TECNIACUSTICA VALENCIA · 2015
46º CONGRESO ESPAÑOL DE ACÚSTICA
ENCUENTRO IBÉRICO DE ACÚSTICA
EUROPEAN SYMPOSIUM ON VIRTUAL ACOUSTICS AND AMBISONICS

LISTENER TRACKING STEREO FOR OBJECT BASED AUDIO REPRODUCTION

Marcos Felipe Simón Gálvez
M.F.Simon-Galvez@soton.ac.uk
Dylan Menzies
Filippo Maria Fazi
Institute of Sound and Vibration Research, University of Southampton, UK

Teofilo de Campos
Adrian Hilton
Centre for Vision, Speech and Signal Processing, University of Surrey, UK

ABSTRACT

This paper introduces a method to provide accurate localisation of audio objects with a stereo reproduction system when the listener is away from the sweet spot. To this end, a formulation is presented which allows to compensate the loudspeaker input feeds so that the soundwaves which propagate from both loudspeakers arrive with the same amplitude and phase independently of the listener position. In order to render audio objects, conventional panning techniques are used, which are also adapted depending on the listener position. The formulation has been implemented in a real time system and an initial set of measurements show that the proposed system is able to pan different objects accurately inside the loudspeaker span. The measurements also show that the panning of virtual audio sources outside of the loudspeaker span is limited to spatial positions close to the loudspeakers.

PACS No:43.38.Vk, 43.60.Fg, 43.66.Lj

1 INTRODUCTION

In a stereo reproduction system, the listener, placed in the symmetry axis between the two loudspeakers, i.e., the *sweet spot*, will perceive one or more virtual *phantom source images* located in the span between the two loudspeakers, whose position is adjusted by varying the amplitude or by applying a delay between the signals feeding the two loudspeakers[1]. If the listener, however, moves away from the sweet spot and leans towards one of the loudspeakers, the position of the phantom sources will shift and finally collapse towards the closest loudspeaker, thus compromising the rendering of spatial sound and the initial intention of the sound producer[2]. Nowadays, thanks to the development of computer vision, an accurate estimation of the listener position with respect to that of the loudspeakers can be obtained, which allows to compensate the input loudspeakers feeds, so that the sound waves impinging from both radiators arrive to the listener at the same time and with the same amplitude. An example of this technology is the *Sweetspotter*[3], wherein the stereo loudspeaker feeds are phase compensated to preserve the position of the virtual sources on the original mix.

The approach presented here allows to render audio objects which adapt to the listener position for a maximum spatial immersion. The output of both loudspeakers is first adjusted to the listener position, so that the incoming waves arrive to the listener at the same time, as in the case of equidistant loudspeakers. Once the loudspeaker inputs are compensated to account for off-axis listening positions, the audio object panning is achieved using vector base amplitude panning (VBAP)[4]. The VBAP formulation has been adapted here to allow for off-axis listening position

rendering. The objects can be plane-wave objects or point-source objects. If the listener moves, he or she will perceive that the incoming direction of the plane-wave objects does not change in the case of point-source objects or that or that the position of the sources is static in absolute coordinates in the space. This provides the listener with *motion parallax* cues, creating a greater 3-dimension awareness, as well as perception of depth.

Section 2 of the paper introduces the practical implementation of the system and the compensation used for extending the sweetspot for listening positions which are not symmetrical with respect to the two loudspeakers (*asymmetrical* listening positions). Section 3 presents the VBAP formulation, which has been adapted to account for asymmetrical listening positions. Objective localisation measurements based on the interaural time difference (ITD) have also been presented. Section 4 summarises the main aspects presented in this work.

2 AN ADAPTIVE STEREO SYSTEM WITH LISTENER COMPENSATION

2.1 Implementation

The formulation which is presented in the following sections has been implemented in a real-time system using a standard PC with the MAX MSP 6 software system. The implementation uses a Microsoft Kinect together with the *dp.kinect* MAX patch to perform the listener tracking[5]. A screenshot of the MAX patch is reported in Fig. 1.

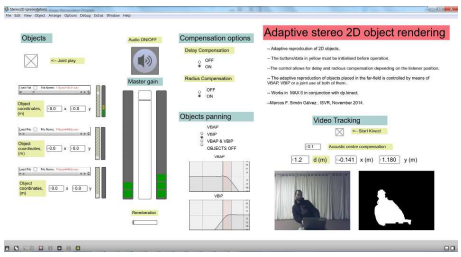


Figure 1: MAX MSP external implementation adaptive stereo object renderer.

2.2 Notation definition

Fig. 2 shows the reference system which is employed for finding the relative position of the listener with respect to the loudspeakers. The vector notation used is as follows: a given vector $\mathbf{x}$ is defined by a magnitude $|x|$ and by an unitary pointing vector $\widehat{\mathbf{x}}$. The vectors $\mathbf{r}_R$ and $\mathbf{r}_L$ define the orientation and distance from each loudspeaker to the centre of the listener's head. These are formulated as

$$\mathbf{r}_L(\mathbf{x}_0) = -\mathbf{x}_0 - \mathbf{d}, \text{ and } \mathbf{r}_R(\mathbf{x}_0) = \mathbf{x}_0 - \mathbf{d}, \tag{1}$$

where $\mathbf{d}$ is the vector representing the right loudspeaker position with respect to the tracking system position.

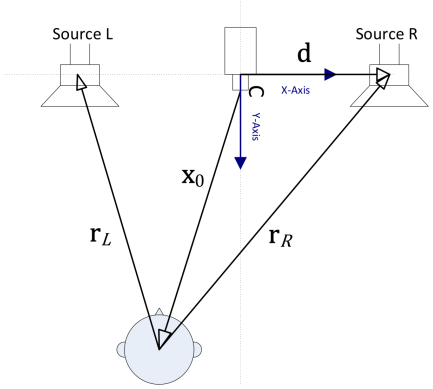


Figure 2: Definition of vectors and quantities used for the adaptive stereo. The Kinect, represented by the camera at C, defines the origin of coordinates ($x = 0$ m,$y = 0$ m).

Using these vectors, the total field generated by both loudspeakers at a listener position $(\mathbf{x}_0)$ is defined as

$$p(\mathbf{x}_0,\omega) = \frac{\tilde{s}_L(\omega)}{4\pi|\mathbf{r}_L(\mathbf{x}_0)|}e^{-jk(|\mathbf{r}_L(\mathbf{x}_0)|)} + \frac{\tilde{s}_R(\omega)}{4\pi|\mathbf{r}_R(\mathbf{x}_0)|}e^{-jk(|\mathbf{r}_R(\mathbf{x}_0)|)}, \tag{2}$$

where ω is the angular frequency, $k = \omega/c_0$ the wavenumber, $c_0 = 344$ m/s is the speed of sound and $e^{-j\omega t}$ represents a time delay. The symbols $\tilde{s}_L(\omega)$ and $\tilde{s}_R(\omega)$ represent the compensated loudspeaker signals.

2.3 Compensation

In order to provide an accurate localisation of virtual sources regardless of the listener position, the first requirement is to ensure that the soundwaves propagating from each loudspeaker arrive at the listener position at the same time instant and with the same amplitude, as it occurs in a symmetric listening situation. This is achieved by applying a gain and a delay to the loudspeaker signals to compensate for the distance attenuation and time of travel of the waves. The gain compensation is performed by multiplying both input signals $s_L(\omega)$ and $s_R(\omega)$ by $|\mathbf{r}_L|$ and $|\mathbf{r}_R|$ respectively. The delay compensation is performed differently depending on the relative sign of the difference $|\mathbf{r}_L(\mathbf{x}_0)| - |\mathbf{r}_R(\mathbf{x}_0)|$. It the listener is closer to the left source, the compensation is applied as

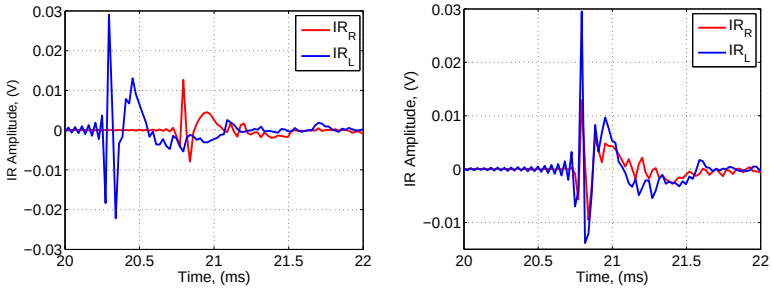
$$\tilde{s}_L(\omega) = s_L(\omega)|\mathbf{r}_L(\mathbf{x}_0)|e^{jk(|\mathbf{r}_L(\mathbf{x}_0)|-|\mathbf{r}_R(\mathbf{x}_0)|)}, \text{and } \tilde{s}_R(\omega) = s_R(\omega)|\mathbf{r}_R(\mathbf{x}_0)|, \tag{3}$$

being s_L the left input signal. If in the other hand the listener is closer to the right source, the right source input feed is given

$$\tilde{s}_R(\omega) = s_R(\omega)|\mathbf{r}_R(\mathbf{x}_0)|e^{jk(|\mathbf{r}_R(\mathbf{x}_0)|-|\mathbf{r}_L(\mathbf{x}_0)|)}, \text{and } \tilde{s}_L(\omega) = s_L(\omega)|\mathbf{r}_L(\mathbf{x}_0)|, \tag{4}$$

being $s_L(\omega)$ and $s_R(\omega)$ the left and right input signals respectively. The frequency independent delays $e^{jk(|\mathbf{r}_L(\mathbf{x}_0)|-|\mathbf{r}_R(\mathbf{x}_0)|)}$ and $e^{jk(|\mathbf{r}_R(\mathbf{x}_0)|-|\mathbf{r}_L(\mathbf{x}_0)|)}$ are implemented in real time by applying a fractional delay[6] of $\delta(t-(|\mathbf{r}_L(\mathbf{x}_0)|-|\mathbf{r}_R(\mathbf{x}_0)|))$ and $\delta(t-(|\mathbf{r}_R(\mathbf{x}_0)|-|\mathbf{r}_L(\mathbf{x}_0)|))$.

The effect of compensation for gain and delay on the impulse responses of each loudspeaker is shown in Fig. 3, for a position which corresponds to ($x = -0.41$ m, $y = 1.15$ m). The impulse responses have been obtained by inverse Fourier transformation of the measured frequency responses between the input to the system and a microphone at the given locations. When no



(a) Pos. 1, OFF (b) Pos. 1, ON

Figure 3: Measured impulse response (IR) of each of the loudspeaker, IR_L for left loudspeaker and IR_R for the right loudspeaker; without compensation, Figure (a), and with compensation, Figure (b), corresponding to a listening position at ($x = -0.41$ m, $y = 1.15$ m).

signal compensation is applied, a large delay between the impulse responses from each loudspeaker can be observed in Fig. 3a. When the compensation is used, which is shown in Fig. 3b, the impulse response peaks coincide. There is, however, a large variability in the peak amplitude, which is not so accurately equalised. This suggests that the position and frequency dependence on the individual radiation patterns of each loudspeaker requires a more accurate compensation than just by compensating according to the distance from the loudspeaker.

3 OBJECT PANNING

Amplitude panning applies different gain values to the input signals of a set of loudspeakers to steer the propagation direction of an original input signal. For a given mono object with associated signal $s_0(\omega)$ and placed in a certain position in the space, it is possible to create a set of loudspeaker signals given by

$$s_L(\omega) = s_0(\omega) g_L(\mathbf{x}_0), \text{ and } s_R(\omega) = s_0(\omega) g_R(\mathbf{x}_0), \tag{5}$$

where $g_L(\mathbf{x}_0)$ and $g_R(\mathbf{x}_0)$ represent the normalised panning gains for the listener position $\mathbf{x}_0$. The loudspeaker signals $s_L(\omega)$ and $s_R(\omega)$ can be further compensated to account for the listener position, as performed in Equations 3 and 4, to give compensated loudspeaker signals $\tilde{s}_L(\omega)$ and $\tilde{s}_R(\omega)$.

3.1 Object panning strategies

The orientation of the different audio objects is adjusted using panning techniques as vector base amplitude panning (VBAP)[4]. The VBAP formulation is used to adjust the amplitude of the loudspeaker feeds so that an object is perceived as being located at a given position, as shown in Fig. 4, where the normalised input feeds $g_L(\mathbf{x}_0)$ and $g_R(\mathbf{x}_0)$ are calculated to keep the position of the audio object static independently of the listener position.

For a given listener position, an object orientation vector $\mathbf{p}_0$ is created according to

$$\mathbf{p}_0(\mathbf{x}_0) = g'_L(\mathbf{x}_0)\hat{\mathbf{r}}_L(\mathbf{x}_0) + g'_R(\mathbf{x}_0)\hat{\mathbf{r}}_R(\mathbf{x}_0), \tag{6}$$

where $g'_L(\mathbf{x}_0)$ and $g'_R(\mathbf{x}_0)$ are the unnormalised gain factors of each loudspeaker and $\hat{\mathbf{r}}_L(\mathbf{x}_0)$ and $\hat{\mathbf{r}}_R(\mathbf{x}_0)$ are the unitary pointing vectors of each loudspeaker. In the original VBAP formulation it is

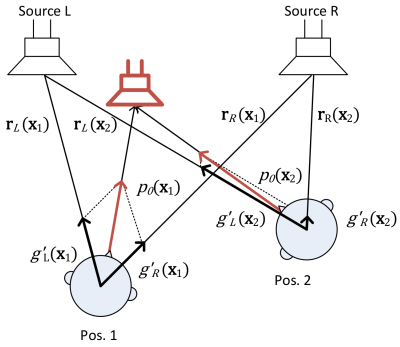


Figure 4: Example of how the VBAP gains are adjusted for two listener position so that the object stays static relative to the listener at a desired 2D coordinate.

assumed that the listener is placed at the same distance from both loudspeakers and hence $\widehat{\mathbf{r}}_L$ and $\widehat{\mathbf{r}}_R$ are only calculated once. In the formulation presented here, however, these account for asymmetrical listening situations. The above equation can be rewritten in matrix notation

$$\begin{bmatrix} p_{0_y}(\mathbf{x}_0) \\ p_{0_x}(\mathbf{x}_0) \end{bmatrix} = \begin{bmatrix} r_{L_y}(\mathbf{x}_0) & r_{R_y}(\mathbf{x}_0) \\ r_{L_x}(\mathbf{x}_0) & r_{R_x}(\mathbf{x}_0) \end{bmatrix} \begin{bmatrix} g'_L(\mathbf{x}_0) \\ g'_R(\mathbf{x}_0) \end{bmatrix}, \tag{7}$$

where the subscripts $_y$ and $_x$ indicate the orthogonal projections on the y and x axis of the pointing vectors $\widehat{\mathbf{r}}_L(\mathbf{x}_0)$ and $\widehat{\mathbf{r}}_R(\mathbf{x}_0)$ and of the virtual source pointing vector $\mathbf{p}(\mathbf{x}_0)$. For a virtual source placed at a position defined by the object pointing vector $\mathbf{p}_0$, the required gains are obtained by[4]

$$\begin{bmatrix} g'_L(\mathbf{x}_0) \\ g'_R(\mathbf{x}_0) \end{bmatrix} = \begin{bmatrix} r_{L_y}(\mathbf{x}_0) & r_{R_y}(\mathbf{x}_0) \\ r_{L_x}(\mathbf{x}_0) & r_{R_x}(\mathbf{x}_0) \end{bmatrix}^{-1} \begin{bmatrix} p_{0_y}(\mathbf{x}_0) \\ p_{0_x}(\mathbf{x}_0) \end{bmatrix}. \tag{8}$$

The unnormalised gains $g'_L(\mathbf{x}_0)$ and $g'_R(\mathbf{x}_0)$ are then converted into normalised gains $g_L(\mathbf{x}_0)$ and $g_R(\mathbf{x}_0)$, so that the acoustic power at the listener position is equal to a certain value, here assumed to be unitary. The normalisation can be performed assuming coherent summation, as it is understood to occur at low frequencies, or assuming incoherent summation, as it occurs at higher frequencies. If coherent normalisation is used, the loudspeaker feeds are normalised so that $g_L(\mathbf{x}_0) + g_R(\mathbf{x}_0)$=1. However, coherent normalisation has the risk of turning the system unstable for phantom source positions in which the loudspeaker feeds are equal in magnitude but with opposite phase, as it happens when objects are panned outside the loudspeaker span. Furthermore, as the system is aimed to be used in a, normal, reverberant room, loudness is best preserved if both gains are normalised with the square root of the powers[7], as also uncorrelated reflections contribute to the general pressure level.

Thus, in order to avoid system instability, incoherent normalisation is preferred throughout the whole frequency band, with the constraint that $\sqrt{g_L^2(\mathbf{x}_0) + g_R^2(\mathbf{x}_0)}$=1. Below 1.5 kHz, it is possible to place objects outside of the stereo span, thanks to the coherent summation of the waves. The panning gains for the low frequency range, $g_L^{(LF)}(\mathbf{x}_0)$ and $g_R^{(LF)}(\mathbf{x}_0)$, are given by

$$g_L^{(LF)}(\mathbf{x}_0) = \frac{g'_L(\mathbf{x}_0)}{\sqrt{g'^2_L(\mathbf{x}_0) + g'^2_R(\mathbf{x}_0)}}, \tag{9}$$

and

$$g_R^{(LF)}(\mathbf{x}_0) = \frac{g'_R(\mathbf{x}_0)}{\sqrt{g'^2_L(\mathbf{x}_0) + g'^2_R(\mathbf{x}_0)}}. \tag{10}$$

Above 1.5 kHz, the gains are normalised in the same way as for the low frequency region when the phantom source lies inside the stereo span. When the phantom source is outside the stereo span, the loudspeaker feed closer to the phantom source is set to 1, with the other loudspeaker feed set to 0. The cut-off frequency of 1.5 kHz has been selected heuristically and may depend on the distance $2|\mathbf{d}|$ between the loudspeakers and the type of room where the system is placed. The high frequency panning gains, $g_L^{(HF)}(\mathbf{x}_0)$ and $g_R^{(HF)}(\mathbf{x}_0)$, are defined by

$$g_L^{(HF)}(\mathbf{x}_0) = \frac{g'_L(\mathbf{x}_0)}{\sqrt{g'^2_L(\mathbf{x}_0) + g'^2_R(\mathbf{x}_0)}} \text{ if } \quad (g'_L(\mathbf{x}_0) > 0, g'_R(\mathbf{x}_0) > 0), \text{or else} \quad 1 \text{ if } g'_R(\mathbf{x}_0) < 0, \tag{11}$$

and

$$g_L^{(HF)}(\mathbf{x}_0) = \frac{g'_L(\mathbf{x}_0)}{\sqrt{g'^2_L(\mathbf{x}_0) + g'^2_R(\mathbf{x}_0)}} \text{ if } \quad (g'_L(\mathbf{x}_0) > 0, g'_R(\mathbf{x}_0) > 0), \text{or else} \quad 1 \text{ if } g'_R(\mathbf{x}_0) < 0,$$

The normalised panning gains for the low and high frequency bands are then compensated depending on the listening position according to Equations 3 and 4 to give the compensated and normalised panning gains, $\tilde{g}_L^{LF}(\omega)$, $\tilde{g}_R^{LF}(\omega)$, $\tilde{g}_L^{HF}(\omega)$ and $\tilde{g}_R^{HF}(\omega)$. For a given audio object $s_0(\omega)$, this is passed through a filtering stage which decomposes the object into a low frequency band and a high frequency band. The output of such filtering stage is afterwards multiplied by the respective low and high frequency compensated and normalised panning gains and then combined to give the compensated loudspeaker signals $\tilde{s}_L(\omega)$ and $\tilde{s}_R(\omega)$.

3.2 Localisation measurements

In order to assess the ability of the real time system, an initial set of measurements has been performed. These have been done in an audio listening room with a low reverberation time using a Kemar dummy head. Impulse responses to both ears of the dummy head have been calculated, and the ITD has been obtained as parameter to assess the performance of the system. The ITD function, $\Psi(t)$, is obtained via the interaural cross-correlation function[8], which is calculated according to

$$\Psi(t) = \underset{\tau}{argmax}\left(E\left\{h_L(t)h_R(t+\tau)\right\}\right), \tag{12}$$

where $h_L(t)$ and $h_R(t)$ are left and right ear impulse responses. These responses are low-pass filtered below 700 Hz, since this frequency range is considered to be of importance for the ITDs[9].

The measurements have been performed at the sweet spot, at an asymmetrical position, ($x = -0.4$ m, $y = 2.0$ m), and at a listening position where the listener is outside of the loudspeaker span, ($x = -1.0$ m,$y = 2.0$ m). The dummy head has been rotated towards the target source direction, which has been panned using the formulation described above. In this case, a 0 ITD would ideally be obtained for a plane wave propagating from that direction. The measured results are shown in Fig. 5, where they are compared against the ITD measured for a single loudspeaker as a function of head orientation. When the dummy head is on the sweet spot, and each loudspeaker is at an angle of 30 ° from the listener position, the results show how the system is able to pan a virtual source with a less than 0.1 ms of ITD difference between -40° and 60°. The ITD difference is lowest between the loudspeaker span 0 at -30° and 30°, where the dummy head is orientated towards the loudspeakers. Between -60° and -40° the ITD difference grows, probably due to errors introduced on the tracking system.

As the dummy head is moved towards the left of the loudspeakers symmetry axis, the ITD error gets larger for sources panned towards negative angles (i.e., towards side where the dummy head

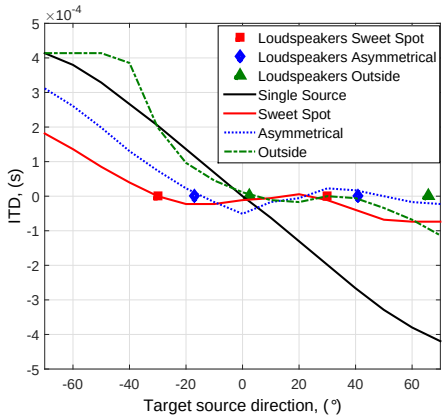


Figure 5: ITD difference obtained from panning a virtual source between -70° and 70°. Results measured at the sweet spot, at an asymmetrical listening position, ($x = -0.4$ m, $y = 2.0$ m), and at a listening position outside of the loudspeaker span, ($x = -1.0$ m,$y = 2.0$ m). The black thick dashed line shown for comparison is the measured ITD from a single loudspeaker.

has been moved to). This, however, remains very close to 0 for the case when the dummy head is at an asymmetrical listening position, ($x = -0.4$ m, $y = 2.0$ m) between -20°, where the closest loudspeaker is placed, and 60 °. If the dummy head is moved further to the left, so that it is now placed outside of the loudspeaker set up, the localisation is still good, i.e., the ITD difference is close to 0, for virtual sources incoming between 0° and 40°. For virtual sources panned between 40° and $60°$ the ITD difference grows to be about 1 ms at $60°$. This indicates that the perceived localisation of virtual sources coming from such direction is affected by small inaccuracies in the tracking system and in the loudspeaker radiation pattern. Nevertheless, it is shown how the ITD difference for target source directions inside the loudspeaker span is small compared to that measured from a single loudspeaker rotated along the dummy head.

4 CONCLUSION

An adaptive stereo system for object based reproduction has been presented. By constantly calculating the position of the listener with respect to both loudspeakers, the proposed formulation adjusts the loudspeaker signals when the listener is outside of the sweet spot, assuring that both loudspeaker feeds arrive with the same amplitude and at the same time. Whilst it is not totally possible to adjust for the loudspeaker directivity, which is frequency and orientation dependent, it is possible to adjust the time arrival of both signals.

The proposed approach is also used for the reproduction of audio objects. Vector base amplitude panning (VBAP) is used for obtaining loudspeaker gains which allow controlling the position of the audio objects, with the loudspeaker feeds also adapted to the listener position. The loudspeaker feeds are calculated differently for low and high frequencies, with a cut off frequency of 1.5 kHz. Below this frequency, the original VBAP formulation is used independently of the listener-object position, and above that frequency, the VBAP formulation is used for objects placed inside the loudspeaker span. This allows to extend the stereo span for the low frequencies, where both pressure waves add coherently. Two normalisation options have been evaluated to weight the panning gains at low frequencies, coherent and incoherent normalisation.

Localisation measurements based on the ITD have shown that this is only accurate when the virtual audio sources are panned inside of the loudspeaker span, even for listener positions outside of the loudspeaker span. The ability to pan audio objects outside of the loudspeaker span requires a very accurate and calibrated system, and it is just obtained for certain listening positions.

5 ACKNOWLEDGEMENTS

The authors of the paper would like to acknowledge the support of the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership.

References

[1] A. D. Blumlein, “British patent specification 394,325 (improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems),” *J. Audio Eng. Soc*, vol. 6, no. 2, pp. 91–98, 130, 1958.

[2] B. B. Bauer, “Broadening the area of stereophonic perception,” *J. Audio Eng. Soc*, vol. 8, no. 2, pp. 91–94, 1960.

[3] S. Merchel and S. Groth, “Adaptively adjusting the stereophonic sweet spot to the listeners position,” *J. Audio Eng. Soc*, vol. 58, no. 10, pp. 809–817, 2010.

[4] V. Pulkki, “Virtual sound source positioning using vector base amplitude panning,” *J. Audio Eng. Soc*, vol. 45, no. 6, pp. 456–466, 1997.

[5] D. Phurrough, “dp.kinect, MAX MSP patch for Microsoft Kinect .” [Online]. Available: https://hidale.com/shop/dp-kinect/

[6] T. Laakso, V. Valimaki, M. Karjalainen, and U. Laine, “Splitting the unit delay [fir/all pass filters design],” *Signal Processing Magazine, IEEE*, vol. 13, no. 1, pp. 30–60, Jan 1996.

[7] M.-V. Laitinen, J. Vilkamo, K. Jussila, A. Politis, and V. Pulkki, “Gain normalization in amplitude panning as a function of frequency and room reverberance,” in *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*, Aug 2014.

[8] B. F. G. Katz and M. Noisternig, “A comparative study of interaural time delay estimation methods,” *The Journal of the Acoustical Society of America*, vol. 135, no. 6, pp. 3530–3540, 2014.

[9] D. Howard and J. Angus, *Acoustics and Psychoacoustics*. Taylor & Francis, 2009.