

Stuart Cunningham, Richard Hebblewhite, Richard Picking and
Wesley Edwards

Centre for Applied Internet Research (CAIR), University of Wales NEWI
Plas Coch Campus, Mold Road, Wrexham, LL11 2AW, UK
{s.cunningham|r.hebblewhite|r.picking|w.f.edwards@newi.ac.uk}

MULTIMODAL INTERACTION AND COGNITION IN 3D MUSIC AND SPATIAL AUDIO ENVIRONMENTS: A EUROPEAN COMPATIBLE FRAMEWORK

1. Introduction

Over previous years there has been an increase in the production of music by commercial artists which is being created specifically for listeners with surround sound listening capabilities. Such examples of these products can be found in DVD, SACD, and high quality CD releases, and are primarily aimed at home users who have 5.1 home theatre systems.

The Pink Floyd album “*Dark Side of the Moon*” was recently re-released in surround sound. A primary concern when re-mixing the album was the effect that this would have on the perception of the music and how the original stereo recording should be, would be, or could be, translated into a multi-channel 3D audio version [1].

When mixing and producing this music on a traditional mixing desk the process is very much unnatural and detached from the audio which is being heard all around.

Traditional methods of audio mixing and production in the recording studio have for many years been limited to two channel (stereo) systems. To cope with the escalating demand for surround sound (such as 5.1 and 7.1) audio mixes, commonly found in personal and home entertainment systems, studios have improved hardware facilities by adding multi-channel amplifiers and speakers to the produc-

tion environment. However, one area which has been slow to respond in this field is the hardware mixing desk. The de-facto mixing desk is clearly designed to be utilised for mono or stereo use, but is poorly equipped, and not user-friendly enough, for use in a multi-channel scenario.

Some kind of multimodal, primarily gesture based, interface would clearly provide sound engineers and producers of multi-channel audio with a form of interaction that is more natural and instinctive.

Recommendations are made into the most effective approach for a system to allow the most suitable method of direction in such an audio space. In addition, the phenomena of gesture localisation and interpretation between different cultures are explored and studied in laboratory exercises, in an attempt to produce a framework that is most suitable for a European-wide group of users, regardless of any linguistic boundaries. Namely, this addresses the issue that gestures often mean different things in different cultures and countries, and that there may be completely different interpretations of what constitutes the correct signal to indicate specific meanings in the domain of sounds and hearing. A generic model that produces suitable compromises between cultural differences across Europe, but still provides the maximum functionality, is suggested.

This paper provides a proposed framework for a standardised system of gestural interaction for such audio control requirements, and also assesses social and cultural differences required to be addressed to produce a model that would be European compatible. That is, a model of gestures that provides intrinsic, maximum accessibility to a range of European users, whilst accommodating any differences caused by cultural localisation.

2. Multimodal Interaction for Spatial Audio

2.1. Rationale

A multimodal, gesture based, system would allow the listener to sit in a static position and control the many dynamic qualities of the audio using various types of gestures. This would be controlled primarily by actions such as arm, hand, head, and feet movement, areas that have long been found useful in the development of innovative interfaces. This would provide ample parameters, and unique methods for controlling the many properties of an audio mixing process, such as volume, equalisation, and panning.

For example, tasks such as panning audio sources from front to right could be achieved by grasping with the hand and arcing the arm round until the hand points in the direction that the audio is desired to be heard coming from.

This type of interaction would require less obtuse physical movement and therefore help to preserve the effects of having spatial audio, by requiring less physical head and body movement by the user, conserving the effect of the user sitting in a 'sweet spot'.

There has been much research carried out into the field of gestural interaction, and the enhancement of the Human Computer Interface (HCI) by employing both haptic motion and audio feedback [2] [3] [4]. Research has also been carried out into the use of such gestural interaction in a 3D audio environment.

2.2. Scope and Sphere of Examination

As the area of haptic interaction is extremely broad, it is beyond the scope of this paper to analyse multiple methods of gestural interaction in this context. It has also been proven that certain parts of the human physiology are more commonly used in non-verbal communication [5]. Therefore, certain parameters have been defined to ensure that, to an extent, qualitative research can be carried out. To this end, the following parameters have been impinged on the studies carried out in the research of this paper. The study will:

- Focus on gestures of the hand and arm
- Concentrate on the basic control of volume, panning, and tonal control from single and multiple sources
- Assume that a 5.1 or similar sound system is employed
- Cover a broad range of test subjects of European Union (EU) nationalities

By adhering to these restrictions a more concentrated framework can be modelled and provide insight and recommendations into how further study of the field can be carried out in future.

3. Audio Control Gestures

3.1. Procedure of Investigation

The tests carried out were set-up in a controlled environment, where a number of static variables were set for all tests. The tests were carried out in a room that was laid out as shown in the top-down, conceptual diagram, of Figure 1. A 5.1 surround sound system was assumed, and used for testing, in this case to keep the trials as simple as possible. As can be observed, the test subject (user) was sat in the central position. All tests were recorded on a digital video camera for further analysis. To keep the tests as consistent as possible, the camera was re-calibrated for each user studied. The user was required to face the direction deemed the ‘front’, which is where the camera was positioned.

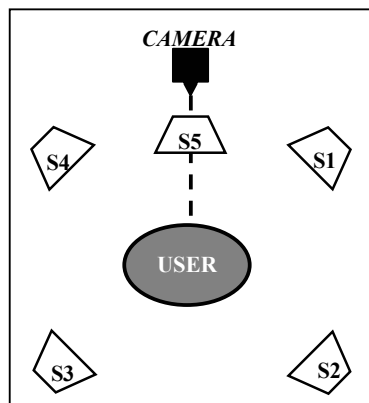


Fig. 1. Positioning of Camera, Speakers, and User for testing

So as not to complicate the test, real speakers and audio sources were not used. Dummy speakers were set-up in place of a real audio system. Rather the user would be posed several questions, and asked to respond to these hypothetical situations, once they had been settled in a suitable position. This setting was explained to the test subject. The test subject was instructed that they could use whatever gesture they felt was most appropriate to each task they were issued. However, subjects were not allowed to issue voice commands, or to get out of the seat they were sitting in when performing a task. Again, to acquire the best quality of results, these scenarios proposed were not of a high degree of complexity. Each user was then asked to perform the following tasks with a purely gestural motion:

- Raising and lowering the volume of the audio coming from speaker S1
- Raising and lowering the volume of the audio coming from speaker S3
- Raising and lowering the volume of the audio coming from speaker S4
- Raising and lowering the volume of the audio coming from the front
- Raising and lowering the volume of the audio coming from the rear
- Raising and lowering the volume of the audio coming from the left
- Raising and lowering the volume of the audio coming from the right
- Pan the audio from S1 to S5
- Pan the audio from the front to the right
- Pan the audio from the front to the left
- Pan the audio from the front to the rear
- Pan the audio from the rear to the right
- Pan the audio from the rear to the left
- Pan the audio from the rear to the front
- Increasing and lowering the treble coming from the front
- Increasing and lowering the treble coming from the rear
- Increasing and lowering the bass coming from the front
- Increasing and lowering the bass coming from the rear

These tests aim to conduct as thorough an investigation as possible, whilst not over-taxing the user. Though simple tests to perform, we can produce a reliable framework and model for a standard by employing a wide variety of sample data, which is of a high quantity, and as tests were observed by researchers, quality is assured.

Test subjects ranged in their nationality from British, Welsh, Spanish, Czech, French, and Syrian-British. As this was an initial pilot study, nine subjects were tested.

3.2. Analysis of Results

Subjects dealt with the majority of the tasks issued without any great difficulty, and understood the scenario presented to them. All of the subjects performed the raising and lowering of sound sources by raising and the lowering either a single hand, or both hands, which were pointed in the general direction of the audio source. The range of movement up and down did vary slightly, but a middle ground was clearly detectible. The majority of subjects directed to sound sources with the hand and arm nearest to the source (so for the right hand side, the right hand and arm were used). However, when dealing with sound which was coming from di-

rectly in front, or from behind the subject, both hands tended to be used to manipulate audio.

Similarly, all subjects indicated towards the source, and then moved the direction of their hand towards the target when instructed to pan audio around the 3D space. Again, when dealing with sound sources at the front, or behind the subject, both hands being used were the most common gestures observed.

One of the most interesting findings of the research was the way in which different test subjects approached the control of the bass and treble properties of an audio source. All test subjects had to pause to think about how to perform such a gesture to represent bass and treble, and one subject was unable to provide any kind of gesture to do this. As another subject commented when contemplating about performing a treble manipulation “*How would I represent that?*”. All test subjects, except of the abstainer, found different approaches to performing the increase and decrease tasks on the bass and treble qualities of an audio source. The raising and lowering elements were performed as before with the volume properties, but all subjects found that some kind of gesture was required before doing this, to indicate that they were now manipulating the bass or treble. The range of gestures used to symbolise this was varied, and an in-depth analysis of each of these gestures is beyond the scope of this paper. These gestures ranged from hand clapping to turning a pointed finger in the air.

3.3. Analysis of Results – European Perspective

In the short series of tests performed, there was little evidence of any localisation specific gestures. However, this was mainly due to the limited quantity of tests performed on a volume of EU students from outside the UK. However, a few observations have been made which warrant further investigation. For example, French subjects preferred to always use an open hand gesture for all tasks, whilst Czech subjects would sometimes use a clenched hand, with a finger or thumb extended. Although all subjects preferred to use their hands and arms as the primary source of interaction, the Syrian-British subject was the only one who preferred to use simple head movements for a select few tasks. British and Czech subjects also tended to use bigger, more dominant gestures, than the French and Spanish subjects, who used much more reserved and controlled signals.

Clearly, it is necessary to conduct much higher quantities of research on a diverse range of subjects from within the EU, to investigate further these observations, and also to determine any new particular requirements related to the nationality of subjects.

4. Recommendations for a Gesture Framework

From the tests performed, we can construct a series of recommendations to be implemented in a gesture recognition system, should it be developed to cope with audio manipulation tasks, such as those required during the tests used in this research. The recommendations are as follows:

Volume manipulation.

Increasing the volume should be carried out by raising the hand and arm from resting (we assume that this is approximately at the waist) to at least shoulder, if not head, height. Such a gesture should incur a fixed percentage of volume increase. This gesture can be repeated to incur further escalation of the volume.

Decreasing the volume should be the opposite of the increasing process, where the hand and arm is dropped from head or shoulder height to at least the waist. Such a gesture should incur a fixed percentage of volume drop. This gesture can be repeated to incur further reduction of the volume.

The direction in which the arm is facing should be taken into account to determine which audio source is being manipulated. If the arm is aimed directly at a speaker in the surround sound system, then only that source should respond. However, if the arm is aimed at the general direction fields (left, right, front and back) then any sources required to produce that audio field should respond synchronically. It is preferred, and indeed, should be allowed by the recognition system, that two hands can be used when dealing with audio from the perceptual front or rear.

Panning Manipulation

Panning a sound source, whether from a direct speaker or from a general perceptual direction should be undertaken as follows: The originating sound source should be indicated by the user directing their hand and arm, nearest to the source, in its direction. A rotation of the arm should then be made until the direction of the target source is established. Upon the dropping of the arm, the gesture is complete.

The panning system should also allow for both hands to be used, either sequentially (to cope with some shifts of 90° or greater) or by moving both hands over the shoulders (to cope with shifts from the perceptual front, to the rear, and vice-versa).

Panning could also realistically be needed to be carried out by moving more than one audio source at the same time. A panning system must be extremely versatile, and because of the nature of the hand and arm gestures used, some kind of indicator would be desirable to specify that the user is now dealing with the manipulation of panning.

Bass and Treble Manipulation

It is imperative that further research is carried out into this field in particular. It is clear that an indicator is required to show that the user is attempting to manipulate these properties of the audio. Again, we assume that the user will direct their hand in the direction of the audio that they wish to manipulate, and that the raising and lowering of bass and treble properties would be carried out in the same way as the volume manipulation.

The main goal is to develop a gesture indicator to represent the properties of bass and treble, which are indicative and natural for any user to be able to use. Research is needed to determine firstly, if such a gesture exists, and secondly how such a gesture could be recognised by the system. There were several interesting approaches to representing the bass and treble properties when the pilot tests were performed, however, as each was unique it is unreasonable at this stage, to assume any of these would be commonly acceptable as a proposed standard method for a interaction. A possible solution would be to use some form of sensor, under the user's foot for example, which could be used to show that the user was intending to manipulate these properties.

Hand Usage

It is suggested that the recognition of the shape of the hand is ignored in the first instance, at least for the manipulation of volume and pan adjustments. Although several test subjects used varied hand gestures, the arm movement for tests remained the same. Therefore, it is suggested that any form of hand shape recognition would currently add to confusion, and result in a poor user-friendly interaction, were such a system implemented.

5. Conclusions

Clearly there are some very common factors and gestures that have been established and observed from this series of tests, and these can contribute to a framework or proposed standard. This can be seen in the similar approaches to manipulating the audio properties of volume and panning. Equally as clearly, however, is the observation that certain tasks require more research and investigation, in particular the bass and treble manipulation requirements.

As the main focus of this research was aimed at developing a system which could be used primarily in the recording studio for mixing and production, this opens up a wide scope for future work. Many aspects of studio mixing are equally as difficult to define as the bass and treble gesture indication. Properties such as a more diverse range of graphic equalisers, gain, and external effect manipulation (such as reverb, distortion, noise gate, compression, and phasing) properties would

also be hoped to be controlled through such a sophisticated method of gestural manipulation.

The development of a European standardised format has been highlighted in this work, but has not been analysed with a broad enough spectrum of research subjects to be able to pose any particular observations or nuances that are required to be adhered to at this stage. Certain aspects of localisation have been indicated at nonetheless, and these may lead to further avenues of investigation, culminating in the development of a truly European compatible system.

A simple framework has been established in this pilot study and the door has been opened for a diverse range of future tests into gestures and recognition, particularly with the focus on building a European compatible system, which is another main goal in the development of such a standard mechanism.

Literature

- [1] Audioworld Online & Today.net, *Producer/Engineer James Guthrie remixes Pink Floyd's Dark Side of the Moon for Surround*. Audioworld, New York, 2003.
Available at:
[http://www.audioworld.com/news/0303/16.james.guthrie.surround.remix.dark.side.shtml\(2003\)](http://www.audioworld.com/news/0303/16.james.guthrie.surround.remix.dark.side.shtml(2003))
- [2] Pirhonen, A., Brewster, S.A. and Holguin, C., *Gestural and Audio Metaphors as a Means of Control for Mobile Devices*. In Proceedings of ACM CHI2002 (Minneapolis, MN), ACM Press Addison-Wesley, pp 291-298, 2002.
- [3] Marentakis, G. and Brewster, S.A. *A Study on Gestural Interaction with a 3D Audio Display*. In Proceedings of MobileHCI2004 (Glasgow, Scotland), Springer LNCS Vol. 3160, pp 180-191, 2004.
- [4] Dix, A., Finlay, J., et al., *Human-Computer Interaction*, 3rd Edition, Pearson Prentice-Hall, 2004.
- [5] Amidon, P., *Non-Verbal interaction analysis*, Minneapolis : Paul Amidon Associates, 1971.

SUMMARY

Traditional methods of audio mixing and production in the recording studio have for many years been limited to two channel (stereo) systems. To cope with the escalating demand for surround sound (such as 5.1 and 7.1) audio mixes, commonly found in personal and home entertainment systems, studios have improved hardware facilities by adding multi-channel amplifiers and speakers to the production environment. However, one area which has been slow to respond in this field is the hardware mixing desk. The de-facto mixing desk is clearly designed to be utilised for mono or stereo use, but is poorly equipped, and not user-friendly enough, for use in a multi-channel scenario. This paper proposes initial responses to this problem, by suggesting a more instinctive mechanism for controlling spatial mixes.

Common actions associated with methods of gestural interaction for audio control and manipulations are explored to produce a study into the most suitable type of system to allow a user to interact with a spatial audio environment. Recommendations are made into the most effective approach for a system to allow the most suitable method of direction in such an audio space. In addition, the phenomena of gesture localisation and interpretation between different cultures are explored and studied in laboratory exercises, in an attempt to produce a framework that is most suitable for a European-wide group of users, regardless of any linguistic boundaries. Namely, this addresses the issue that gestures often mean different things in different cultures and countries, and that there may be completely different interpretations of what constitutes the correct signal to indicate specific meanings in the domain of sounds and hearing. A generic model which produces suitable compromises between cultural differences across Europe, but still provides the maximum functionality, is suggested.