View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Loughborough University Institutional Repository

This item was submitted to Loughborough's Institutional Repository (https://dspace.lboro.ac.uk/) by the author and is made available under the following Creative Commons Licence conditions.

creative commons

COMMONS DEED

Attribution-NonCommercial-NoDerivs 2.5

You are free:

- to copy, distribute, display, and perform the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor.

Noncommercial. You may not use this work for commercial purposes.

No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license).

Disclaimer

For the full text of this licence, please go to:
http://creativecommons.org/licenses/by-nc-nd/2.5/

Convolutive speech separation by combining probabilistic models employing the interaural spatial cues and properties of the room assisted by vision

Muhammad Salman Khan, Ata-ur-Rehman, Yanfeng Liang, Syed Mohsen Naqvi, and Jonathon Chambers

Advanced Signal Processing Group, School of Electronic, Electrical and Systems Engineering, Loughborough University, Leicestershire, LE11 3TU, U.K.
{*m.s.khan2, a.ur-rehman, y.liang2, s.m.r.naqvi, j.a.chambers*}*@lboro.ac.uk*

***Abstract*— In this paper a new combination of the model of the interaural spatial cues and a model that utilizes spatial properties of the sources is proposed to enhance speech separation in reverberant environments. The algorithm exploits the knowledge of the locations of the speech sources estimated through vision. The interaural phase difference, the interaural level difference and the contribution of each source to all mixture channels are each modeled as Gaussian distributions in the time-frequency domain and evaluated at individual time-frequency points. An expectation-maximization (EM) algorithm is employed to refine the estimates of the parameters of the models. The algorithm outputs enhanced time-frequency masks that are used to reconstruct individual speech sources. Experimental results confirm that the combined video-assisted method is promising to separate sources in real reverberant rooms.**

***Index Terms*— Speech separation, reverberation, spatial cues, expectation-maximization, time-frequency masking**

I. Introduction

Humans are experts at focussing on a single source when multiple sources are active. Machines, in contrast, are not as good. Machine audition is required since it would enable multiple applications such as hearing aids, automatic speech recognition, source separation in meeting room and teleconference environments. Different approaches have been proposed for source separation, for instance, frequency-domain convolutive blind source separation (BSS), beamforming, computational auditory scene analysis (CASA). Time-frequency (TF) masking is used for source separation and relies on the assumption that only a single source is active at each TF unit [1]. The TF approach is capable of handling the underdetermined problem where the number of sources is more than the number of sensors.

It has been reported that humans perceive sound as a multimodal process [2], [3]. We propose a source separation algorithm for two-channel reverberant mixtures by using source location information estimated through video. We model the interaural level difference (ILD) and the interaural phase difference (IPD) following the approach in [4] and model the contribution of each source to all mixture channels as in [5] as Gaussian distributions in the time-frequency domain. The parameters of the models are updated through the EM algorithm. In the E-step, the probabilities are calculated using the observations and the initial values of the parameters. In the M-step, the parameters are refined based on the observations and the probabilities from the expectation-step. This model requires knowledge of the properties of the room such as its dimensions, the sensor-to-speaker distances and the wall reflection coefficient which are partly found through video. The proposed combined algorithm outputs soft TF masks for individual sources in the reverberant mixture. The masks are then used to reconstruct the sources. In Section II we discuss the probabilistic models and the video processing. The model parameters are defined in Section III whereas the EM algorithm in Section IV. In Section V we provide simulation results and conclusions follow in Section VI.

II. Algorithm overview

For a two-channel recording the left and right convolutive mixture signals, as shown in Fig. 1, can be written as $l(t) = \sum_{i=1}^{I} s_i(t) * h_{li}(t)$, and $r(t) = \sum_{i=1}^{I} s_i(t) * h_{ri}(t)$, where $s_i(t)$ denote the speech signals, $h_{li}(t)$ and $h_{ri}(t)$ are the room impulse responses (RIR) from source i to the left and right sensors respectively, and $*$ denotes the convolution operation. The time domain signals are then transformed to the TF domain using the short-time Fourier transform (STFT).

A. The ILD and IPD models

The ratio of the STFTs of the left and right channels yields: $\frac{L(\omega,t)}{R(\omega,t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)}$, where $\alpha(\omega,t)$ is the ILD, measured in dB, and $\phi(\omega,t)$ is the IPD. The IPD observations are constrained to be in the range $[-\pi,\pi)$. We model a source with a frequency-dependent interaural time difference (ITD) $\tau(\omega)$, and a frequency-dependent ILD following [4]. The recorded IPD, $\angle(\frac{L(\omega,t)}{R(\omega,t)})$ for each TF unit, cannot always be mapped to the corresponding τ due to spatial aliasing. The model requires that τ and the length of $h(t)$ must be smaller than the Fourier transform window used (64ms). The phase residual error, the difference between the observed IPD and the predicted IPD (by a delay of τ samples), in the interval $[-\pi,\pi)$ is given as, $\hat{\phi}(\omega,t;\tau) = \angle(\frac{L(\omega,t)}{R(\omega,t)} e^{-j\omega\tau})$. The phase residual is modeled with a Gaussian distribution denoted as $p(\cdot)$ with mean $\xi(\omega)$ and variance $\sigma^2(\omega)$ that are dependent on frequency, $p(\phi(\omega,t)|\tau(\omega),\sigma(\omega)) = \mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi(\omega),\sigma^2(\omega))$. The ILD is also modeled with a Gaussian distribution with mean $\mu(\omega)$ and variance $\eta^2(\omega)$, $p(\alpha(\omega,t)|\mu(\omega),\eta^2(\omega)) = \mathcal{N}(\alpha(\omega,t)|\mu(\omega),\eta^2(\omega))$.

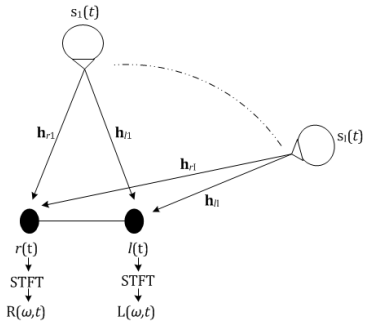


Fig. 1. Signal notations. The left and right sensor convolutive mixtures are transformed to the TF-domain to obtain $\mathrm{L}(\omega, t)$ and $\mathrm{R}(\omega, t)$.

B. The spatial covariance model

The stereo mixture $\mathbf{x}(t)$, obtained by concatenating *l(t)* and *r(t)*, can be expressed as $\mathbf{x}(t) = \sum_{i=1}^{I} \mathbf{img}_i(t)$, where $\mathbf{img}_i(t) = [img_{li}(t), img_{ri}(t)]^T$ is the spatial image of the ith source to the left and right channels. With the assumption that the sources are uncorrelated, we model $\mathbf{x}(\omega, t)$, the STFT of $\mathbf{x}(t)$, as a zero-mean Gaussian distribution with the covariance matrix [5] $\mathbf{R}_{\mathbf{x}}(\omega, t) = \sum_{i=1}^{I} v_i(\omega, t)\mathbf{R}_i(\omega)$, where $v_i(\omega, t)$ is the scalar variance and $\mathbf{R}_i(\omega)$ is the covariance matrix utilizing the spatial properties of the source i. The probability distribution of the model is given by [6]

$$P(\mathbf{x}(\omega,t)|v(\omega,t),\mathbf{R}(\omega)) = \prod_{\omega,t} \frac{1}{\det(\pi\mathbf{R}_{\mathbf{x}}(\omega,t))} \exp(-\mathbf{x}^H(\omega,t)\mathbf{R}_{\mathbf{x}}^{-1}(\omega,t)\mathbf{x}(\omega,t)) \quad (1)$$

where $(\cdot)^H$ is the Hermitian transpose. The spatial covariance $\mathbf{R}_i(\omega)$ of the source i is modeled as the sum of the covariance of the direct path and the covariance of the reverberant part [5] [7]

$$\mathbf{R}_i(\omega) = \mathbf{d}_i(\omega)\mathbf{d}_i^H(\omega) + \sigma^2_{reverb} \begin{bmatrix} 1 & \Omega(d_{lr},\omega) \\ \Omega(d_{lr},\omega) & 1 \end{bmatrix} \quad (2)$$

where $\mathbf{d}_i(\omega)$ is the direct-path direction vector, σ^2_{reverb} is the variance related to the reverberant part and $\Omega(d_{lr}, \omega)$ depends on the distance between left and right sensors d_{lr} and frequency ω. The reverberation observed at both the microphones is assumed to have the same power and its intensity has diffuse characteristics, $\Omega(d_{lr}, \omega) = \frac{\sin(2\pi\omega d_{lr}/c)}{2\pi\omega d_{lr}/c}$. The variance of the reverberant part is given by $\sigma^2_{reverb} = \frac{4\beta^2}{A(1-\beta^2)}$, where A is the total wall area and β is the wall reflection coefficient calculated from the room reverberation time (RT60) using Eyring's formula [7].

1) Estimating the direction vector $\boldsymbol{d}_i(\omega)$: To calculate the approximate positions of the speech sources in a room we use at least two fully calibrated color video cameras. In this work overall complexity is not being considered, rather, proof of concept is the focus. To detect the head of a speaker i in an image, we use the combination of skin color and gradient histograms. We combine the gradient histogram for robustness as the color-based detector alone can fail when a similar colored object is around the target. Further details can be found in our other recent works [8] and [9]. To obtain the 3-D real world Cartesian coordinates the center of the detected head is determined as the approximate position of the lips of the speaker in image coordinates $\boldsymbol{\iota}_i^c = [x_i, y_i]^T$, where c represents the camera index, $c = 1, 2$. In 3-D space each point in each camera frame defines a ray. Intersection of both rays is found by using multi-view geometry, which finally helps in calculation of the location for a speaker $\mathbf{Z}_i = [p_{x_i}, p_{y_i}, p_{z_i}]$ in 3-D real world coordinates [10].

The elevation (θ_i) and azimuth (ϕ_i) angles of arrival to the center of the sensors of each speaker i are calculated as $\theta_i = \tan^{-1}\left(\frac{p_{y_i} - p'_{y_c}}{p_{x_i} - p'_{x_c}}\right)$ and $\phi_i = \sin^{-1}\left(\frac{p_{y_i} - p'_{y_c}}{r_i \sin(\theta_i)}\right)$, where $r_i = \sqrt{(p_{x_i} - p'_{x_c})^2 + (p_{y_i} - p'_{y_c})^2 + (p_{z_i} - p'_{z_c})^2}$, while p'_{x_c}, p'_{y_c} and p'_{z_c} are coordinates of the center of the sensors. The direct-path weight vector $\mathbf{d}_i(\omega)$ for frequency ω and for source $i = 1, ..., I$, can be derived [11] as $\mathbf{d}_i(\omega) = [h_{li}, h_{ri}]^T$, where $h_{li} = \exp(-j\omega/c(\sin(\theta_i).\cos(\phi_i).p'_{x_l} + \sin(\theta_i).\sin(\phi_i).p'_{y_l} + \cos(\theta_i).p'_{z_l}))$, $h_{ri} = \exp(-j\omega/c(\sin(\theta_i).\cos(\phi_i).p'_{x_r} + \sin(\theta_i).\sin(\phi_i).p'_{y_r} + \cos(\theta_i).p'_{z_r}))$, p'_{x_m}, p'_{y_m} and p'_{z_m}, m being the left or right sensor index, are the 3-D positions of the sensors and c is the speed of sound in air at room temperature. The normalized vector $\mathbf{d}_i(\omega)$ is used in the model.

C. Combining the models and generating masks

To form an accurate mask for each source the ILD and IPD models, and the spatial covariance model using the direct-path direction vector obtained with the aid of video are used in conjunction. Estimating the model parameters is a hidden maximum-likelihood problem and thus the expectation-maximization (EM) algorithm is used for its solution as in [4]. Considering the models to be conditionally independent, we combine them given their corresponding parameters as $p(\alpha(\omega,t), \phi(\omega,t), \mathbf{x}(\omega,t)|\widetilde{\Theta}) = \mathcal{N}(\alpha(\omega,t)|\mu(\omega), \eta^2(\omega))\ .\ \mathcal{N}(\hat{\phi}(\omega,t)|\xi(\omega), \sigma^2(\omega))\ .\ \mathcal{N}(\mathbf{x}(\omega,t)|0, \mathbf{R}_{\mathbf{x}}(\omega,t))$, where $\widetilde{\Theta}$ denotes all of the model parameters.

III. Model Parameters

All of the model parameters $\widetilde{\Theta}$ can be collected as a parameter vector

$$\widetilde{\Theta} = \{\mu_i(\omega), \eta_i(\omega), \xi_{i\tau}(\omega), \sigma_{i\tau}(\omega), v_i(\omega,t), \psi_{i\tau}\} \quad (3)$$

where μ_i, $\xi_{i\tau}$, and η_i^2, $\sigma_{i\tau}^2$ are respectively the means and variances of the ILD, IPD models, and v_i is the scalar variance related to the spatial covariance model. The parameter matrix $\mathbf{R}_i(\omega)$ required to calculate $\mathbf{R}_{\mathbf{x}}(\omega, t)$ is found using *a priori* knowledge of the properties of the room and the $\mathbf{d}_i(\omega)$ as explained in Section 2.1. The subscript i indicates that the parameters belong to the source i, and τ and ω show the dependency on delay and frequency. The parameter $\psi_{i\tau}$ is the mixing weight, i.e. the estimate of the probability of any TF point belonging to source i at a delay τ, and is estimated as in [4]. The log value of the likelihood function ($\mathcal{L}$) given the observations can be written as

$$\begin{aligned}\mathcal{L}(\widetilde{\Theta}) &= \sum_{\omega,t} \log p(\alpha(\omega,t),\phi(\omega,t),\mathbf{x}(\omega,t)|\widetilde{\Theta}) \\ &= \sum_{\omega,t} \log \sum_{i,\tau} [\ \mathcal{N}(\alpha(\omega,t)|\mu_i(\omega),\eta_i^2(\omega)) \\ &\qquad . \mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi_{i\tau}(\omega),\sigma_{i\tau}^2(\omega)) \\ &\qquad . \mathcal{N}(\mathbf{x}(\omega,t)|0,\mathbf{R}_{\mathbf{x}}(\omega,t)).\ \psi_{i\tau}\ ]\end{aligned} \tag{4}$$

and the maximum likelihood solution is the parameter vector which maximizes this quantity.

IV. The Expectation-Maximization Algorithm

The algorithm is initialized with the estimated locations of the sources provided by video as explained in Section II-B.1. In the expectation step (E-step) the probabilities are calculated given the observations and the estimates of the parameters as

$$\begin{aligned}\epsilon_{i\tau}(\omega,t) = \psi_{i\tau}\ &. \mathcal{N}(\alpha(\omega,t)|\mu_i(\omega),\eta_i^2(\omega)) \\ &. \mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi_{i\tau}(\omega),\sigma_{i\tau}^2(\omega)) \\ &. \mathcal{N}(\mathbf{x}(\omega,t)|0,\mathbf{R}_{\mathbf{x}}(\omega,t)))\end{aligned} \tag{5}$$

where $\epsilon_{i\tau}(\omega,t)$ is the expectation of the hidden variable $m_{i\tau}(\omega,t)$, which is unity if the TF point belongs to both source i and delay τ and are zero otherwise. In the maximization step (M-step), the parameters are updated using the observations and $\epsilon_{i\tau}(\omega,t)$ from the E-step. The IPD and ILD parameters and $\psi_{i\tau}$ are re-estimated as in [4]. The spatial covariance matrix of the ith source $\mathbf{R}_i(\omega)$ is obtained through assistance from video as discussed previously and $v_i(\omega,t)$ is estimated as [5]

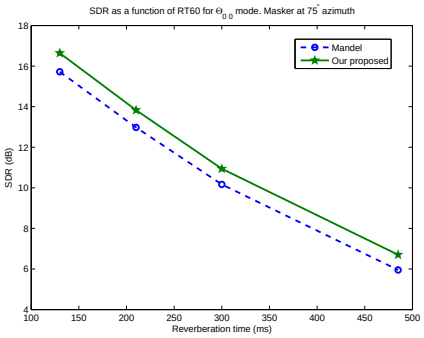
$$v_i(\omega,t) = \frac{1}{2}\text{tr}(\mathbf{R}_i^{-1}(\omega)\hat{\mathbf{R}}_{\mathbf{x}}(\omega,t)) \tag{6}$$

where $\hat{\mathbf{R}}_{\mathbf{x}}(\omega,t)$ is the covariance of the observed mixture and is estimated as $\hat{\mathbf{R}}_{\mathbf{x}}(\omega,t) = \frac{\sum_{\omega',t'} w(\omega'-\omega,t'-t)\mathbf{x}(\omega',t')\mathbf{x}^H(\omega',t')}{\sum_{\omega',t'} w(\omega'-\omega,t'-t)}$ where w is a 2-dimensional TF window.

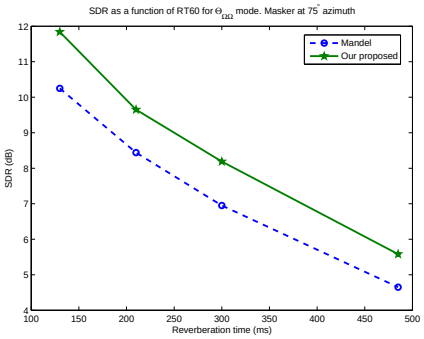
The spatial covariance model contributes once at the second iteration, as in the first iteration the occupation likelihood $\epsilon_{i\tau}(\omega,t)$ is calculated with only the ILD and IPD models. Since $\epsilon_{i\tau}(\omega,t)$ contains the correct order of the sources as in [4] the permutation problem is bypassed. The probabilistic masks for each source can be formed as $M_i(\omega,t) \equiv \sum_\tau \epsilon_{i\tau}(\omega,t)$. In the next section we confirm the effectiveness of the proposed approach experimentally.

V. Experiments and Results

We perform two sets of experiments. Firstly, we simulate sources with varying reverberation conditions with different model complexities. We also provide results for a smaller separation angle. Secondly, we give results for experiments on the AV16.3 audio-visual corpus [12] containing real room recordings. Experiments related to the AV16.3 dataset are discussed in Section V-A. Room dimensions are $(9 \times 5 \times 3.5)$ meters. The speakers are localized in the room through the video processing as explained in Section II-B.1. The speakers' locations and the direction vector $\mathbf{d}_i(\omega)$ obtained are then used in the algorithm. The audio and video observations are manually and independently synchronized. Models with different complexities for the ILD and IPD, Θ_{ildipd}, were evaluated similar to [4]. For instance, the ILD and IPD model complexity of Θ_{00} will have no ILD contribution and an IPD model with zero mean and a standard deviation that varies only by source, whereas $\Theta_{\Omega\Omega}$ uses the full frequency-dependent ILD and IPD model parameters. The desired source was located at 0° azimuth and the interferer was positioned either at 15° or 75°. Speech utterances were randomly chosen from the TIMIT acoustic-phonetic continuous speech corpus [13]. The first (16k$\times$2.5) samples of the utterances were used and were normalized to unity variance before convolving with the RIRs which were generated using the image method [14]. The signal-to-distortion ratio (SDR) as in [15] was used to evaluate the performance of the algorithms. We compare the proposed approach with [4], referred to as Mandel, and initialize it with the source location information found through video so that both algorithms utilize the same resources.



(a)



(b)

Fig. 2. SDR in dB for varying RT60s comparing the proposed method with Mandel for the Θ_{00} model in (a) and $\Theta_{\Omega\Omega}$ model in (b). The interferer was located at 75° azimuth.

In Fig. 2 the two sources were simulated for varying RT60s for the Θ_{00} and $\Theta_{\Omega\Omega}$ models with the interferer positioned at 75°. In Fig. 2(a), the Θ_{00} model, with the minimum contribution of the ILD and IPD models, the spatial covariance model contributes to show an overall improvement of the proposed method over Mandel at all RT60s. While in Fig. 2(b), where the most complex model is considered, the ILD, IPD and the spatial covariance all contribute to improve the performance over the competing method. Table I summarizes results for the Θ_{00} model with a smaller source separation

angle of 15°. This is a difficult case as the interaural cues of the sources are very similar. With the ILD and IPD models not doing very much, the spatial covariance model contributes to give a considerable improvement over its counterpart with average improvement of around 1.2dB over all RT60s.

TABLE I

SDR(dB) AS A FUNCTION OF RT60S COMPARING THE PROPOSED METHOD WITH MANDEL FOR THE Θ_{00} MODEL. THE MASKER IN THIS CASE WAS LOCATED AT 15° AZIMUTH.

RT60 (ms)	130ms	210ms	300ms	485ms
Mandel	5.53dB	3.73dB	1.81dB	0.52dB
The proposed	7.36dB	4.98dB	2.82dB	1.22dB
Improvement	1.83dB	1.25dB	1.01dB	0.70dB

A. Results for the AV16.3 Corpus

The AV16.3 corpus [12] contains real multispeaker recordings. We used the data from the available two-speaker case, where they were seated and simultaneously active as shown in Fig. 3. Speech mixtures from the third and seventh sensor of the microphone array 1 were utilized. To evaluate the performance of the proposed method for the AV16.3 dataset, we conduct listening tests and provide mean opinion scores (MOS tests for voice are specified by ITU-T recommendation P.800).

We extract two mixtures, one from 4.5-7 seconds and the other from 5.5-8 seconds, when both speakers are active and static for the experiments. The direction vector $\mathbf{d}_i(\omega)$ is estimated as explained in Section II-B.1 and is used in the algorithm. The MOSs in Table II (five people participated in the listening tests) highlight the improved performance of the proposed algorithm.

Fig. 3. Image from camera 1 on the top and camera 2 on the bottom. Both the speakers are seated, simultaneously active and static for the time slots under consideration. The third and seventh sensor mixtures of the microphone array 1 were used.

TABLE II

MEAN OPINION SCORES FOR THE STATIC TWO-SPEAKER CASE OF THE AV16.3 CORPUS.

Time slot	Mean Opinion Score (MOS)	
(seconds)	Mandel	The proposed
4.5-7	3.3	3.9
5.5-8	3.1	3.6

VI. CONCLUSION

A new multimodal source separation algorithm was proposed which integrates the model of the interaural parameters and the spatial covariance model. The models together make use of the estimate of the source location information which was derived through video processing. Experimental results indicated that the proposed algorithm can perform well even when the separation angle between the sources is small in contrast to the algorithm utilizing only interaural parameters. Due to the robust nature of the presented video localization scheme, the algorithm can be used in multi-speaker scenarios although this study considered only the two speaker case. Nevertheless the algorithm performs better at higher levels of reverberation than its counterpart, improvement is still needed in these adverse situations.

ACKNOWLEDGEMENT

M. S. Khan would like to thank KPK UET Peshawar and the Higher Education Commission (HEC) of Pakistan for their support and funding.

REFERENCES

[1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[2] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[3] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[4] M. I. Mandel, R. J. Weiss, and D. Ellis, "Model-Based Expectation-Maximization Source Separation and Localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[5] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined convolutive blind source separation using spatial covariance models," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 9 –12, March 2010.

[6] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830 –1840, Sept 2010.

[7] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 791 – 803, Nov 2003.

[8] A. Rehman, S. M. Naqvi, R. Phan, and J. A. Chambers, "MCMC-PF based multiple head tracking in a room environment," *4th UK Computer Vision Student Workshop (*BMVW*)*, 2012.

[9] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 895 –910, Oct. 2010.

[10] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2001.

[11] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part IV, Optimum Array Processing*, John Wiley and Sons, Inc., 2002.

[12] G. Lathoud, J. M. Odobez, and D. G. Perez, "AV16.3: an audio-visual corpus for speaker localization and tracking," *Proceedings of the MLMI'04 Workshop*, 2004.

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM, 1993. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1W.html," .

[14] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America.*, vol. 65, no. 4, pp. 943–950, 1979.

[15] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.