STOCHASTIC ONLINE DICTIONARY LEARNING FOR SPEECH SOURCE LOCALIZATION AND SEPARATION IN SPHERICAL HARMONIC DOMAIN

Vishnuvardhan Varanasi and Rajesh Hegde

Indian Institute of Technology, Kanpur
Email: {vishnuv, rhegde}@iitk.ac.in

ABSTRACT

Frequency and location dependent components in the speech signal can be decoupled by signal processing in the spherical harmonic domain. In this paper, a sparsity based method for joint source localization and separation method using online dictionary learning is proposed. Conventional sparsity based methods utilize an overcomplete dictionary to find a sparse linear combination of dictionary atoms. Online dictionary learning discussed herein, addresses the joint localization and separation problem by learning the dictionary atoms based on stochastic approximation. The location dependent terms present in the dictionary atoms at various frequencies are then clustered to find a robust estimate of number of sources and their locations. Using these estimates, the sources are separated from the mixture. Experiments on speech source localization and separation are conducted at various SNR. Performance evaluation scores like RMSE, log spectral distance and perceptual mean opinion scores indicate reasonable improvement over conventional methods for speech source separation.

Index Terms— Source Separation, Source Localization, Spherical Harmonics, Online Dictionary Learning, Sparse Coding

1. INTRODUCTION

Source localization and separation plays an important role in several applications such as distant speech recognition, music information retrieval [1], automatic camera steering [2] and hence it is an active area of research. Localization and separation in presence of noise is a challenging task. Spherical microphone array(SMA) [3] facilitates signal processing in spherical harmonic domain by capturing the spherical variation of acoustic field. Spherical harmonic domain facilitates the decomposition of received signal into frequency dependent component and location dependent components [4]. This inturn results in the location information available from different frequency components and thus helps in getting a robust estimate of source locations. In this work, we address the problem of source localization and separation by learning the dictionary atoms online without using an over complete dictionary.

A wide range of methods for localization and separation have been developed in the past few decades. In the instantaneous mixing model [5] [6], separation is carried out in time domain by assuming an additive model. Frequency domain separation methods are based on convolutive mixing of sources. Independent Component Analysis is one popular method for solving this problem. Probabilistic source separation methods [7] address this issue by assuming the mixture coefficients to be non-negative and then utilize Non-negative Matrix factorization(NMF). Time frequency masking [8] [9] methods exploit time-frequency sparsity for source separation. It utilizes the knowledge of steering vectors for a set of discrete locations and imposes sparsity for localization and separation of sources. On the other hand learning based approaches for source localization have been very effective in adverse environments.

Learning approaches are able to localize only single source and are not effective for multi source environments. Some methods perform well but are computationally very complex [10] and are not suitable for a real time source separation. The number of sources need to be known apriori in these methods. The contribution of this work is two fold. It proposes a stochastic online learning of the dictionary atoms at each individual frequency. The proposed method also extracts the location dependent components in the dictionary atoms and cluster them to find the number of sources and their steering vectors to jointly localize and separate the sources.

The rest of the paper is organized as follows. Section 2 introduces the data model spherical harmonics domain. Joint source localization and separation using stochastic online dictionary learning is described in section 3. Performance evaluation is discussed in section 4. Section 5 concludes the paper.

2. MODELING AN ACOUSTIC SCENE IN SPHERICAL HARMONIC DOMAIN

Consider an acoustic scene with L point sources and an SMA with I microphones. The location of sources is indicated by Ψ_l for $l = 1, 2, 3.....L$. Location of microphones is indicated by Ω_i for $i = 1, 2, 3.....I$. The following data model captures the pressure observed at the microphones in terms of microphone and source locations, signal strength and noise at microphones.

$$\mathbf{p}(k) = \mathbf{V}(k, \mathbf{\Psi})\mathbf{s}(k) \; + \; \mathbf{n}(k) \tag{1}$$

where k is wave number, $\mathbf{s}(k)$ is source strength vector and $\mathbf{n}(k)$ is noise vector and $\mathbf{V}(k, \mathbf{\Psi})$ is steering matrix containing the steering vectors corresponding to all the source locations. The position of l^{th} source can be represented as Ψ_l =(θ_l, ϕ_l) and corresponding wave vector is $\mathbf{k_l}$=$(k_l cos(\phi_l) sin(\theta_l), k_l sin(\phi_l) sin(\theta_l), k_l cos(\theta_l))^T$, where $k = \|\mathbf{k_l}\| = \frac{\mathbf{2\pi f}}{\mathbf{c}}$ and f is the frequency associated to wave number. The $I \times L$ steering matrix $\mathbf{V}(k, \mathbf{\Psi})$ employs spatial characteristics of the array and represents the room impulse response, it can be expressed as shown below:

$$\mathbf{V}(k, \mathbf{\Psi}) = [\mathbf{v}(k, \Psi_1), \ldots, \mathbf{v}(k, \Psi_L)], \tag{2}$$

where $\mathbf{v}(k, \Psi_l) = \left[e^{\mathbf{k_l^T r_1}}, ..., e^{\mathbf{k_l^T r_I}}\right]^T$ is the $I \times 1$ steering vector corresponding to l^{th} source with each element in the steering vector corresponds to an unit plane wave.

This work was funded by the SERB-DST under project no. SERB/EE/2017242.

Spherical harmonics facilitate decomposition of steering vector into three components, first dependent only on the microphone locations, second dependent only on the source locations and third dependent only on the frequency and radius of the spherical microphone array. Thus, expanding the steering matrix using spherical harmonics [11], Equation 1 can be rewritten as

$$\mathbf{p}(k) = \mathbf{Y}(\mathbf{\Omega})\mathbf{B}(kr)\mathbf{Y^H}(\mathbf{\Psi})\mathbf{s}(k) + \mathbf{n}(k) \tag{3}$$

$\mathbf{Y}(\mathbf{\Omega}) \in \mathbb{C}^{I\times(N+1)^2}$,$\mathbf{Y^H}(\mathbf{\Psi}) \in \mathbb{C}^{(N+1)^2\times L}$ are the spherical harmonics matrices with angular positions corresponding to microphones and sources respectively and they can be expressed as follows.

$$\mathbf{Y}^H(\mathbf{\Psi}) = \left[\mathbf{y}_1^H, \mathbf{y}_2^H, \ldots, \mathbf{y}_L^H\right],$$
$$\mathbf{y}_l = [Y_0^0(\Psi_l), Y_1^{-1}(\Psi_l), \ldots, Y_N^N(\Psi_l)],$$

where Y_n^m is spherical harmonics function of order n and degree m is given by

$$Y_n^m(\theta,\phi) = \sqrt{\frac{(2n+1)(n-m)!}{4\pi(n+m)!}}P_n^m(cos\theta)e^{jm\phi},$$

where P_n^m is the associated Legendre polynomial. $\mathbf{B}(kr)$ is a $(N+1)^2 \times (N+1)^2$ diagonal matrix with each element corresponding to the mode strength $b_n(kr)$ for a rigid spherical microphone array [12] and the matrix is defined as

$$\begin{aligned}\mathbf{B}(kr) &= diag(b_0(kr), b_1(kr), b_1(kr), \ldots, b_N(kr)),\\ b_n(kr) &= 4\pi j^n\left[j_n(kr) - \frac{j_n{}'(kr)}{h_n{}'(kr)}h_n(kr)\right],\end{aligned}$$

where j_n and h_n denote the spherical Bessel and Hankel functions respectively, $j_n^{'}$ and $h_n^{'}$ are their corresponding derivatives. The transformation between spatial and spherical harmonics domain can be done using spherical harmonics matrix as

$$\mathbf{p_{nm}}(k) = \mathbf{Y^H}(\mathbf{\Omega})\mathbf{p}(k) \tag{4}$$

where $\mathbf{p_{nm}}(k)$ is the observation vector in spherical harmonics domain. Using the matrix formulation of orthogonality of spherical harmonics and equation 3, data model takes the following form.

$$\mathbf{p_{nm}}(k) = \mathbf{B}(kr)\mathbf{Y^H}(\mathbf{\Psi})s(k) + \mathbf{Y^H}(\mathbf{\Omega})\mathbf{n}(k) \tag{5}$$

The signal term in Equation 5 becomes independent of frequency by left multiplying it with $\mathbf{B^{-1}}(kr)$. For notational convenience lets use the following representation.

$$\mathbf{x_{nm}}(k) = \mathbf{B^{-1}}(kr)\mathbf{p_{nm}}(k) \tag{6}$$

From Equations 5 and 6, the final data model [13] [14] takes the following form.

$$\mathbf{x_{nm}}(k) = \mathbf{Y^H}(\mathbf{\Psi})s(k) + \mathbf{z(k)} \tag{7}$$

where $\mathbf{z(k)}$ is termed as spherical harmonic noise and defined as $\mathbf{z(k)} = \mathbf{B^{-1}}(kr)\mathbf{Y^H}(\mathbf{\Omega})\mathbf{n}(k)$

3. STOCHASTIC ONLINE DICTIONARY LEARNING FOR SPEECH SOURCE LOCALIZATION AND SEPARATION

Sparsity based methods [15] are very often used for joint localization and separation but the main disadvantage of them is that they require an over complete dictionary with the atoms corresponding to all the possible locations. In case of three dimensional localization, as there are large number of possible locations, an over complete dictionary would be very huge and the computational complexity becomes very high. This also restricts that the environment to be fully known as it needs dictionary atoms for all locations. In online dictionary learning [16], dictionary atoms are learned without the necessity of an over complete dictionary. Problem formulation in spherical harmonic domain is first provided and then online dictionary learning for localization and separation is explained.

3.1. Problem Formulation

Consider the data model presented in Equation 7. For ease of notation, represent $\mathbf{x_{nm}}(k)$ at frame t as $\mathbf{x_t}(k)$. Given an observation $\mathbf{x_t}(k)$ and a dictionary $\mathbf{D}^{(k)}$, the linear combination of the dictionary atoms given by α can found by solving an optimization problem called as l_1 sparse coding problem as given below.

$$\alpha^* = \underset{\alpha}{\operatorname{argmin}} \quad \|\mathbf{x_t}(k) - \mathbf{D}^{(k)}\alpha\|_2^2 + \lambda\|\alpha\|_1 \tag{8}$$

As the dictionary $\mathbf{D}^{(k)}$ is also unknown, cost function need to be framed as a function of dictionary $\mathbf{D}^{(k)}$. Given a set of signals, an empirical cost function can be framed as follows.

$$f_T(\mathbf{D}^{(k)}) = \frac{1}{T}\sum_{t=1}^{T}\left[\|\mathbf{x_t}(k) - \mathbf{D}^{(k)}\alpha\|_2^2 + \lambda\|\alpha\|_1\right] \tag{9}$$

where the summation is done over a loss function which is small if $\mathbf{D}^{(k)}$ is good at representing the observations. The problem of minimizing the empirical cost $f_T(\mathbf{D}^{(k)})$ is that it is not convex with respect to the dictionary D. It can be rewritten as a joint optimization problem with respect to both the dictionary $\mathbf{D}^{(k)}$ and the linear combinations $\alpha_1, ..., \alpha_T$ as given below.

$$\min_{\mathbf{D}^{(k)},\alpha} \quad \|\mathbf{x_t}(k) - \mathbf{D}^{(k)}\alpha\|_2^2 + \lambda\|\alpha\|_1 \tag{10}$$

This joint optimization problem is not convex, but it is convex with respect to each of the two variables $\mathbf{D}^{(k)}$ and α when the other one is fixed. One approach that we followed for solving this optimization problem is discussed in the ensuing sections.

3.2. Solution Using Stochastic Online Dictionary Learning

One way of solving this optimization problem is to alternate between the two variables i.e minimizing the cost function with respect to one variable which keeping the other fixed. But the general interest is in minimizing the expected cost rather than the empirical cost. Hence, the online learning [16] discussed here solves the above mentioned optimization problem based on stochastic approximations, processing one sample at a time.

At any instant t, we are left with observations at that instant $\mathbf{x_t}(k)$ and the estimated dictionary in the previous instant which is $\mathbf{D}_{t-1}^{(k)}$. An optimization problem for this objective can be framed as follows which is convex and hence can be easily solved.

$$\alpha_\mathbf{t} = \min_{\alpha} \quad \|\mathbf{x_t}(\mathbf{k}) - \mathbf{D}_{t-1}^{(k)}\alpha\|_2^2 + \lambda\|\alpha\|_1 \tag{11}$$

Algorithm 1 Algorithm for speech source localization and separation using stochastic online dictionary learning

Require: Observations $\mathbf{p}(t) \in \mathbb{R}^{I\times 1}$, maximum number of sources M, initial dictionary $\mathbf{D}_0^{(k)} \in \mathbb{C}^{(N+1)^2\times M}$, the non-negative regularization parameter λ, parameter β for dictionary purging and the range of frequencies k.

Short Term Fourier Transform (STFT) applied to $\mathbf{p}(t)$ resulting in $\mathbf{p}(k)$

Transformation from spatial domain into spherical harmonics domain by

$$\mathbf{p_{nm}}(k) = \mathbf{Y^H}(\mathbf{\Omega})\mathbf{p}(k) \tag{12}$$

For steering vectors to be independent of frequency

$$\mathbf{x_{nm}}(k) = \mathbf{B^{-1}}(kr)\mathbf{p_{nm}}(k) \tag{13}$$

For ease of notation, represent $\mathbf{x_{nm}}(k)$ at frame t as $\mathbf{x_t}(k)$.

for all k **do**

  Initialization : $\mathbf{A} \leftarrow 0, \mathbf{B} \leftarrow 0$

  for t =1 to T **do** :

    Solve l_1 sparse coding problem using Numerical methods

$$\alpha_\mathbf{t} = \min_{\alpha} \quad \|\mathbf{x_t}(k) - \mathbf{D}_{t-1}^{(k)}\alpha\|_2^2 + \lambda\|\alpha\|_1 \tag{14}$$

    Update the matrices $\mathbf{A}$ and $\mathbf{B}$

$$\mathbf{A}_t \leftarrow \mathbf{A}_{t-1} + \alpha_t\alpha_t^T \tag{15}$$

$$\mathbf{B}_t \leftarrow \mathbf{B}_{t-1} + \mathbf{x}_t(k)\alpha_t^T \tag{16}$$

    Dictionary update :

    repeat

      for j **do**=1 to M

$$\mathbf{u}_j \leftarrow \frac{1}{A_{jj}}(b_j - \mathbf{D}_{t-1}^{(k)}a_j) + d_j \tag{17}$$

$$\mathbf{d}_j \leftarrow \frac{\mathbf{u}_j}{norm(\mathbf{u}_j)} \tag{18}$$

      end for

    until Convergence

  end for

  Purge dictionary atoms that do not contribute to the observation sufficiently. Indicated by β

end for

Collect $\mathbf{D}^{(k)} \quad \forall \quad k$

for i=1 to M **do**

  Dictionary atoms are clustered using k-means algorithm

  Compute Intra and Inter cluster variance

end for

Plot and find number of clusters L and their cluster means

Perform Gram-Schmidt orthogonalization on the cluster means to output optimal dictionary atoms $\mathbf{D}_{opt}$

The resulting dictionary atoms directly correspond to the steering vectors of the source locations and hence, location estimates are readily available.

Separate the speech sources using least squares by using location estimates

$$\min_{\mathbf{s_t(k)}} \quad \|\mathbf{x_t(k)} - \mathbf{D}_{opt}\mathbf{s_t}(k)\|_2^2 \tag{19}$$

Inverse STFT is applied on $\mathbf{s_t}(k)$ to get signal into time domain and thus speech separation is complete

3.2.1. Stochastic Online Dictionary Learning

The problem of finding the optimal dictionary D can be framed as an optimization problem by minimizing the cost function shown in equation 9. Equivalently, the problem can be framed as follows.

$$\mathbf{D}_t^{(k)} = \underset{\mathbf{D}}{\operatorname{argmin}} \quad \left[\frac{1}{2}Tr(\mathbf{D}^T\mathbf{D}\mathbf{A}_t) - Tr(\mathbf{D}^T\mathbf{B}_t)\right] \tag{20}$$

where $\mathbf{A}_t = \mathbf{A}_{t-1} + \alpha_t\alpha_t^T$ and $\mathbf{B}_t = \mathbf{B}_{t-1} + \mathbf{x}_t\alpha_t^T$. The dictionary $\mathbf{D}$ is updated from its previous estimate using block coordinate descent method. Since the algorithm uses $\mathbf{D}_{t-1}$ as its starting point for computing $\mathbf{D}_t$, a single iteration has been empirically found to be enough. Complete algorithm is explained in the ensuing sections.

Dictionary atoms are learnt for each of the frequency. Equation 7 indicates that the dictionary atoms are independent of frequency. Frequency smoothing [17] is done by means of k-means clustering so as to find out the correct number of sources as well as the robust estimates of source locations as the only information available is the maximum number of sources and not the exact number of sources. The correct number of clusters is found by comparing the values of inter and intra cluster variances. These dictionary atoms are nothing but the steering vectors corresponding to the source locations and hence, location estimates are readily available from dictionary atoms.

3.2.2. G-S Orthogonalization for Speech Separation

Gram-Schmidt orthogonalization is performed on the cluster means to get $\mathbf{D}_{opt}$. Speech sources are separated using least squares by using the optimal dictionary atoms $\mathbf{D}_{opt}$. Optimization problem in this case is as mentioned below.

$$\min_{\mathbf{s_t(k)}} \quad \|\mathbf{x_t(k)} - \mathbf{D_{opt}}\mathbf{s_t(k)}\|_2^2 \tag{21}$$

This has a closed form of solution and a solution can be readily found. Solution for this is problem is the set of speech signals in frequency domain. Hence, inverse STFT is applied on $\mathbf{s_t(k)}$ to get signal into time domain and thus speech separation is complete.

4. PERFORMANCE EVALUATION

In this section, the proposed method of speech source localization and separation is evaluated using a spherical microphone array. Performance analysis based on measures such as Root Mean Square Error(RMSE), Log Spectral Distance(LSD) and perceptual mean opinion scores (PESQ) is also presented.

4.1. Experimental Conditions

For these experiments, we utilize the data from GRID corpus [18]. A rigid spherical microphone array [19] of radius 15 cm and consisting of 50 microphones is used. The order of the spherical microphone array [20] used is N=6. The dictionary is initialized randomly and the linear combinations α are found by solving the convex optimization problem using numerical methods [21]. Experiments are conducted for 2 speaker scenario. Maximum number of speakers assumed is 10. Dictionary atoms are purged so that the atoms which contribute at least 30 percent of the total energy are retained. Sensor noise is assumed to be additive white uncorrelated Gaussian noise. The parameters for the algorithm are the non-negative regularization parameter λ and the dictionary purging parameter β. Regarding the dictionary purging parameter β, it purges the dictionary so that the

atoms that do not contribute significantly in representation of the observed signal are not considered. β value that is found empirically to be suitable is 30 percent. Regarding the regularization parameter λ, it indicates a balance between the sparsity of the sources and how close the observed signal is to the combination of separated signals.

4.2. Joint Localization and Separation Experiments

In this section, the performance analysis is carried out based on certain measures. The proposed method is called as ODL(online dictionary learning) method. For analyzing speech source separation, spectrographic analysis and objective evaluation using LSD and PESQ is carried out. For analyzing performance of source localization, RMSE analysis is done by varying SNR and the angular separation between the sources one at a time.

4.2.1. Spectrographic Analysis

One way of performance analysis in case of speech source separation is by analyzing the spectrograms of the original and the reconstructed signals. Separation was carried out for two sources located with an angular difference of 20 degrees. In the figure 1, spectrograms of the original signal and the separated signals are presented and it can be seen that the distortions are minimal in the reconstructed signals.

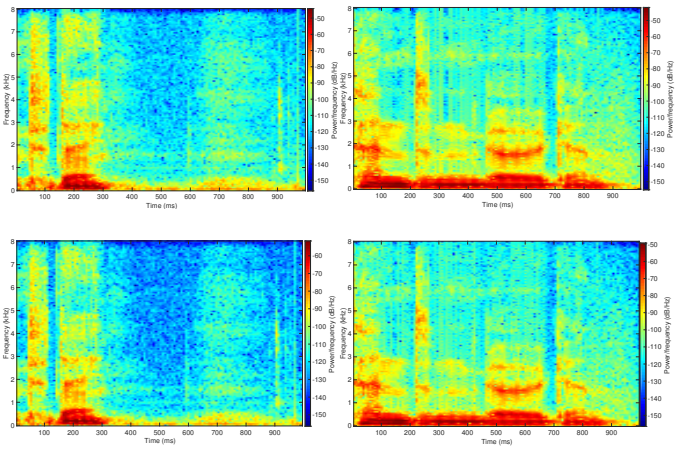

Fig. 1: Spectrographic analysis of the sparsity based method. (a)-(b) spectrogram of the original Source 1 and 2. (c)-(d) spectrogram of the separated Source 1 and 2.

4.2.2. RMSE Analysis

The efficiency of the proposed method is also evaluated using RMSE with respect to localization. RMSE of source localization for the proposed ODL method is compared to MUltiple SIgnal Classification(MUSIC) method. RMSE analysis for source location estimation is illustrated in Figure 2. In this analysis, we observe that error in localization using the ODL method is better than the standard methods of source localization.

RMSE analysis is also performed with respect to the angular separation between the sources in a two speaker scenario. The angular separation is varied from 10 to 50 degrees in steps of 10 degrees and RMSE is calculated with respect to localization. This experiment is carried out at an SNR of 10 dB. It can be seen from the figure 3 that as the angular separation increases between the sources, the localization performance improves.

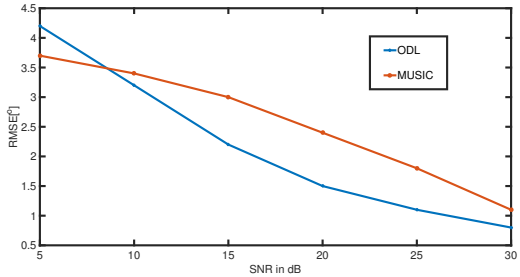

Fig. 2: Variation of RMSE with SNR for proposed sparsity based method

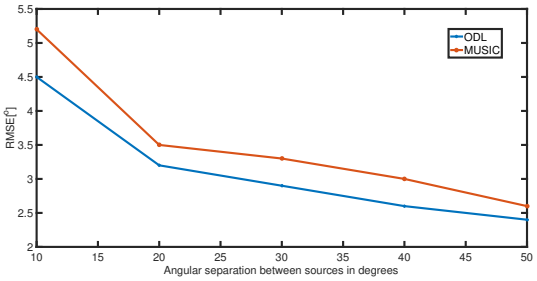

Fig. 3: Variation of RMSE with angular distance between the sources

4.2.3. Objective Evaluation of Speech Source Separation

Objective evaluation of speech source separation is carried out using LSD and PESQ by comparing the clean speech signals and the reconstructed speech signals. Table 1 illustrates the performance comparison of the proposed method to the standard methods existing for speech separation [22]. An improvement in the performance is observed with the proposed method.

Method	SNR (dB)					
	3		5		10	
	LSD	PESQ	LSD	PESQ	LSD	PESQ
ODL	1.32	2.45	1.09	2.32	1.19	2.69
ICA	1.39	1.85	1.23	2.27	1.31	2.22

Table 1: Objective evaluation of source separation using ODL(Online Dictionary Learning) and ICA(Independent Component Analysis) using LSD and PESQ.

5. CONCLUSION

In this work, a novel method for speech source localization and separation based on stochastic online dictionary learning is proposed. Performance analysis is carried out using spectrogram analysis, RMSE and perceptual mean opinion scores. In case of multi source learning environment, this work can be used for computing features corresponding to individual sources from a mixture of sources, which will be addressed in future. Dictionary learning in a reverberative environment will be addressed in future work. Multi source localization in a noisy and reverberative environment is also a topic of future investigation. This can lead to the development of joint source separation and recognition methods in a DNN framework.

6. REFERENCES

[1] J. Traa and P. Smaragdis, "Blind multi-channel source separation by circular-linear statistical modeling of phase differences," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 4320–4324.

[2] Yiteng Huang, Jacob Benesty, and Gary W Elko, "Passive acoustic source localization for video camera steering," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*. IEEE, 2000, vol. 2, pp. II909–II912.

[3] Jens Meyer and Gary Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, May 2002, vol. 2, pp. II–1781–II–1784.

[4] B. Rafaely, "Analysis and design of spherical microphone arrays," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 1, pp. 135–143, Jan 2005.

[5] Te-Won Lee, Michael S Lewicki, Mark Girolami, and Terrence J Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *Signal Processing Letters, IEEE*, vol. 6, no. 4, pp. 87–90, 1999.

[6] Mike E Davies and Christopher J James, "Source separation using single channel ica," *Signal Processing*, vol. 87, no. 8, pp. 1819–1832, 2007.

[7] Tuomas Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1066–1074, 2007.

[8] Emmanuel Vincent, "Musical source separation using time-frequency source priors," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 91–98, 2006.

[9] J. Freudenberger and S. Stenzel, "Time-frequency masking for convolutive and noisy mixtures," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on*, May 2011, pp. 104–108.

[10] Sachin N Kalkur, Sandeep Reddy C, and Rajesh M Hegde, "Joint source localization and separation in spherical harmonic domain using a sparsity based method," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] Thushara D Abhayapala and Darren B Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 2, pp. II–1949.

[12] DP Jarrett, EAP Habets, MRP Thomas, and PA Naylor, "Rigid sphere room impulse response simulation: Algorithm and applications," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1462–1472, 2012.

[13] Nejem Huleihel and Boaz Rafaely, "Spherical array processing for acoustic analysis using room impulse responses and time-domain smoothing," *The Journal of the Acoustical Society of America*, vol. 133, no. 6, pp. 3995–4007, 2013.

[14] Jrgen Hald, "Basic theory and properties of statistically optimized near-field acoustical holography," *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2105–2120, 2009.

[15] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd, "Enhancing sparsity by reweighted 1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[16] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 689–696.

[17] Dima Khaykin and Boaz Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2009, pp. 221–224.

[18] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[19] Finn Jacobsen, Guillermo Moreno-Pescador, Efren Fernandez-Grande, and Jrgen Hald, "Near field acoustic holography with microphones on a rigid sphere (l)a)," *The Journal of the Acoustical Society of America*, vol. 129, no. 6, pp. 3461–3464, 2011.

[20] Boaz Rafaely, *Fundamentals of spherical array processing*, vol. 8, Springer, 2015.

[21] M Grant and S Boyd, "Cvx: Matlab software for disciplined convex programming, version 2.0 beta, sept. 2012," *Available on-line at http://cvxr. com/cvx*.

[22] Nicolas Epain, Craig Jin, and André van Schaik, "Blind source separation using independent component analysis in the spherical harmonic domain," *Proceedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics, Paris*, 2010.