DIRECTION ESTIMATION BASED ON SOUND INTENSITY VECTORS

Sakari Tervo

Helsinki University of Technology
Department of Media Technology
P.O.Box 5400, FI-02015 TKK, Finland
sakari.tervo@tkk.fi

ABSTRACT

The direction of a sound source in an enclosure can be estimated with a microphone array and some proper signal processing. Earlier, in applications and in research the use of time delay estimation methods, such as the cross correlation, has been popular. Recently, techniques for direction estimation that involve sound intensity vectors have been developed and used in applications, e.g. in teleconferencing. Unlike in time delay estimation, these methods have not been compared widely. In this article, five methods for direction estimation in the concept of sound intensity vectors are compared with real data from a concert hall. The results of the comparison indicate that the methods that are based on convolutive mixture models perform slightly better than some of the simple averaging methods. The convolutive mixture model based methods are also more robust against additive noise.

1. INTRODUCTION

Direction or location of a sound source is of interest in several applications that aim to capture or to reproduce the sound [1, 2, 3]. This is of interest for example in teleconferencing [3]. Moreover, when exploring concert hall impulse responses it is of interest from where and when the direct sound and the so-called early reflections arrive [4].

Direction estimation with traditional methods, such as the cross correlation, has been researched widely over several decades (see e.g. [5] and references within). Nowadays, sound intensity vectors are used for direction estimation in increasingly many applications [1, 2, 3, 6]. Not much research, if any, has been published in comparing the different methods used in direction estimation from sound intensity vectors. In this article, five direction estimation methods are compared in a real concert hall environment.

A sound intensity vector has a length and a direction and it is a function of time and frequency. In this work, the focus is on the direction estimation over some set of frequencies during a certain time frame. Direction estimation methods can be broadly classified into two classes: 1) direct 2) and mixture estimation. The first class includes methods such as averaging. The second class contains convolutive mixture models where the direction of the sound source is found by fitting two or more probability distribution of a certain shape to the directional data, as for example in [2] and [7]. Naturally, the methods in the second class are much more complex than in the first, since some optimization method has to be used to obtain the mixtures.

Generally, sound intensity vectors are obtained either with microphone pair measurements or from B-format signals [4]. In practice both of these measurement techniques introduce a bias to the direction of the intensity vector. For example, in Soundfield microphones, the bias is caused by the non-idealities in the directivity patterns of the microphones [4]. In microphone pair measurements the direction estimation is biased, since the gradient of the sound pressure is not constant within the sensor array [6]. In the case of microphone pair measurement, also the non-idealities of the pressure microphones affect the measurement.

The article is organized as follows. In Sec. 2 the calculation of sound intensity vectors as well the bias compensation of the vectors are formulated. Section 3 introduces the direction estimation methods. In Sec. 4 the experimental setup is described and in Sec. 5 the results from the experiments are presented and discussed. In Sec. 6 the final conclusions of this study are given.

2. THEORY

In a room environment the sound $s(t)$ traveling from the sound source to the receiver n is affected by the path $h_n(t)$:

$$p_n(t) = h_n(t) * s(t) + w(t), \tag{1}$$

where $*$ denotes convolution and $w(t)$ is measurement noise, independent and identically distributed for each receiver.

2.1 Sound Intensity

On a certain axis a, the sound intensity is given in the frequency domain as

$$I_a(\omega) = \mathrm{Re}\{P^*(\omega)U_a(\omega)\}, \tag{2}$$

where $P(\omega)$ and $U_a(\omega)$ are the frequency presentations of the sound pressure and of the particle velocity with angular frequency ω. In addition, $\mathrm{Re}\{\cdot\}$ and is the real part of a complex number and * denotes the complex conjugate [4]. Here, Cartesian coordinate system with x -and y coordinate axis, shown in Fig. 1, is used. Corresponding polar coordinate presentation of the Cartesian coordinates is denoted with azimuth angle θ and radius r. The procedure for obtaining the sound intensity proceeds as follows and is similar for both axes.

The sound pressure at the center point between four microphones, shown in Fig. 1, can be approximated as

the average pressure of the microphones [6]:

$$P(\omega) \approx \frac{1}{4}\sum_{n=1}^{4} P_n(\omega). \quad (3)$$

For x-axis, in frequency domain the particle velocity is estimated as

$$U_{\mathrm{x}}(\omega) \approx \frac{-j}{\omega\rho_0 d}[P_1(\omega) - P_2(\omega)], \quad (4)$$

where d is the distance between the two receivers and $\rho_0 = 1.2$ kg/m^3 is the median density of the air and j is the imaginary unit.

Now, the sound intensity in (2) can be estimated with the approximations in (3) and (4). For obtaining the y-component of the sound intensity, the microphones 1 and 2 are replaced in (4) with microphones 3 and 4.

2.2 Bias Compensation for the Square Grid

Since the pressure signals are subtracted from each other in the approximation made in (4), the azimuth angle θ of the intensity suffers from a systematic bias [6]. This bias can be formulated for a four-microphone square grid as [6]

$$\theta_{\mathrm{biased}} = \frac{\sin(\omega\frac{d}{2c}\sin(\theta))}{\sin(\omega\frac{d}{2c}\cos(\theta))}, \quad (5)$$

where $c = 343$ m/s is the speed of sound. The bias can be compensated by finding the inverse of (5). However, this equation does not have any closed form solution. Kallinger *et al.* [6] estimate the inverse function with linear interpolation. In addition, the inverse function is known [6] to have an upper limit at $f_{\max} = c/(d\sqrt{2})$, where $\omega = 2\pi f$, i.e.

$$0 < \omega\frac{d}{2c} < \frac{\pi}{\sqrt{2}}.$$

In this work the linear interpolation for finding the solution to the inverse function of (5) is included and used for compensating the azimuth angles. The unbiased estimate, i.e. the compensated angle at i:th frequency bin is denoted with θ_{i}.

3. METHODS FOR DIRECTION ESTIMATION

In a certain time window, sound intensity vectors are estimated over a set of frequencies. Each intensity vector at frequency bin i, consists of a radial component r_{i} and of an angular component θ_i. In this article, the focus is on finding the direction of a sound source from a set of radial and angular components. Next, five methods for direction estimation are formulated.

3.1 Circular Mean and Median

In direction estimation one has to take into account the fact that the data is circular. Therefore, for example the mean of a set of angles $\theta = [\theta_1, \theta_2, ..., \theta_N]$, $\theta_i \in (-\pi, \pi]$ is defined as the circular mean (CME) [8]:

$$\hat{\theta}_{\mathrm{CME}} = \arg\left\{\sum_{\mathrm{i}=1}^{N} w_{\mathrm{i}} e^{j\theta_{\mathrm{i}}}\right\}, \quad (6)$$

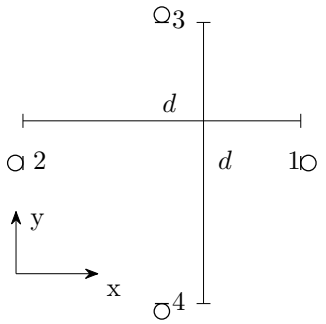

Figure 1: The square grid with four microphones and the coordinate system.

where $\arg\{\cdot\}$ is the argument of a complex number, and $w_{\mathrm{i}} = 1/N$ is the weighting function. In principle, the weighting function can be chosen freely. If the weighting function is chosen as $w_{\mathrm{i}} = r_{\mathrm{i}}$, with $\sum r_{\mathrm{i}} = 1$, where r_i is the corresponding radial component of an intensity vector, then CME is equal to the mean of the Cartesian presentation (MCA) of the sound intensity, i.e.:

$$\arg\left\{\sum_{\mathrm{i}=1}^{N} r_{\mathrm{i}} e^{j\theta_{\mathrm{i}}}\right\} = \arctan\left\{\frac{\mathrm{E}\{\mathbf{I}_{\mathrm{y}}\}}{\mathrm{E}\{\mathbf{I}_{\mathrm{x}}\}}\right\} := \hat{\theta}_{\mathrm{MCA}}, \quad (7)$$

where $\mathrm{E}\{\cdot\}$ is the expected value, $\mathbf{I}_{\mathrm{x}} = [x_1, x_2, ..., x_N]$, and $\mathbf{I}_{\mathrm{y}} = [y_1, y_2, ..., y_N]$ are the Cartesian presentation of the vectors.

Here, the circular median (CMD) is defined analogously to CME:

$$\hat{\theta}_{\mathrm{CMD}} = \arg\left\{\mathrm{Me}\Big\{\mathrm{Re}\{w_{\mathrm{i}} e^{j\theta_{\mathrm{i}}}\}\Big\} + j\mathrm{Me}\Big\{\mathrm{Im}\{w_{\mathrm{i}} e^{j\theta_{\mathrm{i}}}\}\Big\}\right\}, \quad (8)$$

where $\mathrm{Me}\{\cdot\}$ is the median. The mode could also be used instead of the median, but this is not considered in this article. Again, the weighting function can be selected freely. Here, $w_{\mathrm{i}} = 1/N$ is selected as earlier with CME.

3.2 Mixture Models

When inspecting the distribution of the azimuth angle of the intensity vectors, for example in Fig. 2, one can notice that the azimuth angle is a mixture of two distributions rather than one. In the example in Fig. 2 the distribution that has smaller variance (or higher concentration) is caused by the sound source. The second distribution models the noise floor. The shape of these distributions is defined by the impulse response, i.e. the room and the frequency content of the source signal. A new sound source in the room always introduces a new distribution to the total azimuth distribution [2, 7].

Since the azimuth angle is circular, wrapped distributions have to be used for the fitting. Here, two distributions are tested. The first one is the von Mises probability distribution (VM) [2, 8]:

$$f_{\mathrm{VM}}(\theta|\mu,\kappa) = \frac{e^{\kappa\cos(x-\mu))}}{2\pi\mathcal{I}(0,\kappa)}, \quad (9)$$

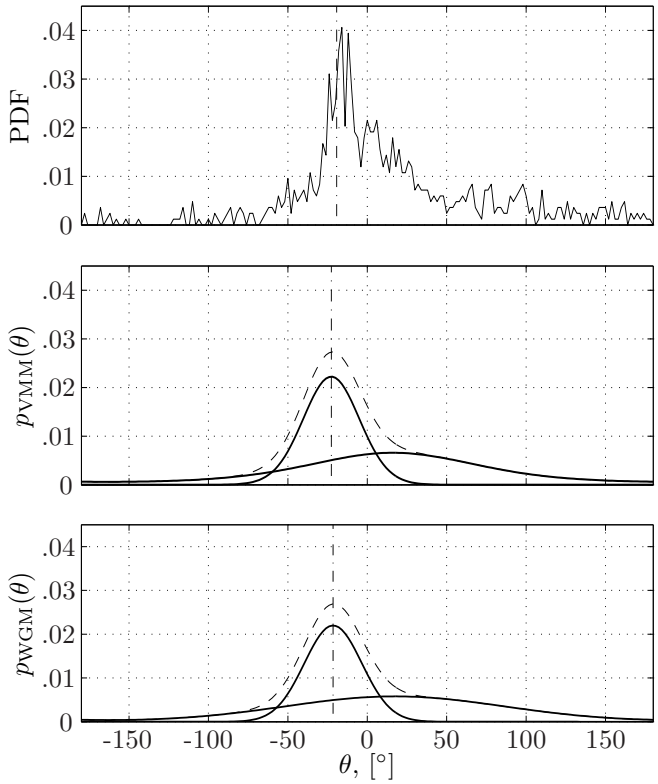

Figure 2: Examples of the normalized histogram of the estimated azimuth angles (top, –), the von Mises mixture model (middle, - -), and the wrapped Gaussian mixture model (bottom, - -), with mixture components shown separately (–). The vertical lines (–·) illustrate the true angle $\theta_{\mathrm{s}} = -19°$ (top), and estimated source directions $\hat{\theta}_{\mathrm{VMM}} = -23°$(middle), $\hat{\theta}_{\mathrm{WGM}} = -22°$ (bottom). Source position S3 and receiver position R2 was used (see Sec. 4).

where κ is a measure of concentration, μ is the mean, and $\mathcal{I}(0, \kappa)$ is the modified Bessel function of order 0. The second tested distribution is the wrapped Gaussian probability distribution (WG)[9]:

$$f_{\mathrm{WG}}(\theta|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{k=-K}^{K} e^{\frac{-(\theta-\mu-2\pi k)^2}{\sigma^2}}, \qquad (10)$$

where σ^2 is the variance, μ is the mean, and $2K+1$ is the number of Gaussian to be wrapped, here $K = 2$.

The mixture model is formulated as the sum of the distributions:

$$p(\theta|\boldsymbol{\mu},\boldsymbol{\rho}) = \sum_{\mathrm{m}\,=\,1}^{\mathrm{M}} a_{\mathrm{m}} f^{(\mathrm{m})}(\theta|\mu_{\mathrm{m}},\rho_{\mathrm{m}}), \qquad (11)$$

where, m indicates the index of a mixture, $f = f_{\mathrm{WG}}$ for the wrapped Gaussian mixture model (WGM), and $f = f_{\mathrm{VM}}$ for the von Mises mixture model (VMM). Parameters $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_{\mathrm{M}}]$ and $\boldsymbol{\rho} = [\rho_1, \rho_2, ..., \rho_{\mathrm{M}}]$ depend also on the model. The weighting factor is here selected as $a_{\mathrm{m}} = 1/\mathrm{M}$ and the number of mixtures is $\mathrm{M} = 2$, since there is only a single sound source. Thus, one of the mixture components models the noise floor and the other one the sound source. The direction of the source is then estimated as the mean parameter μ of the mixture component which has smaller variance (WGM) or higher concentration (VMM). These estimated directions are noted here with $\hat{\theta}_{\mathrm{WGM}}$and $\hat{\theta}_{\mathrm{VMM}}$ according to the used model.

In order to find the parameters of the mixture model, an optimization algorithm has to be used. The search criteria is the same as in maximum-likelihood estimation, where the goal is to maximize the likelihood

$$L(\boldsymbol{\mu},\boldsymbol{\rho}) = \sum_{\mathrm{i}=1}^{N} \log p(\theta_{\mathrm{i}}|\boldsymbol{\mu},\boldsymbol{\rho}). \qquad (12)$$

Here, the parameters are sought with MATLAB's `fminsearch` which uses the Nelder-Mead method [10]. Other optimization algorithms that are more optimal for this problem could be used as well, e.g. the expectation-maximization algorithm. Figure 2 shows examples of distributions estimated with both of the models and an example of the original probability distribution function (PDF), i.e. the normalized histogram of the estimated directions.

4. EXPERIMENTAL SETUP

A microphone array consisting of two four-microphone grids (see Fig. 1) was used. The four-microphone grids share the same center point, having $d = 10$ mm and $d = 100$ mm. The smaller grid is used for the frequencies from 1 kHz to 5 kHz and the larger for frequencies from 100 Hz to 1 kHz. The bias compensation is done separately for both arrays.

The methods were tested with real concert hall data. The impulse response measurement setup in the concert hall (Pori, located in Finland) is given in Fig. 3. This concert hall has 700 seats and a reverberation time of approximately 2.1 seconds. The signals were recorded at 48 kHz with three sound source locations (S), and five receiving (R) locations for the array. Here, the first three receiver positions are used (R1-R3 in Fig. 3). The sound source was an omnidirectional loudspeaker of 26 cm diameter, consisting of 12 driver elements.

Two source signals, 2 seconds of violin playing a signal tone and 2 seconds of white noise, were convolved with the measured impulse responses. Then, signal-to-noise ratio (SNR) was varied by adding white noise ($w(t)$ in (1)) to the signals. Next, sound intensity vectors were calculated and compensated as stated in Sec. 2. The direction was estimated from the intensity vectors with the methods introduced in Sec. 3 on a frame by frame basis. Frames with 1024 samples in length and 50 % overlap were used. For each method and test condition, this leads to $9 \times 187 = 1683$ direction estimates. Moreover, 8192 bins was used in the fast Fourier transform. This implies that the direction was estimated from 837 frequency bins, since the frequency band was from 100 to 5 kHz.

5. RESULTS AND DISCUSSION

In order to compare the methods, the estimated directions are processed as follows. Firstly, the anomalies are

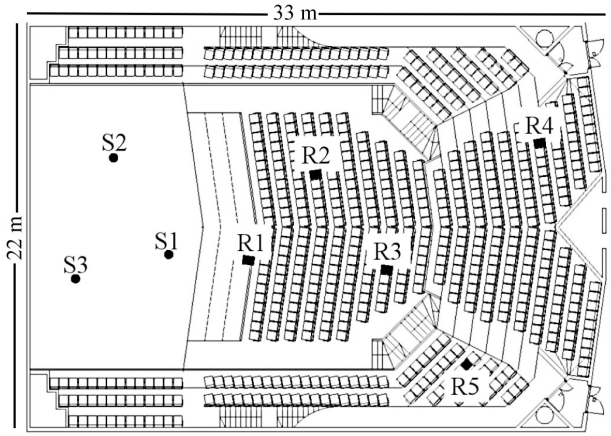

Figure 3: Receiver (R) and source (S) positions in the concert hall of Pori. Receiver positions R1, R2, and R3, and source positions S1, S2, and S3 were used in the experiments.

removed from the direction estimates. Here, the threshold criteria for an anomaly is 30°, i.e. if $|\hat{\theta}_\mathrm{i} - \theta_\mathrm{s}| > 30°$, where θ_s is the true angle of the sound source, the estimate $\hat{\theta}_\mathrm{i}$ is considered to be an anomaly. The goodness of a method is evaluated with three measures. The first measure is the percentage of anomalies P_AN, the second measure is the circular bias of the non-anomalous estimates $\tilde{\theta}_\mathrm{i}$:

$$\mu_\theta = \left| \arg \left\{ \sum_{\mathrm{i}=1}^{\tilde{N}} e^{j(\tilde{\theta}_\mathrm{i} - \theta_\mathrm{s})} \right\} \right|$$

and the third is the circular standard deviation:

$$\sigma_\theta = \left[2 \left(1 - \frac{1}{\tilde{N}} \left\| \sum_{\mathrm{i}=1}^{\tilde{N}} e^{j(\tilde{\theta}_\mathrm{i} - \theta_\mathrm{s})} \right\| \right) \right]^{1/2},$$

where $\tilde{N}$ is the number of non-anomalous estimates, $|\cdot|$ is the absolute value of a real number and $||\cdot||$ is the length of a complex number.

The results of the experiments for a white noise source signal are presented in Fig. 4 and for a violin source signal in Fig. 5. As expected, since white noise has energy in all frequencies it provides more robustness against additive noise and gives more accurate direction estimation results than a violin source signal.

As one can see from Figs. 4 and 5, the mixture model based estimation methods, VMM and WGM, perform the best in all conditions, VMM having lowest number of anomalies in total, and therefore performing the best of all the methods. The best estimator from the first class (introduced in Sec. 3.1) is CME. However, the differences in the performance between the mixture model based estimation methods and two simple averaging methods, CME, and CMD, are not drastical. The worst estimator in all conditions is clearly MCA. Even with the highest SNR the percentage of anomalies is high ($P_\mathrm{AN} \approx 38$ %).

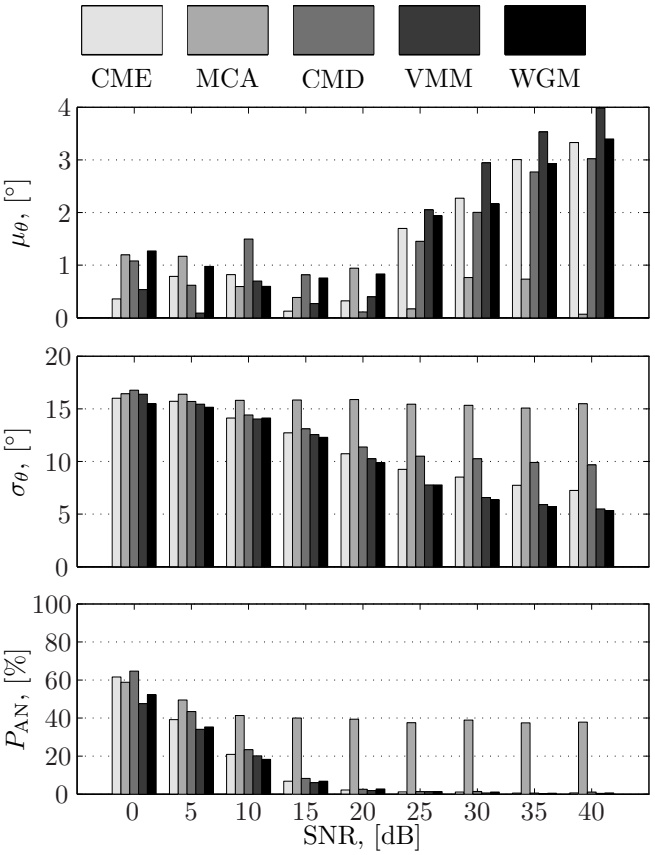

Figure 4: Bias (top) and standard deviation (middle) of the non-anomalous estimates, as well as the percentage of anomalies (bottom) against SNR. White noise was used as the source signal and the reverberation time is about 2.1 seconds.

VMM is the most robust method against additive noise. In Fig. 5, VMM has less than 50 % of anomalies in all the test conditions. Thus, even with high noise level VMM is able to find the distribution caused by the sound source. One reason for this might be that, although the total energy of the noise is higher than the energy of the sound source, the distribution caused by the noise has much smaller concentration than the distribution caused by the sound source.

The bias of all methods is less than 4 ° in all cases. With noise source the bias increases in general when the SNR increases. Thus, the error distribution is biased more with high SNR. The reasons behind this unexpected behaviour are not clear. When the SNR is high the bias is introduced by the reverberation. Perhaps, this is caused by the non-idealities in the microphones. However, with violin source signal the trend of the bias is the opposite. The general trend of the standard deviation is that it decreases when the SNR increases, as expected.

As the results indicated, MCA is not a good method to estimate the direction of arrival from a continuous signal. MCA gives very often anomalous estimates. This is caused by the weighting with the radial components, since when the weighting is not used, as in CME, the

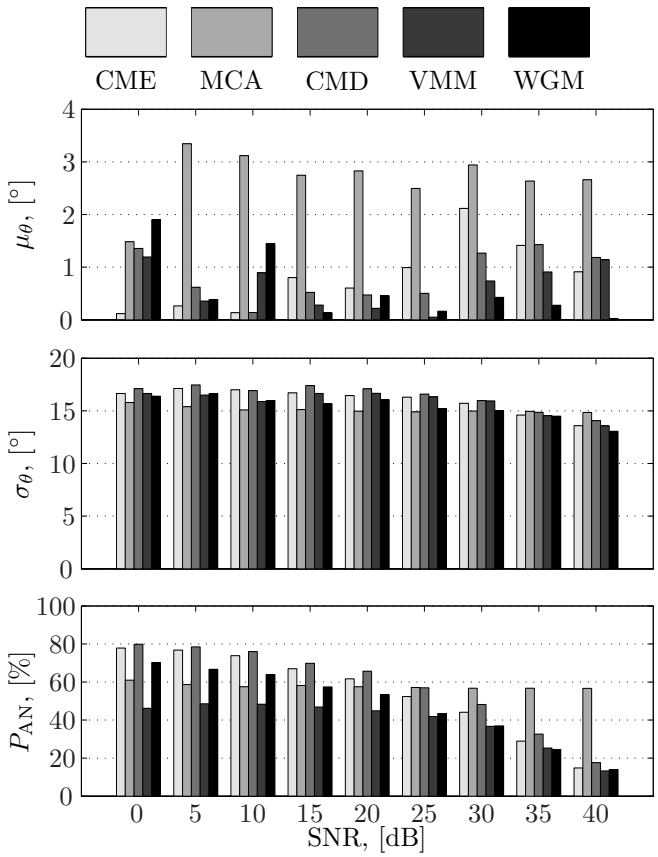

Figure 5: Bias (top) and standard deviation (middle) of the non-anomalous estimates, as well as the percentage of anomalies (bottom) against SNR. The source signal was violin and the reverberation time is about 2.1 seconds.

estimation is close to the actual direction. So, at least in thecase of circular mean it is not a good idea to use the radial components as the weighting function. Perhaps, if one would select only a certain set of the azimuth and of the radial components one would arrive to better result. Also, one could use a maximum-likelihood weighting (with respect to the spectrums of the signal and the noise) for the cross spectral components, as in time delay estimation [5].

6. CONCLUSIONS

Five methods for estimating the direction of a sound source from a set of sound intensity vectors was considered. There were two classes of methods in test. First class includes simple methods such as the circular mean, and the second class contains methods that are based on convolutive mixture models. The methods were tested in a real concert hall environment. The results indicate that the methods from the second class perform better and provide more robustness against additive noise than the methods from the first class. Especially, the von Mises distribution was found to suit well for the problem.

ACKNOWLEDGEMENTS

The author wishes to thank Dr. Tapio Lokki for the supervision of this work. The research leading to these results has received funding from the Academy of Finland, project no. [119092], the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. [203636], and Helsinki Graduate School in Computer Science and Engineering.

References

[1] J. Merimaa and V. Pulkki, "Spatial Impulse Response Rendering I: Analysis and Synthesis," *J. Audio Eng. Soc.*, vol. 53, no. 12, pp. 1115–1127, 2005.

[2] B. Günel, H. Hacıhabiboğlu, and A.M. Kondoz, "Acoustic Source Separation of Convolutive Mixtures Based on Intensity Vector Statistics," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 748–756, 2008.

[3] V. Pulkki, "Spatial Sound Reproduction with Directional Audio Coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, 2007.

[4] J. Merimaa, *Analysis, Synthesis and Perception of spatial sound-Binaural Auditory modeling and multichannel loudspeaker reproduction*, Ph.D. thesis, Helsinki University of Technology, 2006.

[5] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP J. Applied Signal Processing*, vol. 2006, no. 1, pp. 170–170, 2006.

[6] M. Kallinger, F. Kuech, R. Schultz-Amling, G. del Galdo, J. Ahonen, and V. Pulkki, "Enhanced Direction Estimation Using Microphone Arrays for Directional Audio Coding," in *Proc. Hands-Free Speech Communication and Microphone Arrays*, 2008, pp. 45–48.

[7] N. Madhu and R. Martin, "A Scalable Framework for Multiple Speaker Localization and Tracking," in *Proc. Int. Workshop on Acoustic Echo and Noise Cancellation*, 2008.

[8] N.I. Fisher, *Statistical Analysis of Circular Data*, New York: Cambridge University Press, 1993.

[9] Y. Agiomyrgiannakis and Y. Stylianou, "Stochastic Modeling and Quantization of Harmonic Phases in Speech using Wrapped Gaussian Mixture Models," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2007, vol. 4, pp. 1121–1124.

[10] J.A. Nelder and R. Mead, "A Simplex Method for Function Minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308, 1965.