Principal Angles Approach to Time-domain Filter Design for Target Cancelation

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

We consider the design of a pair of time-domain filters to achieve target signal cancelation in a multi-source environment. The problem is formulated as a minimization of a sum squared error cost function with respect to the pair of finite impulse response cancelation filters. Direct minimization is achieved through an alternating gradient descent based method, whereas a novel method based on the method of principal angles is proposed which exploits the singular value decomposition. Simulation studies show that the gradient descent method suffers from slow convergence but this is overcome by the method based on principal angles which also achieves a lower cost than the gradient descent approach. The cancelation filters are then combined with an adaptive filtering scheme to address a video-informed audio source separation problem and preliminary results suggest good performance in terms of objective measures.

Index Terms— beamforming, method of principal angles, speech source separation, cancelation filter estimation

1. INTRODUCTION

Reverberant, noisy and multi-source environments pose a significant challenge in signal processing systems particularly in real-time applications. Often multi-sensor array systems are required to enhance or cancel a target signal source by means of spatial filtering so that the target, or other measured signals, can be processed more efficiently.

A beamformer, which spatially filters measurements from an array of sensors (e.g. microphones) is often employed to achieve such selectivity [1, 2, 3]. With broadband signal sources, such as speech, such beamformers are commonly implemented in the frequency domain. In some applications, however, the size of the array can be limited, so only two sensors can be employed. In this context, in [1], a frequency domain generalized sidelobe canceller (GSC) has been proposed. The processing at each discrete frequency, f, in this GSC is represented in Figure 1. On the left-hand side of the diagram is a lattice structure which at the output of the adder enhances the target signal, whereas at the bottom, due to the subtraction, blocks the target signal so that the input to the adaptive filter nominally contains only other background signals.

Such a frequency domain approach however assumes that the length of the discrete Fourier transform (DFT) used to convert the time-domain sensor measurements into the frequency domain is significantly longer than the impulse responses of the filters used to model the propagation between the sources and the array sensors. In contrast, in order that the adaptive filter in Figure 1 can converge there must be a sufficient number of frequency domain blocks, indexed by t, and this requires the impulse responses modelling the propagation environment to be fixed throughout this period. These assumptions are likely to be violated in many applications therefore our work is focused on two sensor time-domain type GSC processing so we need to design a target signal blocking operation in the time-domain. Operation entirely in the time-domain additionally avoids complex valued signal operations and thereby also has computational advantage in real-time implementation.

In the paper we therefore present two methods to estimate a pair of time-domain finite impulse response filters which suppress any undesired signal components which may pass through the blocking channel due to steering error. This pair of filters helps ensure that energy of the canceled signal, after the blocking vector has been applied, is as small as possible (denoted as $u(t, f)$ in Fig. 1), we refer to this pair of filters as cancelation filters in the paper.

The problem formulation is described in Section 2.1, an alternating gradient descent method is introduced in Section 2.2, afterwards the method of principal angles is introduced in Section 2.3. Then, results for an evaluation of both methods are presented is Section 3.1. Finally, the method of principal angles is applied to a video-informed audio source separation problem which is included in Section 3.2.

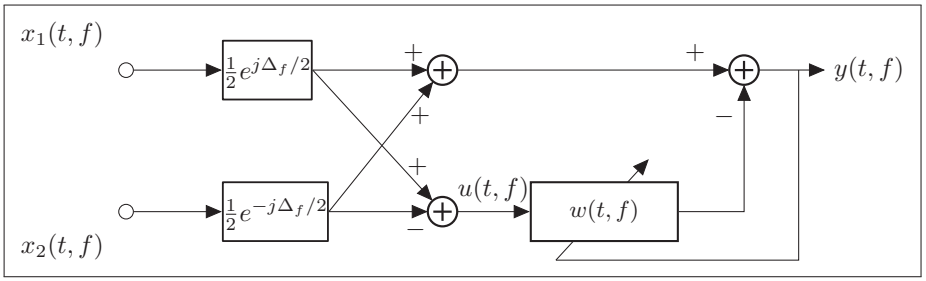

Fig. 1: Two channel generalized sidelobe canceler in the frequency domain. The signal $u(t,f) = \mathbf{b}^H(f)\mathbf{x}(t,f)$ for which the target signal has been blocked, where $\mathbf{b} = 1/2[e^{j\Delta_f/2}, -e^{-j\Delta_f/2}]^H$ is the blocking vector, $\mathbf{x}(t,f)$ is a vector of the short-time Fourier transforms of the time-domain quantities $x_{1,2}$, $w(t,f)$ is the complex parameter in the adaptive filtering stage, Δ_f is the 'uncertainty in angle arrival' which is the time shift to correct for delay in signal arrival, t is the time block index, f is the frequency bin index and $(\cdot)^H$ is the Hermitian (complex conjugate) transpose. Further details can be found in [1].

2. METHOD

2.1. Problem Formulation

The observation at each sensor of a two-sensor array can be modeled in the general case in the time-domain as a convolutive mixture from each source of the form:

$$x_i(k) = \sum_{j=1}^{N} h_{ij}(k) * s_j(k) + n_i(k), \qquad i = 1,2 \tag{1}$$

where s_j is the speech signal generated by the j-th source, h_{ij} is the filter that models the effect of the environment between the j-th source and the i-th sensor, k is the discrete time index, n_i is additive zero mean noise uncorrelated with the speech signals, x_i is the detected signal at the i-th sensor, $*$ denotes convolution and N is the number of sources. For convenience the noise, $n_i(k)$ is dropped for the remainder of the paper. Throughout the paper source number $j = 1$ is the target source that is to be canceled.

In the training phase the sensors are pre-steered so that $\mathbf{h}_{11} \approx \mathbf{h}_{21}$ as this gives the system the best chance of canceling the target by using the raw signals from the sensors. In an acoustic environment application this would be implemented by exploiting the geometery of the acoustic environment, by ensuring the distances between the target signal source and the sensors were equidistant, so that in terms of early reverberation the IRs would be essentially the same. The canceling filters would then correct for the fact that $\mathbf{h}_{11} \approx \mathbf{h}_{21}$.

The core problem formulation is to find a pair of canceling filters ($\hat{\mathbf{g}}_1$ and $\hat{\mathbf{g}}_2$), so that: $\mathbf{h}_{11} * \hat{\mathbf{g}}_1 - \mathbf{h}_{21} * \hat{\mathbf{g}}_2 \approx \mathbf{0}$. An error vector is therefore formulated as:

$$\boldsymbol{\epsilon}_1 = (X_1\mathbf{g}_1 - X_2\mathbf{g}_2) \tag{2}$$

where X_1 and X_2 are the convolution matrices, whose elements are formed from x_1 and x_2 signals as shown in Figure 3, which themselves are convolutions of the target source with h_{11} and h_{21} respectively assuming the other sources are silent during training.

2.2. Alternating Gradient Descent Method

A cost function for the alternating gradient descent method (GD method) is derived from the error vector from Eq. (2), which yields:

$$J_1 = ||\boldsymbol{\epsilon}_1||_2^2, \{\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2\} = \arg\min_{\mathbf{g}_1,\mathbf{g}_2} J_1 \tag{3}$$

The assumption is made that $X_1 \neq X_2$ (i.e. they differ sufficiently so that Eq. (2) cannot be factorized as $X_1(\mathbf{g}_1 - \mathbf{g}_2)$). Taking the partial derivatives of J_1 with respect to the filters to be estimated $\mathbf{g}_1$ and $\mathbf{g}_2$, yields:

$$\frac{\partial J_1}{\partial \mathbf{g}_1} = 2X_1^T X_1\mathbf{g}_1 - 2X_1^T X_2\mathbf{g}_2 \tag{4a}$$

$$\frac{\partial J_1}{\partial \mathbf{g}_2} = 2X_2^T X_2\mathbf{g}_2 - 2X_2^T X_1\mathbf{g}_1 \tag{4b}$$

where $(\cdot)^T$ denotes the vector transpose operator. To minimize the cost function, J_1, the two expressions for the gradient, $\frac{\partial J_1}{\partial \mathbf{g}_1}$ and $\frac{\partial J_1}{\partial \mathbf{g}_2}$ are included in a gradient descent scheme, which updates filter weights according to a change proportional to the gradient of the cost function. Thus, this yields the update equations for the estimated filters:

$$\hat{\mathbf{g}}_1^{\ell+1} = \hat{\mathbf{g}}_1^{\ell} + \mu(X_1^T X_2\hat{\mathbf{g}}_2^{\ell} - X_1^T X_1\hat{\mathbf{g}}_1^{\ell}) \tag{5a}$$

$$\hat{\mathbf{g}}_2^{\ell+1} = \hat{\mathbf{g}}_2^{\ell} + \mu(X_2^T X_1\hat{\mathbf{g}}_1^{\ell+1} - X_2^T X_2\hat{\mathbf{g}}_2^{\ell}) \tag{5b}$$

where ℓ denotes the iteration number and μ denotes the step size. Notice, that in Eq. (5a) $\hat{\mathbf{g}}_2$ is fixed and the update is performed with respect to $\hat{\mathbf{g}}_1$, whereas the reverse applies in Eq. 5b, hence this is an alternating descent. The scale factor of 2 has been factored out and absorbed by μ. The condition $||\hat{\mathbf{g}}_2||^2 = 1$ is applied so that the trivial zero solution is avoided, equally $||\hat{\mathbf{g}}_1||^2 = 1$ could also be applied, though only one condition is used so the remaining filter has more

freedom to reach its optimized value. This is achieved by adding the update $\hat{\mathbf{g}}_2^{\ell+1} = \hat{\mathbf{g}}_2^{\ell+1}/||\hat{\mathbf{g}}_2^{\ell+1}||$ after Eq. (5b).

This constrained optimization corresponds to modifying the cost $J_1 = J_1 + \lambda(||\mathbf{g}_2|| - 1)$, where λ is a Lagrange multiplier. Such an approach to canceler design has been adopted in stereophonic echo cancelation [4], and has been known to exhibit poor convergence due to the correlation between the two signal channels.

2.3. Principal Angles Method

In a similar fashion to the previous method $\hat{\mathbf{g}}_{\{1,2\}}$ are estimated during a training phase. During the training phase only the target signal source is active whilst the other sources are assumed to be silent.

The novelty in our work is that the method of principal angles (PA method) is used to find the filter estimates ($\hat{\mathbf{g}}_1$ and $\hat{\mathbf{g}}_2$), as described in [5], which should overcome the slow convergence in the gradient descent method. To use the method of principal angles we need an orthonormal basis for the convolution matrices; taking the QR decomposition of X_1 and X_2, the error vector is rewritten as;

$$\boldsymbol{\epsilon}_2 = (Q_1\tilde{\mathbf{g}}_1 - Q_2\tilde{\mathbf{g}}_2) \tag{6}$$

where $\tilde{\mathbf{g}}_1 = R_1\mathbf{g}_1$ and $\tilde{\mathbf{g}}_2 = R_2\mathbf{g}_2$. The minimizers of a new cost function are then found as:

$$J_2 = ||\boldsymbol{\epsilon}_2||_2^2,\ \{\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2\} = \arg\min_{\tilde{\mathbf{g}}_1,\tilde{\mathbf{g}}_2} J_2 \tag{7}$$

To find the principal angles and principal vectors of the orthonormal subspaces Q_1 and Q_2, we take the singular value decomposition of $Q_1^TQ_2$, so that $[U, \Sigma, V^T] = SVD(Q_2^TQ_1)$. The constraints $||\tilde{\mathbf{g}}_1||^2 = 1$ and $||\tilde{\mathbf{g}}_2||^2 = 1$ are inherently introduced to the method by taking the SVD of $Q_2^TQ_1$, which avoids the trivial solution $\hat{\mathbf{g}}_1 = \hat{\mathbf{g}}_2 = \mathbf{0}$. Let

$$UQ_1 = [\mathbf{f}_1^1|\ldots|\mathbf{f}_1^p] \tag{8a}$$
$$VQ_2 = [\mathbf{f}_2^1|\ldots|\mathbf{f}_2^q] \tag{8b}$$

be the column partionings of UQ_1 and VQ_2, where $[\mathbf{f}_1^1|\ldots|\mathbf{f}_1^p]$ and $[\mathbf{f}_2^1|\ldots|\mathbf{f}_2^q]$ are the principal vectors of the orthonormal bases Q_1 and Q_2 and, p and q are the dimensions of the matrices X_1 and X_2 respectively (we ensure that X_1 and X_2 are the same size so; $p = q$). Assuming that there are two unit vectors $\mathbf{a}$ and $\mathbf{b}$, so that; $\mathbf{f}_1^1 = Q_1\mathbf{a}$ and $\mathbf{f}_2^1 = Q_2\mathbf{b}$. Thus,

$$\mathbf{f}_1^T\mathbf{f}_2 = \mathbf{a}^TQ_2^TQ_1\mathbf{b} = \mathbf{a}^T(U\Sigma V^T)\mathbf{b}. \tag{9}$$

This function is maximized by setting $\mathbf{a} = \mathbf{u}^1$ and $\mathbf{b} = \mathbf{v}^1$, where the superscript, $(\cdot)^1$, corresponding to the index of the largest diagonal value of Σ (denoted by σ^1 which in turn corresponds to the smallest angle between the orthonormal bases Q_1 and Q_2). To find the principal vectors it follows that; $\mathbf{a} = Q_1\mathbf{u} = \mathbf{f}_1^1$, and $\mathbf{b} = Q_2\mathbf{v} = \mathbf{f}_2^1$. The equalizing filters are the columns of U and V which correspond to σ^1 (as they maximize Eq. (9)), multiplied by the inverse of R_1 and R_2 to allow for the basis change by the QR decomposition, thus:

$$\hat{\mathbf{g}}_1 = R_1^{-1}\mathbf{v}^1 \tag{10a}$$
$$\hat{\mathbf{g}}_2 = R_2^{-1}\mathbf{u}^1 \tag{10b}$$

We next compare the performance of the two approaches.

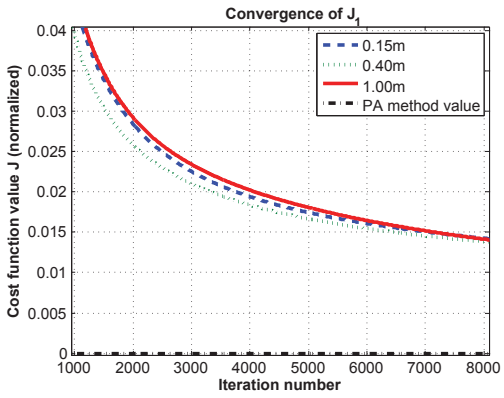

Fig. 2: Convergence performance of the GD method. The strong correlation between both sensor signals x_1 and x_2, as they share a common source convolved with similar IRs, cause slow convergence. Also included is the cost function value achieved by the PA method which is much nearer zero.

3. RESULTS

3.1. Cancelation Filter Performance

The cancelation filter methods were compared by calculating the value of the respective cost functions with the estimated filter vectors for the PA method and the GD method, i.e. $||X_1\hat{\mathbf{g}}_1 - X_2\hat{\mathbf{g}}_2||^2$ for both methods. Binaural room impulse responses (BRIRs) from a classroom were measured with a dummy head between two microphones [6] and then resampled to 8kHz. These BRIRs and a white noise input are used to train the cancelation filters, where the target signal source was positioned at 0°, and at distances of 15cm, 40cm and 1m from the array as marked in Table 1.

Strong correlation between the sensor signals x_1 and x_2 causes slow convergence for the GD method as shown in Fig. 2, normalized values of the cost function J_1 are given after 8100 iterations of the update equations where the cancelation filters' lengths are 810 taps. To train the cancelation filters 10000 samples of white noise were used, this number was chosen to limit the size of the of X_1 and X_2 to save on computational load.

Table 1 clearly shows that the principal angles method offers better performance than the GD method, values are shown for normalized cost, that is to say the raw value from

the cost function divided by the length of the estimated filter. The slow convergence of the alternating GD method also reduces performance, lower normalized cost function values could be achieved if the GD method was run for more iterations, but this would introduce a delay in real-time systems. This is an advantage of the PA method as it finds the optimal filters without the need of update iterations.

Table 1: Values of the cost function with estimated filters. Filters of length 810 were estimated for both methods.

Distance (m)	PA (normalized cost)	GD (normalized cost)
0.15	4.04×10^{-8}	1.42×10^{-2}
0.40	4.04×10^{-8}	1.38×10^{-2}
1.00	4.04×10^{-8}	1.38×10^{-2}

3.2. Video-Informed Source Separation Application

In this section the PA method with an adaptive filtering scheme is used as an alternative to classical higher-order statistics source separation methods (such as independent component analysis [7]) to address the cocktail party problem [8]. An array of two microphones (sensors) are pre-steered towards the target so that IRs between a speaker and the two microphones, which are positioned close together (5cm), are approximately equal, $\mathbf{h}_{11} \approx \mathbf{h}_{21}$, as the microphones are the same distance from the target source.

The microphones are assumed to be pre-steered by video information which provides the location of the target speech source. In practice a microphone array would be orientated towards the target source using a mechanical device. The use of video information is much more robust to background noise than an audio based method for source localization. More emerging robotic human machine interfaces are likely to be equipped with cameras. The extraction of localization information from video information for pre-steering the array is outside the scope of this method, but further details can be found in [9, 10, 11].

In the training phase, the same BRIRs and white noise input as before are used to create mixtures at each microphone, where only the target source is present. The estimated cancelation filters, $\hat{\mathbf{g}}_{\{1,2\}}$, are found for an angle of 0° and distances of 0.15m, 0.40m and 1.00m from the microphone array. After the training phase, the noise reference source (s_{ref}) is then added at 75° and 1m from the array.

At a particular distance, the target source ($\hat{s}_{tar}$) is canceled from the mixture leaving the other source (interference), $\hat{s}_{ref}$. The canceled target source $\hat{s}_{tar}$ is then recovered by using $\hat{s}_{ref}$ as a noise reference in a normalized least mean square (NLMS) adaptive filtering scheme [12]. A diagram of the full system is given in Fig. 3, including the mixing process.

Physically both the target source and reference source are both stationary. Source speech signals were taken from the TIMIT database, where the target source is a male voice and the noise reference voice is female. Six TIMIT files were concatenated to achieve the desired length.

The method is evaluated in the two-microphone two-source scenario, Table 2. Peak and average performance values for signal-to-interference (SIR) [13] and perceptual evaluation of speech quality (PESQ) [14] are given for $\hat{s}_{tar}$.

Table 2: Audio Source Separation: Enhancement improvement for mixtures with BRIRs, when filter lengths for $\hat{\mathbf{g}}_{\{1,2\}}$ are 1300 at 8kHz and the length of $\hat{\mathbf{w}}$ is 6500, source duration of 4 mins 7 secs and $\mu = 0.275$. PESQ values for the unseparated mixtures are 0.96 and 1.00 for x_1 and x_2 respectively.

Distance (m)		SIR (dB)	PESQ (0-5)
0.15	Peak	16.13	4.22
	Mean	8.94	2.47
0.40	Peak	15.82	4.23
	Mean	8.18	2.49
1.00	Peak	14.29	3.21
	Mean	7.13	1.66

The filters $\hat{\mathbf{g}}_{\{1,2\}}$ and the room impulse responses cause the outputs of the algorithm $\hat{s}_{tar}$ (and $\hat{s}_{ref}$) to be filtered versions of the original sources. As expected both performance measures degrade with increasing target source distance. But significant SIR is achieved with a peak improvement of 16.13dB (note that as the target and interfering source have the same variance at the microphones the input SIR is 0dB). The additional filtering on $\hat{s}_{tar}$ causes low average PESQ values, however the effect can be reduced by additional post-processing, as in [15]. The improved SIR ratios also suggest that signal leakage is not a major problem in the operation of the adaptive filter.

4. CONCLUSION

Two methods have been proposed for designing time-domain cancelation filters. The more conventional alternating gradient descent based method was shown to converge slowly and to perform badly in terms of the cost function value, even after a significant number of update iterations. An alternative novel method of principal angles was introduced which minimizes the cost function without the need of iterative updates and gives a much lower cost function value.

Both methods are formulated in the time-domain to ensure that any IR of a particular environment can be adequately covered by the cancelation filters.

In the source separation context, the method may be used as a stand-alone source separation method, for a two-source two-microphone scenario, or can be used as a pre-processing stage for a more conventional blind source separation algorithm in the under-determined case. where there are more sources than sensors.

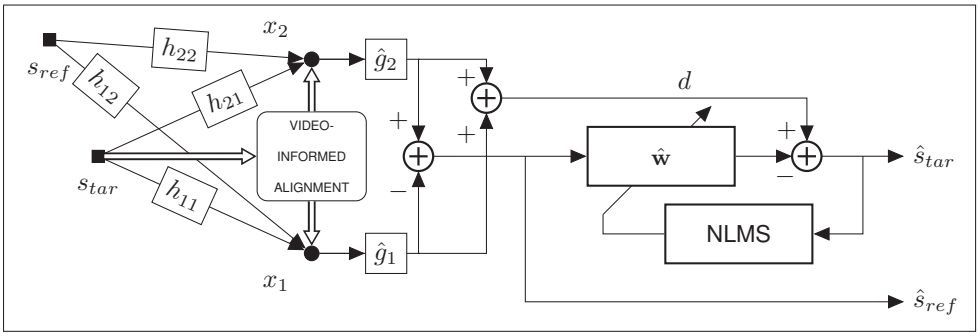

Fig. 3: System overview for the two-sensor configuation, the sensor array is pre-steered towards s_{tar}, the cancelation filters and an NLMS adaptive filtering stage. The overall outputs are $\hat{s}_{tar}$ and $\hat{s}_{ref}$. Thicker hollow arrows indicate information flow in a mechanical/video system which is used to pre-steer the sensor array.

Future work will include implementing the canceling filters from the principal angles method in real-time and expanding this method so that it becomes a pre-processing stage to a conventional source separation algorithm for acoustic sources.

For the audio source separation application, possible changes include removing the training phase and replacing it with a voice activity detector (VAD) [15], which detects silent periods in speech sources.

Acknowledgements: J. Harris is funded by a DGA/DSTL PhD scholarship. This work is partially funded by 2012-ERC-AdG-320684 CHESS. GIPSA-lab is a partner of the LabEx PERSYVAL-Lab (ANR–11-LABX-0025).

5. REFERENCES

[1] I. Cohen, "Analysis of two-channel generalized sidelobe canceller (GSC) with post-filtering," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 684–699, 2003.

[2] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *Antennas and Propagation, IEEE Transactions on*, vol. 30, no. 1, pp. 27–34, 1982.

[3] B.D. Van Veen and K.M. Buckley, "Beamforming: a versatile approach to spatial filtering," *ASSP Magazine, IEEE*, vol. 5, no. 2, pp. 4–24, 1988.

[4] N.T. Forsyth, J.A. Chambers, and P.A. Naylor, "Alternating fixed-point algorithm for stereophonic acoustic echo cancellation," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 149, no. 1, pp. 1–9, 2002.

[5] G.H. Golub and C.F. Van Loan, *Matrix Computations*, vol. 4, Johns Hopkins University Press, 2012.

[6] B. G. Shinn-Cunningham, N. Kopco, and T. J Martin, "Localizing Nearby Sound Sources in a Classroom: Binaural Room Impulse Responses," *J. Acoust. Soc. Am.*, vol. 117, pp. 3100, 2005.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 2002.

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

[14] 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.

[15] 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.