SMALL ARRAY REPRODUCTION METHOD FOR AMBISONIC ENCODINGS USING HEADTRACKING

Dylan Menzies and Filippo Maria Fazi

Institute of Sound and Vibration Research, University of Southampton, UK

ABSTRACT

Compensated Amplitude Panning (CAP) is a spatial audio reproduction method for loudspeakers that takes the listener head orientation into account. It can produce stable images in any direction with as few as two loudspeakers. In its original form CAP is inherently an object-based method, where each image is produced separately. A direct method is presented here for dynamically decoding a first order Ambisonic encoding that is equivalent to using CAP to separately reproduce the constituents of the encoding. Both the stereo and multichannel cases are considered. Ambisonic decoding enables complex scenes to be reproduced with little cost. Ambisonic encodings are now used widely for 360° video, and other applications.

Index Terms— 3D sound, spatial audio, Ambisonics, B-format, 360 video

1. INTRODUCTION

Amplitude panning is a method for producing a spatial audio image in which 2 or more waves combine coherently at the listener position, each carrying the same signal but independent gains. For some choices of plane wave directions and gains the listener perceives an image, or phantom source, from a definite direction, a phenomena known as summing localisation [1]. The direction of the image can be varied continuously by varying the gains.

Below $\approx$ 1000Hz the perception of image direction is mainly determined by the Interaural Time Difference (ITD) cue. In this frequency range, a central stereo image, produced by panning with 2 loudspeakers, is unstable. If the listener faces straight ahead the image is also straight ahead. As the listener turns away from this direction the image moves in the direction of the listener, as illustrated in Fig. 1 [2, 3, 4]. A typical scene contains multiple images in different directions, so at any moment images that are not directly ahead of the listener or inline with a loudspeaker will be distorted. The distortion is greater when the angle between the loudspeakers, viewed from the listener, is increased. For example the listener can approach a stereo pair until the loudspeakers are 180° apart. In this position an image panned to the centre would be completely unstable. Producing consistent ITD cues when the head rotates, otherwise known as *dynamic ITD cues*, is important for localisation [1, 5, 6, 7].

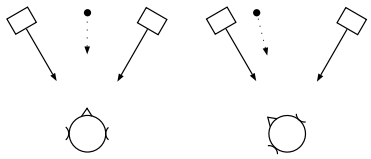

Fig. 1. The black dot indicates the direction of the image when 2 loudspeakers each have the same signal, for different head directions.

The change in the panned image direction when the head is rotated is caused by the ITD cue not matching that of a static source for each head angle. Compensated Amplitude Panning (CAP), is an extension of conventional panning methods in which the ITD cues are corrected by modifying the gains to take account of the head orientation of the listener [8]. Tracking the listener accurately in real-time with low latency is a challenging requirement for this system. However such tracking technology is progressing rapidly, driven by a wide range of applications.

CAP was initially developed for 2 loudspeaker reproduction (Stereo-CAP), which produces more stable images than conventional stereo across the front stage. Further more, the method can produce images in any direction, because the ITD is reproduced accurately for nearly all head orientations. In particular dynamic ITD cues generated by small head movements allow the resolution of front-back ambiguities, and provide elevation cues.

To cover the full bandwidth CAP can be combined with high frequency reproduction methods. CAP requires only 2 loudspeakers that are capable of driving the ITD frequency range, while the high frequency range can be driven using smaller and lighter loudspeakers, that are practical to use in higher numbers. Energy based panning, or *Vector Base Intensity Panning (VBIP)* [9] can be combined with Stereo-CAP to provide a very stable full bandwidth front stage. Stereo-CAP provides low frequency coverage elsewhere, which is useful for immersive ambience and reverberation. High frequency coverage can also be provided in all directions using *crosstalk cancellation* [10, 11].

An extension to Stereo-CAP for near-field images has been made by matching the low frequency ILD (Inter-aural Level Difference) to that of a near source. This is possible using complex panning gains realized with a 1st order filter [12].

For a low frequency spherical head model, [8], the condition that the ITD and ILD cues match with the target plane wave can be formulated as

$$\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_I - \boldsymbol{r}_V) = 0 \tag{1}$$

where $\hat{\boldsymbol{r}}_I$ is the direction of the image, $\hat{\boldsymbol{r}}_R$ is the inter-aural axis, and $\boldsymbol{r}_V$ is the Makita localisation vector that represents the sound field at low frequencies [13, 8]. If the field is produced by panning, the waves at the listener can be approximated as plane waves provided the listener is not so close to the loudspeakers that near-field cues are significant. In this case the Makita vector is given by

$$\boldsymbol{r}_V = \frac{\sum g_i \hat{\boldsymbol{r}}_i}{\sum g_i} \tag{2}$$

where g_i are the gains of the source signal *at the listener*, and $\hat{\boldsymbol{r}}_i$ are the direction vectors of the loudspeakers relative to the listener [8]. The gains at the loudspeakers are compensated for the variable distance to the loudspeakers. Since the wave amplitude falls by $1/r$ the compensated loudspeaker gains are $r_i g_i$. Also delays are introduced to the loudspeaker feeds so that the signals at the listener are in phase. These compensations depend on accurate knowledge of the ambient speed of sound, as well as the distances.

Non-trivial solutions for the Stereo-CAP gains can be found using the constraint (1) and an additional constraint normalising the total gain, which fixes the overall level,

$$g_1 = \frac{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_I - \hat{\boldsymbol{r}}_2)}{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_1 - \hat{\boldsymbol{r}}_2)} \quad g_2 = \frac{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_I - \hat{\boldsymbol{r}}_1)}{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_2 - \hat{\boldsymbol{r}}_1)} \tag{3}$$

These panning laws were tested objectively by calculating the resulting cues at different frequencies for a KEMAR dummy head [8]. The perceived directional error was then calculated and found to be within a Minimum Audible Angle (MMA) [14] for a wide range of target images and head orientations. Subjective tests were carried out to evaluate the stability of images in all directions. Dynamic head tracking was used to allow natural unrestricted listening. The tests showed that images between loudspeakers were improved, and further more stable images could now be created in all other directions.

It is helpful to visualise the 3-dimensional vectors in the solution. Fig. 2 shows a plan view of these vectors. This is called a *Makita diagram* here since each point on this diagram corresponds to a value of $\boldsymbol{r}_V$, rather than a position in 3-dimensional space. The dotted circle is a cross section through a sphere of radius 1. A point $\boldsymbol{r}_V$ on the circle or sphere corresponds to a plane wave, such as that from a distant loudspeaker or source. The dotted line represents a plane perpendicular to the page containing all the values of $\boldsymbol{r}_V$ of sound

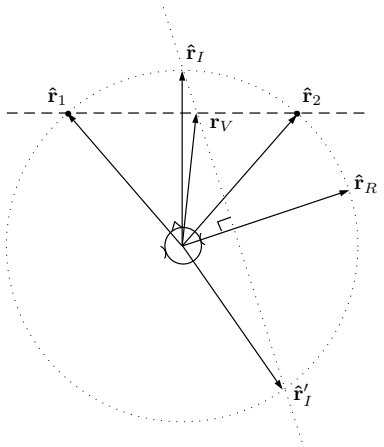

Fig. 2. Makita diagram for Stereo CAP, in plan view, for a listener facing towards left of centre of the stereo array. The Makita vector is to the right of centre in order to keep the image central. Shown are loudspeaker directions $\hat{\boldsymbol{r}}_\mathbf{1}$, $\hat{\boldsymbol{r}}_\mathbf{2}$ the inter-aural direction $\hat{\boldsymbol{r}}_R$, image direction $\hat{\boldsymbol{r}}_I$ and Makita vector $\boldsymbol{r}_V$

fields that produce an image $\hat{\boldsymbol{r}}_I$. The image is not unique, since there is a circle of consistent images, where the plane intersects with the sphere, the *cone of confusion*. The dashed line shows the values of $\boldsymbol{r}_V$ that can be produced by panning using the 2 loudspeakers. Where the plane and line cross is the single value of $\boldsymbol{r}_V$ that can produce the image using stereo panning. The method is valid whatever the direction of the image, even if it is behind or above.

The panning gains are positive for values of $\boldsymbol{r}_V$ between $\hat{\boldsymbol{r}}_\mathbf{1}$ and $\hat{\boldsymbol{r}}_\mathbf{2}$. Outside this region, one of the gains is negative, and there is cancellation of the pressure at the listener. The cancellation implies the sum of gain magnitudes $\sum |g_i|$ is greater than the sum of gains $\sum g_i$. Since the reproduction error due to each gain generally accumulates, then for given $\sum g_i$ the total error increases as the sum of gain magnitudes $\sum |g_i|$, and degree of cancellation. Reproduction error is due to inaccuracies in the head model, the audio hardware, and the tracking of the listener and loudspeakers.

If the listener faces towards the side of the loudspeakers, the plane and line become close to parallel, and the denominators vanish. The gains become large and polarised, and the error increases. The common gain due to the denominators can be limited, however the perceived image level will fade. This issue has been solved by extending CAP for 3 or more loudspeakers [15].

The gain encoding equations (3) are inherently objected-based. A pair of gains is produced for a single image in the direction $\hat{\boldsymbol{r}}_I$. The loudspeaker signals for multiple images can be summed to produce a scene containing these images. In the next section a decoding method is derived that converts a channel-based Ambisonic signal directly to stereo

loudspeaker signals, which are equivalent to those produced by summing stereo-CAP signals.

2. AMBISONIC DECODING

The loudspeaker signals $L_1,\ L_2$ for several images can be formed by summing the separate image signals. The image signals are written I_n where the index n is over the set of images. The gain for the ith loudspeaker and nth image is $g_{i,n}$. Then the first loudspeaker signal is

$$L_1 = \sum_n g_{1,n}\, I_n \tag{4}$$

Substituting for the gain definition in (3), and separating summed terms,

$$L_1 = \frac{1}{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_1 - \hat{\boldsymbol{r}}_2)} \sum \hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_{I_n} - \hat{\boldsymbol{r}}_2)\, I_n \tag{5}$$

$$= \frac{\hat{\boldsymbol{r}}_R}{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_1 - \hat{\boldsymbol{r}}_2)} \cdot (\sum \hat{\boldsymbol{r}}_{I_n} I_n - \hat{\boldsymbol{r}}_2 \sum I_n) \tag{6}$$

The terms $\sum \hat{\boldsymbol{r}}_{I_n} I_n$ and $\sum I_n$ identify with components of a 1st order Ambisonic signal, or *B-format* signal, (W, X, Y, Z) that encodes all the sources, [16]. The components of $\hat{\boldsymbol{r}}_{I_n}$ are direction cosines of the image direction vectors, and so coincide with the B-format figure-of-eight encoding directivities for the signals X, Y, Z. Then,

$$L_1 = \frac{\hat{\boldsymbol{r}}_R}{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_1 - \hat{\boldsymbol{r}}_2)} \cdot \left(\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} - \sqrt{2}\, W \hat{\boldsymbol{r}}_2 \right) \tag{7}$$

Similarly for the other loudspeaker,

$$L_2 = \frac{\hat{\boldsymbol{r}}_R}{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_2 - \hat{\boldsymbol{r}}_1)} \cdot \left(\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} - \sqrt{2}\, W \hat{\boldsymbol{r}}_1 \right) \tag{8}$$

The $\sqrt{2}$ factor is included to match the weightings used in the original B-format definition. Other normalisations are used, and require different weightings.

A more general solution for two or more loudpseakers, is given by Multichannel CAP (MCAP) [15]. This solution minimises the total radiated energy, which is usually the main source of error in a reverberant space:

$$g_i = \frac{(\eta\phi - \beta)\alpha_i + \gamma - \beta\phi}{r_i^2(\gamma\eta - \beta^2)} \tag{9}$$

where,

$$\alpha_i = \hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_i\ ,\quad \phi = \hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_I \tag{10}$$

$$\eta = \sum \frac{1}{r_i^2}\ ,\quad \beta = \sum \frac{\alpha_i}{r_i^2}\ ,\quad \gamma = \sum \frac{\alpha_i^2}{r_i^2} \tag{11}$$

The expression for the gains (12) can be rearranged to isolate the dependence on image direction contained in ϕ,

$$g_i = \frac{\phi(\eta\alpha_i - \beta) + (\gamma - \beta\alpha_i)}{r_i^2(\gamma\eta - \beta^2)} \tag{12}$$

This can be written, for convenience, using two parameters $a_i,\ b_i$,

$$a_i = \frac{\eta\alpha_i - \beta}{r_i^2(\gamma\eta - \beta^2)} \quad b_i = \frac{\gamma - \beta\alpha_i}{r_i^2(\gamma\eta - \beta^2)} \tag{13}$$

so that,

$$g_i = a_i\phi + b_i \tag{14}$$

The gains for multiple images, indexed by n, can be written

$$g_{i,n} = a_i\phi_n + b_i \tag{15}$$

since a_i and b_i depend only on the loudspeaker directions, not the images. The loudspeaker signals are the sum over the signals for each image,

$$L_i = \sum_n g_{i,n} I_n \tag{16}$$

$$= \sum_n (a_i\phi_n + b_i) I_n \tag{17}$$

$$= \sum_n (a_i\, \hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_{I_n} + b_i) I_n \tag{18}$$

$$= a_i \hat{\boldsymbol{r}}_R \cdot \sum_n \hat{\boldsymbol{r}}_{I_n} I_n + b_i \sum_n I_n \tag{19}$$

As for the stereo case, the sums can be identified with B-format signals,

$$L_i = a_i \hat{\boldsymbol{r}}_R \cdot \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + \sqrt{2}\, b_i W \tag{20}$$

The decoding formula (7), (8) and (20) have been checked by substituting in the initial conditions. The method will be referred to as B-format CAP (BCAP) for short. The derivation shows that the decoding formulae apply to any B-format signal composed of discrete image signals. Furthermore, since any field can be decomposed with arbitrary precision using plane waves, they apply to a B-format signal derived from any scene, recorded or synthesized, possibly containing near or diffuse sources, or reverberation.

CAP is valid in the low frequency range up to $\approx$ 1000 Hz. The decoding formula show that 1st order B-format encoding in this frequency range is sufficient to encode a scene that is then decoded with BCAP. This is expected because for conventional Ambisonic reproduction 1st order encoding is sufficient in this frequency range.

Discrete CAP allows the listener to change position while keeping the desired image directions. This allows images to be located in the near or far-field using parallax cues. BCAP,

however produces fixed image directions that are independent of the listener's position, so there is no parallax variation and the overall scene appears distant, providing there are no conflicting near-field cues. For the case of a moving listener BCAP is useful for encoding a background scene or *bed*. Additional foreground images can be produced using the original discrete panning CAP formulation.

BCAP and other CAP variants have been implemented in a real-time C++ / Python framework for spatial sound rendering, called the *Versatile Interactive Software Rendering framework (VISR)* [17]. Software plugins have been produced that run on digital audio workstation software. This environment allows different test cases to be created quickly and compared side by side.

3. PERFORMANCE

By construction, BCAP produces identical loudspeaker signals to those produced by applying CAP to the component signals and summing. In particular objective and subjective performance for discrete images is identical to that reported previously for CAP reproduction [8, 18]. The process is linear and filterless and the so transients are not smeared (A crossover filter maybe introduced if used to separate low frequencies from broadband source signals). The ill-conditioning, when the listener faces to the sides, remains in the stereo BCAP case, and can be managed by limiting the gains, as before. For the multichannel case the ill-conditioning is removed, and the listener can change orientation without any artefacts.

BCAP is particularly useful for rendering complex diffuse backgrounds, including reverberation, either synthesized or recorded using 3D microphones. In principle these could be created using instances of CAP, but this would be very inefficient.

4. ACKNOWLEDGMENT

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) "S3A" Programme Grant EP/L000539/1, and the BBC Audio Research Partnership. No new data was created in this work.

5. REFERENCES

[1] Jens Blauert, *Spatial hearing*, Cambridge, MA: MIT Press, 1997.

[2] Benjamin Bernfeld, "Attempts for better understanding of the directional stereophonic listening mechanism," in *Audio Engineering Society Convention 44*, March 1973, number C-4.

[3] Michael Anthony Gerzon, "General metatheory of auditory localisation," in *92nd Audio Engineering Society Convention, Vienna*, 1992, number 3306.

[4] Ville Pulkki, "Compensating displacement of amplitude-panned virtual sources," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, Jun 2002.

[5] Hans Wallach, "On sound localization," *J. Acoust. Soc. Am*, vol. 10, pp. 270–274, 1939.

[6] Hans Wallach, "The role of head movements and vestibular and visual cues in sound localization.," *Journal of Experimental Psychology*, vol. 27, no. 4, pp. 339, 1940.

[7] Bosun Xie and Dan Rao, "Analysis and experiment on summing localization of two loudspeakers in the median plane," in *Audio Engineering Society Convention 139*. Audio Engineering Society, 2015.

[8] Dylan Menzies, Marcos F. Simon Galvez, and Filippo Maria Fazi, "A low frequency panning method with compensation for head rotation," *IEEE Trans. Audio, Speech, Language Processing*, vol. 26, no. 2, February 2018.

[9] Jean-Marc Jot, Veronique Larcher, and Jean-Marie Pernaux, "A comparative study of 3-d audio encoding and rendering techniques," in *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*, Mar 1999.

[10] Bishnu S. Atal and Manfred R Schroeder, "Apparent sound source translator," Feb. 22 1966, US Patent 3,236,949.

[11] Ole Kirkeby, Philip A. Nelson, and Hareo Hamada, "Virtual source imaging using the stereo dipole," in *Audio Engineering Society Convention 103*, Sep 1997.

[12] Dylan Menzies and Filippo Maria Fazi, "Spatial reproduction of near sources at low frequency using adaptive panning," in *Proc. TecniAcustica, Valencia*, October 2015.

[13] Y Makita, "On the directional localization of sound in the stereophonic sound field," *E.B.U Review*, vol. A, no. 73, pp. 102–108, 1962.

[14] Allen William Mills, "On the minimum audible angle," *The Journal of the Acoustical Society of America*, vol. 30, no. 4, pp. 237–246, 1958.

[15] Dylan Menzies and Filippo M. Fazi, "Surround sound without rear loudspeakers: Multichannel compensated amplitude panning and ambisonics," in *Proc. DAFX18*, 2018.

[16] David G Malham and Anthony Myatt, "3-d sound spatialization using ambisonic techniques," *Computer music journal*, vol. 19, no. 4, pp. 58–70, 1995.

[17] Andreas Franck and Filippo Maria Fazi, "Visr—a versatile open software framework for audio signal processing," in *Audio Engineering Society Conference: 2018 AES International Conference on Spatial Reproduction-Aesthetics and Science*. Audio Engineering Society, 2018.

[18] Dylan Menzies and Filippo Maria Fazi, "A complex panning method for near-field imaging," *IEEE Trans. Audio, Speech, Language Processing*, 2018, accepted for publication.