A Low Frequency Panning Method with Compensation for Head Rotation

Dylan Menzies, Marcos F. Simón Gálvez and Filippo Maria Fazi

***Abstract*—Amplitude panning produces Inter-aural Time Difference (ITD) cues that help localise images in directions between loudspeakers. However, if the panning gains are static the ITD cues produced in this way vary inconsistently as the listener's head rotates, compared with a real source, and so the dynamic ITD cues are inaccurate. This effect destabilises the perception of the image and overall scene, and is worse for loudspeakers that are more widely spaced relative to the listener. Based on a simple head model that is accurate in the low frequency ITD regime, the ITD is calculated for a general field, including those produced by panning. A simple formula is derived relating head orientation, image direction, and a field description vector. Panning functions are then found that compensate for head orientation, and are valid for any image direction. For the special case when the listener is facing the image, the functions are equivalent to Vector Base Amplitude Panning (VBAP). The performance is first assessed objectively using measured binaural responses, rather than the simple head model. Subjective comparison is then made with pre-existing listening tests, and new listening tests in which the listener's head is tracked to control the panning gains in real-time. These show that images can be stabilised as predicted, and, furthermore, that with the same panning functions images can be produced in all directions using two loudspeakers placed in front.**

***Index Terms*—IEEE, IEEEtran, journal, LaTeX, paper, template.**

MATHEMATICAL SYMBOLS

k	wavenumber
ω	angular frequency
Z_0	characteristic impedance of sound in air
P	pressure of the incident free field at the head centre position
$P_L,\ P_R$	resultant pressures at the ears
$\boldsymbol{V}$	velocity vector of the incident free field at the head centre position
$\boldsymbol{V}_\Re$ and $\boldsymbol{V}_\Im$	real and imaginary components of $\boldsymbol{V}$
$\boldsymbol{r}_L,\ \boldsymbol{r}_R$	displacement vectors from the head centre to each ear
$\bar{\boldsymbol{r}}_R,\ \bar{\boldsymbol{r}}_L$	$\bar{\boldsymbol{r}}_R = \frac{3}{2}\boldsymbol{r}_R,\ \bar{\boldsymbol{r}}_L = \frac{3}{2}\boldsymbol{r}_L$
$\hat{\boldsymbol{r}}_I$	direction to the image
$\boldsymbol{r}_V$	Makita localisation vector ($\hat{\boldsymbol{r}}_V = -\hat{\boldsymbol{V}}$)
$\theta_{VN}, \theta_{IN}, \theta_{VI}$	azimuth change from direction indicated by 1st subscript to 2nd. V - $\boldsymbol{r}_V$, N - nose, I - image
θ_L	$2\theta_L$ is the stereo loudspeaker separation from the listener
θ_N	directed angle from the mid point between stereo loudspeakers to the direction the listener is facing in
θ_I	directed angle from the mid point between stereo loudspeakers to the image
g_1, g_2	stereo panning gains
$\hat{\boldsymbol{r}}_1,\ \hat{\boldsymbol{r}}_2$	direction vectors to the stereo loudspeakers from the listener
h_{1L}, h_{1R} h_{2L}, h_{2R}	head related impulse responses from each loudspeaker (1,2) to each ear (L,R)

ABBREVIATIONS

ITD	Inter-aural Time Difference
ILD	Inter-aural Level Difference
VBAP	Vector Base Amplitude Panning
CAP	Compensated Amplitude Panning
KU100	Neumann KU100 binaural microphone
KEMAR	KEMAR binaural microphone
HRIR	Head-Related Impulse Response
MAA	minimum audible angle

I. INTRODUCTION

Amplitude panning is a method of producing a spatial audio image in which two or more plane waves are combined coherently at the listener position, each carrying the same signal and with separate gain. 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.

If a listener faces a panned image then turns away, the position of the image is perceived to move significantly relative to its initial direction, towards the listener[2;3;4], as illustrated in Fig. 1. A typical scene contains multiple images in different directions, so at any time most of the scene appears distorted. Image movement is particularly evident in sound with low

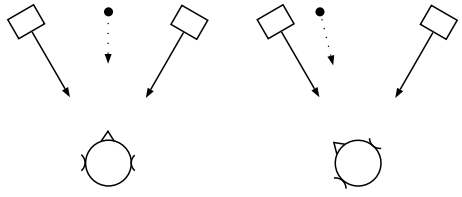

Fig. 1: The black dot indicates the direction of the image when two loudspeakers each have the same signal, for different head directions.

D. Menzies, M. F. Simón and F.M. Fazi are with the University of Southampton. e-mail: d.menzies@soton.a.cuk

frequency content, for which the ITD cue is prevalent, and varies according to the image direction, emphasizing instability in the perceived scene and its plausibility. Under normal listening conditions head rotation plays an important role in localisation, by enabling *dynamic ITD cues*[5;6;7;1] as well as other dynamic cues. Accurate reproduction of dynamic ITD cues can be expected to improve imaging. As will be shown, it also enables images to be produced in all directions, beyond the region possible using conventional panning.

The aims of this work are to investigate the distortions produced by panning, and develop improved methods, by applying acoustic analysis to existing understanding about spatial perception. The hope is to produce a system that can be of practical use, making use of the rapidly advancing technologies for head tracking. The specific contributions are:

- A spherical head model is used to calculate and characterise the ITD and ILD produced by general low frequency sound fields.
- The analysis is applied to sound fields produced by panning, and compared with cues produced by a real source to produce a relationship between the head orientation, panned image direction, and a vector describing the sound field.
- From the vector relationship, stereo panning functions are derived that take compensate for head orientation. These are shown to be equivalent to existing panning laws when the listener is facing the image. Furthermore they produce the correct ITD for an image in any direction, not only within the stereo loudspeaker span.
- The compensated stereo panning functions are tested objectively with measured head responses, and subjectively with existing listening tests and new tests in which the listener's head is tracked in real-time. The compensation is extended to include listener position compensation, so that the listener can move freely while keeping images fixed in position, near or far. The tests confirm that the functions improve image stability, and also allow images to be produced in other directions. Separate processing for high frequency content gives further improvement.

This article is organised as follows. In Section II a low frequency approximation is derived for binaural signals, based on the incident sound field. These are then used in Section III to find expressions for the inter-aural Time Difference (ITD) and inter-aural Level Difference (ILD), in terms of a general field description and the listener's head orientation. By comparison with the cues for a single plane wave, a relation is derived between the image direction, the head orientation and the field description. In Section IV stereo panning laws are derived, by finding a panned field that satisfies the relation for the desired image. The approach is referred to as *Compensated Amplitude Panning* (CAP). In Section V, CAP is evaluated using measured HRTFs. These test the robustness of the method when applied with a realistic head model, rather than the ideal model originally assumed. In Section VI a real-time implementation is described. This is applied in a listening test, described in Section VII, comparing different methods of stereo reproduction, with and without compensation. The conclusion in Section VIII summarises the findings, and gives some outlook.

Earlier related work has been presented by the authors[8;9]. An extension for near-field images has been introduced in[10], but is not developed here.

II. SOUND FIELD REPRESENTATION

A sound field over a source free region can be expanded in a Taylor series about any point in the region[11]. The first order approximation of the pressure P at a point $\boldsymbol{x}$, expanded about point $\boldsymbol{x_0}$, can be given in the frequency domain in terms of pressure and pressure gradient ∇P by

$$P(\boldsymbol{x}) \approx P(\boldsymbol{x_0}) + \nabla P(\boldsymbol{x_0}) \cdot (\boldsymbol{x} - \boldsymbol{x_0}) \tag{1}$$

The approximation is good provided the wavelength is considerably larger than the distance $\|\boldsymbol{x} - \boldsymbol{x_0}\|$, and higher order derivative terms are small compared with $P(\boldsymbol{x_0})$ and $\nabla P(\boldsymbol{x_0})$, which is usually the case when there are no close sources. Below around 700 Hz, the radius of the adult human head is less than $1/4$ of one wavelength. The relative error of the expansion pressure for a plane wave is at most 50% for any point in a region that can just enclose the head, and decreases rapidly with decreasing frequency. The sound field in this region contains all the information needed to approximate the scattering caused by placing the head in the region. The ITD cue operates strongly below 700 Hz, and is increasingly less prevalent above. It is not a coincidence that the ITD frequency limit is similar to linear approximation frequency limit, since both are determined by the head size: At higher frequencies ITD becomes an increasingly ambiguous cue for direction. This provides encouragement that the linear approximation is useful in deriving results that are valid for ITD cues, in particular producing panning laws. The success of this approach will be judged by the accuracy of the laws, which are eventually measured under realistic listener scattering conditions.

The binaural signals are not equal to the pressures at the corresponding locations in the incident field, even at low frequency, due to the non-vanishing scattered field at the head surface. Using a spherical model for the head, with ears at antipodal locations, an analytical approximation can be found for the binaural signals[12;11;13]. The resultant pressure field on the sphere, radius r, is equal to the pressure of the incident free-field plane wave on the corresponding sphere at radius $\frac{3}{2}r$[14;11]. Further more, any field in a free-field region can be approximated arbitrarily well by a plane wave expansion[15]. This implies that for any incident field at low frequency, the resultant pressure at the surface of a rigid sphere of radius r can be approximated by the incident field, evaluated at radius $\frac{3}{2}r$, provided there are no sources within this radius, which is already an assumption for the first order approximation.

With the above considerations applied to a general first order field, the binaural signals at the right and left ears are approximated by

$$P_R \approx P + \bar{\boldsymbol{r}}_R \cdot \nabla P \tag{2}$$

$$P_L \approx P + \bar{\boldsymbol{r}}_L \cdot \nabla P \quad , \tag{3}$$

where P and ∇P are the pressure and gradient of the incident field at the central point between the ears, in the centre of the

head, see Fig. 2. $\bar{\boldsymbol{r}}_R, \bar{\boldsymbol{r}}_L$ are defined for convenience, $\bar{\boldsymbol{r}}_R = \frac{3}{2}\boldsymbol{r}_R,\ \bar{\boldsymbol{r}}_L = \frac{3}{2}\boldsymbol{r}_L$. The gradient is related to particle velocity

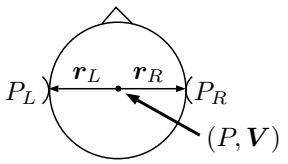

Fig. 2: Sound field variables and vectors in relation to the listener's head.

$\boldsymbol{V}$ by Euler's equation in the frequency domain[16],

$$\nabla P = -jkZ_0\boldsymbol{V} \quad , \tag{4}$$

using the positive frequency convention $P(\boldsymbol{x}, t) = P(\boldsymbol{x})e^{j\omega t}$, for angular frequency ω. Z_0 is the characteristic impedance of sound in air and $k = \frac{\omega}{2\pi}$ is the wavenumber[11]. The low frequency approximation condition is $kr_R << 1$.

In the following discussion only the relative phases of the binaural signals are of interest, so without loss of generality and in order to simplify calculation the pressure and velocity phases are rotated by the same amount so that pressure is real-valued,

$$(P, \boldsymbol{V}) \rightarrow \frac{\bar{P}}{|P|}(P, \boldsymbol{V}) \tag{5}$$

where $\bar{P}$ is the complex conjugate of P. In the following it is assumed P, $\boldsymbol{V}$ and derived quantities P_L, P_R have been phase rotated. The real and imaginary vector components of $\boldsymbol{V}$ are written $\boldsymbol{V}_\Re$ and $\boldsymbol{V}_\Im$ where $\boldsymbol{V} = \boldsymbol{V}_\Re + j\boldsymbol{V}_\Im$. Using $\boldsymbol{r}_L = -\boldsymbol{r}_R$ and Euler's equation (4), the binaural signals, (2,3), can be written as

$$P_R = P + kZ_0(\bar{\boldsymbol{r}}_R \cdot \boldsymbol{V}_\Im - j\bar{\boldsymbol{r}}_R \cdot \boldsymbol{V}_\Re) \tag{6}$$

$$P_L = P - kZ_0(\bar{\boldsymbol{r}}_R \cdot \boldsymbol{V}_\Im - j\bar{\boldsymbol{r}}_R \cdot \boldsymbol{V}_\Re) \tag{7}$$

Fig. 3 illustrates the general case using the complex plane. Both $\boldsymbol{V}_\Im$ and $\boldsymbol{V}_\Re$ are non-zero, and in different directions. As the listener's head rotates around any axis, P_R and P_L move around on opposite sides of an ellipse, shown with the dashed line. This is because the terms $\bar{\boldsymbol{r}}_R \cdot \boldsymbol{V}_\Im$ and $\bar{\boldsymbol{r}}_R \cdot \boldsymbol{V}_\Re$ each vary as the cosine of the head rotation angle relative to $\bar{\boldsymbol{V}}_\Im$ and $\bar{\boldsymbol{V}}_\Re$, which are the projections of $\boldsymbol{V}_\Im$ and $\boldsymbol{V}_\Re$ in the plane in which $\bar{\boldsymbol{r}}_R$ rotates. When $\bar{\boldsymbol{V}}_\Im$ and $\bar{\boldsymbol{V}}_\Re$ are not identical there a phase difference between $\bar{\boldsymbol{r}}_R \cdot \boldsymbol{V}_\Im$ and $\bar{\boldsymbol{r}}_R \cdot \boldsymbol{V}_\Re$, and so an ellipse is traced.

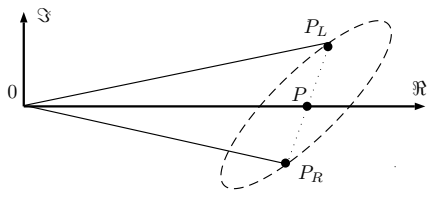

Fig. 3: P_R and P_L in the complex plane for non zero and non-aligned $\boldsymbol{V}_\Re$ and $\boldsymbol{V}_\Im$

III. LOCALISATION CUES AND REPRODUCTION

The inter-aural Time Difference (ITD) is the time difference between the arrival of a sound at the two ears. ITD is an important cue for the auditory system to localise sounds. Low frequency ITD cues from content below 1500Hz are very robust, and dominate directional cues at higher frequencies. ITD cues from higher frequencies are present but less effective. It is not known exactly how the auditory system forms ITD cues internally. In the low frequency region the binaural signals are approximated by the simple expressions (6) and (7). For a harmonic field at a single frequency, the binaural signals are then sinusoids phase shifted relative to each other. The inter-aural Phase Delay (IPD)[1], is ϕ_{RL}/ω, where $\phi_{RL} = \arg(P_R/P_L)$ is the phase shift. So the ITD is

$$\text{ITD} = \arg(P_R/P_L)/\omega \tag{8}$$

The ILD is given by the magnitude of the pressure ratio,

$$\text{ILD} = |P_R/P_L| \tag{9}$$

A more general problem is considered first: to identify all possible incident fields at the listener that produce the same cues as a given *target* plane wave. This is equivalent to finding fields with the same ITD and ILD as the target plane wave. The target plane wave field itself is a trivial solution. Another solution is provided by matching pressure and velocity, since ITD and ILD depend only on pressure and velocity, and head orientation in the low frequency limit, from (6), (7). The Ambisonic reproduction method employs this principle[3].

There are also pressure and velocity solutions that give the correct ITD and ILD but do not match the target plane wave pressure and velocity. To find these solutions first observe that according to (8) and (9) the ITD and ILD depend on the shape (but not the size) of the triangle $\triangle 0 P_L P_R$ in Fig. 3. So if two given fields have triangles with the same shape, then they produce the same image.

For an incident plane wave $\boldsymbol{V}_\Im = 0$, so the ILD = $|P_R/P_L|$ = 1 = 0 dB, see Fig. 4. This is consequence of the low frequency approximation. The ILD of a real head rises slightly with frequency over the ITD range[1]. In this case $\boldsymbol{V} = \boldsymbol{V}_\Re$. $\boldsymbol{V}$ can be decomposed as

$$\boldsymbol{V} = V\hat{\boldsymbol{V}} = \frac{P}{Z_0}\hat{\boldsymbol{V}} \tag{10}$$

where $\hat{\boldsymbol{V}}$ is the velocity direction unit vector. A hat is used to denote a vector of length 1. For convenience $\hat{\boldsymbol{r}}_I$ is defined as the direction vector to the perceived image. For the case of a plane wave the image is in the direction of the velocity since this is the direction of the source,

$$\hat{\boldsymbol{r}}_I = -\hat{\boldsymbol{V}} \tag{11}$$

Fields are sought that have the same ITD and ILD as the target plane wave field. If $\tilde{\boldsymbol{V}}_\Re$, $\tilde{\boldsymbol{V}}_\Im$ and $\tilde{P}$ are the velocity and pressure components of such a general field, then an immediate consequence is that $\tilde{\boldsymbol{V}}_\Im = 0$ in order to match the ILD. The ratio $|P_L - P|/P = \tan(\phi_{RL}/2)$ then determines the shape of any triangle of the form shown in Fig. 4. Evaluating

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2775860, IEEE/ACM Transactions on Audio, Speech, and Language Processing

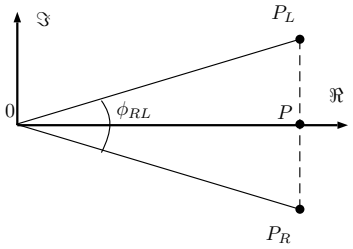

Fig. 4: P_R and P_L in the complex plane for $\boldsymbol{V}_{\Re} > 0$ and $\boldsymbol{V}_{\Im} = 0$

this quantity for a plane wave field, and a general field using (7), (10) and (11), gives

$$\frac{|P_L - P|}{P} = \begin{cases} -jk\bar{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_I, & \text{plane wave} \\ jkZ_0\bar{\boldsymbol{r}}_R \cdot \tilde{\boldsymbol{V}}_{\Re}/\tilde{P}, & \text{general field} \end{cases} \tag{12}$$

Equating the expressions on the right side in (12) implies the corresponding triangles will have the same shape, and consequently the general field will have the same ITD and ILD as the plane wave, and hence the image direction perceived in the presence of the general field will also be also be $\hat{\boldsymbol{r}}_I$,

$$-jk\bar{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_I \quad = \quad \frac{jkZ_0\bar{\boldsymbol{r}}_R \cdot \tilde{\boldsymbol{V}}_{\Re}}{\tilde{P}} \quad , \tag{13}$$

which can be written more simply as

$$\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_I - \boldsymbol{r}_V) = 0 \tag{14}$$

where $\bar{\boldsymbol{r}}_R$ has been reduced to the normal direction vector $\hat{\boldsymbol{r}}_R$, and

$$\boldsymbol{r}_V = -\tilde{\boldsymbol{V}}_{\Re}Z_0/\tilde{P} \tag{15}$$

is defined for convenience, and is equal to the *Makita Localisation Vector* given by Gerzon[3]. $\boldsymbol{r}_V$ is in the opposite direction to the velocity, and for a plane wave field is the unit direction vector to the source. $Z = P/V_{\Re} = Z_0/r_V$ can be viewed as the local specific impedance of the field. Equation (14) is a central result in this article, and will be used to derive compensated panning laws with a variety of interesting behaviour. It relates the head orientation, of which $\hat{\boldsymbol{r}}_R$ is a function with the image direction $\hat{\boldsymbol{r}}_I$ and the field description given by the Makita vector $\boldsymbol{r}_V$. Note that $\hat{\boldsymbol{r}}_R$ is a function only of head orientation, but it is not equivalent, and contains less information.

In the previous discussion the ITD is unambiguous for a simple harmonic source, since only one independent timing parameter is available, the IPD. Similarly for a single source with energy across the low frequency region, the binaural signals are offset in time, by virtue of the resultant at $\bar{\boldsymbol{r}}$ being equal to the freefield at $\boldsymbol{r}$, and so the ITD is unambiguous. More generally for a low frequency signal with $\boldsymbol{V}/P$ that is constant across the band then in the linear approximation, given by (6) and (7), then there is a linear phase relation between P_R and P_L, implying again that ITD is unambiguous. $\boldsymbol{V}/P$ is constant across band in the case of panned reproduction, as discussed in Section IV. Since only one parameter is required for ITD, this simplifies the design of processes that exploit ITD. Detailed knowledge about how the auditory system produces its own internal representation of ITD is not required. As a side remark, the simplicity and robustness of the low frequency ITD may help explain why the auditory system is biased towards it.

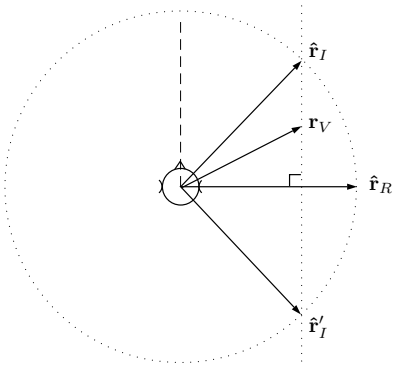

Fig. 5: Vector diagram in plan view, showing the inter-aural direction $\hat{\boldsymbol{r}}_R$, image direction $\hat{\boldsymbol{r}}_I$ and Makita vector $\boldsymbol{r}_V$

Equation (14) provides a general relationship in 3-dimensions between image direction, inter-aural direction, and the field description. This is visualised in a vector diagram, Fig. 5, showing a projection from above the listener. The image direction shown is in the horizontal plane. Given an image direction $\hat{\boldsymbol{r}}_I$ and a inter-aural direction $\hat{\boldsymbol{r}}_R$, possible values of $\boldsymbol{r}_V$ satisfying (14) lie on a plane normal to $\hat{\boldsymbol{r}}_R$, represented by the dotted line, and which contains $\hat{\boldsymbol{r}}_I$. Conversely, given a field vector $\boldsymbol{r}_V$, a consistent image can lie anywhere on the *cone of confusion* about $\hat{\boldsymbol{r}}_R$ and containing $\hat{\boldsymbol{r}}_I$. In Fig. 5 $\hat{\boldsymbol{r}}'_I$ is the rear image on this cone that is in the horizontal plane and is consistent with $\boldsymbol{r}_V$. The alternative images manifest as front-back confusion. In normal listening conditions spectral and dynamic cues help to isolate the perceived image to a particular direction on the cone at each time.

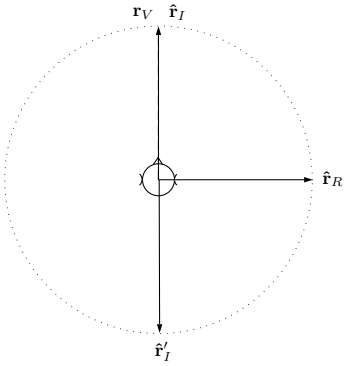

Fig. 6: Vector diagram in plan view, showing the inter-aural direction $\hat{\boldsymbol{r}}_R$, image direction $\hat{\boldsymbol{r}}_I$ and Makita vector $\boldsymbol{r}_V$ in the case where they are aligned.

If the listener's head points in the direction $\boldsymbol{r}_V$, so that $\hat{\boldsymbol{r}}_R$ is perpendicular to $\boldsymbol{r}_V$, then the ITD is 0s and a possible image direction $\hat{\boldsymbol{r}}_I$, is in the direction of $\boldsymbol{r}_V$, as shown in Fig. 6. It is then natural to ask how much the image deviates from $\boldsymbol{r}_V$ when the head direction deviates from $\boldsymbol{r}_V$. The constraint (14) is formulated with vectors, which provide the natural and most efficient way to express and derive formula

in 3-dimensions, including the panning formula in Section IV. However, in order to plot the image deviation in the 2-dimensional plane, some angle variables are introduced here, with mnemonic subscripts: θ_{VI} is the azimuth angle change from the field vector $\boldsymbol{r}_V$ to the image direction $\hat{\boldsymbol{r}}_I$, and θ_{VN} is the angle change from $\boldsymbol{r}_V$ to the forward head, or median, direction (N stands for *nose*). θ_{VI} is the angular deviation of the image when the head is turned away from the image direction by angle θ_{VN}. θ_{IN} is the angle change from $\hat{\boldsymbol{r}}_I$ to $\hat{\boldsymbol{r}}_N$. The sign convention chosen is positive for anti-clockwise change, seen from above. The angles are illustrated in Fig. 7. The separation between the dotted and dashed lines can be

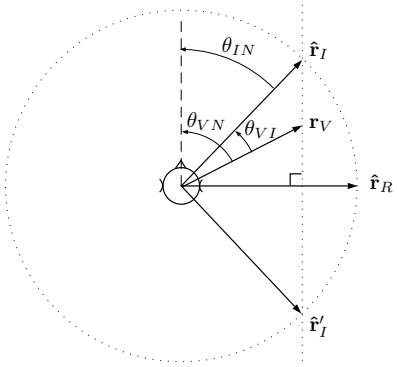

Fig. 7: Vector diagram in plan view, showing the inter-aural direction $\hat{\boldsymbol{r}}_R$, image direction $\hat{\boldsymbol{r}}_I$ and Makita vector $\boldsymbol{r}_V$, in 2-dimensions with azimuthal angles.

calculated in two ways. Equating these,

$$|\boldsymbol{r}_V| \sin\theta_{VN} = |\hat{\boldsymbol{r}}_I| \sin\theta_{IN} \tag{16}$$

which in terms of only θ_{VN} and θ_{VI} is

$$r_V \sin\theta_{VN} = \sin(\theta_{VN} - \theta_{VI}) \tag{17}$$

so,

$$\theta_{VI} = \theta_{VN} - \arcsin(r_V \sin\theta_{VN}) \tag{18}$$

θ_{VI} is plotted against θ_{VN} in Fig. 8. Each line was calculated for a constant r_V. The specific impedance is $Z = Z_0/r_V$, and θ_L is referred to later in Section IV. For $r_V = 1.0$ the pressure and velocity match that of a plane wave field, and no deviation in image direction from the field vector $\boldsymbol{r_V}$ occurs as the head rotates. For $r_V < 1$, the ratio P/V is greater compared with that of a plane wave. When the listener turns away from $\boldsymbol{r_V}$ the image also moves away from $\boldsymbol{r}_V$ in the direction the head is moving, but by a smaller angle. The deviation θ_{VI} reaches a maximum when $\theta_{VN} = 90^o$ (^{o} are used to indicate calculation in degrees rather than radians). Conversely, when $r_V > 1$, P/V is less than that for a plane wave. As the listener turns away from $\boldsymbol{r_V}$ the image moves away from $\boldsymbol{r_V}$ in the opposite direction the head is moving, shown by the dashed lines. In this case for a value of $\theta_{VN} < 90^o$ depending on r_V, $\hat{\boldsymbol{r}}_I$ is aligned with $\hat{\boldsymbol{r}}_R$. For greater values of θ_{VN} there is no strict solution for the image as the ITD is greater than is possible from a plane wave. For this reason the dashed lines extend less than $\theta_{VN} = 90^o$. In practice the image becomes fixed at $\theta_{NI} = 90^o$ for these excessive ITDs.

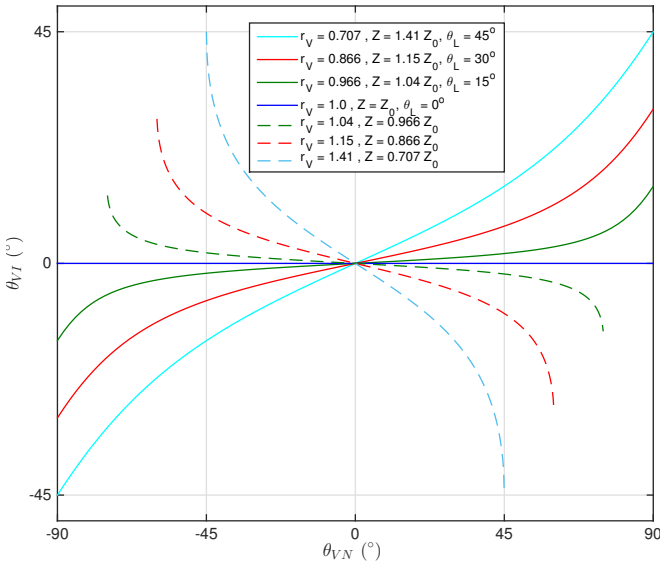

Fig. 8: The shift of image angle relative to the velocity vector, θ_{VI}, is plotted as a function of the angle between head direction and the velocity vector, θ_{VN}, for several values of the magnitude of the Makita Vector, r_V.

To finish this section a link is made between the field description given by the Makita vector and the *B-format* description used in Ambisonics. B-format consists of 4 signals (W, X, Y, Z). The pressure and velocity are proportional to the components, $P = \gamma\sqrt{2}W$, $\boldsymbol{V}Z_0 = \gamma(X, Y, Z)$, where γ is a calibration constant[17]. Then from (15) the Makita vector is given by

$$\boldsymbol{r}_V = \frac{(X, Y, Z)}{\sqrt{2}W} \quad , \tag{19}$$

from which the image stability can be assessed according to Fig. 8. The B-format signals (W, X, Y, Z) can be measured directly with a *Soundfield microphone*[18].

IV. PANNING WITH COMPENSATION FOR HEAD ROTATION

Amplitude panning is a spatial audio reproduction method in which several loudspeakers produce plane waves converging at the listener position in phase. For each image, the signals driving the loudspeakers are produced by multiplying the source signals by a set of real-valued gains.

In the following, g_i is defined as the gain for the i-th loudspeaker. Also, $\boldsymbol{V}_i$ and P_i are defined as the velocity and pressure at the central head position, and $\hat{\boldsymbol{r}}_i$ is the direction vector from the centre, for to the i-th loudspeaker. $\hat{\boldsymbol{r}}_i = -\hat{\boldsymbol{V}}_i$. Possibly additional delay compensation is required for some driving signals to compensate for different values $\boldsymbol{r}_i$, so that the waves all arrive in phase. The pressure and velocity at the

listener are then given by the sum of the contributions from each loudspeaker, so that

$$\boldsymbol{r}_V = -Z_0 \frac{\boldsymbol{V}}{P} = -Z_0 \frac{\sum \boldsymbol{V}_i}{\sum P_i} = -\frac{\sum P_i \hat{\boldsymbol{V}}_i}{\sum P_i} = -\frac{\sum g_i \hat{\boldsymbol{V}}_i}{\sum g_i} = \frac{\sum g_i \hat{\boldsymbol{r}}_i}{\sum g_i} \tag{20}$$

This is in agreement with the original definition of Makita Localisation Vector in terms of panning gains[3].

To discuss existing stereo panning methods, the following angles will be used, θ_L, θ_I and θ_N, as defined in Fig. 9. The Tangent Law[2;23] for stereo loudspeakers is based on the

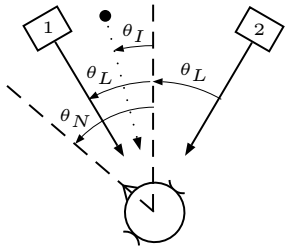

Fig. 9: Stereo reproduction setup showing conventional angle variables. θ_I is the angle from the loudspeaker centre line to the image, which is positive in the direction shown, and negative in the opposite. θ_N the angle from the line to the listener direction. $2\theta_L$ is the separation between the loudspeakers.

observation that when the listener faces an image the ITD is 0s. From this the panning law can be found relating the image direction to loudspeaker gains,

$$\frac{\tan\theta_I}{\tan\theta_L} = \frac{g_1 - g_2}{g_1 + g_2} \tag{21}$$

The panned field is not generally a plane wave but it can produce the same ITD cue as a plane wave comming from the direction of the image, like the general fields considered in Section III. Generalising for more than 2 loudspeakers, first observe that $\boldsymbol{r}_V$ and the velocity $\boldsymbol{V}$ are aligned with the direction of $\sum g_i \hat{\boldsymbol{r}}_i$, from (20). So a listener facing this direction will experience an ITD of 0s, and the image will be directly ahead. This is the basis for Vector Base Amplitude Panning (VBAP)[19]. When the listener moves their head away from the initial image direction then $\hat{\boldsymbol{r}}_V$ is no longer aligned with $\hat{\boldsymbol{r}}_N$, and VBAP and the Tangent Law no longer reproduce the desired image correctly. The method presented here takes into account the head alignment to give gains that reproduce a given image correctly, at least according to low frequency localisation cues.

Fig. 8 includes image shift plots for different stereo configurations described by angle θ_L. In these cases $r_V < 1$ and r_V is calculated using (20). For wider loudspeaker separations, r_V is smaller, and the image deviation θ_{VI} increases more rapidly as the listener turns away from the central direction.

The compensated panning problem is to find panning gains $\{g_i\}$ that produce a stable image in the direction $\hat{\boldsymbol{r}}_I$ for a given inter-aural axis $\hat{\boldsymbol{r}}_R$ determined by the head orientation. A solution is first sought for two loudspeakers. This is the simplest but also the most restrictive case. Besides compensating for image instability, an additional aim is to produce stable images in any direction, not only between the loudspeakers, as for conventional stereo panning.

To simplify calculation the following additional constraint is imposed,

$$g_1 + g_2 = 1 \tag{22}$$

This does not restrict the solution search in practice since given any solution (g_1, g_2), the solution $(g_1, g_2)/(g_1 + g_2)$ satisfies (22), and the original solution is simply a multiple of this, by the factor $(g_1 + g_2)$. With this constraint the Makita vector is, from (20),

$$\boldsymbol{r}_V = g_1 \hat{\boldsymbol{r}}_1 + g_2 \hat{\boldsymbol{r}}_2 = g_1 \hat{\boldsymbol{r}}_1 + (1 - g_1)\hat{\boldsymbol{r}}_2 \tag{23}$$

This expresses the sound field in terms of the loudspeaker layout geometry. The right-hand side is the expression of a straight line passing through $\hat{\boldsymbol{r}}_1$ and $\hat{\boldsymbol{r}}_2$, parametrised by g_1. $\boldsymbol{r}_V$ lies on this line, independently of constraint (22). Taking the scalar product of both sides of (23) with $\hat{\boldsymbol{r}}_R$ and substituting for $\hat{\boldsymbol{r}}_R \cdot \boldsymbol{r}_V$ using (14) leads to an expression for g_1, and likewise for g_2,

$$g_1 = \frac{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_I - \hat{\boldsymbol{r}}_2)}{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_1 - \hat{\boldsymbol{r}}_2)} \qquad g_2 = \frac{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_I - \hat{\boldsymbol{r}}_1)}{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_2 - \hat{\boldsymbol{r}}_1)} \tag{24}$$

These panning gains are valid except where the loudspeakers are symmetrically at the side of the listener, in which case $\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_1 - \hat{\boldsymbol{r}}_2) = 0$, and the image is not directly at either of the loudspeakers. The image instability in this configuration is known[3;20], and was one of the motivations in the early development of Ambisonics. The formulation here shows explicitly there is no solution.

In the special case where the listener is facing the image $\hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_I = 0$. Then from (14) $\hat{\boldsymbol{r}}_R \cdot \boldsymbol{r}_V = 0$, and the image direction, listener direction and velocity are aligned, and the ITD is 0 s. Using the variables defined here for the two loudspeaker case, VBAP is based on the condition[19],

$$\hat{\boldsymbol{r}}_I = g_1 \hat{\boldsymbol{r}}_1 + g_2 \hat{\boldsymbol{r}}_2 \tag{25}$$

From (20) the right hand side vector in (25) is in the direction of velocity. So VBAP is equivalent to compensated panning in the case of the listener facing the image, up to an overall gain scale factor. As shown by Pulkki[19], the Tangent Law is equivalent to VBAP for two loudspeakers.

In VBAP the gains are scaled to satisfy a normalisation condition, in order to control the perceived level. This does not effect the ITD or localisation. The normalisation used is

$$\hat{g}_1^2 + \hat{g}_2^2 = 1 \tag{26}$$

where $\hat{g}_1$ and $\hat{g}_2$ are the normalised gains, proportional to g_1 and g_2. This normalisation is appropriate in the high frequency region, where incident waves mix incoherently, and is acceptable for broadband signals where most of the energy is in high frequency region. However for low frequency signals the initial condition (22) already provides an appropriate normalisation. This is because at low frequency the binaural amplitudes approximate to the free field pressure P, and (22) implies that P is kept constant.

Fig. 10 shows how $\boldsymbol{r}_V$ can move to the right to compensate for head movement to the left in order to keep the image central. Possible values of $\boldsymbol{r}_V$ that can be produced by panning between the two loudspeakers lie on the dashed horizontal line between $\hat{\boldsymbol{r}}_1$ and $\hat{\boldsymbol{r}}_2$, by virtue of (20). Values of $\boldsymbol{r}_V$ which produce the desired images $\hat{\boldsymbol{r}}_I$ or $\hat{\boldsymbol{r}}'_I$ lie on the dotted line. The intersection of the two lines gives the one value of $\boldsymbol{r}_V$ that produces the desired image by panning. Fixing $\hat{\boldsymbol{r}}'_I$ rather than $\hat{\boldsymbol{r}}_I$ shows that the correct ITD for rear images can be produced using only the front pair of loudspeakers. Wallach[5] achieved a related effect using a dense array of loudspeakers at the front. Depending on the head rotation a loudspeaker was selected to produce a source directly in the direction $\hat{\boldsymbol{r}}_I$, so that $\hat{\boldsymbol{r}}'_I$ was kept fixed direction behind. This and other experiments support the idea of dynamic ITD cues, which were shown to be significant for all image directions. Related experiments have been reported[6;7;1], and Wallach's experiments have been repeated with modern equipment[21]. The CAP system produces the correct ITD for any target image, and so supports dynamic ITD cues.

For $\boldsymbol{r}_V$ outside the circle the gain of the furthest loudspeaker is negative. When the dashed and dotted lines are parallel and distinct there is no solution. The panning gains

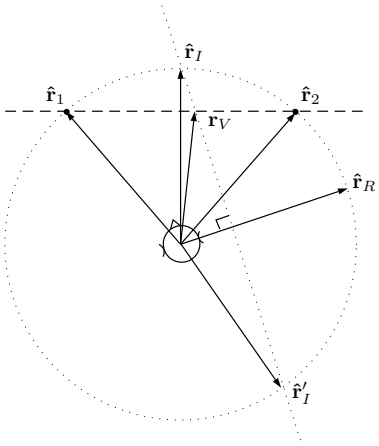

Fig. 10: Vector diagram in plan view, for a listener facing towards left of centre of the stereo array. The Makita vector is to the right of centre in order to keep the image central. Shown are loudspeaker directions $\hat{\boldsymbol{r}}_\mathbf{1}$, $\hat{\boldsymbol{r}}_\mathbf{2}$ the inter-aural direction $\hat{\boldsymbol{r}}_R$, image direction $\hat{\boldsymbol{r}}_I$ and Makita vector $\boldsymbol{r}_V$

(24) are equally valid for target image positions that are outside the horizontal plane of the loudspeakers, either above or below, since in every case constraint (14) is satisfied, and the correct ITD cue is produced according to the sphere head model. This raises the possibility of height synthesis using CAP with only two loudspeakers. If the listener is off the line of the loudspeakers, then the general picture is like Fig. 10, except that $\hat{\boldsymbol{r}}_I$ will appear inside the circle in the plan view, since it is a projected view.

There is an interesting case where the listener is on the line between the loudspeakers, and the image is directly above the line, illustrated in Fig. 11. Possible values of $\boldsymbol{r}_V$ are given by the intersection of the line between the loudspeakers, represented by the dashed line, and the plane normal to $\boldsymbol{r}_R$,

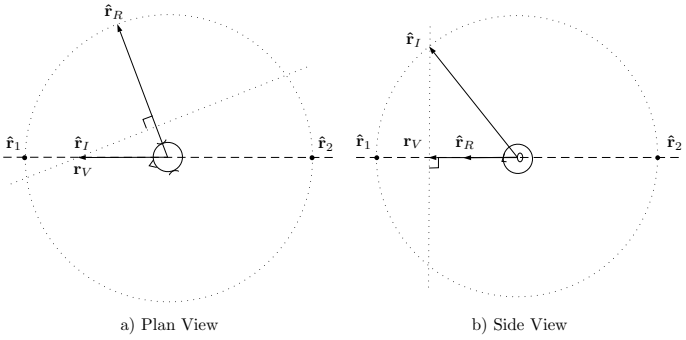

Fig. 11: Vector diagrams in plan and side views, for a listener positioned between two loudspeakers, with the image directly above he line of the loudspeakers. Shown are loudspeaker directions $\hat{\boldsymbol{r}}_\mathbf{1}$, $\hat{\boldsymbol{r}}_\mathbf{2}$ the inter-aural direction $\hat{\boldsymbol{r}}_R$, image direction $\hat{\boldsymbol{r}}_I$ and Makita vector $\boldsymbol{r}_V$.

represented by the dotted line. If the listener is not looking in the direction of the loudspeakers then there is only one possible value $\boldsymbol{r}_V$, which is the vertical projection of $\hat{\boldsymbol{r}}_I$ downwards on to the dashed line. This solution is unusual because it does not depend on the head azimuth, and so does not require orientation tracking (It is nonetheless compensated for head orientation, because $\boldsymbol{r}_R$ is assumed to be horizontal). If the listener is looking along the line of the loudspeakers, then $\boldsymbol{r}_V$ can be anywhere in the median plane, as the ITD is 0s.

To calculate $\boldsymbol{r}_V$ and the panning gains g_1 and g_2 observe that in this case $\boldsymbol{r}_V$ is given by the projection of $\hat{\boldsymbol{r}}_I$ in $\hat{\boldsymbol{r}}_1$

$$\boldsymbol{r}_V = (\hat{\boldsymbol{r}}_I \cdot \hat{\boldsymbol{r}}_1)\ \hat{\boldsymbol{r}}_1 \tag{27}$$

Equating the RHS with $\boldsymbol{r}_V$ expressed in terms of panning gains, (23), leads to

$$g_1 = \frac{1 + \hat{\boldsymbol{r}}_I \cdot \hat{\boldsymbol{r}}_1}{2} \qquad g_2 = \frac{1 - \hat{\boldsymbol{r}}_I \cdot \hat{\boldsymbol{r}}_1}{2} \tag{28}$$

When the image is overhead, $g_1 = g_2 = 1/2$, as expected. The gains can also be derived directly from (24):

$$g_1 = \frac{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_I - \hat{\boldsymbol{r}}_2)}{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_1 - \hat{\boldsymbol{r}}_2)} = \frac{\hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_I + \hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_1}{2\hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_1} \tag{29}$$

$$= \frac{(\hat{\boldsymbol{r}}_I \cdot \hat{\boldsymbol{r}}_1 + 1)\ \hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_1}{2\hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_1} \tag{30}$$

$$= \frac{1 + \hat{\boldsymbol{r}}_I \cdot \hat{\boldsymbol{r}}_1}{2} \tag{31}$$

and similarly for g_2. The step to (30) uses the identity $\hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_I = (\hat{\boldsymbol{r}}_I \cdot \hat{\boldsymbol{r}}_1)(\hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_1)$, which follows from the orthogonality of the two planes described by $(\hat{\boldsymbol{r}}_I, \hat{\boldsymbol{r}}_1)$ and $(\hat{\boldsymbol{r}}_R, \hat{\boldsymbol{r}}_1)$. The panning gains (28) are very similar to those produced by *W-Panning*[22] , described previously in the context of Ambisonic reproduction, and which provides a convincing way to pan an image through the listening position by controlling the ratio of W signal representing pressure to the X, Y, Z signals, representing velocity. For a horizontal array this has the effect of producing an enveloping image. There is also a noticeable sense of elevation, which is lost if the listener's head is pitched upwards. In its original form W-panning uses an energy normalisation, rather than pressure normalisation.

Improved equalisation can be achieved with the frequency dependent normalisation used in CAP. The dynamic height cue for ITD is ambiguous in this case, it could equally indicate an image below the horizontal. The ITD analysis here provides some insight into the perceptual effects of W-Panning. Section VII includes some tests for height.

In order to directly compare the compensated panning functions with the existing Tangent Law method, the gain ratio is calculated, since this contains all information apart from the normalisation.

$$\frac{g_1}{g_2} = \frac{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_I - \hat{\boldsymbol{r}}_2)}{\hat{\boldsymbol{r}}_R \cdot (\hat{\boldsymbol{r}}_1 - \hat{\boldsymbol{r}}_I)} = \frac{\cos\theta_{RI} - \cos\theta_{R2}}{\cos\theta_{R1} - \cos\theta_{RI}} = \frac{\sin\theta_{IN} - \sin\theta_{2N}}{\sin\theta_{1N} - \sin\theta_{IN}} \tag{32}$$

where θ_{IN}, θ_{1N} and θ_{2N} are the azimuth angle changes to the direction the listener is facing, from respectively: the direction of the image, the direction of loudspeaker 1, and the direction of loudspeaker 2. It is natural to use these angles based on the vector formulation, but they are less familiar in stereo terminology. In terms of these variables (32) can be written as $\theta_{IN} = 0$

$$\frac{g_1}{g_2} = \frac{\sin(\theta_N - \theta_I) - \sin(\theta_N + \theta_L)}{\sin(\theta_N - \theta_L) - \sin(\theta_N - \theta_I)} \tag{33}$$

When the listener faces the image, $\theta_{IN} = 0$, $\theta_I = \theta_N$, and (33) simplifies to

$$\frac{g_1}{g_2} = \frac{\sin(\theta_L + \theta_I)}{\sin(\theta_L - \theta_I)} \tag{34}$$

The right hand side can be rewritten in terms of tangent functions by expanding the numerator and denominator using trignometric addition formulae and then dividing both by $\cos\theta_L \cos\theta_N$, giving,

$$\frac{g_1}{g_2} = \frac{\tan\theta_L + \tan\theta_I}{\tan\theta_L - \tan\theta_I} \tag{35}$$

which is the Tangent Law (21) rearranged for the gain ratio g_1/g_2, showing directly that CAP is an extension of this law. This was expected earlier because both methods are based on matching ITD to the target image.

Pulkki has investigated the panning of off-median images[4]. In an experiment subjects adjusted the panning gains for two loudspeakers separated by 60^o so that the image aligned with real reference sources, for various fixed head directions. The gain ratio g_1/g_2 was recorded. His subjective results are compared with results predicted using CAP (32).

The test results[4] are reproduced in Fig. 12, with annotations added to show the predicted results. Four different broadband sound sources were tested, shown at the bottom of the figure. The angles shown measure horizontal azimuth relative to the central direction between the loudspeakers. The loudspeakers with gains g_1 and g_2 were displaced by -30^o and $+30^o$ azimuth respectively from the centre. The reference sources were located in one of three directions shown as Ref -15^o, Ref 0^o, Ref +15^o. The box plots show the measured gain ratios for each reference source. The dashed green lines show the predicted gain ratio required to pan the image to the reference direction with the listener also facing the reference direction, according to the Tangent Law (this is also indicated by the scale on the right side). The solid red lines indicate the gain ratio predicted using (32). There is agreement with subjectively adjusted gains for gain ratios within -5 to 5 dB. Also the compensation in gain ratio from the Tangent Law case to the compensated case is considerable in some cases: For example for reference -15° there is approximately 12 dB compensation between listener direction -60°, shown by the red line, and -15° shown by the dashed line.

Some cases do not match so well. For the listener at 0° the observed ratio is expected to be further from 0 dB than the Tangent Law case shown by the dashed line. The sources in the experiment are all broadband, so cues from the high frequency content may be acting against the low frequency ITD cues. Also, since the head is fixed, dynamic ITD cues are not present, which may weaken the impression based on ITD.

The angular scale, shown on the right side, is very non-uniform, so that graphically errors appear exaggerated particularly for the last two cases where the reference is at 15°. In the last case the listener is 75° from the reference. The greater uncertainty in adjusted gain values is expected because localisation accuracy is reduced towards the inter-aural axis, as noted by Pulkki. On the other hand for the case with the listener at -60° and the reference at -15° the angular separation is 45° and there is agreement. This experiment only covers a few cases, and does not focus on the range of image directions about the listener direction that would be most important in practice. However it provides some agreement, within the stated limitations.

Pulkki found a compensated panning formula that fits the data for specific loudspeaker separations, and horizontal images[4]. The panning formula derived here, (32), can be applied to any loudspeaker separation, image direction, and head orientation, and for images out of the loudspeaker-listener plane. This is useful under normal conditions where the listener has unrestricted head movement and the apparent loudspeaker separation may vary continuously as the listener moves.

V. TESTS WITH A MEASURED HEAD RESPONSE

The formulae that have been derived for compensated stereo panning, (24), are exact for a spherical head in the low frequency limit, by construction. They are tested here for a more typical head shape, in the low frequency ITD range. Head related impulse responses (HRIRs) measured from a KEMAR binaural microphone[24], were used to simulate the binaural response for a panned sound field produced by two loudspeakers. If the binaural impulse responses from the two loudspeakers are (h_{1L}, h_{1R}) and (h_{2L}, h_{2R}), then the simulated binaural responses for a signal panned with gains (g_1, g_2) are formed by adding the response from each loudspeaker,

$$(g_1 h_{1L} + g_2 h_{2L}, \quad g_1 h_{1R} + g_2 h_{2R}) \tag{36}$$

The simulations accurately predict the measurements that would be made in a real panning experiment using a KEMAR head. The only assumption is that mutual loudspeaker scattering is negligible. The MIT KEMAR HRIR set used consists

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2775860, IEEE/ACM Transactions on Audio, Speech, and Language Processing

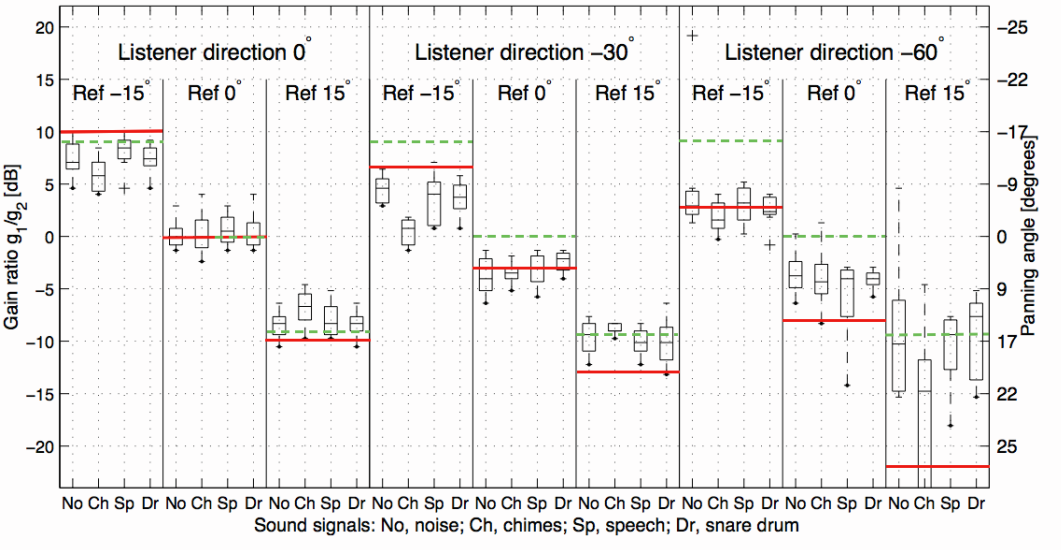

Fig. 12: Results reproduced from Pulkki[4] with CAP predictions added. The box plots show stereo gain ratios adjusted by the subjects to match the stereo image to a reference loudspeaker. There are three references, three listener directions and 4 different sounds. The dashed lines show the gain ratios using the Tangent Law, for a listener facing the reference. The solid lines show the gain ratio using the CAP.

of two different symmetrical sets, each defined for one ear. To recover one set for both ears the left ear response was calculated from the right ear response, $h_L(\theta) = -h_R(-\theta)$, for azimuth θ.

The ITD is calculated by cross-correlating the binaural signals. A low pass filter is applied first to extract the low frequency parts. This approximates the front end of the auditory system which behaves like a filter bank, and so IPD variations across the pass band are averaged. The calculation is repeated for several filter cutoffs, in order to assess over what frequency range the panning method is effective. The filters are designed with stopband level of -100 dB, passband ripple of 5 dB, and transition region of $(400\alpha$ Hz, 700α Hz$)$. Values of α used are (1.0, 2.0, 3.0). The cross-correlated signals are up-sampled by a factor of 32, filtered using a linear phase low pass FIR filter to suppress aliasing, and finally the peak is detected. The time of the peak indicates the ITD. Upsampling is necessary because the resolution of human ITD, around $10\mu S$, is superior to the sampling resolution used for the HRIRs[1].

A representative range of cases is considered in order to assess the objective performance of CAP. Corresponding plots shown in Fig. 13. Each case is specified using the angle variables θ_L, θ_I and θ_N defined previously in Fig. 9. There are three graphs in each plot. *mono* shows the ITD for a single source in the target image direction, calculated by processing the binaural response taken directly from the KEMAR data. The response is selected according to the relative angle between the listener's head and the target image. *stereo* shows the ITD when the image is panned statically using the Tangent Law, so this is expected to be near 0 s when the listener is facing the target image direction. It is found by processing responses calculated with (36), using Tangent Law gains. These gains are calculated using (24) and (40), with $\hat{\boldsymbol{r}}_R$ set for a listener pointing in the target image direction, which is equivalent to the Tangent Law / VBAP as shown earlier. *stereo compensated* shows the ITD when the image is panned with compensation, calculated as for the stereo case but with $\hat{\boldsymbol{r}}_R$ varied according to the listener direction indicated by the horizontal axis.

The angle of a stereo image cannot be read from the ITD plots. However this and the absolute error compared with the target image can be found indirectly, as follows. θ_N is the direction the head is facing for the stereo field in question, compensated or not. For clarity in this derivation we define $\theta''_N = \theta_N$, the number of primes indicating stereo or mono. This produces an ITD that could also be produced if the same listener were facing in some angle θ'_N with a mono source positioned at θ_I. Again for clarity, $\theta'_I = \theta_I$ is defined. θ'_N can be found by reading backwards through the mono ITD plot. The angular separations between the head and the image direction are the same in the mono and stereo cases since the respective ITDs are the same. θ'_I is the known target image direction. If θ''_I is the unknown image direction produced by the stereo case (with head direction θ_N), then

$$\theta'_N - \theta'_I = \theta''_N - \theta''_I \tag{37}$$

The perceived image error is defined by

$$\theta_E = \theta''_I - \theta'_I \tag{38}$$

which cannot be evaluated directly because θ''_I is unknown. However, it can be found by rearranging (37),

$$\theta''_I - \theta'_I = \theta''_N - \theta'_N \tag{39}$$

Hence the error θ_E can be read directly from the horizontal gap between the mono plot and the stereo plot in question. Fig. 13a illustrates the error for the uncompensated stereo

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2775860, IEEE/ACM Transactions on Audio, Speech, and Language Processing

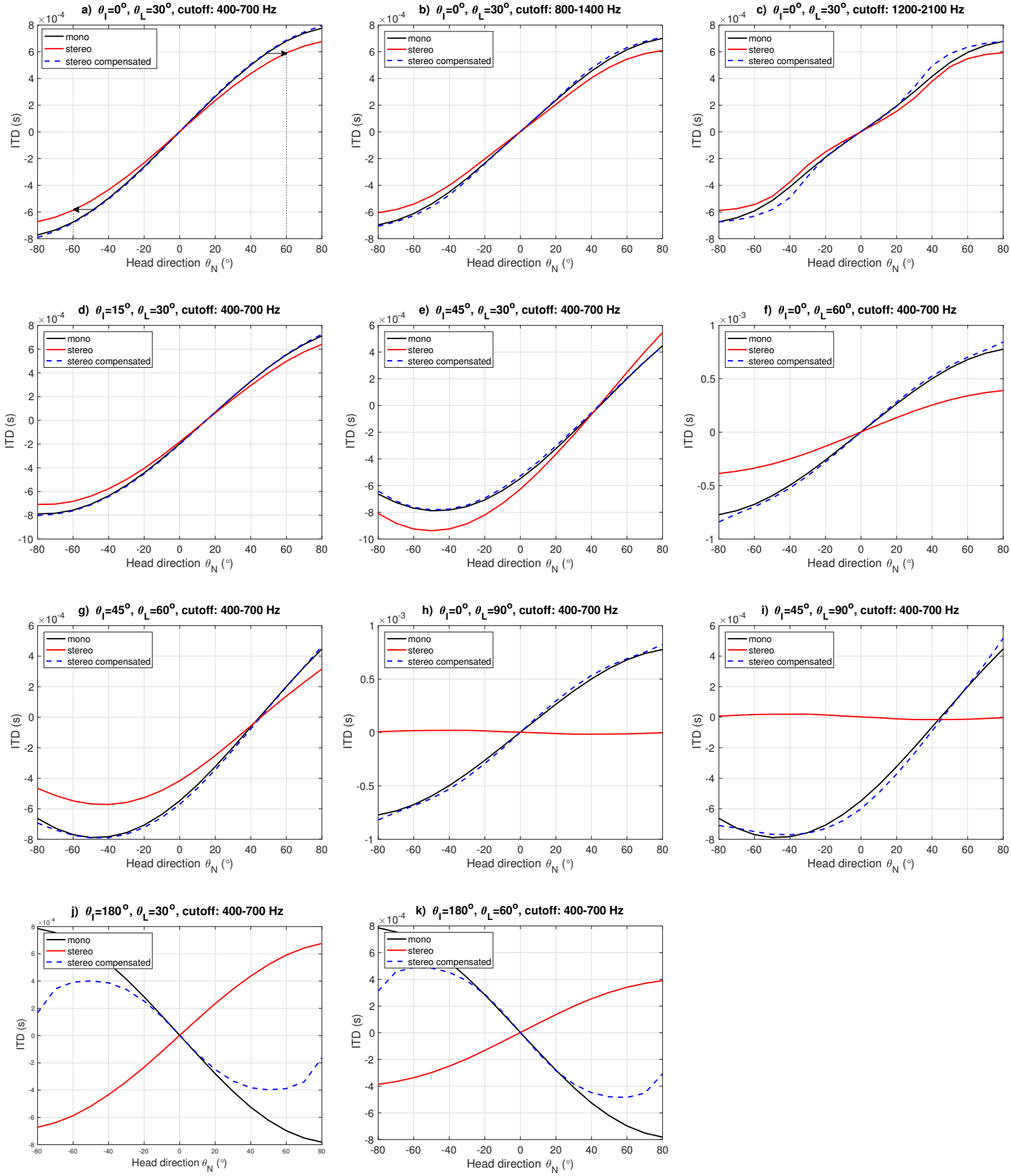

Fig. 13: Calculated ITD vs head direction, for uncompensated and compensated stereo reproduction using measured responses from a KEMAR binaural microphone, and for variations of target image θ_I, loudspeaker separation $2\theta_L$, and low pass filter.

(solid red line) at $\theta_N = -60°$ and $\theta_N = 60°$. A black arrow indicates the image error magnitude and sign: right positive, left negative. The arrow points from the mono to the stereo plot. θ_E can be read directly from the ITD plots to the accuracy needed, and so additional θ_E plots have not been included. The ITD plots also provide insight about the underlying cues, including cases where the imaging is unclear.

CAP is only useful where the uncompensated error is significant, which the case when the error exceeds the Minimum Audible Angle (MAA)[25]. For lateral horizontal image displacements of $0°$, $30°$, $60°$ and $75°$, representative MAAs in the low frequency ITD range are respectively $1°$, $1.5°$, $3°$,$7°$. MAA is measured under static head conditions, which suppresses the ITD localisation process. Variants of MAA under moving source and self moving head conditions show greater sensitivity to angular change,[26;27]. Even these may not capture the full sensitivity of the ITD localisation process, and therefore underestimate the extent to which uncompensated error is significant.

The first case, Fig. 13a, is for a standard stereo setup with a frontal image and a cutoff region 400-700 Hz. The uncompensated image error exceeds the MAA for head deviations greater than $5°$, while the compensated image error is well within the MAA over the whole range of head directions. For a cutoff region 800-1400 Hz the compensated ITD error is slightly outside the MAA, as shown in Fig. 13b. For 1200-2100 Hz the compensated error is much more significant, although there remains a region $|\theta_N| < 20°$ where error is within the MAA. This case is on the outer limit for low frequency ITD cues, and shows that the contribution to ITD from the additional frequencies rapidly degrades the compensated error.

There is a similar picture when the image is moved off centre to $\theta_I = 15°$ in Fig. 13d, and when moved outside the loudspeaker range to $\theta_I = 45°$, Fig. 13e. For an image outside of the loudspeaker span, the gain of the loudspeaker furthest from the image is negative. Cancellation at the listener makes the ITD more sensitive to the departure of KEMAR from the spherical model, which is reflected by the increased compensated error.

Repeating with a wider loudspeaker separation $\theta_L = 60°$ (Fig. 13f, and 13g), the uncompensated error is much greater and exceeds the MAA over nearly the whole range. The compensated error is within or slightly exceeds the MAA over the whole range. For example in Fig. 13f the uncompensated error is $12°$ for a head deviation of $20°$. When the loudspeaker separation is increased so they are facing opposite to each other (Fig. 13h and 13i), there is further increase in the compensated error, but still close to the MMA over the whole range. So even though the loudspeakers are at the sides of the listener, these plots indicate that compensated images are stable for low frequency sources over a remarkably wide range of image directions (-45°, +-45°), and listener directions (-90°, +-90°), whereas the uncompensated stereo image is completely unstable. Clearly it is not possible to pan high frequency content effectively in this arrangement and so it will generate conflicting directional cues. The high frequency content should be either reduced or reproduced by an alternative approach. The separation of image from loudspeakers can be pushed further by reversing the image to the rear, with a standard stereo arrangement, see Fig. 13j. The error is within the MMA in the region $|\theta_N| < 20°$, but diverges rapidly beyond. The panning gains become negative for less head deviation, compared with the frontal image case. Increasing the loudspeaker separation (Fig. 13k) widens the region where the error is within the MMA. Rear images offset to the side show similar behaviour, except that the region that is compensated well is biased towards the direct rear, as shown in Fig. 13l.

VI. REAL-TIME IMPLEMENTATION

A panning system with adaptation for head rotation was implemented using Max/MSP. This also includes compensation for listener position described previously[28]. Position compensation is achieved by updating the relative image direction, $\hat{\boldsymbol{r}}_I$, and the loudspeaker directions, $\hat{\boldsymbol{r}}_1$, $\hat{\boldsymbol{r}}_2$, according to the listener position. Additional gain factors compensate for the varying distances between the listener and the loudspeakers. The loudspeakers used were Genelec 8010, which are nearly omni directional in the low-frequency ITD range, so that no compensation was needed for off-axis listening. An overview is shown in Fig. 14, omitting details relating to position tracking. A Microsoft Kinect-based video tracking tracking system was later used for position sensing, in combination with a wireless Sparkfun inertial measurement unit mounted on a cap for head orientation tracking. This works well over an area 4 x 4 m provided there are no significant distortions of the local magnetic field, which effect the compass in the Sparkfun unit. Each time the sensor cap put on, the head orientation measurement is calibrated by having the listener look at a central mark. Systems based on optical markers and video offer high spatial resolution, low latency tracking. However these are also much more expensive and less portable than the system used. For many practical applications the tracking system should be as non-invasive as possible. One possible approach is to use only external video cameras with image recognition to identify the head. Such systems of sufficient quality are not currently readily available. However, this situation is likely to continue improving rapidly due to the demand in a variety of areas.

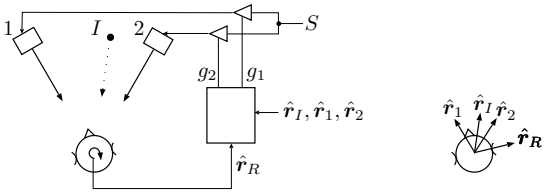

Fig. 14: Overview of a system implementing CAP for stereo, including adaptation for head rotation, shown for one object image. The signal from the head track is shown, and the input audio signal S.

The system inputs are 3-dimensional vectors represented with cartesian co-ordinates: the directions of the image, $\hat{\boldsymbol{r}}_I$, loudspeaker directions, $\hat{\boldsymbol{r}}_1$, $\hat{\boldsymbol{r}}_2$ and head alignment, $\hat{\boldsymbol{r}}_R$. The loudspeaker signal gains are calculated using (24) for each

reproduced object image. The denominators can be pre-calculated since they are the same for all images. Whenever the head changes direction or moves the gains (g_1, g_2) are updated.

An option is provided to pan the high frequency component of the source signals separately, according to an energy localisation model. The Gerzon Energy Vector $\boldsymbol{r}_E$ provides an estimate of the image direction cue produced by panning high frequency signals, and is independent of head direction,[3]. It is defined by

$$\boldsymbol{r}_E = \frac{\sum \bar{g}_i^2 \hat{\boldsymbol{r}}_i}{\sum \bar{g}_i^2} \tag{40}$$

where the hat is used to indicate the gains are for high frequency. By comparison with (20) the gains can be found by first calculating the low frequency gains using the Tangent Law / VBAP then applying the mapping $\bar{g}_i = \sqrt{g_i}$. The gains can then be modified to satisfy the normalisation $\sum \bar{g}_i^2 = 1$. This technique is known as Vector Base Intensity Panning (VBIP)[29]. A convenient way to find the stereo gains is to use (24) setting $\hat{\boldsymbol{r}}_R \cdot \hat{\boldsymbol{r}}_I = 0$ because the listener is facing the image. The normalisation is already correct in this case.

The two frequency regions are processed separately to give better overall localisation, and each with optimal normalisation to give equalisation that is more spatially uniform. In the implementation a 2nd order crossover filter is used with a cutoff frequency of 1500 Hz. This frequency can be tuned interactively while listening to the reproduction. Ambisonic decoders have a similar dual decoding process, realised with shelf filters acting on the input B-format signal.

VII. LISTENING TEST

To assess the stereo CAP system subjectively, a listening test was performed. One possible approach, described in Section IV, is to fix the listener's head and ask them to describe off median images. However these are very unnatural listening conditions, and it is not clear how representative the results are either in terms of directional accuracy, or overall impression. Certainly it does not allow for dynamic ITD cues, which turn out to be essential for image formation in some cases. On the other hand, if an image location task is set with unrestricted head movement, then it is hard to associate a head direction with the image estimation. There is a natural tendency for the listener to turn towards the image, which undermines the test.

Informal listening indicated that listeners can reliably compare the *positional stability* of an image for different reproduction conditions. This is described as the degree to which an image that is intended to be stationary appears fixed in space, when the listener moves or rotates their head. Distant objects are fixed in direction relative to the listener. This can be extended to the overall stability of a reproduced scene, containing several objects. A listening test was devised around this idea. Each subject listener was asked to rate the stability of a single image under 5 different reproduction conditions. The subjects were initially standing so that the loudspeakers were separated by 60^o ($\theta_L = 30^o$), and then encouraged to move around the vicinity and rotate their heads while listening. The subjects could choose to listen to the conditions in any order as many times as they wished. The order of conditions was randomised for each trial. The scores were recorded for each using sliders on a tablet device. For each subject the slider values were scaled to a score from 1 to 10 so that the minimum score was 1 and maximum 10. The loudspeakers were at standing head height, separated by 2m, and positioned 1m from the wall in a listening room meeting the requirements of IEC 60268-13 and BS 684, with dimensions 6.5m x 4.2m x 2.7m. A 3rd identical reference loudspeaker was positioned 1m behind the central point between the loudspeakers. This was used to reproduce the MONO condition. For the other conditions the target image was set to the location of the reference loudspeaker. The source signal was a repeating loop of a Cello with strings being plucked repeatedly. Approximately half of the energy was in the frequency range above 1500 Hz. The sample was selected because representative of sounds that can be localised well with good transient properties and a balance of energy in the low and high frequency regions.

The reproduction condition labels and descriptions are as follows: STEREO - uncompensated stereo reproduction, CAP-P - stereo Tangent Law panning with compensation for head position, CAP-PR - stereo with compensation for head position and rotation. Amplitude panning is applied across the whole frequency range, CAP-PRF - compensated stereo with energy panning over 1500 Hz.

14 subjects were tested. All had prior experience with audio testing, reported normal hearing, and were involved as students or professionals with audio. Fig. 15 shows box plots of the scores for each condition. They show an increase in median scores across the conditions towards the most advanced reproduction method. The statistical significance of the results was tested using paired t-tests and binomial tests[30], comparing pairs of conditions. The results are shown in Table I.

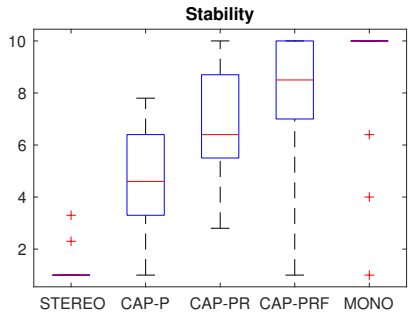

Fig. 15: Box plots of the positional stability score for the reproduction of a Cello sound sample, for 5 conditions.

The comparison of greatest interest is CAP-P / CAP-PR, since this shows the improvement made by adding compensation for rotation to the existing positional compensation system. $p < 0.05$ for both tests, strongly supporting improvement. Support for improvement from CAP-PR to CAP-PRF is marginal however. There is good evidence that MONO is preferred to CAP-PR, although some subjects preferred CAP-PR to MONO. Based on their comments this may be because the CAP-PR image has less directivity compared with the real MONO image, effecting the overall stability judgement.

TABLE I
STATISTICAL COMPARISONS BETWEEN PAIRS OF STATIC AND COMPENSATED REPRODUCTION CONDITIONS

	Paired t-test	**Binomial**
STEREO / CAP-P	p = 0.000	p = 0.007
CAP-P / CAP-PR	p = 0.014	p = 0.007
CAP-PR / CAP-PRF	p = 0.160	p = 0.090
CAP-PRF / MONO	p = 0.503	p = 0.395
CAP-PR / MONO	p = 0.140	p = 0.090

TABLE II
DISTRIBUTION OF SELECTED CATEGORIES FOR TARGET IMAGE CONDITIONS FRONT, BEHIND, UP AND DOWN

	FRONT	**BEHIND**	**UP**	**DOWN**
F	14	0	0	0
B	0	9	0	0
U	0	1	9	1
D	0	0	0	4
?	0	4	5	9
p	1	0.64	0.64	0.29
σ/n	0	0.13	0.13	0.12

The same setup was used, with CAP-PRF reproduction, to test how well the subjects could localise images away from the loudspeakers. The test source was a tuba sound, chosen to minimise conflicting cues from high frequencies, with 99% energy below 1000 Hz, and 50% of energy below 280 Hz. There were 4 conditions: FRONT - control target 2m in front of the loudspeaker centre, BEHIND - target 5m behind the centre, UP - 2m above the centre, and DOWN - 2m below the centre. The subjects were asked to categorise the image location - front (F) / behind (B) / up (U) / down (D) / unclear (?). The total number reported in each category is shown in Table II. p is the estimated probability that a subject will guess correctly, and σ/n is the estimated standard deviation of p, where σ is the standard deviation of a binomial random variable $B(n,p)$, $\sigma/n = \sqrt{p(1-p)/n}$. Achieving any responses B, U or D is significant, because the FRONT condition produces only the result F.

While it is desirable to create reliable images in all directions, studies show that the ability to localise real sources away from the frontal region varies significantly across the population, with some 30% unable to localise overhead sources[5;6]. The vertical results should be seen in this light. The listener is free to reorient their heads to face any image, however they may not do this if the initial image is far from the correct location.

The downward images were the least likely to be recognised. The reason for this is not clear, but may be because of conflicting prior expectation, or other conflicting cues. If head pitch and roll sensing is turned off then no downward images are reported. Upward images are reported part of the time, but these are lost if the head is pitched, agreeing with previous observations about W-panning.

The tests have been designed to make the listening experience natural in some ways, by allowing free movement. In other ways the experience is remains unnatural, including conflicting visual stimulus, lack of complexity in the images and scene, and lack of movement in the scene.

VIII. CONCLUSION

The ITD and ILD for a general low frequency field were calculated, knowing the orientation of the listener's head, and applied to find the image produced by a general coherent field. Compensated panning gains were then found, depending on the head orientation, that produce a stable image in a given target direction. The objective assessment using binaural head measurements, and subjective listening tests verify the initial aims of stabilising images, and furthermore show that images can be produced where it is not possible with passive panning. CAP provides stable images for wider loudspeaker separations, which means that more of the space in front of the loudspeakers is useable by the listener, which would be valuable for example in CAVE virtual reality systems[31].

CAP can be applied to broadband signals, due to the dominance of the low frequency ITD cue. This can be improved by processing the high frequencies separately using an energy based panner. For cases where the high frequency cues will conflict excessively, such as for reverse images or where the loudspeakers are widely space, then high frequency content should be reduced. Alternatively CAP could be complemented by other methods that treat high frequency content separately, such as transaural reproduction[32;33]. Unlike direct binaural methods, CAP has the benefit of being independent of the listener's head size, and does not require independent control of the binaural signals.

The spherical head model could be refined according to individual measurements, including ear position and elliptical parameters. However, the objective assessment of CAP using a measured head response shows that the compensated image error is already within the MAA in many cases, and such improvements are likely to be marginal.

CAP is being extended for reproduction with 3D arrays. This is expected to produce improved imaging using a sparse array when compared with VBAP on 3D arrays. An extension to control ILD for near source cues has been published separately, and will be followed with a more detailed study. The data from the subjective study are available via DOI: 10.5258/SOTON/D0152.

IX. ACKNOWLEDGEMENTS

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. Thanks to Daniel Shaw for assisting with the subjective experiment.

REFERENCES

[1] J. Blauert, *Spatial hearing*. Cambridge, MA: MIT Press, 1997.
[2] B. Bernfeld, "Attempts for better understanding of the directional stereophonic listening mechanism," in *Audio Engineering Society Convention 44*, no. C-4, March 1973. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=1743
[3] M. A. Gerzon, "General metatheory of auditory localisation," in *92nd Audio Engineering Society Convention, Vienna*, no. 3306, 1992.
[4] V. Pulkki, "Compensating displacement of amplitude-panned virtual sources," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment*

Audio, Jun 2002. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=11138

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

[6] ——, "The role of head movements and vestibular and visual cues in sound localization." *Journal of Experimental Psychology*, 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," in *Audio Engineering Society Convention 139*. Audio Engineering Society, 2015.

[8] D. Menzies and F. M. Fazi, "A theoretical analysis of sound localisation, with application to amplitude panning," in *Proc. AES 138th Convention, Warsaw*, May 2015.

[9] F. M. Fazi and D. Menzies, "Estimation of the stability of a virtual sound source using a microphone array," in *Proc. 22nd International Congress on Sound and Vibration (ICSV22), Florence*, July 2015.

[10] D. Menzies and F. M. Fazi, "Spatial reproduction of near sources at low frequency using adaptive panning," in *Proc. TecniAcustica, Valencia*, 2015.

[11] P. Morse and K. Ingard, *Theoretical Acoustics*. New York, NY: McGraw-Hill, 1968.

[12] R. S. Woodworth and H. Schlosberg, *Experimental psychology*. Oxford and IBH Publishing, 1954.

[13] G. F. Kuhn, "Model for the interaural time differences in the azimuthal plane," *The Journal of the Acoustical Society of America*, vol. 62, no. 1, pp. 157–167, 1977.

[14] B. Xie, *Head-related transfer function and virtual auditory display*. Plantation, FL: J Ross, 2013.

[15] F. M. Fazi, M. Noisternig, and O. Warusfel, "Representation of sound fields for audio recording and reproduction," *Acoustics 2012 Nantes*, 2012.

[16] E. Williams, *Fourier Acoustics: sound radiation and nearfield acoustical holography*. Cambridge, MA: Academic Press, 1999.

[17] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc*, vol. 55, no. 6, pp. 503–516, 2007. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=14170

[18] P. G. Craven and M. A. Gerzon, "Coincident microphone simulation covering three dimensional space and yielding various directional outputs," Aug. 16 1977, uS Patent 4,042,779.

[19] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc*, vol. 45, no. 6, pp. 456–466, 1997. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=7853

[20] G. Theile and G. Plenge, "Localization of lateral phantom sources," *J. Audio Eng. Soc*, vol. 25, no. 4, pp. 196–200, 1977. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=3376

[21] W. O. Brimijoin and M. A. Akeroyd, "The role of head movements and signal spectrum in an auditory front/back illusion," *i-Perception*, vol. 3, no. 3, pp. 179–182, 2012.

[22] D. Menzies, "W-panning and o-format, tools for object spatialisation," in *Proc. AES 22nd International Conference*, 2002.

[23] J. C. Bennett, K. Barker, and F. O. Edeko, "A new approach to the assessment of stereophonic sound system performance," *Journal of the Audio Engineering Society*, vol. 33, no. 5, pp. 314–321, 1985.

[24] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR dummy-head microphone," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.

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

[26] D. R. Perrott and K. Marlborough, "Minimum audible movement angle: marking the end points of the path traveled by a moving sound source," *The Journal of the Acoustical Society of America*, vol. 85, no. 4, pp. 1773–1775, 1989.

[27] W. O. Brimijoin and M. A. Akeroyd, "The moving minimum audible angle is smaller during self motion than during source motion," *Frontiers in neuroscience*, vol. 8, no. 273, 2014.

[28] M. Simon, D. Menzies, F. M. Fazi, T. de Campos, and A. Hilton, "A listener position adaptive stereo system for object based reproduction," in *Proc. AES 138th Convention, Warsaw*, May 2015.

[29] 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," in *Proc. Conf. Digital Audio Effects (DAFx-98)*. Citeseer, 1998, pp. 76–80.

[30] S. Bech and N. Zacharov, "Perceptual audio evaluation-theory, method and application," 2002.

[31] C. Cruz-Neira, D. J. Sandin, and T. A. DeFanti, "Surround-screen projection-based virtual reality: the design and implementation of the cave," in *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*. ACM, 1993, pp. 135–142.

[32] S. M. R. Atal Bishnu S, "Apparent sound source translator," Feb. 22 1966, uS Patent 3,236,949.

[33] O. Kirkeby, P. A. Nelson, and H. Hamada, "Virtual source imaging using the stereo dipole," in *Audio Engineering Society Convention 103*, Sep 1997.

Dylan Menzies Dr 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 PhD in electronics from the University of York, an BA in mathematics from Cambridge University, and has worked as a research engineer for several companies including Sony Professional Audio.

Marcos F. Simón Galvez Marcos Simón graduated in 2010 from the Technical University of Madrid with a BSc in telecommunications. In 2011 he joined the Institute of Sound and Vibration Research, where he has worked in personal audio with loudspeaker arrays for improving speech intelligibility in hard of hearing people and also in the modelling of cochlear mechanics. He obtained his PhD title in 2014, and is currently working on the S3A spatial audio program. He has interests in spatial audio synthesis and reproduction the and interaction between computer vision systems and audio reproduction.

Filippo Maria Fazi Filippo Maria Fazi graduated in Mechanical Engineering from the University of Brescia (Italy) in 2005. He obtained his PhD 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, multi-channel systems, virtual acoustics, microphone and loudspeaker arrays.