Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Blind source separation and directional audio synthesis for binaural auralization of multiple sound sources using microphone array recordings

B. Gunel, H. Hacihabiboglu and A. Kondoz

I-Lab Multimedia and DSP Research Group, Centre for Communication Systems Research,
University of Surrey, GU2 7XH Guildford, UK
b.gunel@surrey.ac.uk

Microphone array signal processing techniques are extensively used for sound source localisation, acoustical characterisation and sound source separation, which are related to audio analysis. However, the use of microphone arrays for auralisation, which is generally related to synthesis, has been limited so far. This paper proposes a method for binaural auralisation of multiple sound sources based on blind source separation (BSS) and binaural audio synthesis. A BSS algorithm is introduced that exploits the intensity vector directions in order to generate directional signals. The directional signals are then used in the synthesis of binaural recordings using head related transfer functions. The synthesised recordings subsume the indirect information about the auditory environment conveying the source positions and the acoustics similar to dummy head recordings. Test recordings were made with a compact microphone array in two different indoor environments. Original and synthesized binaural recordings were compared by informal listening tests.

1 Introduction

Auralization systems differ in their way of obtaining information about the room and presenting it. Auralization systems may also aim at making audible a real acoustic space or a virtual one. If a virtual environment is to be auralized, a 3-D model of the space is created and the room impulse responses of the environment are obtained with acoustical simulations [1, 2, 3, 4, 5]. These room impulse responses are then convolved with anechoic recordings for presentation by headphones or loudspeakers. For auralizing a real environment, other possibilities exist. Room impulse responses can be directly measured within the environment such as with the MLS technique to save from the computing power required for modelling [6]. Alternatively, the recording of the source material can be done directly in the room considering the difficulty of obtaining anechoic recordings. This direct approach requires further attention, because, when the recording method does not match the reproduction method, auralisation could not be achieved. To enable reproduction with different methods, the source direction and the acoustical information should be preserved by the recording technique that enables extraction by the processing.

Microphone arrays are used heavily for source localisation, separation and acoustical analysis [7, 8]. As these are information extraction methods by nature, they are suitable for auralisation applications as well. This paper proposes an auralisation technique based on blind source separation applied on the signals captured by a microphone array in an environment to be auralised. Since the target application is auralisation, the source separation technique should be able to deal with convolutive mixtures. As the reflections are also needed to be reproduced, the separation should produce more channels than the sources, which can be considered as the under-determined case. It is desirable that the technique works in real-time, so that the recordings can be auralised directly. Finally, the quality of the recordings should be high. These requirements prevent the usage of some the well-known BSS techniques, such as those based on independent component analysis (ICA) [9, 10] and adaptive beamforming (ABF) [11, 12]. The scaling and permutation issues related to frequency domain techniques [13], which run faster, may result in decrease in sound quality. Moreover, most of these techniques require arrays that are made up of physically separated microphones. Such recordings are not useful for auralisation as the sound field observed at each sensor position differs.

The source separation technique employed in this paper uses a compact microphone array and provides a closed-form solution, which is desirable from the computational point of view [14]. This deterministic method depends solely on the determinist aspects of the problem such as the source directions and the multipath characteristics of the reverberant environment [15, 16]. Multiple sound sources are recorded simultaneously with the microphone array, which are then separated based on the analysis of intensity vector directions. The separated sources are then filtered with corresponding head related transfer functions (HRTFs) to obtain the binaural signals. Although the technique is used for binaural auralisation, it can be modified to work with multichannel loudspeaker systems.

This paper is organized as follows. In Section 2, the closed-form source separation technique is explained based on the formulation of the signals captured by a coincident array. Section 3 describes the processing of the separated channels for obtaining binaural signals. Section 4 details the experimental test conditions and provides the results of comparisons between the original and synthesized binaural room impulse responses. Section 6 concludes the paper.

2 Directional Separation

2.1 Intensity Vector Calculation

Four microphones closely spaced to form a plus sign on the horizontal plane can be used to obtain signals which are known as B-format signals, p_W, p_X, p_Y [17]. The p_W is similar to an omnidirectional microphone, and p_X and p_Y are similar to two bi-directional microphones that approximate pressure gradients along the X and Y directions, respectively.

In the time-frequency domain, the B-format signals can be written as the sum of plane waves coming from all directions:

$$p_W(\omega, t) \simeq \int_0^{2\pi} 2s(\theta, \omega, t) d\theta, \quad (1)$$

$$p_X(\omega, t) \simeq \int_0^{2\pi} j2kd \cos\theta s(\theta, \omega, t) d\theta, \quad (2)$$

$$p_Y(\omega, t) \simeq \int_0^{2\pi} j2kd \sin\theta s(\theta, \omega, t) d\theta. \quad (3)$$

where $s(\theta, \omega, t)$ is the pressure of a plane wave arriving from direction θ, k is the wave number related to the

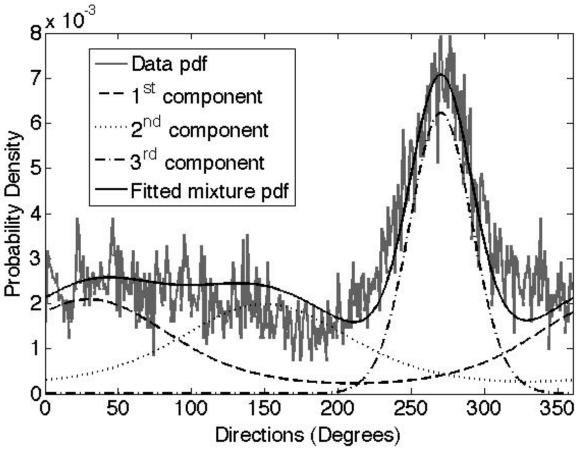

Figure 1: The probability density function of the intensity vector directions, individual mixture components and fitted mixtures for three sources at 30°, 150° and 270°.

wavelength λ as $k = 2\pi/\lambda$, j is the imaginary unit and $2d$ is the distance between the microphones.

Using these pressure signals, the direction of the intensity vector, $\gamma(\omega, t)$ can be calculated as [18]:

$$\gamma(\omega, t) = \arctan\left[\frac{Re\{p_W^*(\omega, t)p_Y(\omega, t)\}}{Re\{p_W^*(\omega, t)p_X(\omega, t)\}}\right]. \quad (4)$$

where $*$ denotes conjugation and $Re\{\bullet\}$ denotes taking the real part of the argument.

2.2 Spatial Filtering

For a single sound source at direction μ with respect to the array, the statistical distribution of the intensity vector directions can be modelled as von Mises for a circular random variable θ. Von Mises distribution is the circular equivalent of the Gaussian distribution which is observed due to the effect of reverberation [19].

$$f(\theta; \mu, \kappa) = \frac{e^{\kappa\cos(\theta-\mu)}}{2\pi I_0(\kappa)}, \quad (5)$$

where, $0 < \theta \leq 2\pi$, $0 \leq \mu < 2\pi$ is the mean direction, $\kappa > 0$ is the concentration parameter and $I_0(\kappa)$ is the modified Bessel function of order zero.

Figs. 2.2 shows the probability density functions of the intensity vector directions, individual mixture components and the mixture of von Mises functions for three sound sources fitted to the data by expectation maximisation, respectively. The sources are at 30°, 150° and 270°. The intensity vector directions were calculated for a 0.37 s recording at 44.1 kHz in a room with reverberation time of 0.83 s.

Th von Mises functions can be used for beamforming in the direction of μ, where κ is selected according to the desired beamwidth θ_{BW} of the spatial filter as

$$\kappa = \ln 2/\left[1 - \cos(\theta_{BW}/2)\right]. \quad (6)$$

Then, the signal corresponding to the estimate of the plane wave arriving from the direction μ is obtained by spatial filtering the pressure signals with this directivity function:

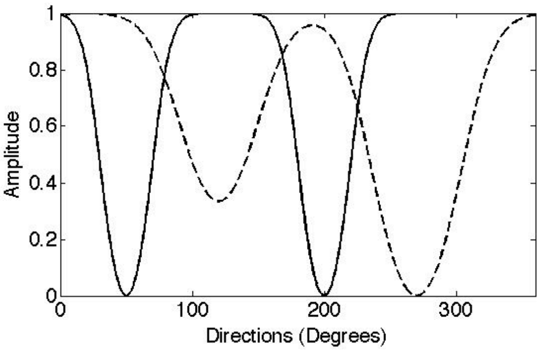

Figure 2: Two spatial filter examples based on von Mises functions for suppression of sounds at 50° and 200° with a beamwidth of 40° and at 120° and 270° with different suppression levels with a beamwidth of 70°.

$$\tilde{s}(\mu, \omega, t) = p_W(\omega, t)\, f\big(\gamma(\omega, t); \mu, \kappa\big). \quad (7)$$

2.3 Suppression of specific sounds

The separation of signals in all directions before the binaural processing enables modifications to the acoustic scene by removing some sounds. Unwanted sounds can be filtered out based on their directions using a spatial filter $g(\theta)$;

$$\tilde{s}_{new}(\mu, \omega, t) = \tilde{s}(\mu, \omega, t)g(\gamma(\omega, t)). \quad (8)$$

The level of suppression can also be chosen. Two example filters based on von Mises functions defined in Eq. (5) can be found in Fig. 2.3. The first filter suppresses sounds at 50° and 200° directions with a beamwidth of 40°. The second filter suppresses sounds at 120° and at 270° directions with a beamwidth of 70°, while the latter direction is suppressed more than the former.

3 Binaural Processing

For auralisation, the separated signals corresponding to the plane waves arriving from all directions need to be auralised. These signals are multiplied with the corresponding HRTFs in the frequency domain and summed to obtain the left ear and the right ear binaural signals, b_L and b_R, respectively:

$$b_L(\omega, t) = \frac{1}{2\pi}\int_0^{2\pi} \tilde{s}(\mu, \omega, t)h_L(\mu, \omega)d\mu \quad (9)$$

$$b_R(\omega, t) = \frac{1}{2\pi}\int_0^{2\pi} \tilde{s}(\mu, \omega, t)h_R(\mu, \omega)d\mu \quad (10)$$

where h_L and h_R are the left ear and right ear HRTFs in the frequency domain.

3.1 Head movements

Head movements can also be incorporated in this model. When the head rotates along its axis, the horizontal arrival directions of the direct sound and early reflections with respect to the listener also rotate. For a rotation of α degrees in the horizontal plane, the separated signals in Eqs (9) and (10) are replaced with

$$\tilde{s}_{new}(\mu, \omega, t) = \tilde{s}(\mu - \alpha, \omega, t). \qquad (11)$$

As the processing is done for each time-frequency block, compensation for head movements can easily be included in the applications.

3.2 Distortion

Due to the spatial filtering applied on each time-frequency block, the separated signals $\tilde{s}$ contain distortion, albeit to a limited extend. This distortion, however, is alleviated by binaural processing in Eqs (9) and (10) as the summation restores the missing time-frequency blocks. The suppression however, introduces additional distortion, which increases with increasing beamwidth.

As the distortion levels have been found to be very low, which were also confirmed by informal listening tests, no further investigation of the distortion levels were carried out.

4 Results

4.1 Test recordings

The recordings used in the testing of the algorithm were obtained by exploiting the linearity and time-invariance assumptions of the linear acoustics. The array recordings of convolutive mixtures were obtained by first measuring the B-format room impulse responses for individual sound sources, convolving anechoic sound sources with these impulse responses and summing the resulting reverberant recordings. Similarly, binaural recordings were obtained by first measuring binaural room impulse responses, convolving, anechoic sound sources with these and summing the results.

The impulse responses were measured in two different rooms. The first room was an ITU-R BS1116 standard listening room with a reverberation time of 0.32 s. The second one was a meeting room with a reverberation time of 0.83 s. Both rooms were geometrically similar (L = 8 m; W = 5.5 m; H = 3 m) and were empty during the tests.

For both rooms, impulse response recordings were obtained at 44.1 kHz both with a SoundField microphone system (SPS422B) and a Neumann KU100 dummy head at the same recording position using a loudspeaker (Genelec 1030A) and playing a 16th-order maximum length sequence (MLS) signal [20]. A set of binaural room impulse responses and B-format room impulse responses were obtained for six source directions of 0°, 60°, 120°, 180°, 240° and 300°. Each of the 6 measurement positions were located on a circle of 1.6 m radius for the first room, and 2.0 m radius for the second room. The recording points were at the center of the circles, and the frontal directions of the recording setup were fixed in each room. At each measurement position, the acoustical axis of the loudspeaker was facing towards the array location, while the orientation of the microphone system was kept fixed. The source and recording positions were 1.2 m high above the floor. The loudspeaker had a width of 20 cm, corresponding to the observed source apertures of 7.15° and 5.72° at the recording positions for the first and second rooms, respectively.

Anechoic sources sampled at 44.1 kHz were used from the Music for Archimedes CD [21]. The 5-second long portions of male English speech (M), female English speech (F), male Danish speech (D), cello music (C) and guitar music (G) sounds were first equalized for energy, then convolved with the impulse responses of the required directions and the recording setup. Combinations of different sound sources were then obtained by summing the results, which provided the binaural and array recordings of real acoustic environments containing multiple sound sources.

4.2 Preliminary listening test results

A small informal listening test was designed where two subjects were presented with synthesized and original binaural recordings. The number of sound sources in the recordings were ranging from three to five. The subjects were asked to comment on any differences on the perceived source locations between the synthesized and original recordings. As no differences were detected in the test runs, no further tests were carried out.

The subjects also mentioned the lower level of high frequencies in the synthesized recordings than the original recordings, which is due to the difficulty of calculating intensity vector directions accurately for high frequencies. The subjects could clearly identify the rooms where the recordings were made by listening to either the synthesized recordings, or the original recordings, indicating that the reverberant characteristics of the rooms were preserved in the synthesized recordings.

5 Conclusions

An algorithm based on the exploitation of intensity vector directions has been introduced for direct binaural coding of microphone array recordings. It has been shown that directional recordings provide detailed information about a sound field which can be used to synthesize BRIRs with inclusion of head rotation compensation. Analysis results are then used together with an HRTF database for synthesizing binaural recordings. The method also enables suppression of unwanted sounds by spatial filtering prior to binaural synthesis. Since the room impulse response characteristics and the spectral shaping of the pinnæ, head and torso are processed separately in the binaural synthesis, different HRTF databases or the individualized HRTFs can be employed to increase realism.

The method eliminates the need to make recordings with a mannequin or a human test subject. Comparisons of measured and synthesized binaural room impulse responses show that the method can be employed for virtual collaboration to provide immersive aural communication.

Future work will include the analysis and processing of elevated sources and reflections and will investigate reproductions on multichannel systems. The perceptual effects and artifacts will be determined by formal listening tests.

6 Acknowledgments

The work presented was developed within VISNET II, a European Network of Excellence, funded under the European Commission IST FP6 programme.

References

[1] U. P. Svensson and U. R. Kristiansen, "Computational modelling and simulation of acoustic spaces," in *Proc. of the* 22^{nd} *AES Conference on Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, June 2002, pp. 1–20.

[2] J. H. Rindel, "The use of computer modeling in room acoustics," *Journal of Vibroengineering*, vol. 3, no. 4, pp. 41–72, 2000.

[3] M. Kleiner, B. I. Dalenbäck, and P. Svensson, "Auralization - An overview," *J. Audio Eng. Soc.*, vol. 41, no. 11, pp. 861–875, November 1993.

[4] D. G. Malham, "Sound spatialisation," in *Proc. of the International Conference on Digital Audio Effects (DAFx-98)*, Barcelona, Spain, November 1998.

[5] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, April 1979.

[6] J. Vanderkooy, "Aspects of mls measuring systems," *J. Audio Eng. Soc.*, vol. 42, no. 4, pp. 219–231, April 1994.

[7] B. Günel, H. Hacıhabiboğlu, and A. M. Kondoz, "Wavelet-packet based passive analysis of sound fields using a coincident microphone array," *Appl. Acoust.*, vol. 68, no. 7, pp. 778–796, July 2007.

[8] M. Brandstein and D. Ward, Eds., *Microphone Arrays.* New York: Springer-Verlag, 2001.

[9] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[10] J.-F. Cardoso, "Blind source separation: statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, October 1998.

[11] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrainted adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, January 1982.

[12] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 352–362, September 2002.

[13] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[14] B. Günel, H. Hacıhabiboğlu, and A. M. Kondoz, "Acoustic source separation of convolutive mixtures based on intensity vector statistics," *IEEE Trans. Audio, Speech Language Process.*, vol. 16, no. 4, pp. 748–756, May 2008.

[15] A.-J. van der Veen, "Algebraic methods for deterministic blind beamforming," *Proc. IEEE*, vol. 86, no. 10, pp. 1987–2008, October 1998.

[16] J. Yamashita, S. Tatsuta, and Y. Hirai, "Estimation of propagation delays using orientation histograms for anechoic blind source separation," in *Proc. 2004 IEEE Int. Joint Conf. on Neural Networks*, vol. 3, Budapest, Hungary, July 2004, pp. 2175–2180.

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

[18] F. J. Fahy, *Sound Intensity*, 2nd ed. London: E&FN SPON, 1995.

[19] K. V. Mardia and P. Jupp, *Directional Statistics.* London and New York: Wiley, 1999.

[20] M. R. Schroeder, "Integrated-impulse method measuring sound decay without using impulses," *J. Acoust. Soc. Am.*, vol. 66, no. 2, pp. 497–500, August 1979.

[21] Bang & Olufsen, "Music for Archimedes," CD 101, 1992.