New Zealand Acoustics, vol. 24, no. 4, pp. 4-11, 2011

Environmental Monitoring Using Arbitrary Microphone Arrays

Wyatt Page[1], Peter Driessen[2], Pranav Sakulkar[3] and Farook Sattar[4]

[1]Institute of Food Nutrition and Human Health, Massey University, New Zealand

[2]School of Engineering & Advanced Technology, Massey University, New Zealand

[3]Department of Electrical Engineering, Indian Institute of Technology, Kanpu, India

[4]Department Of Computer System & Technology, University of Malaysia, Kuala Lumpur, Malaysia

A paper previously presented at ISSA 2010, 29-31 August 2010, Auckland

Abstract

Augmented audio reality (AAR) is a mixture of virtual reality and the natural environment. AAR enhances the acoustics of a real world environment with additional audio information. A hearing aid is a simple example of AAR. We describe an experiment in an indoor environment with up to 16 simultaneous voices (such as in a crowded room) and up to 48 microphones in arbitrary configurations. We develop array signal processing algorithms to train the system, and separate, localize, track, enhance and zoom in on the voices. Effective audio zoom algorithms and natural gesture control will be useful in many contexts in of themselves. Their combination will enable a "super" hearing aid which can be applied to restore impaired hearing, guide visually impaired people, as well as enhance natural hearing. The longer-term objective is to develop algorithms to control the location on which the array is focused using head and eye gestures. The listener can simply look at the location from where s/he wants to hear the audio, thus controlling the audio zoom via head and eye gestures.

Introduction

In this paper we report on the early phases of building a system using microphone arrays of arbitrary geometry to zoom in on desired audio in a noisy environment. Ultimately a listener will be able to simply look at the location from where s/he wants to hear the audio, controlling the audio zoom via head and eye gestures.

The microphone arrays may be a permanent part of a venue, embedded in the walls and ceiling. Or using low cost wireless devices, deployed ad-hoc at the time of use (e.g. stuck to the walls and ceiling with removable adhesive).

Applications of this new audio zoom system include a "super" hearing aid for people in a crowded noisy environment. Such a hearing aid will have performance far exceeding any standard hearing aid with microphones near the ears. Audio zoom may be very useful for the film industry, to capture better quality audio during on-location filming, and reduce the amount of rerecording and post-production required. It may also be useful for the computer games industry where zooming on natural sounds may be desired as part of the game-play. Audio zoom will be a very useful research tool for studying bird communications, providing detailed spatial information on territorial birdsong, which may help decipher the song function.

Background

Augmented audio reality (AAR) [1][2][3] is a mixture of virtual reality and the natural environment. AAR enhances the acoustics of a real world environment with additional audio information. A hearing aid is a simple example of an AAR system. AAR is a subset of Augmented Reality (AR), but in practice AR systems are focused on visual augmentation with little or no emphasis on any auditory elements. AR is part of the wider mixed reality continuum where the aim is to present the real and virtual elements in a way that they are perceived as one.

One of the most commonly quoted, commercially successful implementations of an AAR system are the Audio Guides such as the Sennheiser guidePORT (www.guideport.com). Although these audio guides typically include noise cancelling technology, they are not really an AAR system but simply location aware audio.

Härmä et al [2] describe an augmented audio environment as:

> "The concept of augmented reality audio characterizes techniques where a real sound environment is extended with virtual auditory environments and communication scenarios. An augmented audio environment is produced by superimposing a virtual sound environment onto the psycho-acoustic environment."

Literature Review

Microphone array signal processing is reviewed in [6][7]. For our application, the source signals are wideband and mostly indoors where there will be significant multi-path effects. A thorough review of such MIMO (multiple-input and multiple-output) systems appears in [8]. Relevant theories include the following: a fast and efficient frequency-domain Blind Source Separation (BSS) method using Independent Component Analysis (ICA) for a convoluted mixture of audio signals [9], frequency domain convoluted blind source separation algorithms based on real room recordings [10], a novel approach to directly recover the location of both microphones and

sound sources from time-difference-of-arrival measurements only [11], and a linear closed-form algorithm for source localization from time-differences of arrival (TDoA) [12]. Learning/ training methods are also relevant. A microphone array system is trained using signals from a set of positions and trajectories and subsequently recalls the localization information when presented with new input signals [13]. Because of its learning nature this method provides practical advantages in setting up a microphone array, by not requiring favourable room acoustics, careful element positioning or uniformity of sensors. A generative statistical model for speech and noise sources at distinct regions in the soundfield is used in [14], and incremental Bayesian learning is used to track the model parameters over time. A new method for time delay estimation based on the analysis of the cross spectrum between a pair of microphones [15] is also relevant.

A number of the array signal processing methods in the literature make use of, or can benefit from, knowing the location of the microphones in the arrays and the location of the sound sources. This information can be determined from the time difference of-arrival or relative delay between sound sources and microphones. The multiple-measured room impulse response (MMRIR) attempts to capture the MIMO relationships between sound sources and microphones to obtain a more comprehensive representation of the acoustics of a room. For every source-microphone pairing, an impulse response (IR) is measured. The delay from an IR can be used to estimate TDoA for the most direct path while the complete IR captures the multi-path effects, the timing and intensity of the reflections.

There are a range of approaches to measuring room impulse responses [25]. These include sine sweeps and the use of wideband noise. The exponential sine sweep (ESS) approach has evolved over the last decade to become the preferred method for accurate measurement of room IR [26]. However, the wideband noise approach is very useful for non-stationary sound sources and where an estimate of IR is required in near real-time. Kasami sequences are pseudorandom number sequences that can be used to create wideband noise with special correlation properties. A maximal length sequence (m-sequence) has a large autocorrelation at zero lag, with near zero autocorrelation elsewhere, thus enabling the quick determination of the impulse response of a linear time invariant (LTI) system. Maximal length sequences are also the base of sets of sequences with good correlation properties.

The small set of Kasami sequences is one such set that have small off-peak autocorrelations and also small cross correlations between sequences. This property allows accurate determination of time of arrival time of a transmitted sequence, even in the presence of other interfering transmissions. Code for implementing Kasami sequences is available free from MATLAB Central on "The Mathworks" website (www.mathworks.com/matlabcentral).

Some practical applications of the audio zoom will depend upon methods of controlling the location on which the array is focused. The control algorithms and technologies for user tracking are reviewed in [4][5]. We were unable to find good examples of natural transparent zoom control (without external input devices) for visual or audio sensors [22][23][24].

Problem Formulation

We consider a signal model where the source of interest, i.e. the speaker is in a fixed position relative to the microphone array. The noise environment consists of interference, i.e. surround speakers, and ambient noise. The i[th] microphone array received, after sampling, a component from the target speaker, $s(n)$, and a sum of noise sources $v_d(n)$, $d = 1 \ldots D$ together with the ambient noise $v(n)$ as

$$x_i(n) = s_i(n) + \sum_{d=1}^{D} v_{di}(n) + v_i(n) \quad [1]$$

The array data vector received from M point sources, impinging on an I-element array in space, at time n, can be described by

$$\mathbf{x}(n) = \sum_{m=1}^{M} u_m(n)\mathbf{h}_m + v(n) \quad [2]$$

Where $v(n)$ is the received uncorrelated noise, $u_m(n)$ is the signal from the mth point source, and the impulse response (i.e. steering vector)

$$\mathbf{h}_m = [\beta_1 e^{j2\pi f \tau_1}\ \beta_2 e^{j2\pi f \tau_2} \cdots \beta_I e^{j2\pi f \tau_I}] \quad [3]$$

represents the acoustic path (propagation channel) between the mth signal source and the array, where β_i is the attenuation and τ_i is the propagation time delay from point m to the i[th] array element. The output signal $y(n)$ can be written as

$$y(n) = \mathbf{w}(n)^H \mathbf{x}(n) \quad [4]$$

where $w(n)$ is the array filter vector, $x(n)$ defines the corresponding input data vector and H denotes the Hermetian transpose. The purpose is to estimate the filters $w(n)$ such that the target signal $s(n)$ is recovered.

BSS Method Implementation

A fast and efficient blind source separation (BSS) algorithm is implemented in this paper [9]. This frequency domain approach is based on convoluted mixing model, which provides the flexibility to encompass important factors, such as propagation delay as well as multi-path and wideband nature of the sources. The first step of this approach is transforming the data into frequency domain using a block window which can be approximated as:

$$X(k,n) \approx A_k S(k,n) + V(k,n) \quad [5]$$

where $X(k,n), S(k,n), V(k,n)$ are the DFT (Discrete Fourier Transform) of the sensor, source and noise signal vectors for frequency bin k and time frame index n, respectively, and A_k is the mixing matrix of the k^{th} frequency bin.

The next step is to estimate the unmixing matrix by applying the well-known FastICA scheme at each frequency bin. The FastICA is applied iteratively at each frequency bin starting from the lowest frequency to the highest frequency. The FastICA algorithm [28] makes use of an efficient learning rule to maximize the non-Gaussian nature of the projection. It is among the most commonly used algorithms for optimal search of the unmixing matrix **W** that is updated based on a nonlinear contrast function. The optimization techniques like gradient search or Newton optimization are used for updating the contrast function G(**WX**) where **X** is the observed matrix of the mixed source signals. The general form of the gradient search and Newton optimization techniques for updating the unmixing

matrix **W** is given by

$$\mathbf{W}_{n+1} = \mathbf{W}_n + \eta g(\mathbf{W}_n\mathbf{X}) \quad [6]$$

$$\mathbf{W}_{n+1} = \mathbf{W}_n - \eta \frac{g(\mathbf{W}_n\mathbf{X})}{g'(\mathbf{W}_n\mathbf{X})} \quad [7]$$

where g(**WX**) and g'(**WX**) are the first and second derivatives of the contrast function and η is an adaptation step size. The function g() can be any non-quadratic function which must be chosen to provide efficient updates.

For the iterative estimation of the unmixing matrix, the final solution of the previous unmixing matrix is used as initial value for the FastICA iteration in the current frequency bin. In this way, there has been improvement in the inherent permutation problem based on the assumption that the values of the unmixing matrices for the adjacent frequency bins are changing slowly. In addition, the scaling problem has been solved by using the diagonal terms of the inverse of the estimated unmixing matrix considering that the unmixing matrix has full rank.

Figure 1. Part of the recording venue showing some of the boundary placed microphones (wall and ceiling). The white circle indicates the surround-sound microphone array.

Figure 2. One of the random speaker layouts for simulating people having conversations anywhere in the room.

Experimental

As there are no readily available data sets for multi-channel input and output sound in a real venue, we decided to record our own.

Setup

An empty moderately sized (8m x 5m x 3m) internal room with a carpeted floor and a suspected acoustic ceiling was used as the venue for an experimental recording session. The room has windows at one end and a large whiteboard at the other. We hired a mobile recording studio along with a sound engineer, from Sounds Unlimited (www.soundsunlimited.net.nz). The mobile studio used a Yamaha DM2000 digital production console as the front end for Apple Logic Studio, Digital Audio Workstation (DAW) software running on a Macintosh Pro computer. The setup provided eight discrete output channels and 48 microphone inputs. Eight Yamaha MSP7 active studio monitors from our surround sound laboratory were used as the output speakers. A collection of professional microphones were used to make up the 48 inputs. The microphones were a mixture of both omnidirectional and cardioid directional patterns.

The last seven microphones were in a surround sound microphone array assembly [27] (see figure 1). The array has five microphones (all Sennheiser MKH50) equally spaced around a circle on a single plane and up and down pointing microphones (both Sennheiser MKH60). In contrast to this structured microphone array, the other 41 microphones were placed reasonably randomly around the room, initially on or close to the ceiling and close to the walls of the long axis of the room. Because of their close proximity to the surfaces, they behave approximately as boundary microphones.

Experiments

A series of experiments were designed and run in the venue. For each experimental setup, images of the microphone and speaker placement were taken so post calibration of their position could be performed. All data from the microphones was recorded at 24 bits resolution and 48 kHz sampling rate. Almost 50 GB of data was recorded from the sessions and this data is available from the authors for other researchers wishing to collaborate with us.

Table 1. Details of the 11 recordings made, covering the different microphone placement and operational scenarios

#	Title	Description
1	Walk	Test conversation – Two males having a conversation while walking around the room; 12 microphones on the walls + 29 on the ceiling + surround array.
2,3	Conversation 1,2	Group conversation – Up to 12 people having conversation slowly moving around the room; 12 microphones on the walls + 29 on the ceiling + sur-round array.
4, 5, 6,7	Random 1,3,4,5	Simulated group conversation – Up to 16 voices from eight speakers placed randomly around the room; 12 micro-phones on the walls + 29 on the ceiling+ surround array.
8	Table 1	Audio conference – Up to 16 voices from eight speakers placed uniformly around the table; 12 microphones on the walls + 29 on the ceiling + sur-round array.
9	Table 2	Audio conference – Up to 16 voices from eight speakers placed uniformly around the table; 29 ceiling micro-phones + 12 microphones on the table + surround array.
10	Table 3	Audio conference – Up to 16 voices from eight speakers placed uniformly around the table; 41 microphones on the table + surround array.
11	Delay Test	Measurement of recording system internal delay, input to output.

We imitated two microphone placement scenarios. The first is where the microphones are embedded in the walls and ceiling and the second scenario is where they are embedded into an audio/video conference table. In both scenarios we also included the discrete surround-sound array for comparison. Three operational scenarios were imitated in the recording sessions. The first involved real people having conversations anywhere in the room. The second one involved simulating people having conversations anywhere in the room and the third involved simulating people having conversations around an audio/video conference table. For the real-people scenario we invited staff and students to have a conversation amongst the jungle of wires and equipment. Up to 12 different voices could be occurring at the same time. For the simulated conversations, a series of podcasts from Radio New Zealand, the Canadian Broadcasting Corporation and National Public Radio (USA), were downloaded from their website. These podcasts were typically of an interactive interview and included both male and female voices. Each podcast was trimmed to 15 minutes long. The simulated conversations were played-back through the eight studio monitor speakers set on stands 1.2 metres high. Up to 16 different voices could be occurring at the same time in this scenario.

For the second operational scenario, the speakers were placed randomly in the room and with random orientations. Some constraints were applied to the randomization to ensure that the speakers were not hard up against a wall while also being orientated into the wall.

For the simulated audio/video conference operational scenarios, the eight speakers where placed around the table in a regular pattern. Table 1 details the 11 recordings made, covering the different scenarios. Most recording sessions lasted around 19 minutes. For the first four minutes of

each recording, signals were played-back through the speakers to measure the MMRIR. Linear sine-wave sweeps were followed by exponential sweeps from each of the eight speakers. Lastly, two different small set Kasami sequences at different sampling rates were played-back simultaneously through all eight speakers.

Results

Initial testing was done using simulated data with two microphones and two sources (loudspeakers). Two configurations were used: For narrow spacing, the microphones were 4 cm apart. The sources were 1.2 metres apart at a distance of 1 metre from the microphones. For wide spacing, both microphones and sources were arranged in a 3 m x 4 m rectangle with the microphone and speakers 4 m apart.

Blind source separation using the frequency domain method of [9] with two male voices as sources was effective with a two-ray impulse response (direct path plus one reflection), provided that the microphone separation was no more than about 4 cm (half wavelength at 4,250 Hz). The permutation ambiguity problem inherent in ICA technique influences the misalignments for larger separation as frequencies become higher.

Non-blind source separation with the knowledge of IR of the room was tested. As the length of speech signal is not fixed, we cannot perform inverse filtering on the whole signal at the same time. So the signals were windowed and inverse filtering was applied in the frequency domain using STFT (short-time Fourier transform) the and results were recorded for the robustness with changes in parameters.

Figure 5 shows the plot of SIR (Signal Interference Ratio) improvement versus variation in reflection coefficient for wide separation and figure 6 for the narrow separation case. The SIR improvement is better for the narrow separation case and more robust to the changes in reflection coefficient of the wall. This shows that for narrow spacing, we can work with only geometrical information and get good results without the specific knowledge of the wall reflection coefficient. Note that the frequency resolution in the case of mixing is equal to the length of speech signal and in the case of de-mixing is equal to the length of the window.

Table 3 shows the effect of window size on SIR improvement for the wide spacing case. As the size of the window is increased, the inverse filter gives improved results, with the best results for a window size equal to the length of speech. The next step is to use the measured impulse responses using ESS in place of the simulated early reflections model. The experimental data in the recording sessions were done with wide spacing much more than 4 cm. It is expected that the IR measurements will be sufficiently accurate to give good results. FIR approximations to the IIR inverse filters will also be tested.

Conclusions & Future Work

The results presented above are very preliminary. We are working towards a robust audio zoom (source separation) method which is partly blind and partly aided by the measured impulse responses.

Table 2. Parameters used for simulation non-blind source separation

Sampling Rate	16 kHz
STFT frame size	1024 points (64 ms)
STFT frame shift	512 points (32 ms)
Frequency bins	512
Reflection Coefficient of wall	0.5

Table 3. SIR improvement with increasing window size for the wide spacing case.

Window size	1024	2048	4096
SIR Improvement (dB)	8.6	10.9	12.6

Acknowledgements

The author would like to thank Professor Bastiaan Kleijn for his helpful suggestions in planning the recording session. We would also like to thank Neil Maddever from Sounds Unlimited for his recording expertise and Dr. Rajeev Nongpiur for his contributions to discussion of the analysis

References

1 N. Röber, “Interaction with Sound: Explorations beyond the frontiers of 3D Virtual Auditory Environments”. PhD Thesis, Dept of Simulation and Graphics, Otto-von-Guericke-Universität Germany, 2008.
2 A. Härmä, J. Julia; M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, G. Lorhon, “Augmented Reality Audio for Mobile and Wearable Appliances”, JAES Vol. 52 Issue 6, 618-639 (June 2004).
3 M. Cohen, “Augmented Audio Reality: Design for a Spatial Sound GPS PGS”, Proc.of 2nd IEEE International Workshop on Robot and Human Communication (1994).
4 J. P. Rolland, Y. Baillot, and A. A. Goon, “Fundamentals of Wearable Computers and Augmented Reality”. Lawrence Erlbaum Associates (2001).
5 F. Lassabe, P. Canalda, P. Chatonnay, F. Spies, “Indoor Wi-Fi positioning: techniques and systems”, Annals of Telecommunications, Vol. 64, No. 9-10.
6 J. Benesty, J. Chen, and Y. Huang, “Microphone Array Signal Processing”, Springer, 2008.
7 M. Brandstein, D. Ward, “Microphone Arrays”, Springer, 2001.
8 Y. Huang, J. Benesty, and J. Chen, “Acoustic MIMO Signal Processing Springer, 2006.
9 X. Peng and S. L. Grant, “A Fast and Efficient Frequency-Domain Method for Convolutive Blind Source Separation,” IEEE Region 5 Conference, Apr. 2008.
10 S. M. Naqviy, et al, “Evaluation of Emerging Frequency Domain Convolutive Blind Source Separation Algorithms based on real Room Recordings,” 5th IEEE Sensor Array and Multichannel Signal Processing Workshop, Jul. 2008, 345-348.
11 M. Pollefeys, and D. Nister, “Direct Computation of Sound and Microphone Locations from Time-Difference-of-Arrival Data”, IEEE ICASSP, Mar. 2008, 2445-2448.
12 M. D. Gillette and H. R. Silverman,

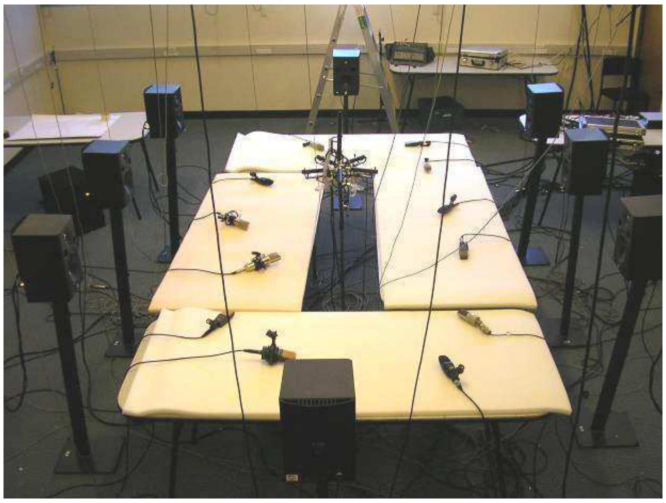

Figure 3. One of the layouts for simulating people having conversations around an audio/video conference table.

Figure 4. The final layout corresponding to recording 10 in Table 1, with all 48 microphones in place.

"A Linear Closed-Form Algorithm for Source Localization from Time-Differences of Arrival," IEEE Signal Processing Letters, vol. 15, 2008.

13 P. Smaragdis and P. Bourounos, "Position and Trajectory Learning for Microphone Arrays", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No. 1, 2007.

14 M. L. Seltzer, I. Tashev and A. Acero, "Microphone Array Post-Filter using Incremental Bayes Learning to Track the Spatial Distributions of Speech and Noise," IEEE Int. Conf. on Acoustics, April 15-20, 2007, I-29-I-32.

15 B. Lee, A. Said, T. Kalker and R. W. Schafer, "Maximum Likelihood Time Delay Estimation with Phase Domain Analysis in the generalized cross correlation framework", Hand-Free Speech Communication and Microphone Arrays, HSCMA 2008.

16 D. Zhang, D. Gatica-Perex, S. Benio and I. McCowan, "Semi-Supervised Adapted HMMs for Unusual Event Detection", Proc. of the 2005 IEEE Comp. Soc. Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, 2005.

17 S. Moncrieff, S. Venkatesh, G. West, "Unifying Background Models over Complex Audio using Entropy", Pattern Recognition, 2006. ICPR 2006.

18 M. Davy, F. Desobry, A. Gretton and C. Doncarli, "An Online Support Vector Machine for Abnormal Events Detection", Signal Processing, vol. 86, no. 8, Aug. 2006.

19 Rabaoui, M. Davy, S. Rossignol, Z. Lachiri and N. Ellouze, "Improved One-Class SVM Classifier for Sounds Classification", IEEE Conf. on Advanced Video and Signal Based Sound Environment Monitoring (AVSS 2007), Sept. 2007, pp. 177-122.

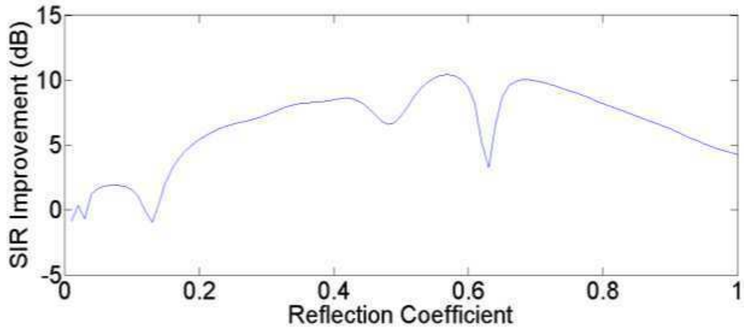

Figure 5. SIR improvement versus reflection coefficient for wide spacing of microphones.

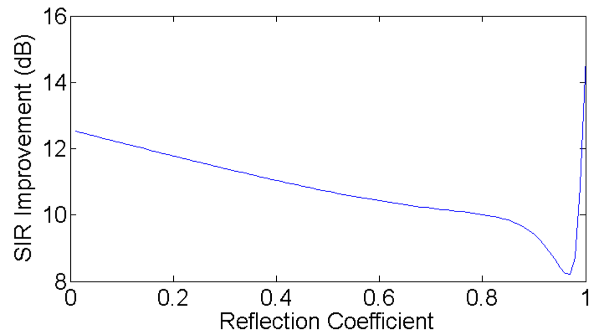

Figure 6. SIR improvement versus reflection coefficient for narrow spacing of microphones.

20 Araki, H. Sawada, R. Mukai, and S. Makino, Underdetermined Blind Sparse Source Separation for Arbitrarily Arranged Multiple Sensors, Signal Process., vol. 87(8)(2007)
21 K.D. Donohue, "Audio Systems Array Processing Toolbox" (for MATLAB), Audio Systems Laboratory, Department of Electrical and Computer Engineering, University of Kentucky,
22 F. Heller, T. Knott, M. Weiss, J. Borchers, "Multi-user interaction in virtual audio spaces". Proc. of the 27th international Conference Extended Abstracts on Human Factors in Computing Systems (Boston, MA, USA, April 04 - 09, 2009). CHI EA '09. ACM, New York, NY, 4489-4494.
23 K. Seungjun, K.D. Anind, "AR interfacing with prototype 3D applications based on user-centered interactivity". Journal of Computer-Aided Design, 2008.
24 J Huopaniemi, "Future of Personal Audio: Smart Applications and Immersive Communication", AES 30th International Conference: Intelligent Audio Environments (March 2007).
25 G. Stan, J.J. Embrechts, D. Archambeau, "Comparison of Different Impulse Response Measurement Techniques", JAES Vol. 50, No. 4, 2002.
26 A. Farina, "Advancements in impulse response measurements by sine sweeps", Presented at the 122nd AES Convention, Vienna, Austria, May 2007.
27 J.D. Johnston, Y.H. Lam, "Perceptual Soundfield Reconstruction", AES Convention 109 (2000).
28 A. Hyvarinen and E. Oja, "A Fast Fixed-point algorithm for Independent Component Analysis", Neural Computation, vol. 9, pp. 483-492, 1997.