A MAXIMUM LIKELIHOOD APPROACH FOR UNDERDETERMINED TDOA ESTIMATION

Janghoon Cho and Chang D. Yoo

Korea Advanced Institute of Science and Technology
Department of Electrical Engineering
LG-Hall 2106, 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Republic of Korea

ABSTRACT

This paper considers the estimation of time difference of arrival (TDOA) of multiple sparse sources when the number of sources is larger than that of the microphones. White Gaussian noise is assumed present at the microphone in addition to the instantaneously mixed sources. The TDOA estimate is obtained based on a maximum likelihood (ML) criteria, and the likelihood is obtained by marginalizing the joint probability over the sources. Explicit marginalization is mathematically intractable, thus the joint probability is approximated as a summation of several Dirac delta functions by assuming the time-frequency component of the source distribution to be a complex-valued super Gaussian, and the global maximum point of the marginalized joint probability is found by Markov chain Monte Carlo sampling. Experimental results show that the proposed algorithm outperforms TDOA estimation using a well-known Gaussian based approximation method in terms of root-mean-square error (RMSE).

Index Terms— Multiple sound source localization, underdetermined TDOA estimation, blind source separation

1. INTRODUCTION

Multiple sound source localization using microphone array has received considerable attention for its potential applications in audio/speech enhancement and recognition, wireless data communication, object tracking and radar signal detection. From the TDOA between a pair of microphones, the directions of sound sources can be easily obtained [1]. This paper considers maximum likelihood estimation of TDOA when the number of sources is more than the number of microphones. This case is often referred to as the underdetermined case.

Various algorithms for TDOA estimation have been studied. One of the most popular methods is based on independent component analysis (ICA) [2, 3]; however, ICA can be used only when the number of sources is less than the number of microphones [4]. Numerous clustering-based algorithms assuming disjoint orthogonality in the time-frequency (TF) domain have been proposed to tackle the underdetermined TDOA problem; however, when you consider the collapse of the disjoint orthogonality assumption in many realistic situations, the limitation of these algorithms in producing accurate estimate of TDOA is not difficult to imagine.

This paper considers a maximum likelihood criterion for estimating the TDOA in the underdetermined case. The joint probability of the mixing parameters, observations and source signals is approximated by a summation of Dirac delta functions assuming sparsity of the sources and then marginalized over the sources. The global optimum point of marginalized joint probability is found by Markov chain Monte Carlo sampling.

This paper is organized as follows. Section 2 describes the problem formulation of underdetermined TDOA estimation. Section 3 presents the proposed maximum likelihood estimation algorithm. Section 4 provides experimental results, and finally Section 5 concludes the paper.

2. PROBLEM FORMULATION

2.1. Problem description

Let $y_{j,t}$ be the observation signal from the j-th microphone at time t and $s_{i,t}$ be the i-th source signal for $j = 1, \ldots, J$ and $i = 1, \ldots, I$, where J and I are the number of sources and microphones, respectively. By anechoic mixing model, $y_{j,t}$ can be described as

$$y_{j,t} = \sum_{i=1}^{I} \lambda_{ji} s_{i,t-t_{ji}} + n_{j,t}, \quad j = 1, \cdots, J \tag{1}$$

where λ_{ji}, t_{ji} and $n_{j,t}$ are the attenuation, the time delay from the i-th source to the j-th microphone and the noise sensed by the j-th microphone, respectively. Here, I is assumed to be more than J ($I > J$).

By applying short-time discrete Fourier transform (STDFT) to the time domain signal $y_{j,t}$ with sampling rate f_s, the following is obtained

$$Y_{j,m}^{(k)} = \sum_{i=1}^{I} H_{ji}^{(k)} S_{i,m}^{(k)} + N_{j,m}^{(k)}, \quad k = 0, \cdots, K-1, \tag{2}$$

and

$$H_{ji}^{(k)} = \lambda_{ji} e^{-j\frac{2\pi k}{K}\tau_{ji}}, \tag{3}$$

where $S_{i,m}^{(k)}$ and $N_{j,m}^{(k)}$ are the complex-valued STDFT coefficients of $s_{i,t}$ and $n_{j,t}$, respectively. Here, $\tau_{ji} = t_{ji} f_s$.

To express the mixing parameters $H_{ji}^{(k)}$ with the TDOA for the i-th source at the chosen microphone pair (j, j'), the ratio component $\tilde{H}_{ji}^{(k)}$ is given as

$$\tilde{H}_{ji}^{(k)} = \frac{H_{ji}^{(k)}}{H_{j'i}^{(k)}} = \frac{\lambda_{ji}}{\lambda_{j'i}} e^{-j\frac{2\pi k}{K}\Delta\tau_i^{(j,j')}}, \quad (4)$$

where $\Delta\tau_i^{(j,j')} = \tau_{ji} - \tau_{j'i}$ for $j \neq j'$, and j' is the index of reference microphone. Then, the Equation (2) can be simply rewritten as vector-matrix form as follows :

$$\mathbf{Y}_m^{(k)} = \tilde{\mathbf{H}}^{(k)} \tilde{\mathbf{S}}_m^{(k)} + \mathbf{N}_m^{(k)}, \quad (5)$$

where $\mathbf{Y}_m^{(k)} = \left[Y_{1,m}^{(k)}, \cdots, Y_{J,m}^{(k)}\right]^T$, $\tilde{\mathbf{H}}^{(k)} = [\tilde{H}_{ji}^{(k)}]_{J\times I}$, $\tilde{\mathbf{S}}_m^{(k)} = \left[\tilde{S}_{1,m}^{(k)}, \cdots, \tilde{S}_{I,m}^{(k)}\right]^T$, $\tilde{S}_{i,m}^{(k)} = H_{j'i}^{(k)} S_{i,m}^{(k)}$ is the spatial image of the i-th source on the j'-th microphone, and $\mathbf{N}_m^{(k)} = \left[N_{1,m}^{(k)}, \cdots, N_{J,m}^{(k)}\right]^T$. Here, the vector of the TDOA for the sources at all microphone pairs including the reference microphone can be given as

$$\mathbf{d} = \left[\Delta\tau_1^{(1,j')}, \cdots, \Delta\tau_I^{(1,j')}, \cdots, \Delta\tau_1^{(J,j')}, \cdots, \Delta\tau_I^{(J,j')}\right]^T, \quad (6)$$

and it should be estimated by a maximum likelihood approach using only microphone observations $\mathbf{Y}^{(k)}$.

2.2. Assumptions

This paper makes three assumptions about the sources :

1. The sources are mutually independent of one another such that

$$p(\tilde{S}_1^{(k)}, \cdots, \tilde{S}_I^{(k)}) = \prod_{i=1}^{I} p(\tilde{S}_i^{(k)}). \quad (7)$$

2. The magnitude of $\tilde{S}_{i,m}^{(k)}$ follows complex-valued super Gaussian distribution and the phase of $\tilde{S}_{i,m}^{(k)}$ is assumed to be uniformly distributed in $[-\pi, \pi)$ as follows :

$$p(|\tilde{S}_{i,m}^{(k)}|) = c\frac{\beta^{1/c}}{\Gamma(1/c)} e^{-\beta|\tilde{S}_{i,m}^{(k)}|^c}, \quad (8)$$

$$p(\angle\tilde{S}_{i,m}^{(k)}) = \frac{1}{2\pi} \quad (9)$$

where the parameters c and β are the shape and the scale of the distribution, respectively. Also, $c, \beta > 0$. When $c = 1$, the distribution is Laplacian and when $c = 2$, it is Gaussian. With decreasing c, sparsity increases. In this paper, the source prior is assumed to follow the super-Gaussian, $c < 1$. Note that the sequences of the source coefficients are independent and identically distributed.

3. Here, $\mathbf{N}_m^{(k)}$ is assumed to be a complex-valued independent and identically distributed Gaussian noise with zero mean and covariance $\sigma^2\mathbf{I}_J$, where $\mathbf{I}_J$ denotes the $J \times J$ identity matrix.

2.3. Objective

The objective is to estimate TDOA $\mathbf{d}$ via ML criteria as follows :

$$\hat{\mathbf{d}}_{\text{ML}} = \arg\max_{\mathbf{d}} p(\mathbf{Y}|\mathbf{d}). \quad (10)$$

3. PROPOSED ALGORITHM

3.1. Joint probability

For achieving the objective mentioned in Section 2.3, the joint probability of random variables $\tilde{\mathbf{H}}, \tilde{\mathbf{S}}, \mathbf{Y}$[1] should be defined. It can be expressed as follows :

$$p(\tilde{\mathbf{H}}, \tilde{\mathbf{S}}, \mathbf{Y}) = p(\tilde{\mathbf{H}}) p(\tilde{\mathbf{S}}) p(\mathbf{Y}|\tilde{\mathbf{H}}, \tilde{\mathbf{S}}), \quad (11)$$

where, $p(\tilde{\mathbf{H}})$, $p(\tilde{\mathbf{S}})$ are prior of $\tilde{\mathbf{H}}$, $\tilde{\mathbf{S}}$ and $p(\mathbf{Y}|\tilde{\mathbf{H}}, \mathbf{S})$ is likelihood. A super-Gaussian prior on $\tilde{\mathbf{S}}$ is assumed as mentioned in Section 2.2. Under the white Gaussian noise assumption, the likelihood can be written as

$$p(\mathbf{Y}|\tilde{\mathbf{H}}, \tilde{\mathbf{S}}) = \prod_{k=0}^{K-1} \prod_{m=1}^{M} \mathcal{N}(\mathbf{Y}_m^{(k)}|\tilde{\mathbf{H}}^{(k)}\tilde{\mathbf{S}}_m^{(k)}, \sigma^2\mathbf{I}_J), \quad (12)$$

where $\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma})$ denotes the multivariate Gaussian distribution with a mean vector μ and a covariance matrix $\mathbf{\Sigma}$.

3.2. Marginal likelihood approximation

Since the mixing matrix $\tilde{\mathbf{H}}$ is a function of $\mathbf{d}$, the objective function in Equation (13) can be rewritten as

$$\hat{\mathbf{d}}_{\text{ML}} = \arg\max_{\mathbf{d}} p(\mathbf{Y}|\tilde{\mathbf{H}}). \quad (13)$$

With predefined joint probability in Equation (11), the objective function, the marginal likelihood can be computed as

$$p(\mathbf{Y}|\tilde{\mathbf{H}}) = \int p(\tilde{\mathbf{S}}) p(\mathbf{Y}|\tilde{\mathbf{H}}, \tilde{\mathbf{S}}) d\tilde{\mathbf{S}}. \quad (14)$$

Since $p(\tilde{\mathbf{S}})p(\mathbf{Y}|\tilde{\mathbf{H}}, \tilde{\mathbf{S}})$ is very complicated in underdetermined case, the integration in the Equation (14) is intractable. In [7], $p(\tilde{\mathbf{S}})p(\mathbf{Y}|\tilde{\mathbf{H}}, \tilde{\mathbf{S}})$ is approximated as a multivariate Gaussian around the posterior mode. This posterior mode can be solved by maximizing a posteriori (MAP) criteria using various algorithms. When $\tilde{\mathbf{S}}$ is real-valued, linear programming [8] or shortest path [9] algorithm can be used, and when $\tilde{\mathbf{S}}$ is complex-valued, second order cone programming

[1] Henceforth, the superscript (k) and the subscript m will be omitted for simplicity.

[10] or combinatorial algorithm [11] can be adopted. However, when $c < 1$, $p(\tilde{\mathbf{S}})p(\mathbf{Y}|\tilde{\mathbf{H}}, \tilde{\mathbf{S}})$ has ${}_I\mathrm{C}_J$ non-differentiable local maximums and these maximums are located at ${}_{(b)}\tilde{\mathbf{S}}$ for $b = 1, \cdots, {}_I\mathrm{C}_J$ where $(I - J)$ components of $\tilde{\mathbf{S}}$ are zeros where $\mathbf{Y} = \tilde{\mathbf{H}}\tilde{\mathbf{S}}$[2]. Here, ${}_{(b)}\tilde{\mathbf{S}}$ can be obtained by multiplying the inverse of the $J \times J$ submatrix of $\tilde{\mathbf{H}}$ to $\mathbf{Y}$. Rather than approximating $p(\tilde{\mathbf{S}})p(\mathbf{Y}|\tilde{\mathbf{H}}, \tilde{\mathbf{S}})$ as a single multivariate Gaussian, the approximation considering non-differentiable local modes is used. In [12], $p(\tilde{\mathbf{S}})p(\mathbf{Y}|\tilde{\mathbf{H}}, \tilde{\mathbf{S}})$ is approximated as a summation of the Dirac delta functions, and those Dirac delta functions are located at ${}_{(b)}\tilde{\mathbf{S}}$ for $b = 1, \cdots, {}_I\mathrm{C}_J$,

$$p(\tilde{\mathbf{S}})p(\mathbf{Y}|\tilde{\mathbf{H}}, \tilde{\mathbf{S}}) \approx \sum_{b=1}^{{}_I\mathrm{C}_J} p(\tilde{\mathbf{S}} = {}_{(b)}\tilde{\mathbf{S}})\delta(\tilde{\mathbf{S}} - {}_{(b)}\tilde{\mathbf{S}}). \quad (15)$$

Substituting Equation (15) into Equation (14), the marginal likelihood can be written as

$$p(\mathbf{Y}|\tilde{\mathbf{H}}) \approx \prod_{k=0}^{K-1} \prod_{m=1}^{M} \sum_{b=1}^{{}_I\mathrm{C}_J} p(\tilde{\mathbf{S}}_m^{(k)} = {}_{(b)}\tilde{\mathbf{S}}_m^{(k)}). \quad (16)$$

3.3. Markov chain Monte Carlo sampling

The maximum likelihood estimate of TDOA, $\hat{\mathbf{d}}_{\mathrm{ML}}$ can be estimated by solving the optimization problem by maximizing the approximated objective function in Equation (16). Proposed algorithm uses Markov chain Monte Carlo (MCMC) sampling to find the global optimum solution.

The solution is solved iteratively, and for initialization, ${}^{(0)}\mathbf{d}$ is randomly chosen. In the l-th iteration, the candidate sample ${}^{\star}\mathbf{d}$ is generated from proposal distribution $q(\mathbf{d}|{}^{(l-1)}\mathbf{d})$. Here, $q(\mathbf{d}|\mathbf{d}')$ is set to be the multivariate Gaussian distribution that the mean vector is $\mathbf{d}'$ and the covariance matrix is $\sigma_l^2\mathbf{I}_{I\times(J-1)}$, where σ_l decays exponentially as the iteration index l increases. Then, the acceptance probability r of ${}^{\star}\mathbf{d}$ is calculated as

$$r = \min(1, \frac{p(\mathbf{Y}|{}^{\star}\mathbf{d})q({}^{(l-1)}\mathbf{d}|{}^{\star}\mathbf{d})}{p(\mathbf{Y}|{}^{(l-1)}\mathbf{d})q({}^{\star}\mathbf{d}|{}^{(l-1)}\mathbf{d})}), \quad (17)$$

and ${}^{\star}\mathbf{d}$ is accepted with probability of r. ${}^{(l)}\mathbf{d}$ is set to be ${}^{\star}\mathbf{d}$ when it is accepted or ${}^{(l-1)}\mathbf{d}$ when it is rejected. This iteration procedure is repeated until ${}^{(l)}\mathbf{d}$ converges.

4. EXPERIMENTS

To evaluate the performance of TDOA estimation algorithm, the root-mean-square error (RMSE) defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{I \times (J-1)} \sum_i \sum_{j \neq j'} |\Delta\tau_i^{(j,j')} - \hat{\Delta}\tau_i^{(j,j')}|^2} \quad (18)$$

is adopted.

[2] ${}_I\mathrm{C}_J = \frac{I!}{(I-J)!J!}$

Table 1. Experimental conditions

Number of microphones	$J = 2$
Number of sources	$I = 3$ or 4
Mic spacing	0.05m
Source types	Female, male speeches, audio sources with drums, without drums
Reverberation time	0ms or 130ms or 250ms
Observation noise (SNR)	0dB $\sim$ 30dB
Sampling rate	16kHz
Signal length	10 sec
STFT frame size	2048samples (128ms)
STFT frame shift	256samples (16ms)

Experiment is performed on synthetically generated data. Given virtual room environment illustrated in Fig. 1, the channel impulse responses are generated by the room impulse response generator [13] with various reverberation times and directions of the sources. Detailed experimental conditions are enumerated in Table 1, and four cases of different directions of the sources are listed in Table 2.

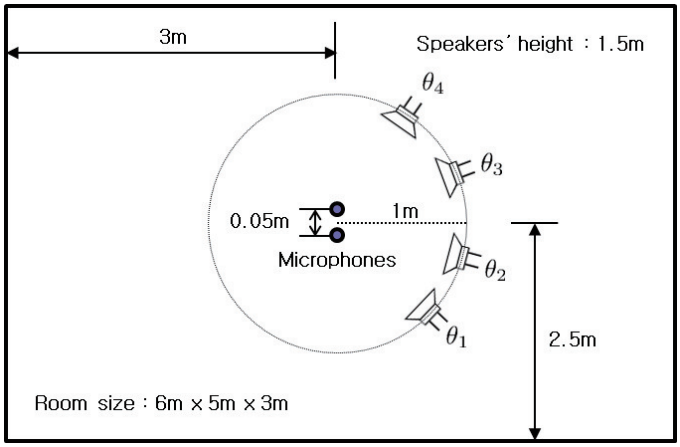

Fig. 1. Virtual room setup

At each frequency bin, whitening the observation vectors is conducted as a pre-processing step. The shape and scale parameters of the source distribution are fixed to be $c = 0.4$ and $\beta = 1$, respectively. The attenuation factor of $\tilde{H}_{ji}^{(k)}$, $\frac{\lambda_{ji}}{\lambda_{j'i}}$ is assumed to be 1.

The performance of proposed algorithm is compared to the algorithm which approximates $p(\tilde{\mathbf{S}})p(\mathbf{Y}|\tilde{\mathbf{H}}, \tilde{\mathbf{S}})$ as a single Gaussian. RMSE measures of two algorithms with various source types, reverberation times, and angle of the sources are enumerated in Table 3. Note that listed values are average values of 10 Monte Carlo runs in each case. As shown, the proposed algorithm estimated TDOA more accurately than the

Table 3. RMSE measure of proposed method and single Gaussian approximation method (SG) varying source types, reverberation time and angle of sources.

RT_{60}		0ms				130ms				250ms			
Case		1	2	3	4	1	2	3	4	1	2	3	4
Female speeches	Proposed	**0.009**	**0.006**	**0.006**	**0.006**	**0.063**	**0.030**	0.044	**0.047**	**0.513**	0.124	**0.147**	**0.124**
	SG	0.014	0.006	0.008	0.007	0.073	0.031	**0.043**	0.053	0.517	**0.122**	0.148	0.153
Male speeches	Proposed	0.012	0.008	**0.008**	**0.005**	**0.077**	**0.047**	**0.033**	**0.040**	**0.486**	**0.149**	**0.322**	**0.206**
	SG	**0.011**	**0.006**	0.009	0.010	0.157	0.050	0.036	0.048	0.495	0.151	0.359	0.224
Audio without drum	Proposed	**0.032**	**0.012**	**0.012**	**0.051**	**0.245**	**0.045**	**0.100**	**0.090**	**0.483**	**0.087**	**0.264**	**0.425**
	SG	0.035	0.015	0.013	0.111	0.260	0.046	0.103	0.175	1.048	0.089	0.269	0.681
Audio with drum	Proposed	**0.135**	**0.010**	**0.009**	**0.015**	**0.122**	**0.135**	**0.057**	**0.156**	**0.839**	**0.199**	**0.194**	**0.431**
	SG	0.221	0.011	0.016	0.020	0.226	0.135	0.062	0.158	0.848	0.201	0.203	1.061

Table 2. Direction angles generated in the four cases studied

	θ_1	θ_2	θ_3	θ_4
Case 1	$30°$	$45°$	$60°$	x
Case 2	$-60°$	$15°$	$75°$	x
Case 3	$-75°$	$40°$	$60°$	x
Case 4	$-60°$	$-30°$	$15°$	$60°$

algorithm with a single Gaussian approximation in noiseless cases. The experiment with various observation signal-to-noise ratios (SNR) is also conducted, and the result is shown in Fig. 2. The white Gaussian noise signals are generated and added to the observations. Direction of the sources is set to be 'case 4'($I = 4$, $J = 2$) and 4 different source types in anechoic environment are considered. In each source type, 10 Monte Carlo runs are conducted, and overall 40 Monte Carlo runs are averaged in each observation SNR. As shown, the TDOA estimate obtained by the proposed algorithm is also more accurate than the TDOA estimate obtained by the algorithm using a single Gaussian approximation in all cases of observation SNRs.

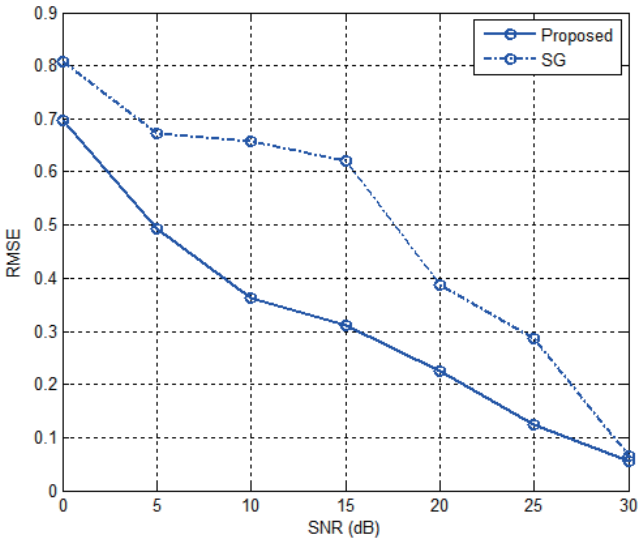

Fig. 2. RMSE measure of proposed method and single Gaussian approximation method (SG) varying observation SNR.

5. CONCLUSIONS

This paper considers the estimation of TDOA of multiple sources at the situation that the number of sources is more than that of microphones and the sources are sparse. An algorithm based on maximum likelihood criteria is proposed and the likelihood is obtained by marginalizing the joint probability over the sources. Mathematically intractable integration included in marginalization is solved through approximating the joint probability as a summation of Dirac delta functions. MCMC sampling is adopted to find the global maximum point of marginalized joint probability. The experiment is performed on various environment considering the reverberation and noise. Its results show that proposed algorithm outperforms the algorithm which approximates the joint probability as a single Gaussian.

6. ACKNOWLEDGMENT

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the National Robotics Research Center for Robot Intelligence Technology support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2012-H1502-12-1002) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No.2012-0005378).

7. REFERENCES

[1] S. Makino, T.W. Lee, and H. Sawada, *Blind speech separation*, Springer, 2007.

[2] H. Sawada, R. Mukai, and S. Makino, "Direction of arrival estimation for multiple source signals using independent component analysis," in *Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on*. IEEE, 2003, vol. 2, pp. 411–414.

[3] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Real-time blind source separation and doa estimation using small 3-d microphone array," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2005, pp. 45–48.

[4] S.G. Kim and C.D. Yoo, "Underdetermined blind source separation based on subspace representation," *Signal Processing, IEEE Transactions on*, vol. 57, no. 7, pp. 2604–2614, 2009.

[5] J. Ajmera, G. Lathoud, and L. McCowan, "Clustering and segmenting speakers and their locations in meetings," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. IEEE, 2004, vol. 1, pp. I–605.

[6] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[7] M.S. Lewicki and T.J. Sejnowski, "Learning overcomplete representations," *Neural computation*, vol. 12, no. 2, pp. 337–365, 2000.

[8] S. Winter, H. Sawada, and S. Makino, "On real and complex valued 1-norm minimization for overcomplete blind source separation," in *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*. IEEE, 2005, pp. 86–89.

[9] F.J. Theis, "Mathematics in independent component analysis," in *Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on*. IEEE, 2003, vol. 2, pp. 609–610.

[10] P. Bofill and E. Monte, "Underdetermined convoluted source reconstruction using lp and socp, and a neural approximator of the optimizer," *Independent Component Analysis and Blind Signal Separation*, pp. 569–576, 2006.

[11] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l 1-norm minimization," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 81–81, 2007.

[12] J. Cho, J. Choi, and C.D. Yoo, "Underdetermined convolutive blind source separation using a novel mixing matrix estimation and mmse-based source estimation," in *Machine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on*. IEEE, 2011, pp. 1–6.

[13] EAP Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, pp. 1, 2006.