# PERCEPTUAL SPATIAL-AUDIO CODING

*Craig Jin, André van Schaik*

University of Sydney
School of Electrical and Information Engineering
Bldg. J03
Sydney, NSW 2006
Australia
{craig,andre}@ee.usyd.edu.au

*Virginia Best, Simon Carlile*

University of Sydney
Department of Physiology
Bldg. F13
Sydney, NSW 2006
Australia
{ginbest,simonc}@physiol.usyd.edu.au

## ABSTRACT

A novel technique for the perceptual coding of spatial audio is presented. This coding technique allows individualized 3D audio presentation and exploits the dichotomous roles of the low-frequency interaural timing and level difference cues versus the high-frequency spectral cues in human sound localization. The high-frequency spectral cues are modified to match the acoustics of the listener's outer ears, while preserving the original low-frequency interaural cues. The psychoacoustic principles and theory behind the coding scheme are described and sound localization data are shown demonstrating the fidelity of the coding technique. Based on the coding technique, we develop the notion of directional frequency bands and give some basic requirements for a 3D audio recording and reproduction system.

## 1. INTRODUCTION

Perceptual audio coding techniques primarily exploit the tolerance of the human auditory system with respect to such characteristics as frequency sensitivity, frequency selectivity, loudness sensation, and masking effects in order to achieve further compression than can be achieved with source coding techniques alone. With the recent advances in multi-channel audio formats, the role of perceptual audio coding in multi-channel audio has not yet been made clear. Traditionally, perceptual audio coding techniques have focused on masking effects, but the issue of masking starts to become ambiguous with multiple channels. For example, how does one account for across channel masking? We suggest that it is not really multi-channel audio that should be the issue per se, but spatial audio because the arguable objective of multi-channel audio systems is the generation of a spatial-audio percept. In other words, the listeners should have a sense of presence and be able to localize sounds. In this case, it is reasonable to expect that the psychoacoustics of human spatial hearing will dictate which factors are important for the perceptual audio coding of multi-channel spatial audio. These factors are reviewed below.

The single most difficult issue with 3D audio coding is that human spatial hearing is adapted to a listener's outer ears and that there exist acoustically and perceptually significant differences between different listener's outer ears [1]-[2]. The fact that sound passes through the outer ears to reach the cochlea is the primary acoustic cue informing the brain that the sound is external to the human body. Despite Darwin's suggestion that human outer ears are vestigial organs, evolution has chosen a specific and convoluted shape for the outer ears that acoustically filters the incoming sound leaving directional cues to the location of the sound source. These cues are generally classified as interaural time difference cues, interaural level difference cues, and outer ear spectral cues.

Copious research investigating human sound localization has shown that the interaural time and level difference cues determine the perceived lateral angle of the sound source, i.e., its angle with respect to the midsagittal plane [2]. The set of directions with the same lateral angle roughly defines a cone centered on the interaural axis and is referred to as a "cone of confusion."[1] In order to resolve the ambiguity of the directions within a cone of confusion, the auditory localization system uses the high frequency spectral cues which vary within a cone of confusion.

The acoustic filtering of a listener's outer ears can be measured in the laboratory and are referred to as head-related transfer functions (HRTFs) [1]. The HRTFs describe the gain and attenuation of the outer ear filter as a function of frequency and direction in space. For the purposes of simplifying the measuring process, the HRTFs are often decomposed into two components: the directional transfer function (DTF) and the direction-independent transfer function [3]. The DTFs describe the directional properties of the HRTF and can be used to generate virtual auditory space (VAS), which refers to the simulation of spatial hearing using earphones or a speaker system.

The most fundamental premise behind the 3D audio coding technique described here is the assumption that human spatial hearing is robust to the relatively small variations in the low-frequency interaural cues across different listeners, while being sensitive, on the other hand, to individual differences in the high-frequency spectral cues. The primary focus of the work described here is two-fold: (i) to empirically validate the above viewpoint; (ii) to demonstrate how this viewpoint can be used in terms of 3D audio coding.

## 2. METHODS

### 2.1. HRTF MEASUREMENTS

HRTFs were recorded for three human subjects. The HRTFs were recorded in an anechoic chamber using Sennheiser electret microphones embedded a few millimeters within each ear canal using the blocked-ear recording technique. A semi-spherical robotic arm was used to move a loudspeaker to different source

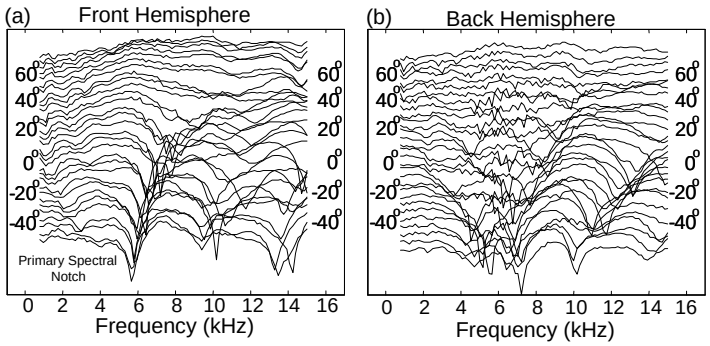


Figure. 1. *A waterfall plot of the DTF magnitude spectra for one subject's left ear is shown for different elevations at (a) -45° and (b) -135° of azimuthal angle . The elevation angle is labeled at the sides of the plot.*

positions within the anechoic chamber. Test stimuli consisted of Golay codes [4] of 1024 pulses in length at an 80 kHz sample rate (magnitude and phase) for 393 directions evenly distributed around the sphere, although not below 45° of the audio-visual horizon due to mechanical constraints during recording.

DTFs were derived from the HRTFs by deconvolving the HRTFs with the average log-magnitude spectra across all directions [3]. The calculation of DTFs may remove some interaural differences that result from the natural asymmetries between the left and right ear and it also removes, of course, the ear canal resonance [5]. In the VAS simulations described below, these acoustic characteristics were not reproduced. Nonetheless, we routinely find that control VAS sound localization of broadband noise is accurate using this method, as will be shown below for the three test subjects (see also [5]).

With respect to the spatial-audio coding technique, the dividing line between the low-frequency interaural cues and the high-frequency spectral cues needs to be clarified. To this end, consider Figure. 1 which shows waterfall plots of the magnitude spectra of the DTFs for the left ear of one subject at approximately -45° and -135° of azimuthal angle with respect to the midsagittal plane (negative angles indicate the left hemisphere of space). The different spectra in the plot correspond to different elevations. What is shown is that the prominent spectral features in the DTFs begin around 5 kHz. Note, for example, that in the front hemisphere of space the center frequency of the primary spectral notch starts at about 5 kHz for low elevations and increases with increasing elevation (Figure. 1a). The center-frequency of the notch is related to the size of the conchal cavity of the outer ear and it appears that a larger conchal cavity corresponds with a lower center-frequency [6]. The DTFs shown in Fig. 1 correspond to a subject with relatively large ears and conchal cavity. Thus, one would expect that the prominent spectral features in DTFs for most listeners would also begin around 5 kHz or higher.

### 2.2. SPATIAL AUDIO CODING

A dummy-head recording using an acoustic mannequin is the simplest method for recording spatial audio. The only problem is that the recording is not individualized. However, we suggest that it may be possible to individualize such recordings by providing appropriate side information together with the dummy-head recording. The side information would have to indicate the and the recordings were averaged over 8 repetitions. For each subject, the HRTFs consist of a complex spectrum direction(s) from which the energy in a given frequency band originates for any given time window. How this may be achieved is not the issue for this paper. Here, we simply posit that such side information can be determined and provided. Furthermore, we assume the directional information is only required for frequency bands above 5 kHz. In addition to the directional frequency band information, it is also required that the DTFs of both the dummy head and listener are known.

Assuming that such information, as described above, is available, the individualized presentation of a dummy-head recorded 3D audio can be achieved by applying gain equalization to the frequencies above 5 kHz to correct for the difference between the listener's and the dummy-head's high-frequency spectral cues. The gain equalization can be achieved using an analysis/synthesis filter bank common to most perceptual audio coders. For this work, the popular modified discrete cosine transform (MDCT) filter bank was used. The MDCT filter bank is a time-domain aliasing cancellation (TDAC) filter bank (a.k.a. a linear orthogonal lapped transform) [7]. The definition for the MDCT is

$$Y(m) = \sum_{k=0}^{n-1} h(k)\,y(k)\cos\left(\frac{p}{2n}\left(2k+1+\frac{n}{2}\right)(2m+1)\right), \quad \text{for } m = 0 \cdots \frac{n}{2}-1 \tag{1}$$

and the definition for the inverse MDCT (IMDCT) is

$$y(k) = \frac{4h(k)}{n}\sum_{m=0}^{\frac{n}{2}-1} Y(m)\cos\left(\frac{p}{2n}\left(2k+1+\frac{n}{2}\right)(2m+1)\right), \quad \text{for } k = 0 \cdots n-1 \tag{2}$$

In order to achieve perfect reconstruction, the window function, $h(k)$, must satisfy the properties:

$$h(n-1-k) = h(k)$$
$$h^2(k) + h^2\left(k+\frac{n}{2}\right) = 1 \tag{3}$$

In this work, the sine window, $h = \sin\left(\frac{kp}{n}\right)$, was used.

The gain equalization applied in this work assumes only a single sound source and was achieved for each time window using the MDCT filter bank as follows. The sound was windowed and an MDCT of length 1024 applied. For each MDCT frequency band, the gain ratio between the DTF magnitude spectrum for the listener and the dummy head was calculated and multiplied with the corresponding MDCT coefficient. The IMDCT was then applied and the usual 50% overlap-and-add algorithm followed. All calculations were made at a sample rate of 48 kHz.

The gain equalization algorithm, described above, can be generalized for a sound field with any number of sources. In this case, a given directional frequency band may appear to originate from multiple directions and a weighted average of the gain ratios for these multiple directions should be calculated using the energy level for each direction as the weighting factor. As this paper focuses on validating the basic assumptions behind the gain equalization technique, the audio coding of multiple sources is not considered further.

### 2.3. AUDITORY LOCALIZATION MEASUREMENTS

The fidelity of simulated VAS can be assessed by examining the accuracy of human sound localization performance. In this work, the localization accuracy of three human subjects was measured using two different types of sound stimuli and three sound conditions. The sound stimuli were: (1) a train of 10 noise bursts (each 40 ms on and 40 ms off); and (2) monosyllable words chosen from phonetically balanced word lists comprising the Harvard speech corpus (see [8] for a detailed description of the speech corpus). These two types of stimuli, i.e., broadband transient sounds and vocalizations, were chosen because they are well-localized by humans. The three sound conditions were: (1) "own ear," using the listener's true DTF filter functions; (2) "different ear," using another listener's DTF filter functions; and (3) "decoded ear," applying the gain-adjusted spatial-audio coding to the stimuli of condition (2). A fourth listener's DTF filter functions were used for the "different ear" sound condition for all three test subjects. For the purposes of sound condition (3), it was implicitly assumed that the stimuli corresponding to the fourth listener were the "dummy-head recordings."

The sound localization experiments were carried out in a darkened anechoic chamber with the subjects standing on a platform in the center of the room. The sound stimuli were presented over in-ear tube-phones (ER2-Etymotic Research). Subjects were instructed to turn and point their nose toward the perceived direction of the sound. The orientation of their head was monitored using an electromagnetic head tracking system (Polhemus FASTRAK). See [9] for a detailed description of the localization paradigm).

## 3. RESULTS

The sound localization data for each type of stimuli and sound condition were combined across all three subjects and then analyzed in terms of the lateral-polar angle coordinate system. This coordinate system provides a simple parameterization of

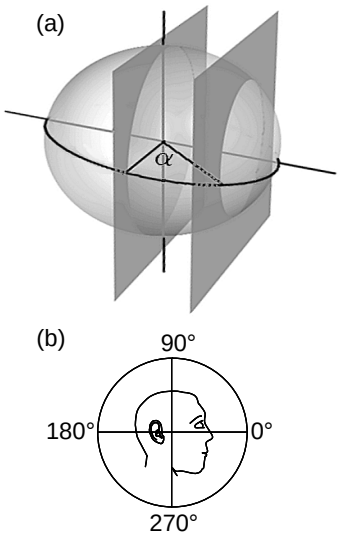


Figure 2. *(a) The lateral angle a from the midline defines a plane which cuts a vertical circle on the co-ordinate sphere. (b) The polar angle is the angle on this circle where 0° is in front and 180° is at the back.*

locations within a cone of confusion. The lateral angle, $\alpha$ ($-90^o < \alpha < 90^o$), is the horizontal angle away from the midline, where negative and positive lateral angles correspond to the left and right hemispheres of space, respectively (Fig. 2a). The polar angle ß ($0 < ß < 360^o$) is the angle on the circle defined by $\alpha$. A polar angle of 0º describes the front-most location on this circle and increases to 90º at the top, 180º at the rear aspect, and 360º again at the front (Fig. 2b).

The correspondence between the subjects' response directions' lateral and polar angles with the target directions' lateral and polar angles are shown in Fig. 3. The lateral angle was accurately perceived across all sound conditions for both the noise burst and speech stimuli (Fig. 3a,b). The polar angle data show that the perceived polar angle was significantly more accurate in the "own ear" and "decoded ear" sound conditions compared with the "different ear" sound condition. The mean polar angle error for the "own ear" sound condition is 26° and 41° for the noise train and speech stimuli, respectively. The corresponding values for the "decoded ear" sound condition are 25° and 44° compared with 56° and 65° for the "different ear" sound condition. ANOVA on the polar angle error showed that the differences between the "decoded ear" and "different ear" sound conditions are statistically significant ($p<0.005$), while those between the "own ear" and "decoded ear" are not ($p>0.2$). Localization performance was also assessed statistically using the spherical correlation coefficient (SCC) [10], which provides a measure of the correspondence between the perceived and actual directions (1 = perfect correlation, 0 = no correlation). Fig. 4a shows again that the correlation was significantly less in the "different ear" sound condition compared with the other two conditions. Similarly, the percentage of cone of confusion errors with a polar angle error greater than 90º showed the same trend (Fig. 4b).

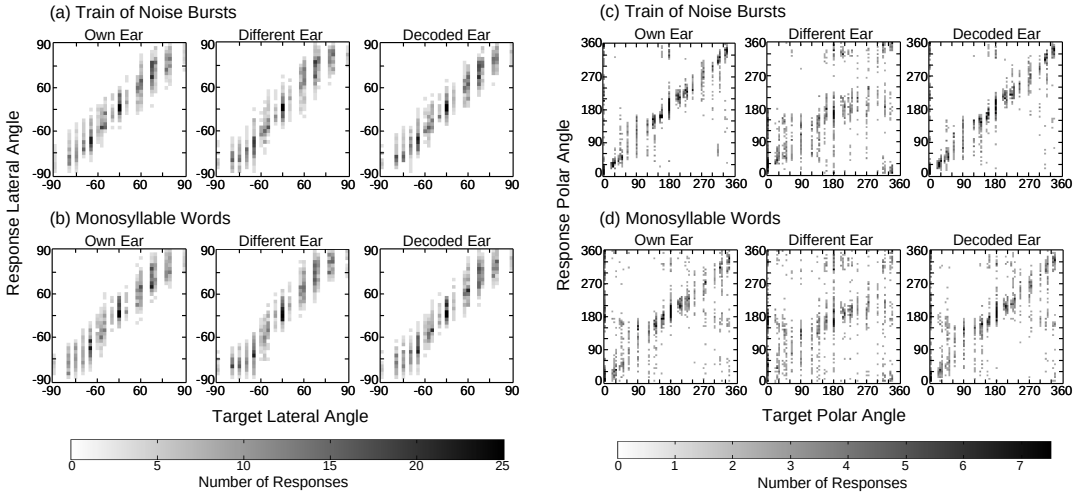


Figure 3. *Correspondence between response lateral and polar angles with target lateral and polar angles is shown in scatter plots (data pooled across all three subjects) for all three sound conditions and stimulus types. The angles were binned in 5º steps and the gray-scale value of the data is indicative of the number of responses that clustered to the given direction.*

## 4. DISCUSSION

The sound localization data shown here confirms the basic premise that human sound localization performance is robust to inter-subject differences in the low-frequency interaural time and level difference cues, but sensitive to the differences in the high-frequency spectral cues. This implies that spatial-audio coding techniques must adapt the high-frequency spectral cues of the sound for each listener. For a sound field with only one source, we have shown that appropriate gain equalization of dummy-head recordings can result in high-fidelity 3D audio. In future work, we plan to investigate generalizations (e.g., as described previously) of the gain equalization technique to more complex sound fields with multiple sources. Further investigation is also required to determine the appropriate frequency band resolution to be used during gain equalization. For this work, a 1024-point MDCT was used, but it is possible that fewer points will work as well.

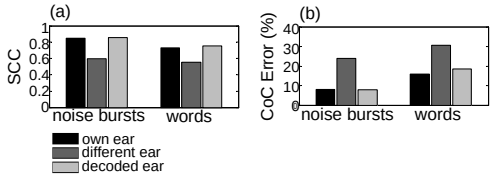


Figure 4. *(a) The spherical correlation coefficient and (b) the percentage of cone of confusion errors with a polar angle error > 90º is plotted for all three sound conditions and both types of sound stimuli.*

## 5. ACKNOWLEDGEMENT

The work described in this paper is patent pending and we would like to thank VAST Audio Pty. Ltd. for permission to publish this work.

## 6. REFERENCES

[1] S. Carlile, *Virtual Auditory Space*. Landes, Austin, 1996.

[2] J. Blauert, *Spatial Hearing. The Psychophysics of Human Sound Localization.* MIT Press, Cambridge, 1983.

[3] J.C. Middlebrooks and D.M. Green, "Directional dependence of interaural envelope delays," *J. Acoust. Soc. Am.*, vol. 87, no. 5, pp. 2149–2162, 1990.

[4] B. Zhou, D.M. Green, and J.C. Middlebrooks, "Characterization of external ear impulse responses using Golay codes," *J. Acoust. Soc. Am.*, vol. 92, pp. 1169-1171, 1992.

[5] S. Carlile, C. Jin, and V. Harvey, "The generation and validation of high fidelity virtual auditory space," In *Proc. 20th Annual Int. Conf. IEEE Engineering in Medicine and Biology Society*, pp. 1090-1095, 1998.

[6] C. Jin, P. Leong, J. Leung, A. Corderoy, and S. Carlile, "Enabling individualized virtual auditory space using morphological measurements," In *Proc. of the First IEEE Pacific-Rim Conference on Multimedia (2000 International Symposium on Multimedia Information Processing*, pp. 235-238, Dec. 2000.

[7] J. Princen, A. Johnson, A. Bradley, "Subband/transform coding using filter bank designs based on time domain aliasing cancellation," In *Proc. of the ICASSP*, pp. 2161-2164, 1987.

[8] C. Jin, V. Best, S. Carlile, T. Baer, and B. Moore, "Speech localization," In *Proc. of the 112th Convention of the Audio Engineering Society*, May 2002.

[9] S. Carlile, P. Leong, and S. Hyams, "The nature and distribution of errors in the localization of sounds by humans," *Hearing Research*, vol. 114, pp. 179-196, 1997.

[10] I.N. Fisher, T. Lewis, and B.J.J. Embleton, *Statistical Analysis of Spherical Data.* Cambridge University Press, Cambridge, 1987.