from the Listen HRTF Database for each speaker location which represent how to transform that speaker signal to a stereo binaural representation. Each speaker output signal is convolved with these impulse responses to produce a stereo output suitable to playback over headphones. The Vambu Sound application also provides the option to record the speaker outputs or the binaural rendering, if that has been selected for monitoring.

IMPLEMENTATION

The focus behind the implementation was to require as little user input as possible in the Vambu Sound application, so that the majority of the controls were stored within the digital audio workstation project. The Vambu Sound application automatically loads the audio driver, MIDI driver, sample rate, buffer size and vector size. Speaker locations are stored in a setup text that is automatically read by the Vambu Sound application. Preset information used for a convolution VST are saved in the standard system location for VST presets, but are given specific names so they can easily be loaded from within the digital audio workstation controls. The Vambu Sound application main screen shows audio level meters for the sixteen audio outputs and two meters for the binaural rendering. When the application is first loaded it takes several seconds to load the default settings and to read the setup text file. Once this has finished the Vambu Sound logo fades in to indicate that the load time has elapsed. Overall, the Vambu Sound application can be fairly CPU heavy when using all the available audio tracks, reverbs [8-10] and binaural rendering due to lots of convolution and audio multipliers.

Jack OS X audio connection kit was used to send audio between Nuendo and the Vambu Sound application, and also between the Vambu Sound application and the sound card. For connecting MIDI signals between applications, the Inter Application Connection (IAC) MIDI driver built into OSX was used.

The physical setup of the system was done in BBC Television Centre at BBC R&D's studio. 14 PMC DB1-A speakers were used, in addition to a PMC SB-100 subwoofer and a ButtKicker system as previously mentioned. A 50" plasma television was used as the main display, with a projector and hemispherical mirror used to provide the peripheral projection around the television (see Figure 1).

DAW CONTROL DESIGN

A feature of Nuendo is its 'user panel' technology which allows custom control panels to be created for controlling MIDI tracks. This can be used to provide an intuitive interface for sending correctly formatted data to the Vambu Sound application for controlling the parameters for audio tracks.

The author designed a set of user panels for interacting with Vambu Sound, and created a template project in Nuendo. Each audio track in the project has its control information stored as a MIDI track. For ease of use, the audio and MIDI tracks were grouped together using folders.

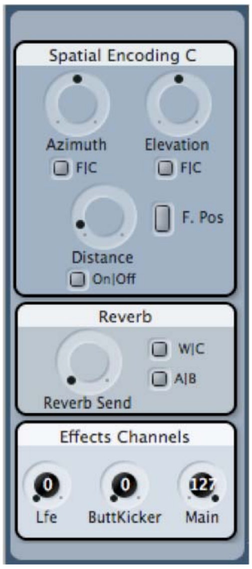

Figure 3 – A Nuendo User Panel that controls a single audio track

Figure 3 displays the user panel for an audio track. In the top section the encoding controls are presented: azimuth and elevation with fine and coarse control options, distance with on/off option and an option for fixed position (the position is not affected by scene rotations). The middle section is the reverberation controls where the main focus is the reverb send control and there are option buttons for algorithmic or convolution reverb and then bank A or B of the chosen reverb type. Finally the bottom section controls, for want of a better term, effects. There is control for a dedicated LFE signal, ButtKicker and main speaker output (so that the audio can be for LFE or ButtKicker only).

Figure 4 displays the User Panel for master controls. In the top section of the display the user has control over LFE (dedicated bass), Subwoofer (bass management) and ButtKicker master levels. In the middle section

(Spot Movement) are the rotation controls: Rotate, Tilt and Tumble, all with fine/coarse control buttons. Finally in the bottom section of the panel are the movement parameters: Up/Down, Forward/Backward and Left/Right. These movement controls alter the reverberation sound field based on Anderson's 'press' transform [11].

There were several more User Panels that are not shown here. They controlled binaural setup, master settings, convolution reverb controls, algorithmic reverb controls and control of B-format audio, where the latter was never used.

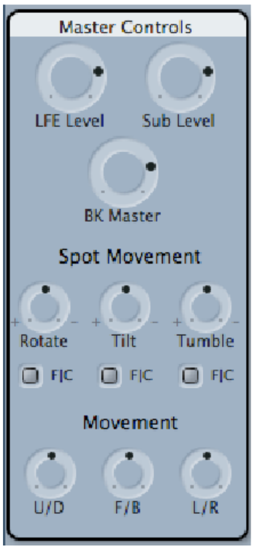

Figure 4 – A Nuendo User Panel for the master controls

The User Panel feature of Nuendo allowed for a somewhat elegant solution to sending MIDI commands to external software (intended for hardware control).

DAW USAGE EXPERIENCE

A three-dimensional mixing system has more creative options available, but also means that there are more things to consider and take into account when mixing. It is difficult to control the three position parameters (azimuth, elevation and distance) available at one time, let alone alter multiple sources' attributes at once. Each parameter is controlled on a different channel/controller number variation, altering each takes considerably more time than a stereophonic two channel mix and this extra effort to accomplish the same task can lead to stifling creativity.

One of the fundamental flaws in having the control and audio data separate in Nuendo is that when the audio is moved on the timeline the MIDI data is not; again each parameter used needs moving individually. This again becomes very time consuming and an extremely tedious activity for the sound engineer.

When considering two-channel stereophony only a few background sounds are needed to set the location atmosphere and too many sounds can lead to severe masking affects of the final mix. However, in a three-dimensional system there is more space and hence masking effects are reduced considerably. This extra space proved the ability to add extra sound sources to create a realistic background environment.

Although there are no rules for use of the ButtKicker, since it usually derives its signal from the subwoofer signal of a mix, it becomes quite intuitive to use. Effects such as rumbles, door bangs and the main character breathing and heartbeat become obvious and effective use of the dedicated 4D effects channel.

EVALUATION

Following the Surround Video project [1], an animatic (storyboard) was commissioned by BBC R&D to act as test material. The author produced a soundtrack to accompany the first five minutes of the animatic using the described system in order to demonstrate the technology. Initial impressions by the authors were that the resulting experience was immersive, and that the audio sources in the front section had strong localisation. As expected, the surrounding sound sources were hard to localise, but provided a good atmosphere.

A problem was identified with the video, where the field of view of the source material was not as wide as expected, limiting the field of view to approx. 90º rather than the expected 160º. Additionally, the authors noticed a difference in tonality of audio sources as they moved between the VBAP-encoded region and the Ambisonics-encoded region. The Ambisonics encoding sounded slightly low-pass filtered, but a review of the encoding equations did not identify any issues.

The animatic was used to demonstrate the system to 21 people, consisting of 4 females and 17 males of a wide age range. At the end of a demonstration each subject was asked to fill in a brief questionnaire to gain feedback on their experience of the system. The questionnaire consisted of five ranking questions where the subject answered from 0 to 10 in whole increments and a final comments section.

Figure 5 shows the results of the first question. Participants were asked to compare the tonality between the front and surround sections. This was to gauge whether they heard a difference between the VBAP and Ambisonic reproduction techniques. The scale given from 0 – "nothing alike" to 10 – "the same". The mean result of 7.1 indicates that there are some perceivable differences when an individual sound changes from being reproduced in VBAP to Ambisonics. This is mainly detrimental to sound sources which move between the front and surround sections, whether through panning or scene rotations.

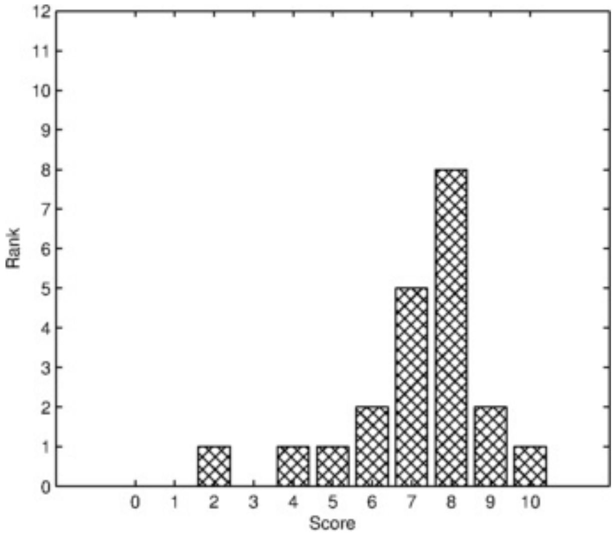

Figure 5 – Participant Questionnaire Answers to 'How did the frontal section stage tonality compare to the surrounding (sides, rear, above and below) sections?'

The second question posed was with regards to the immersive properties of the system. They were asked how "engulfed in the scene" they felt. The scale went from 0 – "not at all" to 10 "I was there". The results are displayed in Figure 6. The mean answer was 7.3, which shows an encouraging result for the system as all participants felt at least half way to being in the programme world presented to them. It is worth noting that the results reflect a combination of both the system and the produced content. One of the subjects commented "Not enough sounds to feel realistic", which indicates that the perceived immersion is biased towards the content more than the technology.

In order to give a rough measure of the localisation performance of the system, participants were asked whether they felt the video and audio locations matched. The scales went from 0 – "nothing alike" to 10 – "the same". The results are shown in Figure 7. The mean result was 6.9, which gives a positive indication, but individual results went as low as 4 and as high as 10. As the animatic has a very low frame rate, the visual position of characters leaves room for interpretation. In addition to the human error involved in positioning sound sources, this may account for the wide range of opinion in this result.

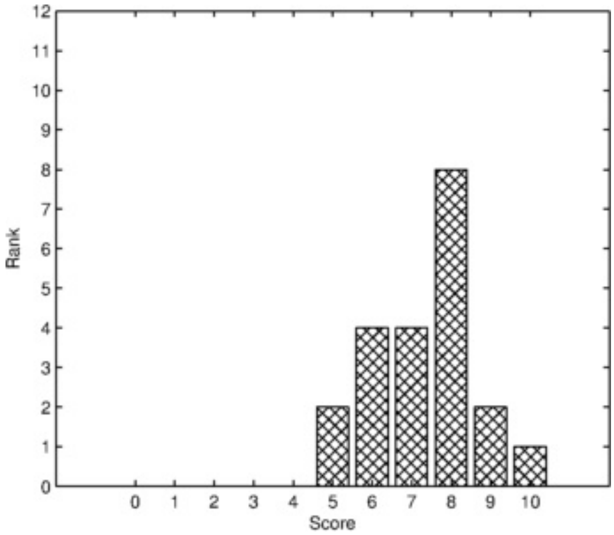

Figure 6 – Participant Questionnaire Answers to 'How engulfed (in the scene) did you feel for the audio?'

Participants were asked "How engulfed (in the scene) did you feel for the video?". The mean result was 4.8. This low score was due to the problems encountered with the field of view of the video content. As most of the participants had already experienced surround video with a full field of view, the score reflects how important it is to fill the viewer's peripheral vision to produce a good sense of immersion.

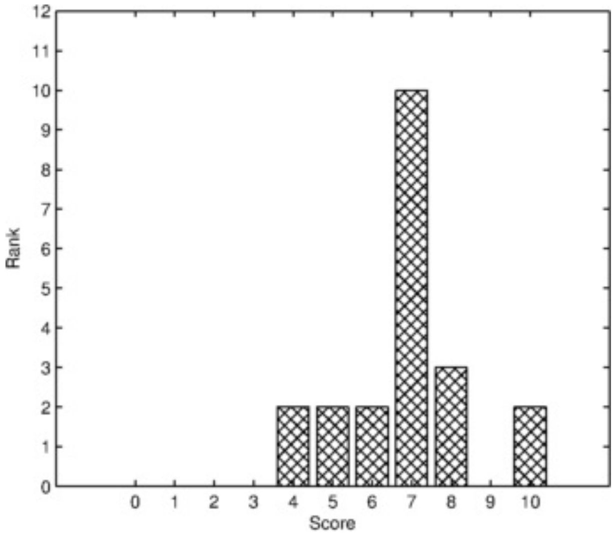

Figure 7 – Participant Questionnaire Answers to 'Did you feel the video and audio locations matched?'

Finally, as the front and surround sections were encoded using different technologies, the authors wanted to understand how scene rotations were subsequently affected. Participants were asked "Was the turning of the sound scene natural?" with a scale from 0 – "not at all" to 10 – "the same". The mean result was 7.0, which shows that the effect of scene rotation did not fall apart under the system, but leaves room for improvement.

Overall, the results of the informal questionnaire were mostly positive but indicate the need for improvement. The problem with the video content is a simple fix, and full frame rate video will help significantly in localisation, but issues around tonal differences between VBAP and Ambisonics for example, will need more attention.

One of the recurring themes, to the authors' surprise, within the subject comments was the use of the ButtKicker as the 4D effect. "ButtKicker was very effective", "Added immersiveness", "LFE in the chair was a little distracting at first, but was good overall." and "the ButtKicker was really effective for the heartbeat". Some participants commented that the ButtKicker has a delay compared to the audio, but this is inherent from its design. The comments indicate that an immersive experience can be significantly enhanced by stimulating the touch sense, in addition to the visual and auditory senses.

CONCLUSION

The authors have presented a system that uses a mixture of both Vector Base Amplitude Panning and Ambisonics to create a system that has a higher level of directionality in a frontal section. This attribute, whilst not a conventional goal, matches the localisation/resolution attributes of Surround Video. Such a system may also be beneficial for reproduction of musical concerts where there is a main focus of an ensemble within a frontal section and atmospheric sounds from everywhere else around the central listener.

The use of Nuendo's user panels feature was an elegant solution to implementing a user interface within the DAW. However, not being able to move the audio clips and audio parameters at the same time was a notable inconvenience. This may be remedied by creating a VST plugin which would allow parameters to be stored within the audio track.

An informal questionnaire on a demonstration of the system indicated that there is a good sense of immersion, but with room for improvement. It showed that the difference in tonality between VBAP and Ambisonic encoding is noticeable, which is an issue which should be addressed. Most comments were positive towards the system, and many people said that the vibration effect of the ButtKicker enhanced the experience.

Overall the system shows that the use of technologies that use all our field of vision and hearing are well received and go a long way to immerse the end user. Further to this the authors found that the stimulation of other senses alongside hearing can add to the sense of immersion.

ACKNOWLEDGEMENT

This research was supported by the Engineering and Physical Sciences Research Council [grant number EP/H500162/1].

REFERENCES

[1] Mills, Peter and Sheikh, Alia and Thomas, Graham and Debenham, , Paul. "Surround Video." *NEM2011*. Torini, 2011.

[2] Pulkki, Ville. "Spatial Sound Generation and Perception By Amplitude Panning Techniques." PhD, Helsinki University of Technology, Helsinki, 2001.

[3] Pulkki, Ville. "Virtual Sound Source Positioning Using Vector Base Amplitude Panning." *Journal of the Audio Engineering Society* 45, no. 6 (June 1997): 456-466.

[4] Gerzon, Michael A. and Barton, Geoffrey J. "Ambisonic Decoders for HDTV." *Audio Engineering Society Convention 92*. Audio Engineering Society, March 1992.

[5] Gerzon, Michael A. "Periphony: With-Height Sound Reproduction." *Journal of the Audio Engineering Society* 21, no. 1 (1973): 2-10.

[6] Gerzon, Michael A. "Practical Periphony: The Reproduction of Full-Sphere Sound." *Audio Engineering Society Convention 65*. Audio Engineering Society, February 1980.

[7] The Guitammer Company. *Buttkicker*. http://www.thebuttkicker.com/ (accessed 09 26, 2011).

[8] Murphy, Damian T. and Shelley, Simon. "OpenAIR: An Interactive Auralization Web Resource and Database." *Audio Engineering Society Convention 129*. November 2010.

[9] Stewart, Rebecca and Sandler, Mark. "Database of Omnidirectional and B-Format Room Impulse Responses." *ICASSP 2010*. 2010.

[10] Wiggins, Bruce. "Has Ambisonics Come of Age?" *Proceedings of the Institute of Acoustics*, vol 30, Pt. 6. 2008.

[11] Anderson, Joseph. "Introducing… The Ambisonic Toolkit." *Ambiosnics Symposium 2009*. Graz. 2009.

[12] Heller, Aaron and Benjamin, Eric. "Is MY Decoder Ambisonic?" *Audio Engineering Society Convention 125*. 2008.

[13] Daniel, Jerome. "Representation De Champs Acoustiques, Application a la Transmission et a la Reproduction de Scenes Sonores

Complexes dans un Contexte Multimedia." PhD, l'Universite Paris, Paris, 2000.

[14] Warusfel, Olivier. "Listen HRTF Database." *http://recherche.ircam.fr/equipes/salles/listen/index.html*. 2003.