2013 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 22–25, 2013, SOUTHAMPTON, UK

VIDEO-INFORMED APPROACH FOR ENHANCING AUDIO SOURCE SEPARATION THROUGH NOISE SOURCE SUPPRESSION

Jack Harris[⋆†]*, Bertrand Rivet*[⋆]*, Syed Mohsen Naqvi*[†]*, Jonathon A. Chambers*[†]*, Christian Jutten*[⋆]

[⋆]GIPSA-Lab, CNRS UMR 5216, Université de Grenoble, France.
{jack.harris, bertrand.rivet, christian.jutten}@gipsa-lab.grenoble-inp.fr
[†]School of Electronic, Electrical and Systems Engineering, Loughborough University, UK
{s.m.r.naqvi, j.a.chambers}@lboro.ac.uk

ABSTRACT

This paper describes a method where an interference noise source within an audio source separation scenario is suppressed from a mixture. The principal idea of the proposed method is to use a video camera array for locating a interference noise source whose 3D position will be used to estimate a matrix of frequency responses (FRs) by linearly combining a series of previously known FRs. A filter is calculated to remove the contribution of the noise source from a convolutive mixture at each microphone, through the estimated FRs. The proposed method is assumed to implemented in a 'block-wise' manner in time domain and has been tested on mixtures created by impulse responses generated by the image method for small room acoustics.

Index Terms— audio-visual source separation, source enhancement, transfer function estimation, non-stationary sources

1. INTRODUCTION

The cocktail party problem was first proposed in [1], which describes a situation where there are multiple human speakers talking simultaneously within a room environment and it is required that each speaker's voice is isolated (separated) from the other voices which are present, similar to the manner in which a human sensory system can identify individual speakers in a situation such as a party, hence the name of the problem.

Blind source separation (BSS) problems such as this are normally addressed using higher order statistics, a common approach is to perform independent component analysis (ICA) [2]. In some applications, particularly with non-stationary sources in real time, it is desirable to use a more efficient method as higher order methods can consume large amounts of system resources and require too much time to produce accurate estimates.

In this paper a different approach is taken to audio-visual source separation, it relies on a given set of previously calculated FRs and an unwanted noise source's 3D location from video information provided by an array of video cameras. The proposed method can be viewed as a pre-processing stage before a more conventional BSS algorithm that suppresses a noise source with a known location; in addition it can be used by itself in the 2-microphone 2-source scenario to extract a filtered version of one of the sources.

Previous work into audio-visual source separation for moving and non-stationary sources can be found in [3, 4], where a 3D position tracker is used to identify the location of a source. Then based on this information, an appropriate BSS algorithm is selected depending on the movement of the source; the method proposed in this paper could be used within such a framework. In [5] a time-frequency masking approach is described that exploits direction of signal arrival. A more complex environment is described in [6] where the number of speakers is not fixed and move in and out of the environment. The work in this paper seeks to build upon previous work [7] by providing a method to suppress a noise source within a mixture, based on source localization information from video signals. Although this paper does not directly deal with the identification and tracking of sources it is assumed that this information is available (eg. by implementing one of the video tracking methods found in [3, 4, 5]).

Throughout the paper the method is considered by itself in the general case and there is an equal number of microphones and sources (N). When $N = 2$, the method can be used by itself, although further processing could be used if it was desired to extract the suppressed noise from the mixture. However, if the number of microphones and sources was increased, for example when $N = 3$, after noise source suppression two sources would be left in the mixture and a further BSS algorithm could be used such as [8]. It is assumed that any movement of the sources is slow, so that a quasi-static assumption can be made, and that the method uses block-wise processing, as a result changes in the room impulse response due to movements of the sources, or other persons or objects, that affect the acoustic environment are considered to be neg-

978-1-4799-1180-6/13/$31.00

ligible between time blocks. One appreciates in a real room environment this assumption is only an approximation, but it is sufficient for this proof of concept work.

2. METHOD

The observation at each microphone can be modelled in the general case in the time domain as a convolutive mixture of the sources:

$$x_i(k) = \sum_{j=1}^{N} h_{ij}(k) * s_j(k) + n_i(k), \qquad i = 1, \ldots, N \quad (1)$$

where s_j is the speech signal generated by each source, h_{ij} is the filter that models the effect of the room between the j-th source and the i-th microphone, k is the discrete time index, n_i is additive noise and x_i is the detected signal at the i-th microphone. The set of equations in (1) can be transformed in the time-frequency domain (assuming the noise $n_i(k)$ is negligible) and expressed as:

$$\mathbf{x}(t,f) = \sum_{j=1}^{N} \mathbf{h}_j(f) s_j(t,f) \quad (2)$$

where, $\mathbf{x}(t,f) = [x_1(t,f), ..., x_N(t,f)]^\mathsf{T}$ is the short time Fourier transform (STFT) of the detected audio signals at each microphone as a vector ($(\cdot)^\mathsf{T}$ is the vector transpose), and $\mathbf{h}_j(f) = [h_{1,j}(f), ..., h_{N,j}(f)]^\mathsf{T}$ is a vector of frequency domain terms, $s_j(t,f)$ is the STFT of the audio signal generated by the j-th source, t is the time block index and f is the frequency bin index.

The algorithm is divided into two principal stages. The first stage consists of estimation of FRs between the noise source and multiple microphones (Section 2.1). In a second stage these FRs are used to find a suppression filter to remove the effect of the noise source on the mixture at each microphone (Section 2.2). Finally, an optional post-processing stage is used to improve the quality of the remaining sources in the mixture by source amplitude recovery (Section 2.3).

The transforms of known impulse responses (IRs) which have typically been measured over a spatial grid are calculated. From these FRs a weighted linear combination is calculated to estimate an FR at the point where the noise source is measured. It is only necessary to know the FRs around the noise source to remove it from the mixture. Note that the number of microphones always needs to be equal to the number of sources including the noise source, as the related transfer functions are used to create a suppression filter.

2.1. Frequency Response Estimation

The room is divided into cubes known as voxels which are arranged into a non-overlapping 3D spatial grid pattern. Based on work in [9] and [10], part of the motivation calculating FRs in such a manner is to correct inaccuracies in measured IRs (Fig. 1). An FR is estimated by calculating a weighted average of previously known FRs at each corner of the voxel that contains the noise source, calculated by the Fourier transform of an IR [11]. The weighted average depends on the noise source position (p_n), which is provided by video information, as represented in Figure 2. The purpose of taking averages of FRs is to avoid storing an IR and a transfer function (TF) for every possible point within the room which would be impractical. Weights are assigned to each FR at each corner ($a = 1, \ldots, 8$) and are calculated by:

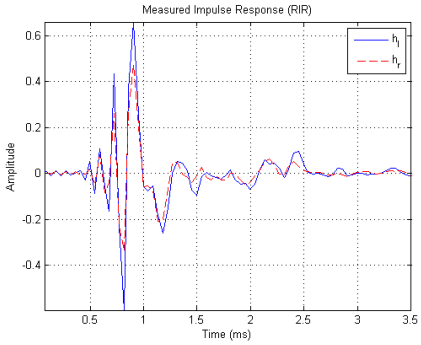


Fig. 1: Example of a measured IRs (MIRs) between a source in position 2 in Fig. (4) and both microphones within an anechoic room. Despite being 'geometrically symmetrical' within the room it shows (possibly due to placement errors) differences in the impulse responses that need to be accounted for.

$$w_a = \left(1 - \frac{x_a}{A}\right)\left(1 - \frac{y_a}{A}\right)\left(1 - \frac{z_a}{A}\right) \quad (3)$$

where w_a is the weight at corner (a), A is the edge length of the voxel and x_a, y_a and z_a are the distances in each dimension between each corner and p_n. The linear combination is then:

$$\hat{h}_{ij}(f) = \sum_{a=1}^{8} w_a h_{ij}^a(f) \quad (4)$$

where $h_{ij}^a(f)$ is the previously calculated FR at each corner and $\hat{h}_{ij}(f)$ is the estimated FR between p_n and each microphone (i).

2.2. Noise Source Suppression

This method principally exploits the property of two orthogonal vectors (Figure 3), so that when the dot product between two vectors is calculated the result is 0.

A new filter $\hat{\mathbf{G}}_n(f) \in \mathbb{C}^{(N-1)\times N}$ is calculated from the estimated FR vector that removes the source n from the mixture, thus $\mathbf{G}_n(f)\mathbf{x}(t,f) = \mathbf{f}(s_1, ..., s_{n-1}, s_{n+1}, ..., s_N)$. Dropping the f index for convenience, using (2) this implies: $\hat{\mathbf{G}}_n\hat{\mathbf{h}}_n = 0$ and $\hat{\mathbf{G}}_n\mathbf{h}_n \approx 0$, where $\mathbf{h}_n$ represents the vector of actual frequency responses which are unknown.

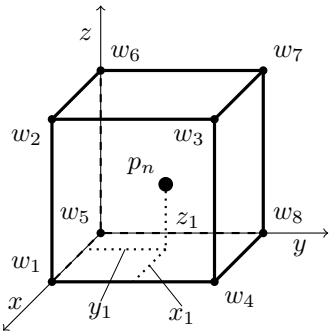


Fig. 2: An example of a voxel within a room, distances are given for corner $a = 1$. p_n is the noise source position within the voxel.

In a general setup, including the estimated suppression filter $\hat{\mathbf{G}}_n$ in (2), yields:

$$\hat{\mathbf{G}}_n(f)\mathbf{x}(t,f) = \sum_{j=1}^{N} \hat{\mathbf{G}}_n(f)\mathbf{h}_j(f)s_j(t,f) \qquad (5)$$

The result of (5) would be a distorted version of the unsuppressed sources, possibly with a small contribution from the suppressed noise source due to a mismatch between the estimated filter ($\hat{\mathbf{G}}_n$) and its ideal value ($\mathbf{G}_n$). To reverse the potential effects of $\hat{\mathbf{G}}_n$ on the unsuppressed sources, an optional post processing stage is used (Section 2.3).

To find the filter ($\hat{\mathbf{G}}_n$) one constructs an orthogonal projection [12] by:

$$\hat{\mathbf{G}}_{n,proj}(f) = (I - \hat{\mathbf{h}}_n(f)(\hat{\mathbf{h}}_n^{\mathsf{H}}(f)\hat{\mathbf{h}}_n(f))^{-1}\hat{\mathbf{h}}_n^{\mathsf{H}}(f)) \qquad (6)$$

where $\hat{\mathbf{h}}_n$ is the steering vector of the interference source and $\hat{\mathbf{G}}_{n,proj}$ is the projection matrix. Singular value decomposition (SVD) is then performed on $\hat{\mathbf{G}}_{n,proj}$, so that; $\hat{\mathbf{G}}_{n,proj} = \mathbf{U}_n\mathbf{\Sigma}\mathbf{V}_n^{\mathsf{H}}$, where $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose. To find the filter, all non-zero values of the diagonal matrix (denoted $\mathbf{\Sigma}$) are used to identify the corresponding columns in the unitary matrix (denoted $\mathbf{U}$), these columns in the unitary matrix correspond to the non-zero values of $\hat{\mathbf{G}}_{n,proj}$. $\hat{\mathbf{G}}_n$ is the concatenation of singular vectors associated with the non-null singular values, so $\hat{\mathbf{G}}_n = [\mathbf{U}_1, \ldots, \mathbf{U}_{N-1}]^{\mathsf{H}}$. Given a mixture, the output of the suppression stage is $\mathbf{x}'(t,f) = \hat{\mathbf{G}}_n(f)\mathbf{x}(t,f)$.

The method, upto this point can be summarized in the following tables (Algorithms 1 and 2).

2.3. Post-Processing

The purpose of the optional post-processing stage is to reverse any adverse effects on the remaining sources in the mixture after the noise source has been suppressed. The post-processing stage is based on the idea of amplitude recovery where a new scaling factor is applied using known variables at each frequency bin:

$$x''(t,f) = \frac{E[x(t,f)x'(t,f)^*]}{E[|x'(t,f)|^2]}x'(t,f) \qquad (7)$$

where $\frac{E[x(t,f)x'(t,f)^*]}{E[|x'(t,f)|^2]}$ is the scaling factor, $(\cdot)^*$ denotes the complex conjugate and x'' is the output of the overall method, including the post-processing stage. Further explanation of the post-processing stage can be found in [13].

3. RESULTS

3.1. Experimental Setup

Simulated IRs generated by the image method (IM) [11] in an almost anechoic simulation ($T_{60} = 37ms$) (Test 1) and a reverberant simulation ($T_{60} = 100ms$) (Test 2) were used when $N = 2$ to create mixtures using two 15 second speech utterances in a variety of source positions (to simulate the 3D location of a source provided by video camera information). $N = 2$ is used for set of results for the sake of brevity. The FR at each corner of a voxel is calculated by the Fourier transform of IRs generated by the IM for all tests. The voxel size used in all tests was 0.03m. Utterances from the TIMIT database from a male and female speaker were used. One utterance is a 'desired' source and the other acts as a noise source. Sources were positioned in an arc around a two-microphone array at 0.6m, the source is at 0° when it is equidistant to both microphones, negative angles are to the left and positive ones to the right of the array (Fig. 4).

To measure the performance the signal to distortion ratio (SDR) [14] and 'Perceptual Evaluation of Speech Quality'

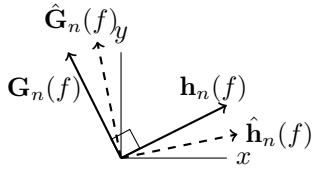
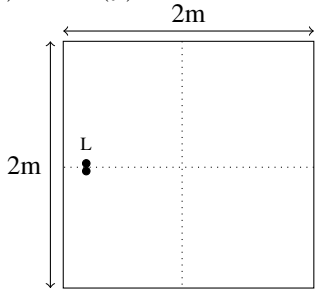


Fig. 3: Orthogonality property, when $N = 2$, where $\mathbf{G}_n(f)$ is the suppression filter vector at a particular frequency and $\mathbf{h}_n(f)$ the FR vector at a particular frequency for the noise source n. $\hat{\mathbf{G}}_n(f)$ and $\hat{\mathbf{h}}_n(f)$ are the estimated equivalents.



Algorithm 1 Stage 1: Database Training

Input: Room dimensions, acoustic properties of the room, voxel grid parameters.
Output: TF database.

1: **Calculate** IRs ($\mathbf{H}(k)$) using Allen & Berkley's image method.
2: Take the FFTs of the estmimated IRs, so: $\mathbf{H}(f) \leftarrow FFT(\mathbf{H}(k))$

where $\mathbf{H}(f)$ denotes the databases of room TFs, $\mathbf{H}(k)$ denotes the databases of room IRs and $\leftarrow$ denotes an assignment.

Algorithm 2 Stage 2: Calculate weights and find the suppression filter matrix.

Input: 3D location of source to be suppressed, TF database.
Output: Suppression filter matrix for source n ($\hat{\mathbf{G}}_n$)

1: **for** each corner of a voxel, ie. $a \leq 8$ **do**
2: Find the weight for each voxel corner, $w_a \leftarrow (1 - \frac{x_a}{A})(1 - \frac{y_a}{A})(1 - \frac{z_a}{A})$
3: **end for**
4: Calculate the a TF estimation for source n by the weighted mean of the TFs, $\hat{h}_n \leftarrow \sum_{a=1}^{8} w_a h_n^a(f)$.
5: Find an orthogonal matrix (at each frequency bin) by a projection matrix, $\hat{\mathbf{G}}_{n,proj} \leftarrow \mathbf{I} - \hat{h}_n(\hat{h}_n^{\mathsf{H}}\hat{h}_n)^{-1}\hat{h}_n^{\mathsf{H}}$.
6: Calculate the SVD decomposition, $\{\mathbf{U}_n, \mathbf{\Sigma}, \mathbf{V}_n^{\mathsf{H}}\} \leftarrow svd(\hat{\mathbf{G}}_{n,proj})$.
7: Sort (in descending order) the values of the diagonal matrix ($\mathbf{\Sigma}$), $\{\mathbf{\Sigma}^{sorted}, \mathbf{i}\} \leftarrow sort(\mathbf{\Sigma}^{p,p}) \in p = \{1, \ldots, N\}$, where $\mathbf{i}$ is a vector of the sorted original indices.
8: Rearrange the corresponding columns of $\mathbf{U}_n$ (if necessary) according to $\mathbf{i}$ and remove the column corresponding to the lowest value of $\mathbf{\Sigma}^{sorted}$, $\hat{\mathbf{G}}_n \leftarrow \mathbf{U}_n^q \in q = \{1, \ldots, N-1\}$ (where q is the column index).

(PESQ) [15] are used. The SDR measure takes into account any artefacts introduced by an algorithm, as well as comparing the original source with the recovered estimate, and is calculated by:

$$SDR = 10 \log_{10} \frac{\|\mathbf{s}_{target}\|^2}{\|\mathbf{e}_{interf} + \mathbf{e}_{artif}\|^2} \tag{8}$$

where $\mathbf{s}_{target}$ is a measure of the part of the estimated source which can be attributed to a filtered version of the original source, $\mathbf{e}_{interf}$ is the interference from other sources (which is 0 in this case, as only one source is recovered) and $\mathbf{e}_{artif}$ is anything else that cannot be attributed to the contributions from other sources such as distortion introduced by a BSS algorithm. PESQ is a measure that was originally developed for telephony that compares the quality of a processed signal to the original by predicting the score of subjective listening tests (mean opinion scores) on a scale of 0-5.

3.2. Simulated Mixtures

Table 1 (Test 1) confirms that the concept of the method works well in a simulated environment. The estimated version of the second source is clearly audible, with a small amount of noise and in some cases there is a very small contribution from the suppressed source, in all situations (without the need of a post-processing stage), apart from the positions where both sources are at the same angle (which are highlighted in Tables 1 and 2). Realistically, both sources are unlikely to be exactly at the same position in a room, however they can be close. The method also performed well when the reverberation was increased to $T_{60} = 100ms$ (Table 2), this shows reduced performance in a slightly reverberant environment.

4. CONCLUSION & DISCUSSION

This paper presented an audio-visual noise source suppression method for audio convolutive mixtures and potentially non-stationary sources that can be incorporated into other source separation algorithms to improve overall performance or be used by itself to enhance a source in a 2-microphone 2-source mixture.

Generally, previous audio-visual BSS methods aid the BSS process with video signals to localize a source. Once a source has been identified, an appropriate BSS algorithm is selected [3, 4], the work presented in this paper could be incorporated into the framework of such methods. Other papers use the video information differently such as [13], which exploits pauses in speech to identify silent periods so that one source is silenced. Some methods use audio localization instead of video signals [16, 17]. However, audio localization for simultaneously active speakers in a reverberant room environment is difficult [6, 5]. Video localization is also not always effective, especially if a human face is not visible to at least two cameras [4]. Therefore audio-visual modalities with multiple camera integration is the most suitable choice for source localization, however it is beyond the scope of this paper.

Potential drawbacks, such as time required to provide an accurate estimate, typically associated with some BSS algorithms, prompts the need for more efficient audio-visual

Table 1: Performance of the algorithm with mixtures that use the image method, $T_{60} = 37ms$ (Test 1). NS is the noise source that is to be suppressed from the mixture leaving Source 2. The optional post-processing stage has not been used in these results. Highlighted cells indicate when the sources were in the same position. Results are shown as SDR in dB and PESQ (0-5).

		Source 2									
		Position 1		Position 2		Position 3		Position 4		Position 5	
		SDR	PESQ	SDR	PESQ	SDR	PESQ	SDR	PESQ	SDR	PESQ
NS	Position 1	-7.2	-0.13	19	2	21	2.2	22	2.3	23	2.3
	Position 2	25	2.4	-4.4	0.88	22	2.2	25	2.4	26	2.4
	Position 3	22	2.3	17	2	-5.7	0.099	16	2	20	2.2
	Position 4	19	2.3	16	2	11	1.7	-4.4	1.4	9.6	1.6
	Position 5	22	2.3	19	2.1	17	1.9	12	1.5	-9.4	-0.013

Table 2: Performance of the algorithm with mixtures that use the image method where $T_{60} = 100ms$ (Test 2). NS is the noise source that is to be suppressed from the mixture leaving Source 2. Highlighted cells indicate when the sources were in the same position. Results are shown as SDR in dB and PESQ (0-5).

		Source 2									
		Position 1		Position 2		Position 3		Position 4		Position 5	
		SDR	PESQ	SDR	PESQ	SDR	PESQ	SDR	PESQ	SDR	PESQ
NS	Position 1	-6.1	0.28	6.8	1.6	9.1	1.8	11	2	11	2.1
	Position 2	7.7	1.8	-6.3	1.8	7.5	1.7	8.7	1.9	9.6	2
	Position 3	11	2	7.3	1.6	-5.5	1.4	6.2	1.7	6.8	1.8
	Position 4	10	2	6.6	1.6	4.7	1.5	-6.2	0.45	5.1	1.6
	Position 5	11	2	8.2	1.7	5.8	1.6	6.5	1.6	-6	1.9

methods, that avoid higher-order statistics, such as the one presented in this paper. In this context, [7] uses assumed video information to enhance a source by adaptive filtering, this paper builds on this work, by providing an alternative to the target cancellation method mentioned in [7].

The experimental results in this paper demonstrate a new method that suppresses a noise source with a known location by estimating a filter from estimated FRs, it has been tested on mixtures of speech signals with IRs generated by the image method. Promising results have been found for mixtures using image method IRs. The authors except that to become a practical solution this method has need for expansion, work into more realistic mixtures with longer reverberation times using measured impulse responses is necessary. Also, changes in the acoustic environment would need to be addressed, this could be achieved by adjusting the database of the previously known IRs that are used to calculate an estimated FR in order to allow for changes in the acoustic environment. It is hoped that this could be achieved by using a method such as [18] to alter the database in an adaptive manner and is the subject of ongoing research.

Further research and study will include improving the estimate of the IRs for a more realistic room environment, development of a method to correct suppression filter mismatch, expansion of the method to suppress more than one source, change of the number of sources and different types of background noise.

5. ACKNOWLEDGEMENT

This work has been partly supported by the European project ERC-2012-AdG-320684-CHESS.

6. REFERENCES

[1] E.C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *J. Acoust. Soc. Am*, vol. 25, no. 5, pp. 975–979, 1953.

[2] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press, 2010.

[3] S.M. Naqvi, Y. Zhang, and J.A. Chambers, "Multimodal Blind Source Separation for Moving Sources," *Acoustics, Speech and Signal Processing, ICASSP. IEEE International Conference on*, pp. 125–128, 2009.

[4] S.M. Naqvi, M. Yu, and J.A. Chambers, "A Multimodal Approach to Blind Source Separation of Moving Sources," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 5, pp. 895–910, 2010.

[5] S.M. Naqvi, W. Wang, M.S. Khan, M. Barnard, and J.A. Chambers, "Multimodal (Audiovisual) Source Separation Exploiting Multi-Speaker Tracking, Robust Beamforming and Time-Frequency Masking," *Signal Processing, IET*, vol. 6, no. 5, pp. 466–477, 2012.

[6] D. Gatica-Perez, G. Lathoud, J.M. Odobez, and I. McCowan, "Audiovisual Probabilistic Tracking of Multiple Speakers in Meetings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 601–616, 2007.

[7] J. Harris, S.M. Naqvi, B. Rivet, J.A. Chambers, and C. Jutten, "Visual Based Reference for Enhanced Audio-Visual Source Extraction," *9th IMA Int. Conf. on Mathematics in Signal Processing*, December 2012.

[8] L. Yanfeng, J. Harris, C. Gaojie, S.M. Naqvi, C. Jutten, and J.A. Chambers, "Auxiliary Function Based Indpendent Vector Analysis Using a Source Prior Exploiting Fourth Order Relationships," *Proc. EUSIPCO*, 2013.

[9] S. Cecchi, A. Primavera, F. Piazza, and A. Carini, "An Adaptive Multiple Position Room Response Equalizer," *Proc. EUSIPCO*, 2011.

[10] S.J. Elliott and P.A. Nelson, "Multiple-Point Equalization in a Room Using Adaptive Digital Filters," *J. Audio Eng. Soc*, vol. 37, no. 11, pp. 899–907, 1989.

[11] J.B. Allen and D.A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am*, vol. 65, no. 4, pp. 943–950, 1979.

[12] G.H. Golub and C.F. Van Loan, *Matrix Computations*, vol. 3, Johns Hopkins University Press, 1996.

[13] B. Rivet, L. Girin, and C. Jutten, "Visual Voice Activity Detection as a Help for Speech Source Separation from Convolutive Mixtures," *Speech Communication*, vol. 49, no. 7-8, pp. 667–677, 2006.

[14] C. Févotte, R. Gribonval, and E. Vincent, "BSS EVAL Toolbox User Guide," Tech. Rep. 1706, IRISA Technical Report 1706, Rennes, France, 2005.

[15] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, 2008.

[16] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle Filtering Algorithms for Acoustic Aource Localization," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 826–836, November 2003.

[17] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of Multiple Moving Speakers with Multiple Microphone Arrays," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 5, pp. 520–529, 2004.

[18] T. Ajdler, L. Sbaiz, and M. Vetterli, "The Plenacoustic Function and its Sampling," *Signal Processing, IEEE Transactions on*, vol. 54, no. 10, pp. 3790–3804, 2006.