OPEN ACCESS Freely available online

D. Menzies and and F. M. Fazi, "Multichannel Compensated Amplitude Panning, An Adaptive Object-Based Reproduction Method"
J. Audio Eng. Soc., vol. 67, no. 7/8, pp. 1–8, (2019 July/August.).
DOI: https://doi.org/10.17743/jaes.2019.0029

Multichannel Compensated Amplitude Panning, An Adaptive Object-Based Reproduction Method

DYLAN MENZIES, *AES Member,* **AND FILIPPO MARIA FAZI,** *AES Member*
(d.menzies@soton.ac.uk) (Filippo.Fazi@soton.ac.uk)

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

Conventional approaches for surround sound panning require loudspeakers to be distributed over the regions where images are required. However in many listening situations it is not practical or desirable to place loudspeakers in some positions, such as behind or above the listener. Compensated Amplitude Panning (CAP) is an object-based reproduction method that adapts dynamically to the listener's head orientation to provide stable images in any direction, including behind and above. This is achieved by producing accurate dynamic Inter-aural Time Difference (ITD) cues. Here CAP is extended for more than two loudspeakers. Adding one or more loudspeakers allows the radiated energy and cancellation error to be reduced dramatically for some configurations. The multichannel CAP method is also compared with Ambisonic reproduction.

0 INTRODUCTION

Amplitude panning is a method for producing a spatial audio image in which two or more waves are combined 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 localization [1]. The direction of the image can be varied continuously by varying the gains.

Below ≈1000 Hz, referred from here on as the low band, 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 two 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 [2–4].

A typical scene contains multiple images in different directions, so for any given head direction not all the images can be perceived in the desired directions. This distortion is greater when the angle between the loudspeakers, viewed from the listener, is increased. If the listener is located at the center of a stereo pair, the loudspeakers are then 180° apart. In this position an image panned to the center would be completely unstable.

Dynamic localization cues are those produced by integrating cues over a period in which the listener is moving [1]. Dynamic cues help resolve ambiguities of instantaneous cues. In particular dynamic ITD cues can resolve front-back ambiguities and even determine elevation [5–7].

The change in the perceived image direction when the head is rotated is caused by the ITD cue not matching that of a real source in the target direction. 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]. For this system it is a challenging requirement to track the listener accurately in real-time with low latency and, preferably, without requiring any wearable device. However, suitable tracking technology is progressing very rapidly, driven by a wide range of applications.

Previously we were able to develop CAP for two-loudspeaker reproduction (Stereo-CAP) [8] producing stable images in any direction in the low band. This improves on conventional stereo by providing a more stable front stage, while also adding all around imaging that can produce a fully immersive experience. Head position is tracked, as well as the orientation, which allows image directions to be modified so that images are produced in fixed locations. In this way audio augmented reality applications are possible using loudspeakers.

Listener adaptive energy based panning, or *Vector Base Intensity Panning (VBIP)* [9] can be combined with Stereo-CAP to provide stable full bandwidth imaging in the front stage between the loudspeakers [8]. In the central region between the loudspeakers the low band localization usually dominates to the extent that full bandwidth localization is possible in all directions. Additional loudspeakers could be added to support high frequency coverage. These can be much smaller and lighter than the low band stereo pair. Another possibility is to provide high frequency coverage

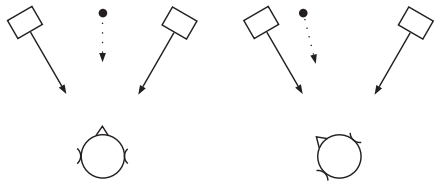

Fig. 1. The black dot indicates the direction of the image when two loudspeakers each have the same signal for different head directions.

using *cross-talk cancellation* [10, 11]. High frequency CTC can be achieved with closely spaced loudspeakers. Cross-talk cancellation (CTC) systems have previously struggled to provide stable low band imaging, including rear imaging. Instead the low band can be provided by CAP and unlike CTC does not require any head parameter information, such as head size.

1 STEREO-CAP OVERVIEW

In this section a brief overview is presented of the main results from the development Stereo-CAP [8, 12]. For a low frequency spherical head model 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 vector* that represents a general incident sound field at low frequencies [13], defined by

$$\boldsymbol{r}_V = -Z_0 \boldsymbol{V}/P \tag{2}$$

where $\boldsymbol{V}$, P are the particle velocity and pressure at the head center position, before the head is introduced. If the incident field is produced by panning, then r_V can be found easily in terms of the panning gains, g_i, assuming the wave from each loudspeaker is approximately planar at the head position,

$$\boldsymbol{r}_V = \frac{\sum g_i \hat{\boldsymbol{r}}_i}{\sum g_i} \tag{3}$$

where $\hat{\boldsymbol{r}}_i$ 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.

Combining Eq. (1) and Eq. (3) for two channels, and normalizing the total gain which determines the total pressure, leads to expressions for Stereo-CAP gains,

$$g_1 = \frac{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_I - \hat{\boldsymbol{r}}_2)}{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_1 - \hat{\boldsymbol{r}}_2)}, \; g_2 = \frac{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_I - \hat{\boldsymbol{r}}_1)}{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_2 - \hat{\boldsymbol{r}}_1)} \tag{4}$$

These panning laws were tested objectively by calculating the resulting cues for a range of configurations at different frequencies using a KEMAR dummy head [8]. The perceived directional error was then calculated and found to be within just noticeable difference 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 steady images could now be created in every direction.

An extension for near-field images is possible by matching the low frequency ILD (Inter-aural Level Difference) to that of a near source [12]. The meaning of near-field here is the range in which listeners are sensitive to range cues caused by level differences between the ears for low band sources, which in turn is related to the head diameter. This is typically within 1.5 m, and sensitivity increases greatly for close range. This leads to a condition like Eq. (1) depending on the real part of the Makita vector and an additional condition depending on the imaginary part,

$$\hat{\boldsymbol{r}}_{\boldsymbol{R}} \cdot (\Re(\boldsymbol{r}_V) - \hat{\boldsymbol{r}}_I) = 0, \; \hat{\boldsymbol{r}}_{\boldsymbol{R}} \cdot \left(\Im(\boldsymbol{r}_V) + \frac{\hat{\boldsymbol{r}}_I}{k r_I} \right) = 0 \tag{5}$$

where $\Re$, $\Im$ denote the real and imaginary parts. For the two-channel case there is a unique solution for the gains. The real parts are

$$\Re(g_1) = \frac{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_I - \hat{\boldsymbol{r}}_2)}{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_1 - \hat{\boldsymbol{r}}_2)}, \; \Re(g_2) = \frac{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_I - \hat{\boldsymbol{r}}_1)}{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_2 - \hat{\boldsymbol{r}}_1)} \tag{6}$$

The imaginary parts of the two gains are opposite one another,

$$\Im(g_1) = -\Im(g_2) = -\frac{\hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_I}{k r_I \hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_1 - \hat{\boldsymbol{r}}_2)} \tag{7}$$

The complex gains can be realized using a single 1st order filter for each image.

It is helpful to visualize the 3-dimensional vectors, and the solution, under the constraint Eq. (1). Fig. 2 shows a plan view of these vectors for the case of real-valued gains given by Eq. (4). This is called a *Makita diagram* here since each point on this diagram corresponds to a value of $\boldsymbol{r}_V$, rather than a point in 3-dimensional space.

The dotted circle is a cross section through a sphere of radius 1. If $\boldsymbol{r}_V$ is on the circle or sphere then it 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 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 two 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 $\hat{\boldsymbol{r}}_1$ and $\hat{\boldsymbol{r}}_2$. 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

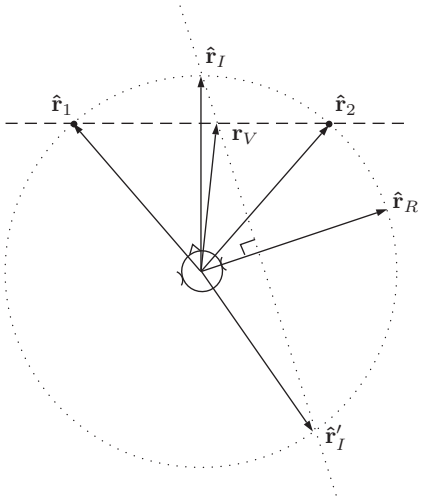

Fig. 2. Makita diagram for Stereo CAP, in plan view, for a listener facing towards left of center of the stereo array. The Makita vector is to the right of center in order to keep the image central. Shown are loudspeaker directions $\hat{\boldsymbol{r}}_1, \hat{\boldsymbol{r}}_2$ the inter-aural direction $\hat{\boldsymbol{r}}_R$, image direction $\hat{\boldsymbol{r}}_I$, and Makita vector $\boldsymbol{r}_V$.

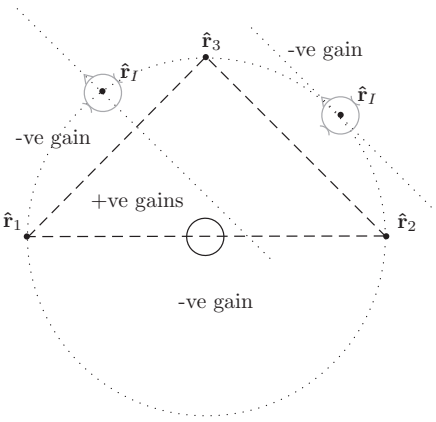

Fig. 3. Makita diagram for CAP with three loudspeakers, in plan view. Shown are loudspeaker directions $\hat{\boldsymbol{r}}_1, \hat{\boldsymbol{r}}_2, \hat{\boldsymbol{r}}_3$, and two images $\hat{\boldsymbol{r}}_I$, each with associated head orientations.

error due to each gain generally accumulate, this means the total error increases as the sum of gain magnitudes $\sum|g_i|$, and degree of cancellation increase, given $\sum g_i$. 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, the plane and line become close to parallel and the denominators vanish. The gains become large and polarized and the error increases. The common gain in the denominators can be limited, however the perceived image level will be reduced.

Introducing another loudspeaker between the existing pair would introduce more freedom for controlling $\boldsymbol{r}_V$, and the singular case can be avoided. Solutions for more than two loudspeakers are developed in the remainder of this article.

2 SOLUTIONS WITH MORE THAN TWO LOUDSPEAKERS

A Makita diagram with three loudspeakers is illustrated in Fig. 3. Provided the loudspeakers' direction vectors are distinct, then the producible values of $\boldsymbol{r}_V$ cover a plane containing $\hat{\boldsymbol{r}}_1, \hat{\boldsymbol{r}}_2, \hat{\boldsymbol{r}}_3$. The corresponding gains are positive for $\boldsymbol{r}_V$ in the triangular region inside these points, the *convex hull* of the points, and at least one gain is negative for each point outside. Two image examples are shown, each with a head superimposed to show the head orientation in order to simplify the picture in Fig. 2. The image direction and head orientation determine the possible $\boldsymbol{r}_V$ values, which lie on the planes indicated by the dotted lines.

If the dotted and dashed planes intersect then there are possible solutions along the line of intersection. There are no solutions only when the planes are parallel and separated, which only happens when the inter-aural axis is perpendicular to the loudspeaker plane, i.e., when one ear is pointing directly up. Different strategies can be considered for selecting from the possible solutions:

Localized energy: It is natural to try and localize loudspeaker energy in the directions where images are. In high frequency panning this reduces image spread and makes images more compatible for multiple listeners in different locations. In the low band, image spread is perceived much less, provided the cues are consistent, because the cues only contain directional information. The image on the left side in Fig. 3 has a localized solution where the dotted line crosses the dashed line between $\hat{\boldsymbol{r}}_1$ and $\hat{\boldsymbol{r}}_3$. The gain is zero for the other loudspeaker $g_2 = 0$. This is similar to a pairwise panning arrangement. However, for the image on the right side there are no positive solutions. Solutions are possible with negative gain and either $g_2 = 0$ or $g_3 = 0$, but they are not localized to the target image. To move continuously between these solutions when the head rotates requires non-zero gain from all loudspeakers.

Least radiated energy: The energy radiated, $\sum r_i^2 g_i^2$, drives room reverberance. Some natural reverberance is usually desirable because it can add authenticity that is difficult to achieve with pure synthetic reverberance. Also the precedence mechanism [1] causes localization to be focused on the direct signal, especially for the low band, making this method more robust than reproduction methods that rely more on the high band.

However, excessive reverberation can interfere with localization of the direct signal. Reducing the radiated energy then improves localization and reduces the maximum power required from the loudspeakers. A minimum energy solution will generally be spread over all the available loudspeakers. However, as explained before, spreading is not a concern in the low band.

Least direct energy: CAP may produce gains with opposite sign and cancellation of pressures at the listener. As with the case of Stereo-CAP, cancellation implies the sum of gain magnitudes $\sum|g_i|$ is greater than $\sum g_i$, and the total reproduction error is increased. The energy sum $\sum g_i^2$ provides a measure of total error that captures the incoherent

addition of errors and is convenient to optimize. Minimizing this quantity will minimize the reproduction error due to the direct signal. The solutions for least radiated energy and least cancellation error could be combined to give partial weight to each strategy. These solutions are the same when the distances r_i are equal. Note that $r_V \gg 1$ implies cancellation and $\sum g_i^2 \gg 1$, however $\sum g_i^2 \gg 1$ is also possible for $r_V = 1$, for example in the case of Ambisonics.

Ambisonic: If the image direction $\hat{\boldsymbol{r}}_I$ is restricted to the plane defined by the loudspeaker directions, then there is a solution $\boldsymbol{r}_V = \hat{\boldsymbol{r}}_I$ that is independent of head orientation. This is equivalent to Ambisonic panning based on mode matching of the sound field to first order [3, 14]. The low frequency cues depend only on the first order approximation. It is unusual to consider mode matching, including rear images, with no loudspeakers behind the listener. Mathematically this is possible, but it is not immediately clear how well conditioned it is and how much cancellation is needed.

2.1 Least Energy Solution

From the above discussion, the most useful solutions for general images are for the least radiated energy and the least direct energy. These solutions can be found analytically. This is shown first for the least radiated energy case. The least direct energy solution is then a special case of this. The solution for far images using real-valued gains is then extended for near-field imaging using complex gains.

Substituting Eq. (3) in Eq. (1) and multiplying by $\sum g_i$ gives the constraint

$$\sum g_i(\hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_i) = \hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_I \tag{8}$$

or,

$$\sum g_i\alpha_i = \phi, \alpha_i = \hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_i, \phi = \hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_I \tag{9}$$

where α_i and ϕ are defined for convenience. The summation range for the index i is omitted here and in the following. A second condition is needed to fix the level of the perceived image to a non-zero value, without which the gains would be minimized to zero. This is achieved by specifying the the incident pressure at the listener, which ensures the binaural signals will match those of a planewave with the same incident pressure. For a normalized level

$$\sum g_i = 1 \tag{10}$$

The least energy problem can be stated by minimizing the total energy radiated by the loudspeakers,

$$\text{argmin}_{\{g_i\}} \sum (r_i g_i)^2 \tag{11}$$

subject to the previous constraints Eq. (9) and Eq. (10). This function and the conditions are smooth, so a closed solution is sought using Lagrange multipliers. The Lagrangian is

$$\mathcal{L} = \sum (r_i g_i)^2 - \lambda_1 \left(\sum g_i\alpha_i - \phi \right) - \lambda_2 \left(\sum g_i - 1 \right) \tag{12}$$

with multipliers λ_1, λ_2. Setting partial derivatives by the unknown parameters to zero, $\partial\mathcal{L}/\partial g_i = 0$, $\partial\mathcal{L}/\partial\lambda_1 = 0$, $\partial\mathcal{L}/\partial\lambda_2 = 0$, produces $n+2$ constraints, including the original 2 constraints, where n is the number of loudspeakers.

$$2r_i^2 g_i - \lambda_1\alpha_i - \lambda_2 = 0, i = 1..n \tag{13}$$

$$\sum g_i\alpha_i = \phi \tag{14}$$

$$\sum g_i = 1 \tag{15}$$

From Eq. (13) the gains can be written

$$g_i = \frac{\lambda_1\alpha_i + \lambda_2}{2r_i^2} \tag{16}$$

Substituting the gains into Eq. (14),

$$\begin{aligned} \sum \frac{\lambda_1\alpha_i + \lambda_2}{2r_i^2}\alpha_i &= \phi \\ \lambda_1 \sum \frac{\alpha_i^2}{r_i^2} + \lambda_2 \sum \frac{\alpha_i}{r_i^2} &= 2\phi \\ \lambda_1\gamma + \lambda_2\beta &= 2\phi \end{aligned} \tag{17}$$

where $\beta = \sum \frac{\alpha_i}{r_i^2}$ and $\gamma = \sum \frac{\alpha_i^2}{r_i^2}$ are defined for convenience. Substituting the gains into Eq. (15),

$$\begin{aligned} \sum \frac{\lambda_1\alpha_i + \lambda_2}{r_i^2} &= 2 \\ \lambda_1\beta + \eta\lambda_2 &= 2 \end{aligned} \tag{18}$$

where $\eta = \sum \frac{1}{r_i^2}$. Eqs. (17) and (18) can be solved for λ_1 and λ_2,

$$\lambda_1 = \frac{2(\eta\phi - \beta)}{\gamma\eta - \beta^2} \tag{19}$$

$$\lambda_2 = \frac{2(\gamma - \beta\phi)}{\gamma\eta - \beta^2} \tag{20}$$

The resulting optimal gains are found by substituting into Eq. (16),

$$g_i = \frac{(\eta\phi - \beta)\alpha_i + \gamma - \beta\phi}{r_i^2(\gamma\eta - \beta^2)} \tag{21}$$

These gains are inexpensive to evaluate, which allows them to be updated frequently when the listener or image moves. The expense is not more than for VBAP, which requires a matrix inversion for each loudspeaker triplet whenever the listener moves.

The compensated loudspeaker gains are $r_i g_i$. A global gain factor can be added to set the reproduction level. The least direct energy solution can be found by setting all the loudspeaker distances $r_i = 1$. The least energy solution using two loudspeakers has to be identical to Stereo-CAP because there can be only one solution. This can also be checked algebraically by simplifying Eq. (21) for the case $n = 2$.

The extension for near-field imaging is made by substituting Eq. (3) in the constraints Eq. (5). The first constraint leads to the real part of the solution already given by

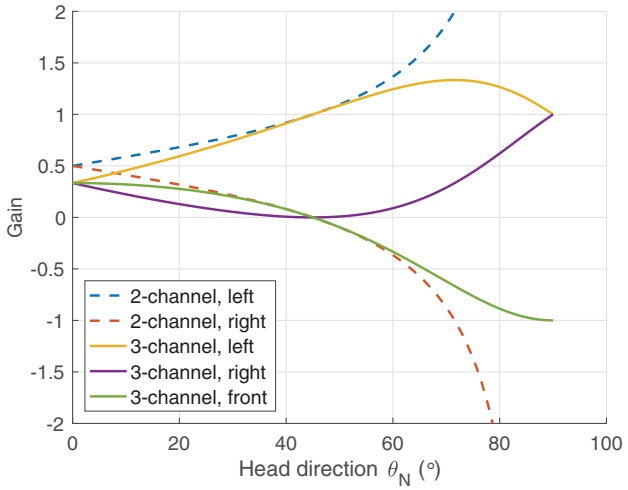

Fig. 4. Gains for Stereo-CAP and 3-way CAP for an image at 180° azimuth and a range of head directions. The left and right loudspeakers are separated by 180°.

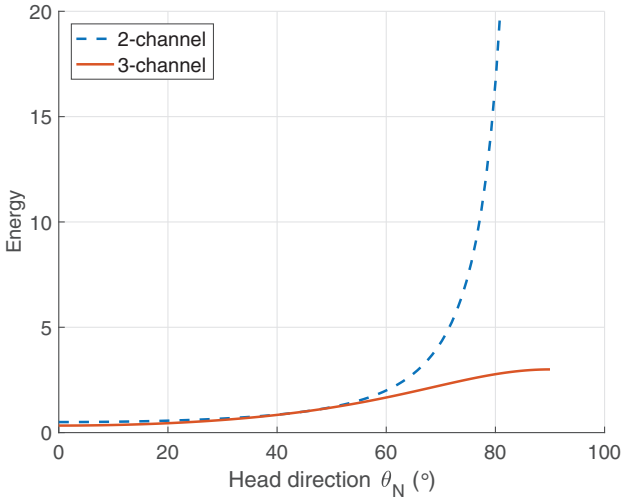

Fig. 5. Total energies for Stereo-CAP and 3-way CAP for an image at 180° azimuth, and a range of head directions. The left and right loudspeakers are separated by 180°.

Eq. (21). Substituting in the second constraint gives a constaint for the imaginary part of the gains,

$$\sum \Im(g_i)\alpha_i = -\frac{\phi}{kr_I} \tag{22}$$

The pressure contraint Eq. (10) implies a second constraint

$$\sum \Im(g_i) = 0 \tag{23}$$

The radiated energy due to the imaginary parts can be minimized in a similar way to the real parts. The resulting gains are

$$\Im(g_i) = \frac{\phi(\beta - \eta\alpha_i)}{kr_I r_i^2(\gamma\eta - \beta^2)} \tag{24}$$

Since the real and imaginary energies can be minimized independently, the minimum total energy is the sum of the minimums for the the two parts separately, so the optimum complex gains are formed by combining the optimum real and imaginary parts.

The complex gains can be implemented with high accuracy in real-time using a single 1st order filter for each image using the same method described for near-field Stereo-CAP [12].

The plots shown in Fig. 4 compare the gains produced by the Stereo-CAP system with the least energy 3-way CAP system for different head directions. The Stereo loudspeakers are positioned directly to the left and right and the image is directly behind. This configuration requires lower gains than for loudspeakers that are less separated in front. The 3-way system has loudspeakers in these positions and an extra one directly in front, the same as Fig. 3. When the listener turns to the side the Stereo-CAP gains become large, whereas the 3-way CAP gains have magnitudes similar to the total gain $\sum g_i = 1$. Fig. 5 plots the total energy of the incident waves at the listener, which is also proportional to the total radiated energy for equidistant loudspeakers. Similar results are obtained for other image directions. From

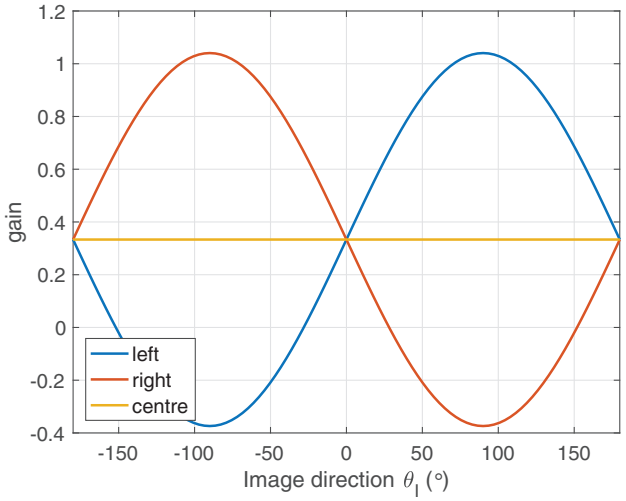

Fig. 6. Gains for 3-way CAP for a forward head direction and a range of horizontal target image directions. The loudspeakers are separated by 90°.

the objective results obtained for the Stereo-CAP, the gains are low enough in the 3-way case so that the resulting image direction errors due to the direct signals are within just noticeable differences, and potential reverberant energy from the indirect radiation is acceptable. Since the subjective results for Stereo-CAP show that accurate ITD predicts high image stability, this implies that the 3-way CAP system will perform subjectively as well in all configurations as best performance using Stereo-CAP.

Fig. 6 shows the 3-way CAP gains as a function of image direction for a fixed forward head direction. The loudspeakers are separated by 90° for comparison with some 3-way static panning functions published previously [15, 16]. The panning function for the new dynamic method looks quite different to the static panning functions, which is because the panning function changes as the head

orientation changes: The static panning functions are a compromise designed to function for a range of head directions, and images are restricted only to the range between the loudspeakers. In the case of Gerzon's 3-way panner [15] the design is explicitly optimized for low and high bands simultaneously, whereas for the dynamic panner band separation is unavoidable to produce full surround images.

A comparison can also be made with VBAP panning. Consider first panning for any triplet of loudspeakers. The VBAP solution corresponds to the intersection of r_I with the plane defined by r_1, r_2, r_3. For MCAP the possible solutions before energy minimization are the intersection of this loudspeaker plane with the plane through r_I and normal to the aural axis r_L. When the listener is looking at the image the MCAP solution set depends on the rotation of the head about the image direction. The only solution common to all sets is the VBAP solution. This is useful theoretical background, however in practice CAP does not require triplets to produce full surround. Horizontal arrays, even stereo, are sufficient.

Spreading three loudspeakers evenly on the horizontal, with 120 degrees separation, allows a reduction of the maximum absolute gains shown in Fig. 4, while retaining the capability for producing images that are outside the horizontal. Adding more loudspeakers further reduces the maximum gain.

Adding a fourth loudspeaker that is not coplanar with the others, for example above the front loudspeaker in the example shown in Fig. 3, increases the space of $\boldsymbol{r}_V$ that can be produced by panning, from a plane to the whole 3-dimensional Makita space. The panning gains are all positive for points inside the convex hull described by the four loudspeaker direction vectors, and at least one gain is negative for each point outside this region. The intersection of the whole space with the plane described by the ITD constraint is always non empty, so there are no singular configurations.

The multichannel solution can be used with any number of loudspeakers. While an advantage of the CAP system is that it requires only a few loudspeakers, loudspeakers can be added to progressively reduce the radiated energy relative to the direct energy: This is because sound is coherently focused on the listener, whereas the radiated sound combines incoherently. The least energy solution is optimal—it gives the lowest radiated energy relative to the direct energy for any given loudspeaker configuration.

2.2 Performance

The multichannel solution satisifies the ITD criteria for all parameter variations in the low frequency limit. In practice the low band extends above the low limit, and real heads are not solid spheres. For Stereo-CAP, previous objective and subjective tests with realistic / real heads show that the ITD errors are perceptually small over a wide range of parameters. For the multichannel case the field never becomes large for any choice of parameters, so the ITD errors towards the high end of the low band remain small. Indeed for some cases as shown earlier the multichannel case has much lower overall energy, and this will result in less error in direct ITD as well as less reverberant interference in the presence of a room. So we can confidently infer that adding more loudspeakers to a given configuration of Stereo-CAP cannot make the reproduction worse, and in some cases greatly improves it, which was one of the original aims. This has been confirmed with informal listening.

2.3 Multiple Listeners

For multiple listeners the direct least energy solution can provide some isolation between listeners due to the focusing effect. Beam-forming arrays could be used to provide better isolation, replacing each loudspeaker in the direct method with an array that produces a beam for each listener. Each beam must be steered dynamically to follow the respective listener. More sophistocated beam-forming is required to isolate listeners that are close.

2.4 Ambisonic Comparison

In the mode-matched Ambisonic approach, an image is formed by reproducing a sound field at the listener that matches the sound field of the associated source, to a given accuracy. To produce an accurate low frequency ITD cue it is enough to reproduce pressure and velocity, which are components of a 1st order Ambisonic encoding. The first order Ambisonic decoding problem can be written in terms of the variables used in this article by combining Eq. (3) and Eq. (10) into a single matrix equation, so that the pressure and velocity components are simultaneously matched,

$$\begin{pmatrix} 1 & 1 & 1.. \\ \hat{\boldsymbol{r}}_1 & \hat{\boldsymbol{r}}_2 & \hat{\boldsymbol{r}}_{3}.. \end{pmatrix} \begin{pmatrix} g_1 \\ g_2 \\ g_3 \\ . \\ . \end{pmatrix} = \begin{pmatrix} 1 \\ \boldsymbol{r}_V \end{pmatrix} \tag{25}$$

Or abreviated,

$$\mathbf{Rg} = \mathbf{s} \tag{26}$$

The least energy solution, where it exists, is given using the pseudoinverse, $\mathbf{R}^+$,

$$\mathbf{g} = \mathbf{R}^+\mathbf{s} \tag{27}$$

$\mathbf{R}^+$ is the Ambisonic decoding matrix. For the example shown in Fig. 4, with three loudspeakers and an image behind,

$$\mathbf{s} = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{R} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{28}$$

$$\mathbf{R}^+ = \begin{pmatrix} 1/2 & -1/2 & 1/2 & 0 \\ 1/2 & -1/2 & -1/2 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \mathbf{g} = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \tag{29}$$

The vector component order is (x, y, z) or (front, left, up), and the loudspeaker order is (left, right, front).

The ratio of radiated energy to direct energy is 9. Depending on the room properties this level may already compromise direct localization. Using CAP the gains in this configuration with listener facing forward are

$$\mathbf{g} = \begin{pmatrix} 0.33 \\ 0.33 \\ 0.33 \end{pmatrix} \tag{30}$$

giving a radiated to direct energy ratio of 1/3. The same ratio for Ambisonics is 3, which in a typically reverberant listening space already degrades the localization. If the loudspeaker angular separation is reduced then this difference rapidly becomes much greater as Ambisonics requires more cancellation: For example if the left and right loudspeakers are positioned at -30°, $+30^\circ$ then the Ambisonic gains required are

$$\mathbf{g} = \begin{pmatrix} 7.46 \\ 7.46 \\ -13.93 \end{pmatrix} \tag{31}$$

while the CAP gains are unchanged, so that the Ambisonic radiation to direct energy ratio is 916 times greater than that for CAP. As the listener turns their head to the side the CAP gains become equal to the Ambisonic gains, so if this configuration is required it is preferable to have the side loudspeakers as wide as possible.

Adding a fourth loudspeaker that is not coplanar with the others, for example above the center loudspeaker in the example shown in Fig. 3, allows gains to be produced for any image direction using the Ambisonic method. Comparatively high gains are required for loudspeakers positioned similarly to the three loudspeaker case.

3 CONCLUSION

Using two loudspeakers and with full six-degrees-of-freedom head tracking, position and orientation, it was previously shown possible to create low band images in any direction, although excessive gain is required for some listener orientations. Here it was shown that with three loudspeakers all images directions can be reproduced with moderate gain except for a small range of orientations that are practically unimportant. Adding more loudspeakers to any stereo configuration does not worsen performance. For comparison, an Ambisonic approach with position tracking and three frontal loudspeakers can, in principle, reproduce horizontal surround images, and four loudspeakers can reproduce full 3D. This does not require orientation tracking. However as loudspeaker separation is reduced Ambisonics suffers from rapidly increasing gains and cancellation for forward head directions and rear images, whereas in this case CAP gains are lower and remain low as separation is reduced. The overall CAP energy can be reduced further by increasing the number of loudspeakers.

The current real-time implementation of the multichannel CAP system is based on an extensive and flexible C++ / Python framework for spatial sound rendering, called the *Versatile Interactive Software Rendering framework (VISR)* [17]. It is planned to make this publicly available in due course. The HTC Vive tracking system is used in experiments to locate the loudspeakers and track the listener with a precision within 1 cm, and .1 degree, over a 5 x 5 x 2 m region, for moderate cost.

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] J. Blauert, *Spatial Hearing* (Cambridge, MA, MIT Press, 1997).

[2] B. Bernfeld, "Attempts for Better Understanding of the Directional Stereophonic Listening Mechanism," presented at the *44th Convention of the Audio Engineering Society* (1973 Mar.), convention paper C-4.

[3] M. A. Gerzon, "General Metatheory of Auditory Localization," presented at the *92nd Convention of the Audio Engineering Society* (1992 Mar.), convention paper 3306.

[4] V. Pulkki, "Compensating Displacement of Amplitude-Panned Virtual Sources," presented at the *AES 22nd International Conference: Virtual, Synthetic, and Entertainment Audio* (2002 Jun.), conference paper 000244.

[5] H. Wallach, "On Sound Localization," *J. Acoust. Soc. Amer.*, vol. 10, pp. 270–274 (1939).

[6] H. Wallach, "The Role of Head Movements and Vestibular and Visual Cues in Sound Localization," *J. Exper. Psych.*, vol. 27, no. 4, p. 339 (1940).

[7] B. Xie and D. Rao, "Analysis and Experiment on Summing Localization of Two Loudspeakers in the Median Plane," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), convention paper 9452.

[8] D. Menzies, M. F. Simon Galvez, and F. M. Fazi, "A Low Frequency Panning Method with Compensation for Head Rotation," *IEEE Trans. Audio, Speech, Language Processing*, vol. 26, no. 2 (2018 Feb.).

[9] J.-M. Pernaux, P. Boussard, and J.-M. Jot, "Virtual Sound Source Positioning and Mixing in 5.1 Implementation on the Real-Time System Genesis," *Proceedings of the Conf. Digital Audio Effects (DAFx-98)*, pp. 76–80 (1998).

[10] B. S. Atal and M. R. Schroeder, "Apparent sound source translator," US Patent 3,236, 949 (1966 Feb. 22).

[11] O. Kirkeby, P. A. Nelson, and H. Hamada, "Virtual Source Imaging Using the Stereo Dipole," presented at the *103rd Convention of the Audio Engineering Society* (1997 Sep.), convention paper 4574.

[12] D. Menzies and F. M. Fazi, "A Complex Panning Method for Near-field Imaging," *IEEE Trans. Audio, Speech, Language Processing*, vol. 26, no. 9 (2018 Sep.).

[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] J. Daniel, "Spatial Sound Encoding including Near Field Effect," presented at the *AES 23rd International Conference: Signal Processing in Audio Recording and Reproduction* (2003 May), conference paper 16.

[15] M. A. Gerzon, "Optimum Reproduction Matrices for Multispeaker Stereo," *J. Audio Eng. Soc.*, vol. 40, pp. 571–589 (1992 Jul./Aug.), URL http://www.aes.org/e-lib/browse.cfm?elib=7038.

[16] M. Poletti, "The Design of Encoding Functions for Stereophonic and Polyphonic Sound Systems, " *J. Audio Eng. Soc.*, vol. 44, pp. 948–963 (1996 Nov.).

[17] A. Franck and F. M. Fazi, "VISR—A Versatile Open Software Framework for Audio Signal Processing," presented at the *2018 AES International Conference on Spatial Reproduction—Aesthetics and Science* (2018 Jul.), conference paper P9-2.

THE AUTHORS

Dylan Menzies

Filippo Maria Fazi

Dylan Menzies is a Senior Research Fellow in the Institute of Sound and Vibration at the University of Southampton. Areas of interest include spatial audio synthesis and reproduction, sound synthesis for virtual environments, and musical synthesis and interfaces. He holds a Ph.D. in electronics from the University of York, a B.A. in mathematics from Cambridge University, and has worked as a research engineer for several companies including Sony Professional Audio.

•

Filippo Maria Fazi is graduated in mechanical engineering from the University of Brescia (Italy) in 2005. He obtained his Ph.D. in acoustics from the Institute of Sound and Vibration Research (ISVR) of the University of Southampton, UK, in 2010, with a thesis on sound field reproduction. In the same year, he was awarded a research fellowship by the Royal Academy of Engineering and by the Engineering and Physical Sciences Research Council. He is currently an Associate Professor at the University of Southampton. Dr. Fazi's research interests include audio technologies, electroacoustics, and digital signal processing, with special focus on acoustical inverse problems, multichannel systems, virtual acoustics, microphone and loudspeaker arrays.