Audio Engineering Society

Convention Paper

Presented at the 128th Convention
2010 May 22–25 London, UK

Design of a Circular Microphone Array for Panoramic Audio Recording and Reproduction: Array Radius

Enzo De Sena[1], Hüseyin Hacıhabiboğlu[1], and Zoran Cvetković[1]

[1] *Centre for Digital Signal Processing Research (CDSPR), Division of Engineering, Kings College London, Strand, London, WC2R 2LS, United Kingdom*

Correspondence should be addressed to Enzo De Sena (enzo.desena@kcl.ac.uk)

ABSTRACT

A multichannel audio system proposed by Johnston and Lam aims at the perceptual reconstruction of the sound field of an acoustic performance in its original venue. The system employs a circular microphone array, of 31cm diameter, to capture relevant spatial cues. This design proved to be effective in the rendition of the auditory perspective, however other studies showed that there is still substantial room for improvement. This paper investigates the impact of the array diameter on the width and naturalness of the auditory images. To this end we propose a method for quantification and prediction of the perceived naturalness. Simulation results support array diameters close to that proposed by Johnston and Lam in the sense that they achieve optimal naturalness in the centre of the listening area, but also suggest that larger arrays might provide a more graceful degradation of the naturalness for listening positions away from the centre.

1. INTRODUCTION

Most multichannel systems achieve desired spatial effects through manual mixing and artificial manipulation of audio material. This requires a careful use of artificial panning and reverberation for rendering of spatial attributes like localisation and envelopment. However, this impairs the consistency of the actual and the reproduced sound fields. Some solutions such as wave field synthesis (WFS) [1] and Ambisonics [5] have been proposed to reconstruct the sound field in a physically accurate way. WFS reproduces the wave front accurately and has a wide optimal listening area, but the number of input and output channels is prohibitively high for home audio systems. Ambisonics is less demanding in terms of the number and the positions of the loudspeakers, however, it provides an accurate reproduction within a narrow listening area. Higher-order Ambisonics configurations provide a larger sweet spot, but, as for first-order Ambisonics, a careful calibration is needed, making them unfit for consumer grade reproduction systems.

A system that achieves a perfect physical reconstruction of the sound field would also deliver a perfect perceptual experience. However the real purpose of multichannel audio systems is *perceptual* fidelity rather than physical accuracy, as they are to be used by humans that have a finite-resolution auditory system. To this end, a sparse multichannel system was proposed by Johnston and Lam [8]. This system uses a circular array of microphones in the horizontal plane. The proposed diameter of the array was 31 cm and the microphones were hypercardioid. The motivation behind this design was to capture and reproduce interaural time delay (ITD) and interaural level difference (ILD) cues that a listener would experience in an actual sound field. A further study by Hall and Cvetković [2] showed that the fidelity of the auditory perspective improves as the diameter of the array is increased. A more recent study by Hacıhabiboğlu and Cvetković [6] revealed that higher order microphones facilitate more accurate rendition of the direction of monochromatic plane waves. Finding optimal microphone array radius and directivity pattern for circularly symmetric multichannel systems are still open problems that need a systematic design approach.

Every surround system that aims at the perceptual reconstruction of the sound field should deliver a convincing auditory experience, ideally coherent with the venue where the recording was made. Such a coherence relies also on the mutual consistency between the ITD and ILD cues reproduced by the system. It has been observed in the literature that natural sound source trigger ITD and ILD values that are highly correlated [4]. A higher difference between the two ear levels (due to the sound shadow of the head) usually corresponds to a higher difference in time arrival (due to the longer distance that the sound wave travels) and vice versa. Phantom images reproduced by any multichannel system should ideally preserve this same property of natural sources. In fact, it has been proven that the auditory system interprets mismatching ITD and ILD as a wider auditory image or even two separate images [4].

The study presented in this paper draws from the observation made during informal listening tests that the radius of the microphone array clearly influences the perceived width and naturalness of the phantom image. In the first part of the study, the natural correlation between ITD and ILD is analysed for single free field sources. This serves as a metric to judge the naturalness of the sound field reproduced by Johnston's system - that is the subject of the second part of this paper. The impact of the microphone array radius and directivity is analysed and conclusions are drawn on their desirable values. Our simulations confirm Johnston's initial intuition of setting the array diameter to 31cm to mimic the sound travel time around the head. An analytical solution based on a simple heuristic observation is also given to support this conclusion.

In Section 2 the setup of Johnston and Lam system is presented. Section 3 describes the auditory model employed in the simulations that follow. In Section 4 the relationship between ITD and ILD is analysed. Section 5 presents the method used to quantify the degree of naturalness of Johnston's system and in Section 6 its results are discussed. Section 7 concludes the paper.

2. JOHNSTON/LAM MULTICHANNEL SYSTEM AND RELATED WORK

In the original design proposed in [8], the recording system is composed of five directional microphones evenly distributed on the horizontal plane on a circle of radius $r_m = 15.5$ cm. Two additional superdirectional microphones aimed vertically up and down are added to improve ambience capturing. The horizontal microphones are hypercardioid[1] whereas the vertical ones are shotgun. Each of the horizontal microphones drives the matching loudspeaker and the two remaining channels are appropriately mixed and played back by all the loudspeakers.

[1] Following the publication of [8], a US patent [9] has been filed by the same author. The proposed microphone had a directivity with the primary lobe down by 3dB at 72° and down to effectively zero at 144°. In the present work this directivity is used - rather than hypercardioid - as it is the most recent specification provided by the author/inventor.

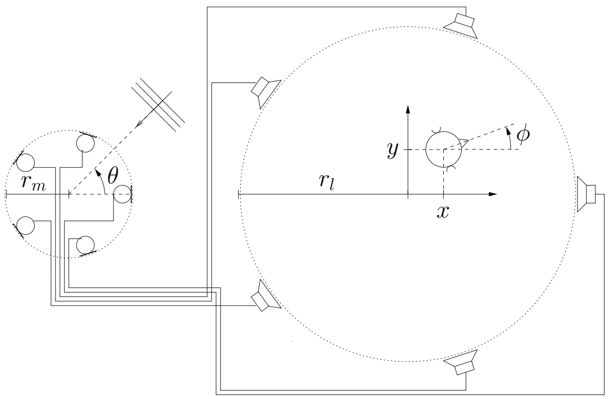

Fig. 1: Horizontal part of Johnston's recording and reproduction system.

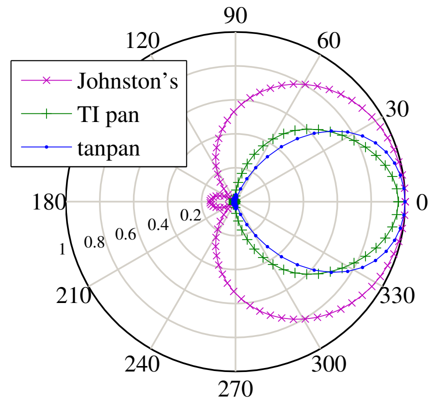

Fig. 2: Comparison between the different directivity patterns analysed in this paper.

The present study is focused on the horizontal part of the system, that is depicted in Figure 1. The recording system has two design variables: the radius of the array of microphones r_m, and their directivity patterns. If the sound source is far enough from the recording array, these two characteristics independently govern the inter-channel delays and the inter-channel levels, that together form the auditory perspective.

It is of interest to investigate how differences in the microphone directivity influence the naturalness of the sound field. To this end, in addition to Johnston's initial proposal, the directivity patterns presented in [6] and in the companion paper [7] are included.

The directivity function introduced in [6] was designed as to have at most two loudspeakers active for any single source direction θ, and to emulate stereophonic tangent panning law between the two active loudspeakers. Objective results in the form of active intensity were also presented, showing that the proposed method provided good directional reproduction of monochromatic plane waves. In the following sections it will be denoted as "*tanpan*" directivity.

In the companion paper [7] a new approach to the microphone directivity design is established within the framework of time-intensity stereophony. The directivity is shaped accordingly to the well known time-intensity panning curves given by Franssen. The design methodology requires the array diameter as an input. In this paper the directivity function obtained for $r_m = 15$ cm is included and will be denoted as "*TI pan*" directivity.

The three directivity functions described above are shown in Figure 2.

3. EMPLOYED AUDITORY MODEL

To date, no analytical model exists for the description of ITD and ILD cues in complex listening situations. As a consequence, research has to be carried out by means of simulations or subjective experiments. In [4] a simple auditory model has been developed to analyse the relationship between ITD and ILD. The auditory model employed in the present paper follows its guidelines.

The spherical head model proposed in [3] is employed to simulate source to ear transfer functions. This choice helps to avoid the problem of HRTF interpolation and provides a good approximation of the diffraction around the head. Where not explicitly indicated in this paper, the head radius is set to 9 cm and the ears are assumed to be placed 180° away from one another. As a first step, the inverse FFTs of the HRTFs are computed to obtain the HRIRs for left and right ear. These impulse responses are then convolved with the source signal in order to obtain the eardrums signals at left and right ear, i.e. $s_L(n)$

and $s_R(n)$, respectively. The signals are then split - via the gammatone filter bank provided in [10] - into 24 critical bands covering the frequency range between 20 Hz and 15.5 kHz. The bandpass signals of the i^{th} critical band will be denoted as $s_{Li}(n)$ and $s_{Ri}(n)$, for the left and right ear, respectively.

The interaural level difference relative to the i^{th} critical band, ΔL_i, is computed as the logarithmic ratio of signal energy in the right and left channel:

$$\Delta L_i = 10 \log_{10} \frac{\sum_{n=0}^{N} s_{Ri}^2(n)}{\sum_{n=0}^{N} s_{Li}^2(n)}. \qquad (1)$$

To obtain the ITDs relative to each critical band, a cross-correlation analysis is carried out. The bandpass signals $s_{Li}(n)$ and $s_{Ri}(n)$ are half-wave rectified and the interaural time difference, τ_i, is taken as the delay at which the cross-correlation achieves its maximum. At low frequencies, the interaural time difference is directly detectable by the auditory system. As the frequency increases, this capability disappears, however - for non-stationary signals - interaural shifts of the envelope are still perceivable. Accordingly, for the critical bands whose central frequencies are above 1.5 kHz, the same cross-correlation algorithm is performed on the envelopes. The envelopes are obtained by means of the discrete Hilbert transform.

The quasi-periodic nature of the cross-correlation function can lead to some errors when the highest peak is used as the only detection rule. In order to correct such errors, in [4] the linear regression of the unwrapped phase of the interaural transfer function is computed, and its angular coefficient is compared to the different local maxima of the cross-correlation function; the location of the peak closest to the normalised angular coefficient is then chosen as τ_i. In the present study these errors have been observed in less than 1% of the cases and only in one critical band at a time, therefore a simpler and faster approach can be followed: after all the τ_i are computed, the errors are corrected by choosing the local maximum that is closer to the average of all the other critical bands values.

4. RELATIONSHIP BETWEEN ITD AND ILD

The first studies that analysed the relationship between ITD and ILD date back to the 1920's when the "trading experiments" were proposed to measure their relative importance: a stimulus with a given time (or level) difference was presented through headphones and the user task was to adjust the opposite level (or time) difference that is needed to return the auditory event to the median plane. In the late 1960's doubts were raised to the validity of this method. It was observed that with slightly mismatching ITD and ILD the auditory image became spread, and, as the mismatch increased, multiple auditory events were detected. As a result, all the earlier studies carried out with trading procedures had to be reviewed because of the uncertainty as to whether the subject was describing the position of one among the multiple images - and which one - or even of a "spatial average" of the different auditory events.

The interaural level difference is mainly due to the diffraction of the sound wave around the listener's head while the interaural time difference is the result of the physical distance of the two ears. These cues are the most relevant in rendering the spatial perspective and are highly correlated. For example, when a natural sound source is directly in front of the user, the ear signals are almost identical and therefore ITD and ILD are both approximately zero. As an opposite case, when the sound source lies exactly to the right (or to the left) of the listener, ITD and ILD are both at their maximum. More in general, when the shadowing effect is strong, it is highly probable that the difference of time arrival at the two ears will be high as well, and vice versa.

This "natural correlation" has been studied in detail in [4] by means of a catalogue of outer ear transfer functions. These were recorded with subminiature microphones placed in the ear canals of three different subjects. Different incidence directions were covered with elevations ranging from $-10° < \delta < 90°$ and azimuths $0° < \theta < 360°$, for a total of 122 directions. For each one of them, $\Delta L_i(\delta, \phi)$ and $\tau_i(\delta, \phi)$ were calculated with an auditory model very similar to the one presented in Section 3 - apart from the HRTF employed. For each critical band, the $(\tau_i, \Delta L_i)$ pairs were plotted in the ITD-ILD plane and it was observed that they mapped into narrow "ribbons". This proved that ITD and ILD are actually highly correlated. A set of functions that associate to a given τ_i (the independent variable) its

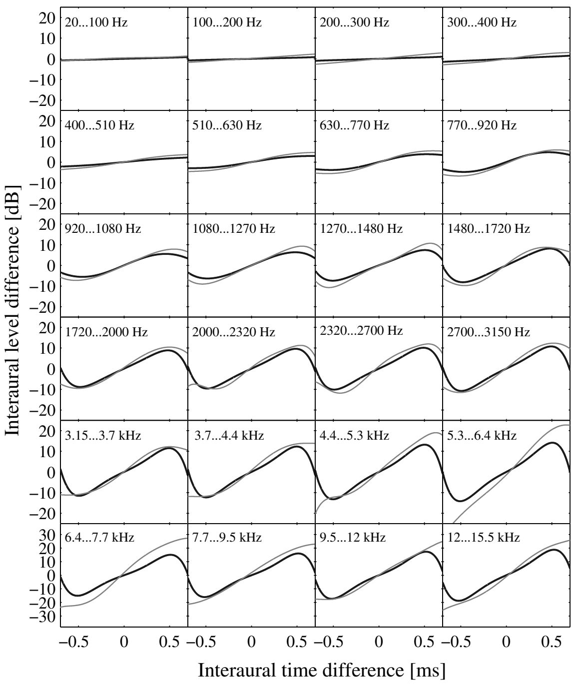

Fig. 3: Comparison between the naturalness functions F_i obtained with our auditory model (black lines), with the ones given by [4] (grey lines) in 24 critical bands.

natural ΔL_i (the dependent variable) were obtained through polynomial approximation of the data. The resulting functions F_i are plotted in figure 3 (grey lines). It can be observed that the function slopes increase with increasing frequency and this is simply due to the frequency-dependency of the shadowing effect. Furthermore the curves appear not exactly symmetrical with respect to the y-axis and this is due to the physical asymmetries of the three subjects and probably to accidental movements occurred during the recording procedure.

In order to support the validity of these naturalness functions on psychophysical grounds, in [4] a listening test had also been carried out. Narrow-band noise signals were presented through headphones with assigned level and time differences. The task of the user was twofold: report split images, and rate the degree of naturalness on a scale between 0 (the auditory event is easy to localise and concentrated on a narrow region) and 10 (hard to localise, very broad or even split up in multiple components). In both the tasks, the user responses were actually in agreement with the simulated results: as the distance between the curves F_i and the supplied $(\tau_i, \Delta L_i)$ increased, the degree of naturalness decreased and more split images were reported. To give a relative reference for the results that are going to be presented in Section 6, it may be useful to report one of the outcome of this subjective experiment: a bandpass noise signal in the critical band $400 - 510Hz$ is presented to the listeners with a $\tau = -600\mu s$ whose natural ILD is $F_5\left(-600\mu s\right) \cong -3dB$. In 16 out of 20 cases, a split image was reported for a $\Delta L = 12dB$, and only 1 out of 20 for a $\Delta L = 0dB$.

The present work focuses on the analysis of natural sources lying on the horizontal plane only. However, in [4] several elevations were taken into account without the possibility to discern the results relative to the elevation $\delta = 0$. Therefore, as part of this study, new naturalness functions F_i are obtained by means of our auditory model. This also ensures the coherence with the simulations run for Johnston's system that are going to be discussed in Section 5. The procedure is followed for 360 directions evenly distributed on the horizontal plane and polynomial approximation of the resulting data is shown in Figure 3 (black lines). These new curves closely match the ones given by [4], thus validating our auditory model for the purpose of this study.

5. METHOD

Every surround system should deliver a coherent and convincing experience. Very broad images or split events should never be produced. As already discussed, the microphone array radius r_m govern the inter-channel delays that in turn have an impact on both the ILD and the ITD. As a consequence r_m heavily influences the naturalness of the reproduced sound field. The main purpose of this paper is to investigate this impact. To this end, the following steps are followed: i) the five microphone signals are obtained for a sound source in a given direction θ and for a given microphone array radius r_m; ii) the reproduction of the system is simulated for a user placed in (x, y) with heading ϕ - refer to figure 1 - and the auditory model described in Section 3 is used to obtain the $(\tau_i, \Delta L_i)$; iii) the $(\tau_i, \Delta L_i)$ pairs are compared with the naturalness curves F_i in each critical band and the final naturalness error ϵ is obtained as:

$$\epsilon \triangleq \frac{1}{M} \sum_{i=1}^{M} \left| \Delta L_i - F_i \left(\tau_i \right) \right| \qquad (2)$$

where M is the number of critical bands. In other words, we define the naturalness error as the average distance between the measured ILD, ΔL_i, and the natural ILD that is associated to the measured ITD, $F_i(\tau_i)$.

In order for this approach to be consistent in all the critical bands, the auditory system has to be capable in detecting shifts of signal envelopes. Therefore the stimuli should be non-stationary and include sharp edges. Accordingly, the chosen source stimuli are impulses, but other non-stationary signals could be used as well.

It is not known how the different critical bands affect the perception of naturalness. Therefore in equation (2), the conservative choice of weighting equally all the critical bands has been made.

In all the simulations, the loudspeaker array radius is $r_l = 2m$ to be consistent with the setup we used in our companion paper [7].

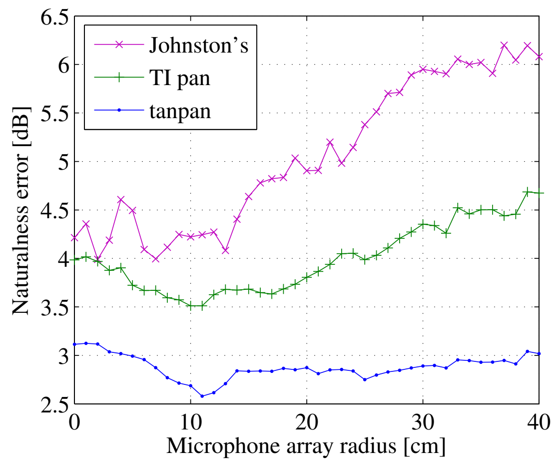

Fig. 4: Naturalness error as a function of the microphone array radius r_m. The user position is $x = 0, y = 0$ and heading $\phi = 0$.

6. RESULTS

6.1. Listener in central position

The system performance is initially analysed for a user placed in the middle of the listening area and facing one of the loudspeakers. The multichannel system considered in this paper aims at a good panoramic rendition and the listener should perceive as natural the phantom images relative to every source direction θ. Therefore the naturalness error ϵ is averaged over 60 directions evenly distributed on the horizontal plane - i.e. $\theta = 0°, 6°, ..., 354°$. The three directivity functions described in Section 2 are tested. Figure 4 shows the result as a function of the microphone array radius. A first remark that can be made is that $\epsilon(r_m)$ is noticeably higher with the directivity proposed by Johnston in the original design. Moreover, *tanpan* directivity appears to deliver the best performance. This result can be justified by considering the width of the different directivity patterns in Figure 2: a wider frontal lobe causes significant levels in three adjacent loudspeaker for any given direction, and this leads to a less "natural" reconstruction. The conclusion that can be made is that narrower directivity functions deliver more natural auditory images. However this result should be taken carefully. In fact it is emphasised here that the reproduced sound field can be highly natural but poorly localised. Consider for example the extreme case of an impulsive directivity. Only the sources that lie in front of one the microphones - i.e. $\theta = 0°, 72°, ..., 288°$ - would be rendered by the system, and only one loudspeaker would be active for each of these source directions. The resulting sound field would be highly natural - one (or no) loudspeaker alone is a natural source - even though the system renders only sources in five exact directions and therefore has no localisation capability at all. So, it is clear that the comparison of different directivity functions should not be carried out with the naturalness as the only metric.

A second and more interesting remark is that $\epsilon(r_m)$ is minimum around 10 cm for all the directivity patterns under study. The question that arises immediately is: is it only by chance that the optimal array radius is a value close to the radius of the head?

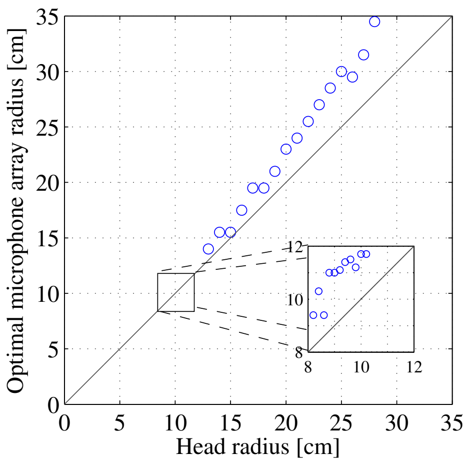

Fig. 5: Optimal microphone array radius in terms of naturalness as a function of listener's head radius.

To answer this question, the same simulation setup is run with different head radii. The result of this study is shown in Figure 5 for *tanpan* directivity, where it is clearly shown that the optimal array radius is approximately a linear function of the head radius and therefore they are highly correlated.

The nature of this correlation is rather complex, but can be explained in a simplified scenario. Consider the case when the user is facing one of the loudspeakers, $\phi = 0$, and the sound source lies in the same direction, $\theta = 0$. Only the three loudspeakers in front of the user will be active, in fact all the directivity functions are approximately null at 144°. Due to the symmetry of this particular scenario, the ear signals are equal, so let us consider the left ear only. It will receive three signals, but the one from the right loudspeaker will be highly attenuated due to the head shadowing. In an ideal situation, only the central loudspeaker should be active: the ITD/ILD combination would be natural (a single loudspeaker is equivalent to a free field source), and the auditory image would lie exactly in front of the user, as the original sound source is. On the other hand, not only two loudspeakers are active, but also the left loudspeaker is slightly closer than the central to the left ear, and its signal arrives earlier. In fact the travel time from the left and central loudspeakers to the left ear are respectively (refer to Figure 6):

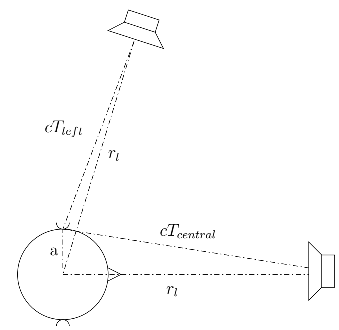

Fig. 6: Distances between left and central loudspeakers to the left ear.

$$T_{central} = \frac{1}{c}\sqrt{r_l^2 + a^2},$$
$$T_{left} = \frac{1}{c}\sqrt{r_l^2 + a^2 - 2r_l a \sin(2\pi/N)},$$

where r_l is the radius of the loudspeaker array, a is the radius of the head, c is the speed of sound and N is the number of microphones/loudspeakers in the system. In all practical situations $r_l \gg a$, so a first order approximation can be used:

$$T_{central} \cong \frac{1}{c}\left(r_l + \frac{a^2}{2r_l}\right), \tag{3}$$

$$T_{left} \cong \frac{1}{c}\left(r_l + \frac{a^2}{2r_l} - a\sin(2\pi/N)\right), \tag{4}$$

therefore the left loudspeaker signal arrives earlier than the central one by $\frac{a}{c}\sin(2\pi/N)$ seconds. Since the auditory image is supposed to be in the central direction, a heuristic condition that should be satisfied is that the signal due to the left loudspeaker arrives *at least* after the one due to the central loudspeaker. This can be achieved by delaying the left loudspeaker signal by a quantity d that satisfies the condition:

$$d \geq \frac{a}{c}\sin(2\pi/N). \tag{5}$$

For a sound source in the far field, the distance between the central and left microphone is such that the signal of the left microphone is delayed by a quantity:

$$d = \frac{r_m}{c}\left(1 - \cos(2\pi/N)\right) \tag{6}$$

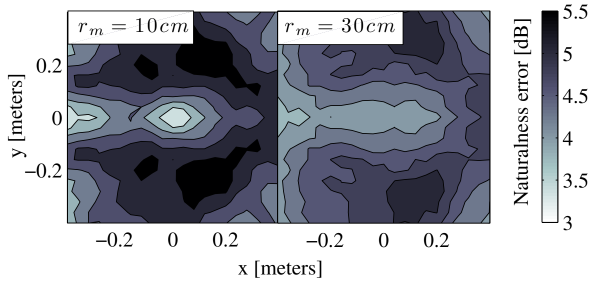

Fig. 7: Average naturalness error as a function of user position for two different microphone array radii.

so, the condition (5) translates into a minimum value for the microphone array radius r_m that is readily obtained as:

$$r_m \geq a \frac{\sin(2\pi/N)}{1-\cos(2\pi/N)} = 1.38a \tag{7}$$

It should now be observed in Figure 4 that, for $N = 5$, the naturalness of the sound field worsens as r_m increases. This is due to an increase in ITD magnitude that is not "followed" by an increase in ILD. Therefore, among the values that satisfy (7), the smallest will achieve the best natural fit between ITD and ILD, i.e. $r_m = 1.38a$. This simple formula confirms the linear relation between optimal r_m and a that has been observed in Figure 5 and ultimately supports Johnston's intuition of choosing the array diameter to mimic the sound propagation time around the head.

6.2. Listener off-centre

The results presented so far were valid for a listener placed at the centre of the loudspeaker array. The system should also provide a large and stable sweet spot, therefore it is of interest to study what happens when the user is slightly off-centre. As the relative distances and angles between ears and loudspeakers change, it is expected that the perceived naturalness will change as well.

The same simulation procedure is run on a grid of 400 points covering an area of 80cm x 80cm around the centre with the listener facing $\phi = 0$. Figure 7 shows the naturalness error $\epsilon(x, y)$ averaged over 60 source directions $\theta = 0°, 6°, ..., 354°$ for two different array radii r_m = 10 cm and r_m = 30 cm. The directivity function used in this example is *TI pan*, however the same patterns are observed for the other directivity functions. A first predictable result is that as the listener moves away from central position the sound field is perceived as less natural. More interestingly, as the radius increases, the naturalness error becomes more uniform. This suggests that an array radius that is optimal in the central position cause the performance to degrade more rapidly as the listener moves, whereas higher radii deliver a non-optimal but more stable sweet spot.

7. CONCLUSIONS

In this paper an analysis of the degree of naturalness for Johnston surround system has been presented. This investigation was motivated by the observation made during informal listening tests that the diameter of the microphone array clearly influences the width and naturalness of the reproduced phantom images. A method for the quantification of the naturalness has been introduced. The rationale behind this method is that natural sound source trigger ITD and ILD cues that are coherent and mutually consistent and that the reproduced phantom images should ideally preserve this property. Two main conclusions can be made: the optimal microphone array diameter is found to be close to the diameter of the human head. However, this result is only valid for a listener close to the centre of the loudspeaker array and it has been observed that a larger array diameter provides a more graceful degradation of the naturalness as the listener moves away from the centre.

ACKNOWLEDGMENTS

The authors are funded by the Engineering and Physical Sciences Research Council (EPSRC) Research Grant, EP/F001142/1, "Perceptual Sound Field Reconstruction and Coherent Emulation".

8. REFERENCES

[1] M.M. Boone, E.N.G. Verheijen, and P.F. Van Tol. Spatial sound-field reproduction by wave-field synthesis. *J. Audio Eng. Soc.*, 43(12):1003–1012, 1995.

[2] Z. Cvetkovic and J. Hall. Coherent Multichannel Emulation of Acoustic Spaces. In *Proc. AES 28th Conference*, July 2006.

[3] R.O. Duda and W.L. Martens. Range dependence of the response of a spherical head model. *J. Acoust. Soc. Am.*, 104:3048, 1998.

[4] W. Gaik. Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling. *J. Acoust. Soc. Am.*, 94:98, 1993.

[5] M.A. Gerzon. Ambisonics in multichannel broadcasting and video. *J. Audio Eng. Soc.*, 33(11):859–871, 1985.

[6] H. Hacıhabiboglu and Z. Cvetkovic. Panoramic recording and reproduction of multichannel audio using a circular microphone array. In *IEEE WASPAA*, 2009.

[7] H. Hacıhabiboglu, E. De Sena, and Z. Cvetkovic. Design of a circular microphone array for panoramic audio recording and reproduction: Microphone directivity. Accepted for presentation at AES 128 Convention, London, UK, May 2010.

[8] J.D. Johnston and Y.H. Lam. Perceptual soundfield reconstruction. AES 109 Convention - Preprint # 5202, September 2000.

[9] J.D. Johnston and E.R. Wagner. Microphone array for preserving sound field perceptual cues, patent number 6845163, 2005.

[10] M. Slaney. An efficient implementation of the Patterson-Holdsworth auditory filter bank. *Apple Computer, Perception Group, Tech. Rep*, 1993.