Ambisonic Decoding for Compensated Amplitude Panning

Dylan Menzies, Filippo Maria Fazi

***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 using only two loudspeakers. In its original form CAP is inherently an object-based method, with each image produced separately. An exact and efficient method is presented here for dynamically decoding a first order Ambisonic encoding, which is equivalent to using CAP to separately reproduce the constituents of the encoding. Both the stereo and multichannel cases are considered.**

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

I. Background

AMBISONIC encoding enables complex scenes to be represented and decoded efficiently using a fixed number of channels, and is used widely for 360° video, and other applications. Compensated Amplitude Panning (CAP) is an object-based spatial reproduction method in which loudspeaker signals depend dynamically on the precise listener head position and orientation [1]. The head tracking allows images to be produced in any 3D direction using as few as 2 loudspeakers, which is of great value in many practical applications. In this article a new reproduction method is derived that decodes Ambisonic input to produce loudspeaker outputs that are the same as those produced by CAP for individual plane wave images. This enables complex scenes to be produced accurately and efficiently on a 2-loudspeaker system with head tracking. In this section the background to CAP is reviewed, in preparation for the derivation in the next section.

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 [2]. The direction of the image can be varied continuously by varying the gains.

Below ∼1000 Hz 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 [3]–[5]. 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 [2], [6]–[8].

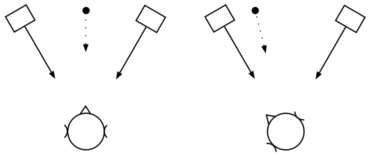

Fig. 1. The black dot indicates the direction of the image when 2 loudspeakers each have the same signal, for different head directions.

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. 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 [1]. It is accurate in the low frequency region up to ∼1000 Hz. 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 and ILD 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 reproduction can also be provided in all directions using *cross-*

D. Menzies and Filippo M. Fazi are with the Institute of Sound and Vibration Research, University of Southampton, UK

talk cancellation [10], [11]. Cross-talk cancellation systems generally perform poorly at low frequencies, and do not adapt accurately where head orientation is tracked. CAP is a simple and low cost method for accurate low frequency imaging, and also allows efficient compensation for listener position as part of the process. The authors are not aware of previous reproduction systems for two loudspeakers that can produce stable rear images, and so provide a full immersive experience.

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].

In the remaining part of this section the key technical details of CAP are reviewed. CAP is based on a low frequency spherical head model. This simplification enables fruitful mathematical manipulation leading to efficient formulae, without, it turns out, significantly compromising localisation performance in the valid frequency band. For a plane wave arriving from a target image direction $\hat{\boldsymbol{r}}_I$, there are, in general, a range of possible incident fields that produce the same ITD an ILD cues for the head model, and so the same image. It was shown [1] that this can be formulated as a vector condition,

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

where $\hat{\boldsymbol{r}}_A$ is the inter-aural axis direction vector, pointing either to the left ear or right ear (replacing $\hat{\boldsymbol{r}}_R$ in previous work), and $\boldsymbol{r}_V$ is the Makita vector representing the incident sound field at low frequencies [13]. 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

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

g_i are the signal gains of the source waves *at the listener*. $\hat{\boldsymbol{r}}_i$ are the direction vectors of the loudspeakers relative to the listener [1]. The signal gains *at the loudspeakers* are $r_i g_i$, the r_i factor compensating for the $1/r$ amplitude decay. In addition, delays are introduced to the loudspeaker feeds so that the signals at the listener are in phase. This compensation depends on accurate knowledge of the ambient speed of sound, as well as the distances. The signal gains in this case are frequency independent. However to allow for frequency dependence that arises in near-field imaging, all gain and signal variables should be interpreted as frequency domain, although the frequency dependence is not explicitly written.

A non-trivial solution for the Stereo-CAP gains can be found using the constraint (1), the equation (2), and an additional constraint normalising the total gain, which fixes the overall level,

$$\sum g_i = 1 \tag{3}$$

The resulting gain solution is

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

These panning laws were tested objectively by calculating the resulting cues at different frequencies for a KEMAR dummy head [1]. The perceived directional error was then calculated and found to be within a Minimum Audible Angle [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 [1]. Dynamic head tracking was used to allow natural unrestricted listening. The tests showed that images between loudspeakers were improved, and furthermore steady images could now be created in directions 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 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.

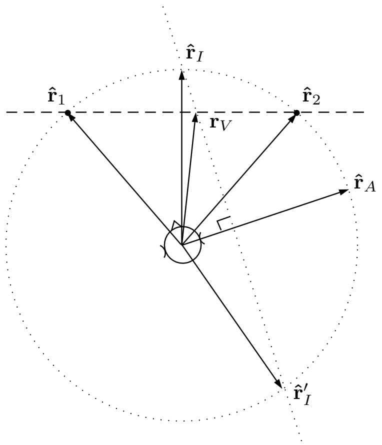

Fig. 2. Makita diagram for Stereo-CAP, in plan view, for a listener facing towards left of centre of the stereo array. The Makita vector is to the right of centre in order to keep the image central. Shown are loudspeaker directions $\hat{\boldsymbol{r}}_{\mathbf{1}}$, $\hat{\boldsymbol{r}}_{\mathbf{2}}$ the inter-aural direction $\hat{\boldsymbol{r}}_A$, image direction $\hat{\boldsymbol{r}}_I$ and Makita vector $\boldsymbol{r}_V$. The vectors are not generally restricted to the plane.

The panning gains are positive for values of $\boldsymbol{r}_V$ on the line 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. Cancellation also implies increasing total reverberant energy due to loudspeaker radiation, relative to the direct signal, which can further degrade the overall signal at the listener.

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, although the perceived image level will fade. This issue has been solved by extending CAP to a more general multichannel method [15].

The gain encoding equations (4) 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.

II. Ambisonic Decoding

The loudspeaker signals L_1, L_2 for several images can be formed by summing the loudspeaker signals for all the images. 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

$$L_1 = \sum_n g_{1,n} \; I_n \tag{5}$$

Substituting for the gain definition in (4), and separating summed terms,

$$\begin{aligned} L_1 &= \frac{1}{\hat{\boldsymbol{r}}_A \cdot (\hat{\boldsymbol{r}}_1 - \hat{\boldsymbol{r}}_2)} \sum \hat{\boldsymbol{r}}_A \cdot (\hat{\boldsymbol{r}}_{I_n} - \hat{\boldsymbol{r}}_2) \; I_n & (6) \\ &= \frac{\hat{\boldsymbol{r}}_A}{\hat{\boldsymbol{r}}_A \cdot (\hat{\boldsymbol{r}}_1 - \hat{\boldsymbol{r}}_2)} \cdot \left(\sum \hat{\boldsymbol{r}}_{I_n} I_n - \hat{\boldsymbol{r}}_2 \sum I_n\right) & (7) \end{aligned}$$

A 1st order Ambisonic signal, or *B-format* signal, consists of 4 components (W, X, Y, Z) [16]. W is the omnidirectional microphone signal at the measurement position, X, Y, Z are cosine directivity (figure-of-eight) signals, with mutually perpendicular axes. If cartesian coordinates for the vectors $\hat{\boldsymbol{r}}$ are chosen to coincide with the X, Y, Z axes, then the term $\sum \hat{\boldsymbol{r}}_{I_n} I_n$ generates the signals X, Y, Z, because $\hat{\boldsymbol{r}}_{I_n}$ are the direction cosines of the nth image component. Writing the signals as a vector,

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \sum \hat{\boldsymbol{r}}_{I_n} I_n \tag{8}$$

Similarly $\sum I_n$ identifies with $\sqrt{2}\; W$. 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. Substituting the identities in (6),

$$L_1 = \frac{\hat{\boldsymbol{r}}_A}{\hat{\boldsymbol{r}}_A \cdot (\hat{\boldsymbol{r}}_1 - \hat{\boldsymbol{r}}_2)} \cdot \left(\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} - \sqrt{2}\; W \hat{\boldsymbol{r}}_2 \right) \tag{9}$$

Similarly for the other loudspeaker,

$$L_2 = \frac{\hat{\boldsymbol{r}}_A}{\hat{\boldsymbol{r}}_A \cdot (\hat{\boldsymbol{r}}_2 - \hat{\boldsymbol{r}}_1)} \cdot \left(\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} - \sqrt{2}\; W \hat{\boldsymbol{r}}_1 \right) \tag{10}$$

The stereo CAP solution was generalised to Multichannel CAP (MCAP) [15], for two or more loudspeakers. The gain solution is given by

$$g_i = \frac{(\eta\phi - \beta)\alpha_i + \gamma - \beta\phi}{r_i^2(\gamma\eta - \beta^2)} \tag{11}$$

where,

$$\alpha_i = \hat{\boldsymbol{r}}_A \cdot \hat{\boldsymbol{r}}_i \;, \quad \phi = \hat{\boldsymbol{r}}_A \cdot \hat{\boldsymbol{r}}_I \tag{12}$$

$$\eta = \sum \frac{1}{r_i^2} \;, \quad \beta = \sum \frac{\alpha_i}{r_i^2} \;, \quad \gamma = \sum \frac{\alpha_i^2}{r_i^2} \tag{13}$$

The expression for the gains (11) can be rearranged to isolate the dependence on image direction contained in ϕ

$$g_i = \frac{\phi(\eta\alpha_i - \beta) + (\gamma - \beta\alpha_i)}{r_i^2(\gamma\eta - \beta^2)} \tag{14}$$

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

$$a_i = \frac{\eta\alpha_i - \beta}{r_i^2(\gamma\eta - \beta^2)} \qquad b_i = \frac{\gamma - \beta\alpha_i}{r_i^2(\gamma\eta - \beta^2)} \tag{15}$$

so that,

$$g_i = a_i\phi + b_i \tag{16}$$

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

$$g_{i,n} = a_i\phi_n + b_i \tag{17}$$

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,

$$\begin{aligned} L_i &= \sum_n g_{i,n} I_n & (18) \\ &= \sum_n (a_i\phi_n + b_i) I_n & (19) \\ &= \sum_n (a_i \; \hat{\boldsymbol{r}}_A \cdot \hat{\boldsymbol{r}}_{I_n} + b_i) I_n & (20) \\ &= a_i \hat{\boldsymbol{r}}_A \cdot \sum_n \hat{\boldsymbol{r}}_{I_n} I_n + b_i \sum_n I_n & (21) \end{aligned}$$

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

$$L_i = a_i \hat{\boldsymbol{r}}_A \cdot \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + \sqrt{2}\; b_i W \tag{22}$$

The decoding formula (9), (10) and (22) can be checked by substituting the gain solution in the initial constraints (1) and (3). 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 plane wave image signals. Furthermore, since any field can be represented with arbitrary precision using plane waves, the formulae apply to a B-format signal derived from any scene, recorded or synthesized, possibly containing point or diffuse sources, or reverberation. This can be shown directly for point sources: The X, Y, Z components for a point source contain an additional factor $(1 - j/(kr_S)$, where r is the distance to

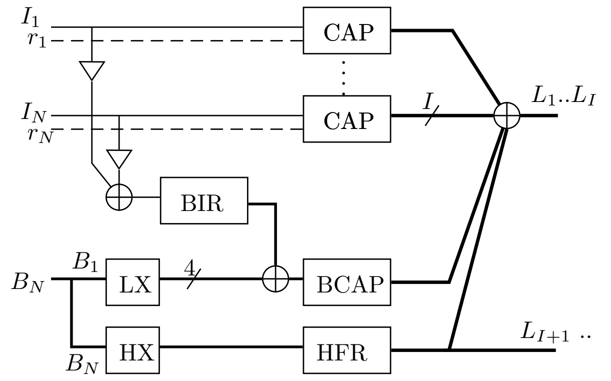

Fig. 3. Example signal paths combining CAP, BCAP, and BIR (B-format impulse response convolution). I_n are mono image signals with locations r_n. High and low crossover filters HX and LX are applied respectively to the high order encoding B_N, and the 1st order part only, B_1. The filtered signals are fed to the BCAP decoder, and HFR, which stands for any high frequency reproduction method, possibly using other loudspeakers. High frequency reproduction of the discrete sources is not shown.

the encoded source [17]. Similarly the MCAP gains for a point image contain an additional imaginary term [15],

$$\Im(g_i) = \frac{\phi(\beta - \eta\alpha_i)}{kr_I r_i^2(\gamma\eta - \beta^2)} \tag{23}$$

such that a_i is multiplied by a factor $(1-j/(kr_I)$. This factor can then be included as part of the X, Y, Z signals, so that (22) remains true.

CAP has been shown to produce good quality images in the low frequency range up to ∼1000 Hz [1]. The BCAP decoding formula show that 1st order B-format encoding in this frequency range is sufficient to encode a scene equivalent to. This is expected because 1st order Ambisonic encoding normally contains nearly all the sound field information in this frequency range.

The direction of images generated discretely using CAP can be made consistent with sources in fixed locations. The resulting parallax cues allow the listener to localise images in the source locations, either in the near or far field. 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. Fig. 3 shows these possible signal paths, including a side chain with a B-format impulse response convolver for adding reverberation, and high frequency reproduction from a high order Ambisonic encoding. For some applications low frequency content may be sufficient for rear imaging, for example dull reverberation. High frequency imaging can be provided in the front stage using simple energy panning methods described previously [1].

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)* [18]. Software plugins have been produced that run on digital audio workstation (DAW) software. This environment allows different test cases to be created quickly and compared side by side. The VISR system has already been released publicly, including CAP source code, and future releases will include plugins and DAW sessions implementing CAP.

III. PERFORMANCE

By construction, BCAP produces identical loudspeaker signals to those produced by applying CAP to the component signals and summing. So the objective and subjective performance for discrete images is identical to that reported previously for CAP reproduction [1], [19]. The process is linear and filterless and so transients are not smeared at all (A cross-over filter can be introduced to separate low and high frequencies in the 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 can be removed, and the listener can change orientation without any artefacts. Multiple instances of CAP have been used to build complex and diffuse scenes. These can be replaced with a single instance of BCAP without any change of reproduction quality. In informal listening BCAP reproduction of pre-existing B-format recordings sound similar to conventional 3D Ambisonic reproduction in the valid frequency range.

IV. 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.

REFERENCES

[1] 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.

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

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

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

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

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

[7] 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.

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

[9] Jean-Marie Pernaux, Patrick Boussard, and Jean-Marc Jot, "Virtual sound source positioning and mixing in 5.1 implementation on the real-time system genesis," in *Proc. Conf. Digital Audio Effects (DAFx-98)*. Citeseer, 1998, pp. 76–80.

[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 *Proc. 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 Maria Fazi, "Surround sound without rear loudspeakers: Multichannel compensated amplitude panning and ambisonics," in *Proc. Digital Audio Effects, Aveiro, Portugal*, September 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] Jérôme Daniel, "Spatial sound encoding including near field effect," in *Proc. AES 23nd International Conference, Helsinger, Denmark*, 2003.

[18] Andreas Franck and Filippo Maria Fazi, "Visr—a versatile open software framework for audio signal processing," in *Proc. Audio Engineering Society Conference on Spatial Reproduction*. Audio Engineering Society, 2018.

[19] Dylan Menzies and Filippo Maria Fazi, "A complex panning method for near-field imaging," *IEEE Trans. Audio, Speech, Language Processing*, vol. 26, no. 9, September 2018.