THE DESIGN OF AMBISONIC REPRODUCTION SYSTEM BASED ON DYNAMIC GAIN PARAMETERS

Bing Bu, Chang-chun Bao, Mao-shen Jia and Rong Zhu

Speech and Audio Signal Processing Laboratory, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China
bubing@emails.bjut.edu.cn, {chchbao, jiamaoshen}@bjut.edu.cn, zhurong123@emails.bjut.edu.cn

ABSTRACT

This paper describes a design approach of Ambisonic reproduction system based on dynamic gain parameters (DGP). In the conventional approaches, the fixed gain parameters are often optimized to minimize the overall objective function for whole 360° sound stage. The proposed approach has an advantage that the gain parameters vary with angles of source objects. The problem of optimization tradeoff among different angles is overcome by DGP, which achieves an optimal solution in each position. Source localizations of the B-Format signals were estimated in frequency bands in order to match the corresponding gain parameters. For the synthesized signals, the process was simplified by the given spatial information. Using the head-related transfer function (HRTF) analysis, the proposed approach was found to be significantly better than reference approaches in interaural time difference (ITD) and interaural level difference (ILD).

Index Terms— Spatial Audio Processing, Ambisonic, Irregular Loudspeakers Reproduction, HRTF Analysis

1. INTRODUCTION

Spatial audio processing aims at reconstructing the sound field perceived as realistically as possible in natural hearing [1]. Ambisonic technology is known as one of the best spatial audio system for capturing and reproducing a sound field, which pioneered by Michael Gerzon [2]. The main advantage of Ambisonic is that the recorded signals are absolutely independent of loudspeakers reproduction process, regardless of the number of loudspeakers and rules of layouts. Ambisonic reproduction systems are applied to transform the recorded signals into a number of sound channels driving loudspeakers. The gain parameters, namely the decoder coefficients, are computed as linear weights of the recorded signals on the basis of different loudspeaker arrays. And each channel signal is derived by combining the recorded signals with relevant gain parameters.

There are two strategies for deducing gain parameters: one is pseudo-inverse matrix for regular loudspeaker arrays [3]; the other is an optimization method for irregular loudspeaker arrays. For regular loudspeaker arrays, such as square array, spatial information can be perfectly recovered by pseudo-inverse matrix. However, for irregular loudspeaker arrays, such as ITU5.1 layout, the gain parameters are harder enough to be derived compared to regular arrays. An effective solution had been proposed by Wiggins [4], where the Tabu search algorithm was used for determining gain parameters to minimize overall objective function according to psychoacoustic criterion, namely Gerzon's localization theory [5]. Furthermore, heuristic genetic algorithm (HGA) [6] has been employed to search the optimal solution in terms of localization and uniform distribution of volume for all surround angles. More recently, Heller [7] explicitly analyzed the implementation, which derived the gain parameters from non-linear optimization (NLopt) software library for arbitrary arrays.

Among the existing approaches, the fixed gain parameters bear a mutual contradiction between different angles. As a result, the gain parameters get an average localization performances for the whole 360° surround. However, the objectives cannot be optimized efficiently for each position of 360° surround sound field. Especially, spatial information inevitably has a poor performance at the sides of listener performance for high-frequency component of sound. Moreover, it produces a huge waste in some directions where sound objects do not exist. For example, compared to frontal loudspeakers, the gain parameters of rear loudspeakers still give equal level values for the sake of balancing overall performance when sound objects are rendered ahead of the listeners. It is not necessary for optimizing objective function corresponding to these directions. In this paper, DGP is introduced to address this problem. When sound objects locate in various positions, different gain parameters should be applied to reproduce the sound field. There is a one-to-one correspondence between DGP and the angles of sound sources. Therefore, DGP provides an optimal localization performance for each audible position.

The remainder of this paper is organized as follows: The proposed method is presented in Section 2. A simplification of the synthesized signals is discussed in Section 3. HRTF analysis results of DGP applied to ITU5.1 layout are given in Section 4, while the conclusions are drawn in Section 5.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

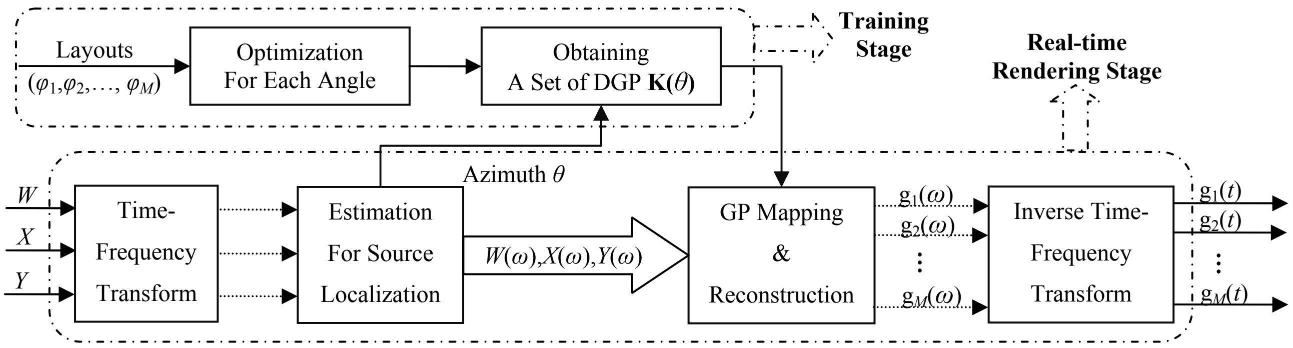
Layouts
$(\varphi_1,\varphi_2,\ldots,\varphi_M)$
Optimization For Each Angle
Obtaining A Set of DGP $\mathbf{K}(\theta)$
Training Stage
Real-time Rendering Stage
Azimuth θ
W
X
Y
Time-Frequency Transform
Estimation For Source Localization
$W(\omega),X(\omega),Y(\omega)$
GP Mapping & Reconstruction
$g_1(\omega)$
$g_2(\omega)$
$g_M(\omega)$
Inverse Time-Frequency Transform
$g_1(t)$
$g_2(t)$
$g_M(t)$

Fig.1. Block Diagram of Ambisonic Reproduction System based on DGP

2. AMBISONIC RENDERING BASED ON DGP

Block diagram of Ambisonic reproduction system based on DGP is shown in Fig.1. The system consists of two stages: one is the training stage which obtains a set of DGP $\mathbf{K}(\theta)$ on the basis of loudspeaker layouts; the other is the real-time rendering stage which combines the B-Format signals with DGP to reconstruct the sound field. In this Section, both two stages are described under the criterion of Gerzon's localization theory. The present work is limited to reproduce the B-format signals in horizontal arrays because the major of Amibisonic recordings are first order. However, there is nothing that limits the approach presented to an arbitrary order.

2.1. Gerzon's localization theory

The crucial models of auditory localization are the acoustic particle velocity model and the acoustic energy-flow model, respectively proposed by Makita [8] and De Boer [9]. Gerzon posits that the two models are able to generalize various aspects of auditory localization. In Ambisonic, they are commonly refered to as the velocity vector ($\mathbf{R}_V$) and energy vector ($\mathbf{R}_E$) models, separately related to low-frequency (<700Hz) and high-frequency signals. The relationship between the two models and the gain parameters can be described as follows:

$$\mathbf{R}_V(\theta) = \frac{\mathbf{UK}}{P} \tag{1}$$

$$\mathbf{R}_E(\theta) = \frac{\mathbf{U}(\mathbf{K}\circ\mathbf{K})}{E} \tag{2}$$

where $\mathbf{K}$ is a vector made up of gain parameters, $\mathbf{K} = [k_1\ k_2 \ldots k_M]^T$, and M is the number of loudspeakers. P is the sum of the elements of $\mathbf{K}$, and E is the sum of the squared $\mathbf{K}$. Besides, the symbol " ∘ " denotes element-by-element multiplication. The azimuth information of loudspeakers $\mathbf{U}$ is given by:

$$\mathbf{U} = \begin{bmatrix} \cos\theta_1 & \cos\theta_2 & \cdots & \cos\theta_M \\ \sin\theta_1 & \sin\theta_2 & \cdots & \sin\theta_M \end{bmatrix} \tag{3}$$

where θ_j denotes the angular position of the j^{th} loudspeaker.

Based on this, an optimum criterion is used to assess accuracy and stability of localization which satisfies the following three equations:

$$|\mathbf{R}_V(\theta)| = |\mathbf{R}_E(\theta)| = 1 \tag{4}$$

$$\theta_V = \arg(\mathbf{R}_V(\theta)) = \theta_E = \arg(\mathbf{R}_E(\theta)) = \theta \tag{5}$$

$$P(\theta) = cons \qquad E(\theta) = cons \tag{6}$$

where $|\mathbf{R}_V|$ and $|\mathbf{R}_E|$ predict the stability and compactness of sound image, and the equation (4) describes the best performance of $|\mathbf{R}_V|$ and $|\mathbf{R}_E|$. θ_V and θ_E are supposed to agree with the direction of orginal sound source. P and E should be constant in order to maintain consistency for volume of all directions.

2.2. Training stage of DGP

The general idea of DGP is that different gain parameters should be applied to reproduce the sound field when sound objects locate in various positions. Each group of DGP only focuses on optimizing one direction. A set of DGP is trained by genetic algorithm [10] and each group corresponds to each angle. Localization blur theory shows that human hearing has a limited resolution in locating sound objects, and the highest perceptual localization resolution is approximately about 1° [11]. Thus, the set of DGP is made up of 360 groups. In order to decrease the complexity of search process, the training times can be reduced to the half, that is, only 180 groups are trained due to $\mathbf{K}(\theta)$=$\mathbf{K}(2\pi$-$\theta)$ if loudspeaker layouts are bilateral symmetry. DGP obtains the optimal objective value for each audible angle, and overall objective value certainly decreases. From this, DGP has better localization performance than the existing approaches, especially at the back side of the listener.

According to Gerzon's localization theory, the six functions of evaluating spatial performance are calculated for each azimuth and determined as follows:

$$\begin{aligned} V_{LF}(\theta) &= \left|1-\frac{P(0)}{P(\theta)}\right| & V_{HF}(\theta) &= \left|1-\frac{E(0)}{E(\theta)}\right| \\ M_{LF}(\theta) &= \big|1-|\mathbf{R}_V(\theta)|\big| & M_{HF}(\theta) &= \big|1-|\mathbf{R}_E(\theta)|\big| \\ A_{LF}(\theta) &= |\theta-\theta_V| & A_{HF}(\theta) &= |\theta-\theta_E| \end{aligned} \tag{7}$$

In order to optimize these functions simultaneously, the weighted sum of six functions should be calculated as ultimate optimization objective, i.e.,

$$O_{objective}(\theta) = w_1 V_{LF}(\theta) + w_2 M_{LF}(\theta) + w_3 A_{LF}(\theta) + w_4 V_{HF}(\theta) + w_5 M_{HF}(\theta) + w_6 A_{HF}(\theta) \quad (8)$$

The six weights [w_1, w_2, ..., w_6] are introduced to address the problem of objectives dominance in case that one of the functions dominates the search. The genetic algorithm optimizes $O_{objective}(\theta)$ for each audible angle because the optimization objective is dependent of angles. A set of DGP obtained from genetic algorithm will be applied to reproduce the sound field.

2.3. Rendering stage for the B-Format signals

The directions of sound objects need to be analyzed from the B-Format signals. The proposed method is based on the assumption that the listener cannot discriminate two sound objects from different directions within a critical band at time instances, and only one of source direction is typically able to be localized by the listener. This is in line with psychoacoustic results made by Perrot [12] and a recent proposed auditory model for source localization [13]. From this assumption, short-time Fourier transform (STFT) has been used to divide signals into frequency bands. The azimuth $\theta(\omega)$ of sound object is estimated as follows:

$$\theta(\omega) = \begin{cases} \arctan\dfrac{Y(\omega)}{X(\omega)} & \text{if } Y(\omega)\cdot W(\omega) \geq 0 \\ \arctan\dfrac{Y(\omega)}{X(\omega)} + \pi & \text{if } Y(\omega)\cdot W(\omega) \geq 0 \end{cases} \quad (9)$$

Given the loudspeakers layouts, such as ITU 5.1 layout, the set of DGP is attained from the training stage. Then, a mapping between gain parameters and $\theta(\omega)$ should be found. The DGP $\mathbf{K}(\theta)$ is applied to reconstruct the sound field in the frequency domain, given as:

$$\begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \\ g_5 \end{bmatrix} = \begin{bmatrix} k_1(\theta) & k_2(\theta) & k_3(\theta) \\ k_1(\theta) & k_2(\theta) & -k_3(\theta) \\ k_4(\theta) & k_5(\theta) & k_6(\theta) \\ k_4(\theta) & k_5(\theta) & -k_6(\theta) \\ k_7(\theta) & k_8(\theta) & k_9(\theta) \end{bmatrix} \begin{bmatrix} W \\ X \\ Y \end{bmatrix} \quad (10)$$

Finally, the resulting frequency representations of loudspeakers are transformed back to time domain based on the inverse STFT. A further investigation of analysis with the B-Format signals can be found in [14].

3. SIMPLIFICATION FOR THE SYNTHESIZED SIGNALS

There are two ways to produce B-Format signals, one is that real sound field is recorded with soundfield microphone or an equivalent microphone array; the other way is that virtual sound field is synthesized using synthetic equation which is equivalent to a model of recording. The directions of sound sources are estimated in frequency domain for the recorded signals, while the synthesized signals artificially provide the directions of sound sources.

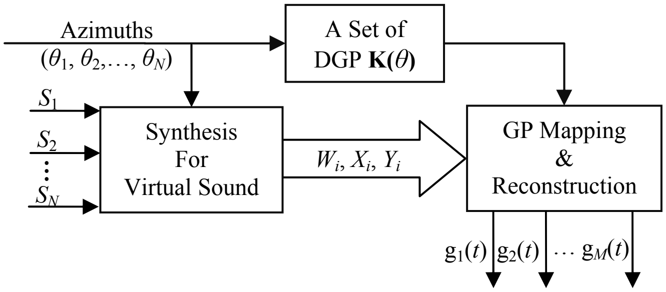
Azimuths
(θ_1, θ_2,…, θ_N)
A Set of DGP $\mathbf{K}(\theta)$
S_1
S_2
S_N
Synthesis For Virtual Sound
W_i, X_i, Y_i
GP Mapping & Reconstruction
$g_1(t)$
$g_2(t)$
… $g_M(t)$

Fig.2. Block Diagram of DGP for Signal Synthesis

Different from the recorded signals, DGP approach is simplified for the synthesized signals, because it implies the azimuths of sound objects. The block diagram of DGP is shown in Fig.2. The azimuths do not need to be analyzed in the frequency domain, and the whole processes are implemented in the time domain.

The B-Format signals are synthesized by using synthetic equation for monophonic signals as follows:

$$\begin{bmatrix} W \\ X \\ Y \\ Z \end{bmatrix} = \frac{1}{N} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & \cdots & 1/\sqrt{2} \\ \cos\theta_1 \cos\varepsilon_1 & \cos\theta_2 \cos\varepsilon_2 & \cdots & \cos\theta_N \cos\varepsilon_N \\ \cos\theta_1 \sin\varepsilon_1 & \cos\theta_2 \sin\varepsilon_2 & \cdots & \cos\theta_N \sin\varepsilon_N \\ \sin\varepsilon_1 & \sin\varepsilon_2 & \cdots & \sin\varepsilon_N \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_N \end{bmatrix} \quad (11)$$

where S_i is the i^{th} monophonic signal, and N is the number of monophonic signals. The direction information (θ_i, ε_i) is artificially given. For horizontal-only reproduction systems, the Z signal is ignored and $\cos(\varepsilon_i)$ is set to 1.

The single sound object is reconstructed from W_i, X_i, Y_i which correspond to the azimuth θ_i. The compromised sound field of all sound objects is rendered by:

$$g_j = \sum_{i=1}^{N} \left[g_{ji}(\theta_i) \right] \quad (12)$$

where N is the number of sound objects, and g_j stands for reproduction signal of the j^{th} loudspeaker.

4. PERFORMANCE EVALUATION

The objective test shows the examples for ITU 5.1 (rear loudspeakers at ±115°), as defined in BS.775 [15], because it is a configuration that others have worked on and therefore provide a good benchmark of DGP approach. HRTF analysis is an effective method which objectively evaluates the localization performance of reproduction systems. Compared to the graphs of velocity vector and energy vector, the results are closer to the sense of auditory experience. In the test, binaural signals are obtained by convoluting reproduction signals g_i with the corresponding head-related impulse responses (HRIR), presented in Fig.3. The HRIR data used are those measured by Gardner and Martin [16].

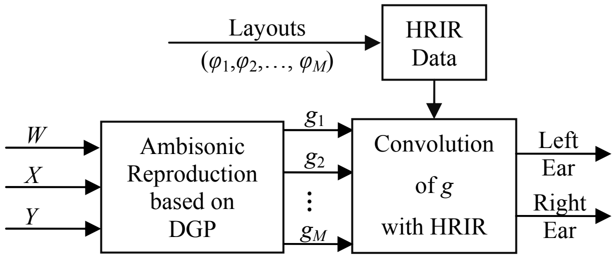
Layouts
$(\varphi_1,\varphi_2,\ldots,\varphi_M)$
HRIR Data
W
X
Y
Ambisonic Reproduction based on DGP
g_1
g_2
⋮
g_M
Convolution of g with HRIR
Left Ear
Right Ear

Fig.3. The Conversion Relationship Between Ambisonic Signals and Binaural Signals

The auditory model is used to estimate the ITD and ILD cues of real sources and Ambisonic reproduction systems [17]. 13 equally spaced angles from horizontal plane were evaluated (i.e. 0°, 30°, 60°, ..., 360°). The proposed approach was compared with three kinds of typical Ambisonic reproduction systems: Basic decoder, Max $\mathbf{R}_E$, and NLopt [7] reproduction systems.

Fig.4 shows the estimated ITD and ILD cues of the four reproduction systems. The red and blue lines respectively mean ITD and ILD of the real source and the Ambisonic reproduction system. The green line stands for the difference between real source and Ambisonic, and the dotted line represents a zero baseline. Corresponding to Fig.4, Table.1 and Table.2 separately display the unsigned ITD and ILD errors for each reproduction system, and μ means the average errors.

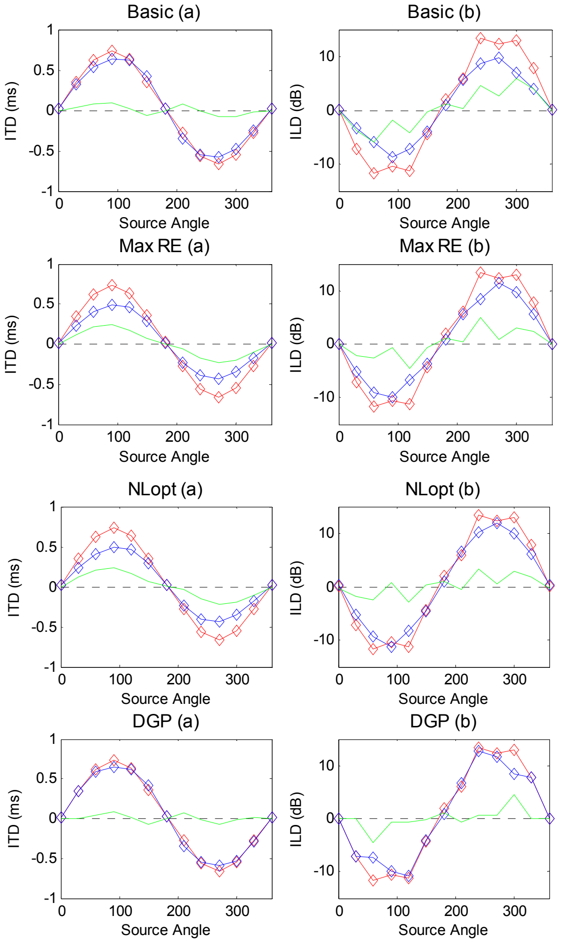

Fig.4. Comparison of ITD and ILD

Table.1. The Unsigned ITD Errors

Systems	0	30	60	90	120	150	180	μ
Basic	0	0.03	0.08	0.09	0.02	0.07	0	0.04
Max Re	0	0.17	0.21	0.25	0.18	0.07	0	0.13
NLopt	0	0.12	0.21	0.24	0.17	0.06	0	0.11
DGP	0	0.01	0.04	0.08	0.02	0.07	0	0.03

Table.2. The Unsigned ILD Errors

Systems	0	30	60	90	120	150	180	μ
Basic	0.07	3.80	5.79	1.76	4.18	0.39	1.10	2.44
Max Re	0.07	2.08	2.67	0.54	4.62	0.58	1.05	1.66
NLopt	0.23	1.88	2.44	0.73	2.97	0.28	0.95	1.35
DGP	0.02	0.09	4.48	0.67	0.58	0.18	1.09	1.02

The results show that DGP performs the best compared to other three systems. Both ITD and ILD are the closest match to the real source, especially at the side and rear of the listener. The ILD of DGP is slightly bigger than Max $\mathbf{R}_E$ and NLopt when the source angle is 60°. This is because DGP has dominance for low-frequency signals in the direction. Appropriate weights should be investigated further.

5. CONCLUSIONS AND FUTURE WORK

A design approach of Ambisonic reproduction system based on DGP is proposed in this paper. On the basis of Gerzon's localization theory, the advantage of DGP is that the objective is able to attain optimal value for each audible angle because the gain parameters vary with the directions of sound objects. Two types of the B-Format signals are addressed respectively to render sound field. In addition, DGP approach can also be applied to encode spatial aspects of sound which are transmitted or stored over a single or several channels carried with spatial information. The design is demonstrated by means of HRTF analysis. Both estimated ITD and ILD have good fit as the results of real sound sources. However, this paper only demonstrates that the B-Format signals could be reproduced based on DGP approach in ITU 5.1 layout. The proposed approach will be extended to assess arbitrary loudspeaker arrays with height for higher order Ambisonic signals in future work.

6. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 61231015, 61201197), Scientific Research Project of Beijing Educational Committee (Grant No. KM201310005008), the Ph.D. Programs Foundation of Ministry of Education of China (Grant No. 20121103120017) and the 12th Postgraduate Science Foundation of Beijing University of Technology (ykj-2013-9322, 2013-9330, 2013-9561).

7. REFERENCES

[1] J. Breebaart and C. Faller, *Spatial audio processing: MPEG Surround and other applications*, John Wiley & Sons, Chichester, UK, 2007.

[2] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, pp. 2–10, Jan. 1973.

[3] J. Daniel, R. Nicol, and S. Moreau, "Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging," *AES 114th Convention*, Preprint 5788, Amsterdam, Mar. 2003.

[4] B. Wiggins, et al, "The design and optimisation of surround sound decoders using heuristic methods," *Proceedings of UKSim*, *Conference of the UK Simulation Society*, pp.106-114, 2003.

[5] M. A. Gerzon, "General metatheory of auditory localization," *AES 92nd Convention*, Preprint 3306, Vienna, 1992.

[6] P.W.M. Tsang, K.W.K. Cheung, "Development of a re-configurable ambisonic decoder for irregular loudspeaker configuration," *IET Circuits Devices Syst.*, vol. 3, no.4, pp. 197–203, 2009.

[7] A. Heller, E. Benjamin, and R. Lee, "Design of ambisonic decoders for irregular arrays of loudspeakers by non-linear optimization," *AES 129th Convention*, vol.1, no.7, 2010.

[8] Y. Makita, "On the directional localization of sound in the stereophonic sound field," *E.B.U. Review*, *Part A -Technical*, no.73, pp.102–108, Jun. 1962.

[9] K. DeBoer, "Stereophonic sound production," *Philips Technical Review*, no.5, pp.107–144, 1940.

[10] C. R. Houk, J. A. Joines, and M. G. Kay, "A genetic algorithm for function optimization: A matlab implementation," *Tech. Rep. North Carolina State Univ.*, 1995.

[11] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*, third print, The MIT Press, 2001.

[12] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, No. 6, Jun. 2007.

[13] C. Faller, J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Audio Eng. Soc.*, vol.116, no.5, pp.3075–3089, Nov. 2004.

[14] V. Pulkki, C. Faller, "Directional audio coding: Filterbank and STFT-based design," *AES 120th Convention*, Preprint 6658, May. 2006.

[15] Rec. ITU-R BS.775-1, "Multichannel stereophonic sound system with and without accompanying picture," Geneva, 1992-1994.

[16] B. Gardner, K. Martin, "HRTF measurements of a kemar dummy-head microphone," Retrieved: Apr. 2007.

[17] D. Moore, "The development of a design tool for 5-speaker surround sound decoders," Ph.D. Thesis, University of Huddersfield, 2009.