A Semantically Motivated Gestural Interface for the Control of Audio Dynamic Range

THOMAS WILSON[1] AND STEVEN FENTON[2]

[1] The University of Huddersfield, Huddersfield, West Yorkshire, HD1 3DH, England
e-mail: thomas.wilson@hud.ac.uk
[2] The University of Huddersfield, Huddersfield, West Yorkshire, HD1 3DH, England
e-mail: S.M.Fenton@hud.ac.uk

September 23rd 2016

Abstract

This paper proposes and tests the efficacy of a 2D gestural interface as a means of controlling audio processing parameters. The process of parameter mapping and subsequent optimisation can be applied within a 3D environment. Highly immersive computer interfaces, such as those found in modern virtual reality systems, offer an alternative platform suitable for 'virtual mixing desk' implementation, using a mixture of familiar controls and novel gestural control. By focusing on a small element of the proposed 'virtual mixing desk', audio dynamic range compression, this paper aims to evaluate the efficacy and practicality of a global gesture set. Following a large scale gesture elicitation exercise utilising a common 2D touch pad and analysis of semantic audio control parameters, a set of reduced multi-modal parameters are proposed which offers both workflow efficiency and a much simplified method of control for dynamic range compression.

1 Introduction

The advent and proliferation of Virtual Reality (VR) and Augmented Reality (AR) technologies has caused an influx of investment and development in the audio industry. Just two recent examples are the acquisition of Thrive Audio by Google [1] and the licensing of RealSpace 3D by Oculus [2]. The popularity of R&D in these areas confirms the importance of audio processing for producing a fully immersive VR experience. Furthermore, some companies have seen the potential of using VR technology as a production tool for mixing audio. 3DCeption Spatial Workstation is a system that allows engineers working on VR or 3D content to place sound sources in an intuitive way using a VR headset and deictic (pointing) gestures [3]. This system is designed to integrate with popular Digital Audio Workstations (DAWs) so that engineers can combine the instinctual, gestural 3D panning tool with traditional, familiar plug-ins. Combining novel and proven technologies in this way helps engineers to comfortably integrate unfamiliar mixing tools into their workflow. This study proposes and tests the efficacy of a 2D gestural interface as a means of controlling audio processing parameters. The process of parameter analysis and subsequent optimisation could be applied to an interface based within a 3D environment. The key advantages of this are the reduction in overhead required to track multiple axis of movement, increased system responsiveness and a reduction in the errors associated with simultaneous, multi-parameter control. In turn, these attributes result in an improved workflow.

An intrinsic characteristic of gestural control could be exploited for the optimisation of modern DAW interfaces. For example, they provide a platform that does not require visual feedback or a mandatory Graphical User Interface (GUI). Instead, engineers could learn to mix through the memorisation of a gesture-set, thus enabling a more immersive and intuitive mix-environment that removes the 'visual barrier' between an engineer and the audio. This kind of mixing platform could be likened to learning and playing a musical instrument, whereby making 'chord shapes' is comparable to performing mix-control gestures.

Visual feedback in audio mixing systems has received some criticism. Mycroft et al [4] suggest that increased complexity of visualisation can risk diverting the engineer's attention from the sonic quality of the mix. Furthermore Schutz and Lipscomb [5] found that the perceived duration of a musical note could be influenced by visual stimulus. Reducing the reliance on visualisation in mixing systems may prove to inspire a more effective audio-interfacing platform. In addition to

the potential visual distractions, it could also be suggested that many modern audio GUIs are unrepresentative of the audio processes and control techniques. Current GUIs often have a tendency to adopt skeuomorphic designs. Skeuomorphism describes a digital design that is based on a real-world product [6]. The most common example of this within DAWs is the 'banks of faders' mix window, a design that is directly based on analogue hardware. The skeuomorphic design methodology is intended to improve user familiarity, but it could be argued that this is at the cost of usability and suitability to the interfacing method. For example, with a hardware compressor, users are able to control two parameters simultaneously using two hands, which might be very useful when setting a balance between threshold and ratio. Simultaneous parameter control is something that is not immediately achievable using traditional Windows Icons Menus and Pointer (WIMP) interfacing methods.

This paper looks to evaluate the suitability of gestural control systems for interfacing with dynamic range compressors. The gestures can be made in 2D or 3D space. Their effectiveness is realised through comparisons with traditional interface methods, testing of various GUIs and the optimisation of higher-level controls using semantic descriptors.

2 Background and Related Work

2.1 2D Gestural Interface Principals

A brief overview of gesture types is presented for reference throughout this paper. Gestures can be fundamentally defined by three classifications [7]:

- Static - motionless gestures such as taps.
- Dynamic - Moving gestures.
- Spatiotemporal - Dynamic gestures that require co-ordinate analysis over time. The most common implementation of this is the drawing of shapes or letters.

Additionally, most gestures can be classified by whether they represent real-world actions or just arbitrary control allocations [8]:

- Mimetic - movements that imitate an action. These are most commonly continuous.
- Semaphoric - "Gestures from a dictionary of abstract symbols" [9]. Derived from semaphores (signalling with flags).
- Deictic - acts of pointing. These are most commonly static.

It is suggested that for the control of audio processors, a gesture set comprised of mimetic gestures would be favourable as memorability and learnability are maximised through cognitive user associations and representations.

2.2 Gestural Audio Controllers

The complexity of audio processor parameter adjustments, such as those of a dynamic range compressor, can produce a bottleneck in the workflow of a fully gestural mixing system. GUIs can be simplified and screen-space can be maximised by replacing 'soft-buttons' with gestures [10]. Despite this, gestural interfaces have been used sparingly within the audio industry, with their deployment usually confined to tasks associated with software navigation, audio selection and transport control rather than control of the audio-processing elements.

Many of the previous studies into the gestural control of audio have focused on spatial elements of mixing [11][12][13]. Spatial mixing seems to lend itself to gestural control; Selfridge and Reiss [14] identify that users find that pointing to a position in the stereo field for panorama is an effective and intuitive way to mix. Furthermore, the 'stage metaphor' GUI, as implemented by Ratcliffe [11], provides clear visual feedback to the user when volume and pan are being controlled simultaneously.

Previous work by the author [15] found that the gestural control of EQ provided a significant improvement to engineer workflow in comparison to traditional interfacing techniques. Conversely, dynamics processor control has only been implemented with mixed success. Selfridge and Reiss [14] found that users had difficulty setting EQ. Likewise, Lech and Kostek [16] found that workflow was disrupted when using their gestural interface because of the requirement of a parameter selection layer. For example, when controlling a compressor, users had to make a semaphoric 'T' gesture to select threshold, thus adding another step to the workflow.

2.3 Semantic Audio Feature Extraction (SAFE) Project

The Semantic Audio Feature Extraction (SAFE) project aims to understand the linguistic associations with parameter settings [17]. For example, they could determine whether a semantic descriptor such as "Punchy" could be attributed to an average compression setting. The project operates by offering free downloads of a plug-in suite, which allows engineers to contribute their settings for each semantic descriptor. An average

of these contributions is then taken and offered as a 'model setting' for the corresponding descriptor.

An issue to consider with this elicitation process is that the settings are source-dependent and engineers will be using a range of varying sources. However, it is hoped that a large enough average would tend towards an 'ideal' setting. Furthermore, the study includes more specific presets such as 'warm vocal compression' and 'rock kick drum compression' that would help to make the settings more contextually accurate. By offering 'semantically motivated presets' the SAFE project aims to improve workflow for less experienced engineers.

The SAFE presets were chosen to form the basis of parameter rationalisation into 'gestural shortcuts'.

2.4 Automation of Compression Controls

The ambiguity and source-dependence of compression settings can prove to be disruptive to the workflow of mix engineers. This has been discussed by Giannoulis et al [18], who attribute the difficulty of working with a compressor to its non-linear, time dependent operation. They suggest that automating the parameters, through the analysis of the input signal, will reduce the required amount of user interaction and the number of control parameters, thus simplifying and improving the interface. Similarly, Cartwright et al [19] propose the combination of parameters into a single control. Their 'Mixploration' interface is operated by moving a ball in a two-dimensional plane, where movements are mapped to changes in spatial characteristics of the mix.

3 Methodology

Testing was carried out in multiple phases; these were the elicitation of the gesture set (Sub-section 3.1), testing to evaluate the workflow improvement of the proposed gesture set (Sub-section 3.2) and testing to determine the effectiveness of the proposed 'semantically rationalised' gestural interface at matching reference compression settings (Sub-section 3.3). Following this, speed and accuracy of all interfaces is evaluated and compared (Section 4).

24 participants took part in the initial gesture elicitation study and 20 participants took part in subsequent the interface testing phases.

The two gestural interface tests took place three months apart, thus reducing any familiarity with the reference sample. The reference sample used was a standard 'rock kick drum' from the Logic sample library. The Logic 9 DAW was used, with stock plug-ins to implement the compression settings. The compression reference settings were consistent between tests. Testing took place in a semi-anechoic chamber with a pair of Genelec 8040a reference monitors set to a comfortable listening level.

The order of test stages was randomised between subjects. For example, in Sub-sections 3.2 and 3.3 some participants began with the mouse and keyboard interface, while others started with a gestural interface, as determined by a random number generator.

3.1 Initial Gesture Elicitation

The gestural elicitation study was carried out to determine the most intuitive gestural associations for given compression controls. Test participants were asked to describe and draw gestures that best represented the controls found on a typical dynamic range compressor [20]. The derived gesture-set is presented in Figure 1.

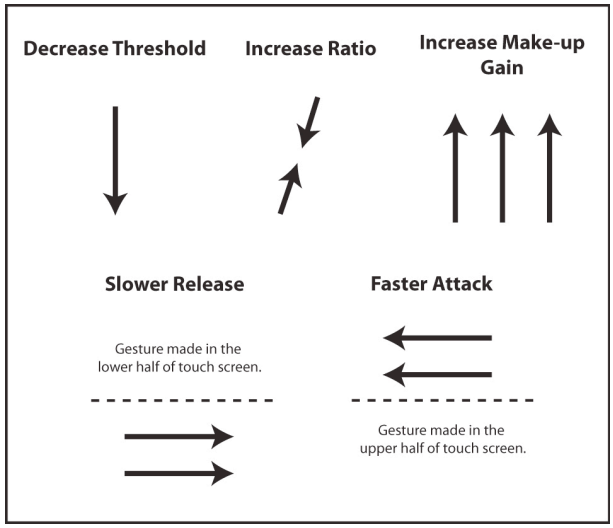

Figure 1 - Elicited Gesture Set

The majority of gestures presented in Figure 1 are arbitrary, semaphoric allocations. Only 'increase ratio' could be described as mimetic, where an inward pinch represents a 'squashing of the audio'. These results are indicative of a fundamental problem that can arise when associating gestures with compression parameters: Direct, or discrete, control mappings have no immediately intuitive or mimetic gestures. The concept of a motion or shape representing a change in the dynamic characteristics of a sound seemed quite abstract to most of the test participants. Therefore, changes to the GUI and rationalisation of controls

should be investigated as a way of making gestural control more intuitive and practical.

3.2 Gesture Set Workflow Test

An objective 'target matching' study observed the time taken for participants to match compression settings with a gestural interface against a traditional mouse and keyboard interface. The gestural controller offered some combined controls, such as the ability to perform both threshold and ratio gestures simultaneously (however, it was observed that the majority of subjects chose to control these independently). The gestures under test were based upon those elicited in Sub-section 3.1.

The underlying parameter set for each continuous gesture was determined using linear interpolation. The start and end point of this range was chosen based on the two most extreme SAFE settings for each control. For example, the parameter settings for 'Hard Compression' and 'Soft Compression' were first chosen as the range and all associated control settings would be derived from this.

The test was devised to assess the influence of different GUIs in a gestural system, in addition to the performance of the gestural controller itself. Participants were asked to match the compression settings of a reference sample. The sample used was a single repeating kick drum. Four interfacing methods were tested: three gestural controllers with differing GUIs and one WIMP (traditional mouse and keyboard) method. The three GUIs were:

• Plug-in GUI - gestural interface was used with the original compressor plug-in GUI as a visual reference.

• Novel GUI - a process-representative GUI was displayed on the touch pad. Examples of two different compressor settings using this GUI are illustrated in figure 2.

• Blind - no visual feedback - engineers were required to remember gestures.

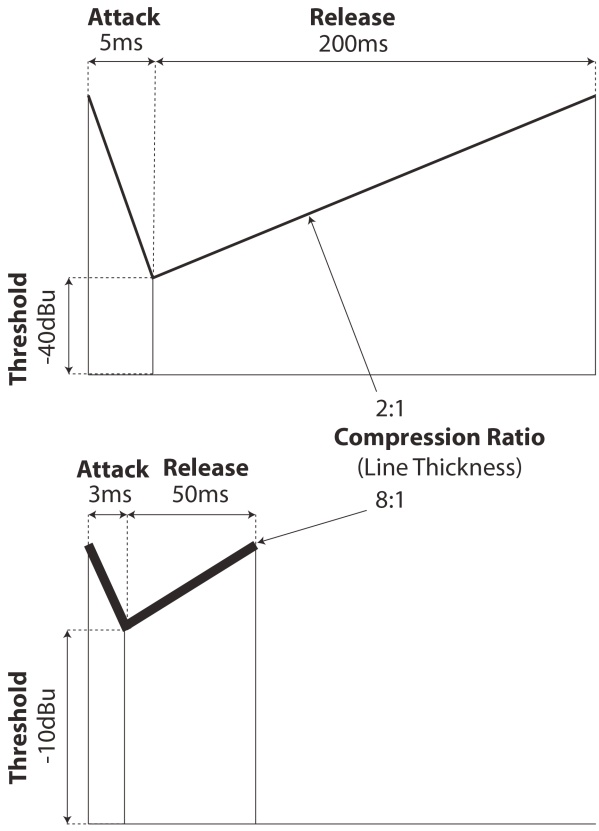

Figure 2 - Novel Gesture Representative GUI

Times were normalised for each participant to eliminate individual performance factors. Therefore, a value of 1.00 represents the slowest time out of all interfacing methods for each participant, with the remaining times given as a proportion of this. Figure 3 shows the Reference Matching Times (RMTs) for each interface.

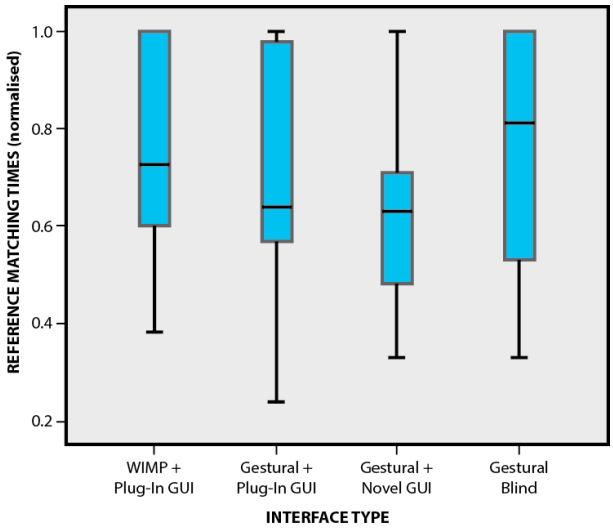

Figure 3 - Reference Matching Times

The mean values appear to show that the novel 'representative GUI' had the quickest reference matching time, however this should be assessed for statistical significance. The data-set is evenly distributed (as determined by skewness and kurtosis values), therefore a repeated measures ANOVA was used. Returned p values were > 0.05, therefore statistical significance cannot be deduced. This reveals that the 'blind mixing' test did not significantly impair the mix times. Subsequently, the GUI could be considered a relatively superfluous design consideration when trying to improve engineer mix times in a gestural system. Furthermore, this suggests that engineers are capable of learning and implementing a gesture-set without visual stimulus.

3.3 Semantically Rationalised Gestural Control Test

Through observations made during the test described in Sub-section 3.2, it is suggested that the difficulty in controlling compression with a gestural interface was down to two contributing factors:

• Inter-parameter influence and dependence - with compression processing, similar results can be achieved with differing settings. For example, a setting with high threshold, high ratio can be perceived as sonically similar to a setting with low ratio, low threshold. This can cause confusion to a mix engineer, especially when mixing to a reference, as it creates an element of uncertainty when experimenting with combinations of settings. Make- up gain is also a contributor to the interdependence of parameter values, particularly in relation to perceived loudness.

• Time and source dependent perception of envelope settings - when an engineer is making adjustments to the envelope of a compressor, the changes are not immediately noticeable as they are dependent on transients within the source. For example, the release of a compressor might be adjusted before the audio source has crossed the threshold, which would have no audible effect. If the source was not at a transient part of the audio, the effects of parameter changes could not be immediately perceived. Often it is the purpose of the GUI to present visual feedback for less obvious parameter changes.

In light of this, it is proposed that the combination of threshold, ratio, gain, attack and release into a single control will offer a much simpler 'gestural shortcut' to the engineer.

It has been identified that there are two immediate design flaws with the 'continuous preset'. Firstly, its effectiveness is source dependent and can be impaired by inconsistent loudness levels between samples, especially when setting attack and release times. Secondly, it was reported during pilot tests that engineers felt that they did not have a sufficient level of control to accurately reach their desired compression settings. A test was devised to asses whether the elicited 'continuous preset' values could allow engineers to sonically match a range of varying compression references. The significance of this test is that the parameter settings in two of the three references are impossible to match precisely with the 'continuous preset'. The three references included in the test were produced using the following compression settings, as detailed in Table 1:

	Reference 1	**Reference 2**	**Reference 3**
Threshold	-25.5dB	-32.5dB	-34.5dB
Ratio	5.0:1	5.2:1	13:1
Attack	5.5mS	10.5mS	8mS
Release	110mS	1200mS	120mS
Gain	3.0dB	8.5dB	7.0dB

Table 1 - Reference Compression Settings

- Reference 1 - The 'continuous gestural preset' is unable to match these settings precisely.
- Reference 2 - The parameter settings can be matched exactly by the 'continuous gestural preset'.
- Reference 3 - This reference represents a compression setting that is far from the boundaries of the 'continuous gestural preset'.

The resulting normalised Reference Matching Times for each participant are presented in Table 2:

Participant #	Reference 1	Reference 2	Reference 3
1	1.00	0.99	0.70
2	0.35	0.53	1.00
3	0.18	0.73	1.00
4	0.34	1.00	0.79
5	1.00	0.54	0.92
6	0.86	0.85	1.00
7	0.41	0.39	1.00
8	0.77	1.00	0.72
9	0.88	0.49	1.00
10	0.63	1.00	0.60
11	1.00	0.77	0.48
12	1.00	0.75	0.86
13	0.68	0.64	1.00
14	0.33	1.00	0.39
15	0.73	1.00	0.55
16	0.30	0.34	1.00
17	0.63	1.00	0.60
18	0.47	0.43	1.00
19	0.39	0.88	1.00
20	0.69	0.95	1.00
Mean	0.63	0.76	0.83
SD	0.27	0.24	0.21

Table 2 - Normalised Reference Matching Times

The Friedman test reported that there was no statistically significant difference between the NRMTs presented in Table 2, with a value of $p = .522$. Therefore, each reference was matched with equal ease. Additionally, participants were asked to listen back to their compressed samples at the end of the test and rate them out of 5 for closeness to the references. It was reported that users felt that they matched all references with equal accuracy.

4 Comparison of Results

Sub-Section 4.1 compares the reference matching speed results for all of the results presented in Section 3. Sub-Section 4.2 analyses the same set of data to determine the reference matching accuracy for each interface.

4.1 Gestural Interface Speed

It was predicted that the implementation of a single high-level control would improve mix times. Figure 4 presents the average normalised speed results for all interfacing methods.

On average, participants were able to suitably mix the reference with the semantically rationalised 'continuous preset' interface in less than half the time of any other interface method. However, if these settings were less accurate than other interfacing methods, then the proposed combination of parameters could be deemed unsuitable.

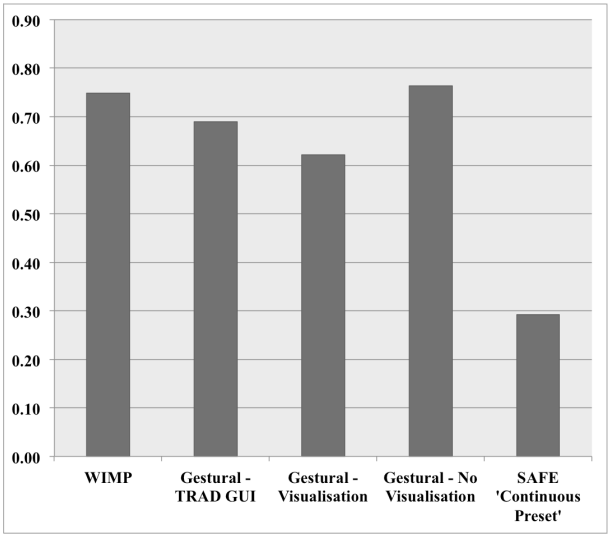

Figure 4 - Average NRMTs for Each Interface Method

4.2 Gestural Interface Accuracy

Cross-correlation analysis was used to assess the accuracy of a user's compression settings in relation to the test reference sample. This result was normalised to give a value between 0 and 1 that represented the similarity between two waveforms (where 1 is identical and 0 is no similarity). Using this technique, the similarity between each participant's compression settings and the reference sample can be used to provide an objective measure of accuracy. Table 3 shows the mean of the cross-correlation values for each interface.

Details of interface numbers:

1. Mouse and Keyboard (WIMP), Plug-In GUI
2. Individual Parameter Gestural Interface, Plug-In GUI
3. Individual Parameter Gestural Interface, Novel GUI
4. Individual Parameter Gestural Interface, Blind Mixing
5. Rationalised 'Gestural Shortcut' Interface, Blind Mixing

	Interface Number 1	2	3	4	5
Mean	0.9850	0.9862	0.9816	0.9767	0.9871
SD	0.0128	0.0107	0.0178	0.0196	0.0190

Table 3 - Average Reference Matching Accuracy

The accuracy measurement datasets are unevenly distributed, as assessed by the Shapiro-Wilk test, therefore the non-parametric Friedman Test is used to assess statistical significance of the data. The Friedman Test concluded that there is a statistically significant difference between the Interface accuracy values, $X2\ (4) = 11.980, p < .05$.

Figure 5 illustrates the improved reference matching accuracy of the simplified 'continuous preset' gestural interface.

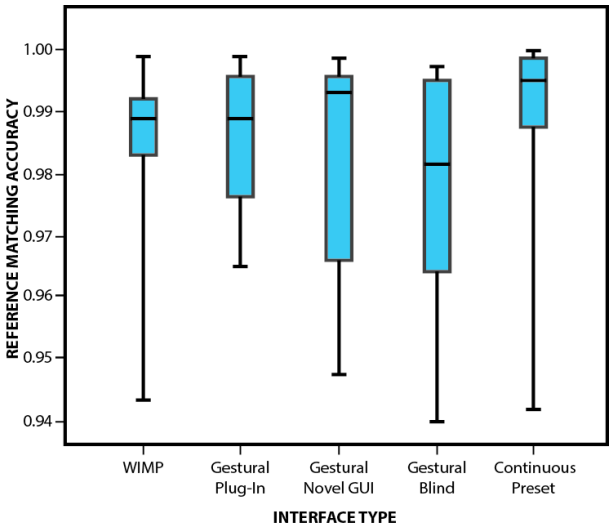

Figure 5 - Interface Accuracy Measurements

5 Conclusion

Testing revealed that the 'continuous preset' gesture allowed test participants to match the reference sample both more quickly and more accurately than other interfacing methods. Furthermore, the 'continuous preset' achieved this performance without the use of a GUI. This suggests that an ideal 'blind mixing tool' could be produced through the use of gestures to control semantically rationalised parameter settings. However, more testing is required with varying audio sources to test the effectiveness completely. The 'continuous preset' may require more contextual information about the source such as the instrumentation or musical genre.

In addition, control parameter information could be derived through analysis of the input signal itself leading to automation of some parameters, as with the system by Giannoulis et al [18].

The 'continuous preset' gestural mix controller might be particularly suited to applications where a quick mix is essential, such as live music production. Studio engineers might feel uncomfortable giving up the lower-level controls, but the gestural mixer could serve as an additional interface that can be used as a starting point for mix sessions.

Within a 3D mix system, perhaps where a VR headset is used to place sound sources in a 360 degree sound field, the less complex and reduced parameter set afforded by 2D gestures could be implemented to allow simplified and intuitive, multi-parameter control of audio effects processing.

References

[1] Lang, B. (2015) *Google Acquires 'Tilt Brush' Developer and Thrive Audio to Add to VR Team*, April 2015 [online] available at: http://www.roadtovr.com/google-acquires-tilt-brush-developer-and-thrive-audio-to-add-to-vr-team/ [Accessed 2nd August 2016]

[2] VisiSonics Corporation (2014) *VisiSonics' RealSpace 3D Audio Software Licensed by Oculus for Virtual Reality* [online] available at: http://www.prnewswire.com/news-releases/visisonics-realspace-3d-audio-software-licensed-by-oculus-for-virtual-reality-278413231.html [Accessed 2nd August 2016]

[3] Two Big Ears (2016) *3DCeption Spatial Workstation* [online] available at: http://www.twobigears.com/spatworks/ [Accessed 2nd August 2016]

[4] Mycroft, J. Reiss, J, D. Stockman, T. (2013) *The Influence of Graphical User Interface Design on Critical Listening Skills*. In: Proceedings of the Sound and Music Computing Conference 2013, Stockholm

[5] Schuts, M. Lipscomb, S. (2007) *Hearing Gestures, Seeing Music: Vision influences*

perceived tone duration. In: Perception, volume 36(6), pp. 888- 897.

[6] Judah, S. (2013) *What is Skeuomorphism? BBC News Magazine*, June 2013 [online] available at: http://www.bbc.co.uk/news/magazine-22840833 [Accessed 4th August 2016]

[7] Nielson, M. Störring, M. Moeslund, T. B. Granum, E. (2008) Gesture Interfaces. In: Kortum, P. HCI Beyond the GUI. London: Kaufmann, pp. 75-106.

[8] Westerman, W. Elias, J, G. (2001) *Multi-Touch: A New Tactile 2-D Gesture Interface For Human - Computer Interaction* In: Proceedings of the Human Factors And Ergonomics Society 45th Annual Meeting, Minneapolis, Minnesota, 8 - 12th October.

[9] Balin, W. Loviscach, J. (2011) *Gestures to Operate DAW Software* In: Audio Engineering Society 130th Convention, London, UK, 13 - 16 May

[10] Bragdon, A. Nelson, E. Li, Y. Hinckley, K. (2011) *Experimental Analysis of Touch-Screen Gesture Designs in Mobile Environments* In: Proceedings of the Conference for Human-Computer Interaction (CHI 2011), Vancouver, Canada, 7 - 12 May.

[11] Ratcliffe, J. (2014). *MotionMix: A Gestural Audio Mixing Controller*. In: Audio Engineering Society 137th Convention, Los Angeles, USA, 9 - 12th October.

[12] Carrascal, J, P. Jordá, S. (2011) Multitouch Interface for Audio Mixing. In: Proceedings of New Interfaces for Musical Expression (NIME 2011). Oslo, Norway.

[13] Gelineck, S. Overholt, D. Büchert, M. Andersen, J. (2013) *Towards an Interface for Music Mixing based on Smart Tangibles and Multitouch*. In: Proceedings of New Interfaces for Musical Expression (NIME 2013), Daejeon, Korea, 27 - 30 May

[14] Selfridge, R. Reiss, J. (2011) *Interactive Mixing Using Wii Controller*. In: Audio Engineering Society 130th Convention, London, UK, May 13 - 16.

[15] Wilson, T. (2015) *The Gestural Control of Audio Processing*. Master's Thesis, University of Huddersfield, UK.

[16] Lech, M. Kostek, B. (2013) *Testing A Novel Gesture-Based Mixing Interface*. In: Journal of the Audio Engineering Society, vol.61, no.5, 2013 May.

[17] Stables, R. Enderby, S. Man, B. D. Fazekas, G. Reiss, J. (2014) *SAFE: A System for the Extraction and Retrieval of Semantic Audio Descriptors* In: 15th International Society for Music Information Retrieval Conference (ISMIR 2014) Taipei, Taiwan, 27 - 31 October.

[18] Giannoulis, D. Massberg, M. Reiss, J, D. (2013) Parameter Automation in a Dynamic Range Compressor. In: *Journal of the Audio Engineering Society, vol.61, no.10, October 2013*

[19] Cartwright, M. Pardo, B. Reiss, J. D. (2014) *MIXPLORATION: Rethinking the Audio Mixer Interface*. In: Proceedings of The International Conference on Intelligent User Interfaces (IUI 2014). Haifa, Israel, 24 - 27 February.

[20] Wilson, T. Fenton, S. (2014) *Two Dimensional Gestural Control of Audio Processing*. In: Audio Engineering Society 136th Convention, Berlin, Germany, 26-29 April.