Over-complete Blind source separation by applying sparse decomposition and information theoretic based probabilistic approach

A. Ossadtchi[1] and S. Kadambe

[1]Dept. of ECE, Univ. of Southern California, Los Angeles, CA

HRL Laboratories, LLC, 3011 Malibu Canyon Road, Malibu CA 90265[1]

E-mail: aos@usc.edu and skadambe@hrl.com

ABSTRACT

Both in the case of cellular communication and in the case of spoken dialogue based information retrieval systems on mobile platforms there exist a number of interference signals. Therefore, it is essential to separate these interference signals from the intended signal(s) in order to have clear communication in the case of cellular phone and to improve the speech recognition accuracy in the case of spoken dialogue based information retrieval system. Since the number and nature of source signals (intended + interference signals) change, it is not practical to know them *a priori*. Therefore, it is not always practical to apply signal separation techniques that work well when the number of source signals is equal to the number of sensors. In addition, since how the signals get mixed is unknown, we need to apply blind techniques for the separation. This paper is concerned with a blind source separation (BSS) technique for the over-complete case (#signals > #sensors) that is based on the sparse decomposition and, the joint estimation of mixing matrix and the separated source signals by applying information theoretic based probabilistic approach. Experimental results of signal separation using various real speech and noise signals indicate that the quality of separated source signals are 4 dB better than the current techniques.

1. INTRODUCTION

Both in the case of cellular communication and in the case of spoken dialogue based information retrieval systems on mobile platforms there exist a number of interference signals. The signal that is received at a sensor (e.g., antenna in the case of cellular communication or microphone in the case of spoken dialogue system) is a mixed signal that consists of interference signals and the intended signal(s). Note that here afterwards the interference signals and the intended signal(s) together is referred as source signals. It is essential to separate these interference signals from the intended signal(s) in order to have clear communication in the case of cellular phone and to improve the speech recognition accuracy in the case of spoken dialogue based information retrieval system. Since the number and nature of source signals change, it is not practical to know them *a priori*. Therefore, it is not always practical to apply signal separation techniques that work well when the number of source signals is equal to the number of sensors. In addition, since how the signals get mixed which is referred here, as the mixing matrix is unknown, we need to apply blind techniques for the separation. This paper is focused on a BSS technique for the over-complete case that is based on the sparse decomposition and, the joint estimation of mixing matrix and the separated source signals by applying information theoretic based probabilistic approach. In references [1-3] separation of mixed signals in over-complete case is discussed. In [4] an approach based on the concept of estimating the spatial directions of mixing matrix is applied for BSS. However, in [4] the authors separate source signals from the mixed signals only when number of sensors is equal to number of sources. This paper differs from the previously published papers as follows: (a) The direction estimation of the mixing matrix in the Fourier sparse domain by applying the proposed "dual update" algorithm. (b) The application of information theoretic approach for the initial direction estimation of the mixing matrix. (c) The restoration of source signals in the wavelet sparse domain by applying the probabilistic approach. (d) The derivation of the theoretical Cramer-Rao bound (CRB) for the estimation of source signals and the mixing matrix. The details of the proposed approach are provided in the following section. The simulation details and the experimental results are provided in section 3. In section 4, we summarize and indicate future direction of our research in this area.

2. A BRIEF DESCRIPTION OF THE PROPOSED APPROACH

2.1 Problem formulation:

Let the mixed signal $\boldsymbol{X}$ that is received by an array of sensors be: $\mathbf{X} = \mathbf{AS} + \mathbf{V}$. Where $\boldsymbol{X}$ is an M x T matrix corresponding to the output of M sensors at times t = 1,2,……., T, $\boldsymbol{S}$ is the N x T matrix of underlying source signals, $\boldsymbol{A}$ is the unknown M x N mixing matrix that correspond to the environment effect in mixing the signals and $\boldsymbol{V}$ is an M x T noise matrix. The problem of BSS is then to recover $\boldsymbol{S}$ from $\boldsymbol{X}$ without the prior knowledge of source signals and the mixing matrix $\boldsymbol{A}$. This paper is focused on a BSS technique that can handle $N > M$ - over-complete BSS. Due to additive noise and rectangular mixing matrix $\boldsymbol{A}$, the problem of separating source signals from the mixed signals can be solved more efficiently, by applying techniques based on probabilistic approach. Hence, here, we are focusing on such an approach. Probabilistic approaches mainly correspond to minimizing the log of *a posterior* likelihood function $\log\left(P\left(\mathbf{S}\middle|\mathbf{X},\mathbf{A}\right)\right) \propto \log\left(P\left(\mathbf{X}\middle|\mathbf{A},\mathbf{S}\right)P\left(\mathbf{A},\mathbf{S}\right)\right) = L\left(\mathbf{X}\middle|\mathbf{A},\mathbf{S}\right) + L\left(\mathbf{A}\right) + L\left(\mathbf{S}\right)$ with respect to $\boldsymbol{S}$. Here L corresponds to *log*(*P*()), and L(A) and L(S) correspond to log prior probabilities of $\boldsymbol{A}$ and $\boldsymbol{S}$, respectively. The minimization of log likelihood function then corresponds to minimizing $L\left(\mathbf{X}\middle|\mathbf{A},\mathbf{S}\right) + L\left(\mathbf{S}\right)$ with respect to $\boldsymbol{S}$

since there is no prior information on ***A***. Since the separation of ***S*** depends on ***A***, we propose that by jointly optimizing the above log likelihood function with respect to ***A*** and ***S*** we can separate the source signals from ***X*** more efficiently. We also have noticed this dependency when we derived the theoretical Cramer-Rao bound for the estimation of ***A*** and ***S*** [6]. For this joint optimization, we have developed a novel algorithm that is described in the next section.

2.2 Description of joint optimization algorithm

In the joint optimization problem, for mathematical simplicity, we assume that the (a) the source signals are statistically independent of each other and follow Laplacian probability distribution in the sparse Fourier and wavelet domains and, (b) noise ***V*** is Gaussian. To reduce the complexity of the problem of separation of mixed signals, we first transform it to the sparse domain by applying the wavelet/short-time Fourier transform. This has another advantage of reducing the noise effect – by thresholding the wavelet/Fourier coefficients we can achieve de-noising. We then apply the proposed probabilistic approach of BSS in the sparse domain. The observed mixed signals in the transformed domain has the same form as that of the time domain - $W(\mathbf{X}) = \mathbf{A}W(\mathbf{S}) + W(\mathbf{V})$ where W is either Fourier or wavelet transform. Therefore, the general probabilistic approach mentioned in the previous section is applicable here. To get the separated source signals back from the transformed domain to the time domain, the inverse wavelet or Fourier transform is applied. For the joint optimization, we start with $L\left(W(\mathbf{S})\middle|W(\mathbf{X}),\mathbf{A}\right)$. In the next subsection, this joint optimization is described briefly.

2.2.1 Proposed "dual update algorithm"

With the assumption of Laplacianity of source signals in the sparse domain the prior probability $P\left(W(\mathbf{S})\right) = \frac{\lambda}{2} e^{-\lambda \mathbf{c}^T |W(\mathbf{S})|}$ where $\mathbf{c}^T = [1,1,\cdots 1]$ a unit vector. By applying the "Laplacianity" of signals, "Gaussianity" of noise and no prior information on ***A***, it can be shown that

$$L\left(W(\mathbf{S})\middle|\mathbf{A}, W(\mathbf{X})\right) = \left(W(\mathbf{X}) - \mathbf{A}W(\mathbf{S})\right)^T \mathbf{R}_{W(\mathbf{v})}^{-1} \left(W(\mathbf{X}) - \mathbf{A}W(\mathbf{S})\right)$$

$$+ \lambda c^T W(\mathbf{S}) \text{ where } \mathbf{R}_{W(\mathbf{v})} \text{ is the noise covariance matrix.}$$

The above equation can shown to be:

$$L\left(W(\mathbf{S})\middle|W(\mathbf{X}),\mathbf{A}\right) = \sum_{t=1}^{T}\left(W(\mathbf{x}_t) - \mathbf{A}W(\mathbf{s}_t)\right)^2 + \lambda \mathbf{c}^T W(\mathbf{s}_t)$$

where $\mathbf{x}_t$ & $\mathbf{s}_t$ are the column vectors of $\mathbf{X}$ & $\mathbf{S}$.

This is obtained with the unit covariance assumption and using individual components of ***X*** & ***S***. By differentiating the above equation with respect to ***A*** and setting it to zero**,** using the individual components of W(**S**), ***A***, and replacing the summation with the expectation operation this equation can be written as:

$$E\left\{W(\mathbf{x}_t)^T W(\mathbf{s}_t^i)\right\} = E\left\{\left(\sum_{j=1}^{M} \mathbf{a}^T{}_j W(\mathbf{s}_t^j)^T\right) W(\mathbf{s}_t^i)\right\}.$$

$E\left\{W(\mathbf{s}_t^i)^T W(\mathbf{s}_t^j)\right\} = 0$ for $i \neq j$ since the source signals are statistically independent in the sparse domain. Using this we can write the above equation as: $\Sigma_{\mathbf{XS}} = \left[\mathbf{a}_1, \mathbf{a}_2, \cdots \mathbf{a}_M\right]\Sigma_{\mathbf{S}}$ where $\Sigma_{\mathbf{S}}$ is the $\mathrm{cov}(W(\mathbf{S})) = E\left\{W(\mathbf{S})^T W(\mathbf{S})\right\}$ and $\Sigma_{\mathbf{XS}} = E\left\{W(\mathbf{X})^T W(\mathbf{S})\right\}$

$= [\sigma_{\mathbf{Xs}_1}, \sigma_{\mathbf{Xs}_2} \cdots \sigma_{\mathbf{Xs}_N}]$

Then the estimated ***A*** matrix is:

$$\hat{\mathbf{A}} = \left[\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2 \cdots, \hat{\mathbf{a}}_M\right] = \Sigma_{\mathbf{XS}}\Sigma_{\mathbf{S}}^{-1}. \qquad \text{Equation (1)}$$

There is no closed form solution to solve this set of equations but can be solved iteratively by applying the Linear Equality Constraints (LEC) optimization technique [5]. The LEC optimization in turn corresponds to application of line search together with the projection gradient method. For the line search we apply the Armijo rules. By applying the LEC, we can solve Equation (1) iteratively by using the following two steps:

1. Find $W(\hat{\mathbf{S}})$ that min $\lambda \mathbf{c}^T \left|W(\hat{\mathbf{S}})\right|$
2. Use $W(\hat{\mathbf{S}})$ from 1 and estimate $\hat{\mathbf{A}} = \Sigma_{\mathbf{XS}}\Sigma_{\mathbf{S}}^{-1}$

From the above, it can be seen that a good initial estimate of ***A*** is needed to get good initial estimate of $\hat{\mathbf{S}}$ that is used in step 1 above. Note that the above problem is not convex in ***A*** thus even though splitting of the optimization into two parts is not theoretically justified, we found that given a good initial estimate of the directions of ***A*** the proposed dual update algorithm can improve the accuracy of estimation of ***A*** and ***S***. Also since we are interested in finding directions of lines (angles of lines) we consider ***A*** matrix as a function of θ. In the next section we describe a method that we have developed which is based on [3,4] but augmented with the newly introduced adaptive thresholding technique to get a good initial estimate $\hat{\mathbf{A}}$.

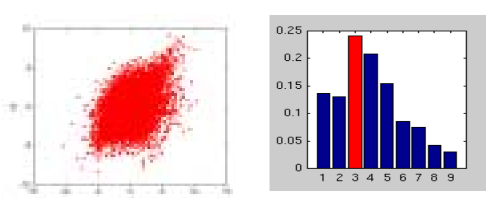
2.2.2 Information theoretic based initial estimate of *A*

For the initial estimate of ***A*,** the Fourier sparse domain is considered as opposed to wavelet domain since finer frequency resolution can be obtained and hence, more accurate initial estimate of ***A***. The observed signals are transformed to the sparse Fourier domain by applying the spectrogram. The spectrogram was computed using the Hamming window of length 16 samples. The window was shifted by 8 samples. We then choose a frequency sub-band for the initial estimation of ***A*** by applying the maximum mutual information criterion. In essence, by applying this criterion we are finding a frequency sub-band in which the directions of spread of observed mixtures are as well resolved as possible. To further improve the resolution of directions of spread of observed mixtures in the chosen frequency sub-band, we apply a novel thresholding technique. To explain this approach let us define a random variable $ang = \arctan\frac{x_1(band)}{x_2(band)}$. Spatially white additive noise appearing in the mixtures transforms into uniformly distributed random variable (RV) in *ang*-domain. On the contrary, spatially localized sources transform into RV with well defined means and relatively small variances. The resulting distribution of variable *ang* can be described as multi-modal distribution corresponding to spatially local sources "masked" by a uniform distribution.

This masking effect prohibits from resolving all local maxima of the histogram in order to obtain the first estimate of the directions of mixing matrix. It should be clear from a scatter plot (see Figure 1d) that thresholding operation applied to the observed mixtures should reduce the masking of the maxima of a histogram of *ang*. The proposed approach to determine the threshold value is based on measuring the entropy of *ang* obtained from the thresholded mixture. The masking of uniform distribution tends to increase the entropy. Let us define a function *E*(*ang*, ANG) equal to the entropy of *ang* obtained from the observed thresholded mixture with threshold value set to ANG. We keep increasing the value of ANG within a chosen range until the rate of descent of function *E*(*ang*, ANG) is minimum. We then choose the value ANG that minimized this function as the threshold value to threshold the mixtures in the Fourier domain. In this study the range of ANG is set to $\left[0.1 * Axb \;\; 2 * Axb\right]$ where Axb = cov(spectral values in the chosen frequency sub-band). The probability values needed for the entropy computation are obtained by estimating the distribution of spectral values in the chosen sub-band by using the histogram approach.

To make the concepts described in this section more clear an example is provided below. For this example five sources (three speech sources and two noise sources) and two sensors were considered in the Fourier sparse domain. In the following figure, first the scatter plot (Figure 1a) of these two observed mixtures is provided. From this plot it can be seen that it is very hard to resolve the direction of spread of source signals. Next, the bar plot (Figure 1b) of mutual information of different frequency sub-bands is provided. The red colored bar in this plot corresponds to the band that has the maximum mutual information. This band - *Nband* is chosen for the initial estimate of ***A***. The first step in this estimation is to obtain the probability distribution function of ***A*** – mixing matrix of angles. This is obtained by computing $\arctan \dfrac{\mathbf{x}_1(Nband)}{\mathbf{x}_2(Nband)}$. The histogram of this angle matrix is plotted next (Figure 1c) and also the scatter plot (Figure 1d) corresponding to this chosen sub-band. From this scatter plot it can be seen that it is possible to resolve the directions of spread of source signals as marked by the black lines. However, this resolution can be improved further if we could threshold the spectral values that are in the center (the "masking" area marked by a blue circle in this figure). For the threshold selection the entropy based technique described above was applied. In Figure 1e, the entropy versus threshold values is plotted. The threshold value corresponding to the minimum rate of descent of the entropy function *E*(*ang*, ANG) was chosen automatically. This is marked in this plot by a red line. After thresholding the spectral values, the angles were recomputed. The histogram of these angles are plotted next (Figure 1f). From this, it can be seen that the three local maxima that correspond to angle of spread of three source signals are well pronounced. These local maxima were automatically detected and were used in the initial estimate of ***A***.

(a)



(b)

(c) (d) (e) (f)

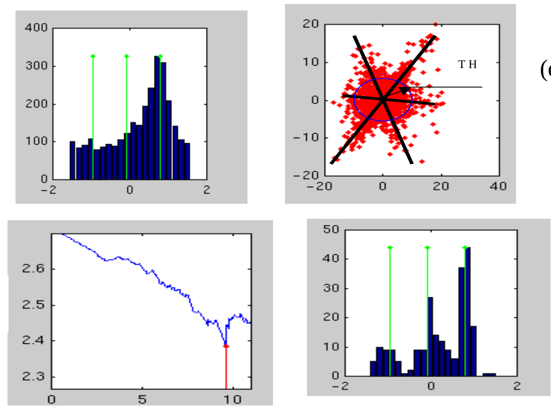


Figure 1. (a) Scatter plot of original mixed data, (b) Bar graph of Mutual information vs. frequency sub-bands, (c) Angle histogram of chosen sub-band, (d) Scatter plot of spectral values of the chosen sub-band, (e) Entropy plot with threshold selection (red line indicates the minimum entropy) and (f) the angle histogram of the chosen band after thresholding.

Using this initial estimate of ***A***, initial estimate of $\hat{\mathbf{S}}$ is obtained. Then the "dual update" algorithm described above was applied. When the stopping criterion was reached the final estimate of ***A*** matrix was obtained.

2.2.3 Time-courses restoration of source signals

The next step in the proposed over-complete BSS is the restoration of separated source signals. For this, we use the final estimate of ***A*** obtained from the technique described in the last section and we transform the observed signals in to wavelet sparse domain. This domain is used for the reconstruction of source signals because in the case of the spectrogram the phase information is lost and the restored separated source signals will not be accurate. One could argue then why not apply the proposed "dual update" algorithm in the wavelet sparse domain and estimate both ***A*** and restore the separated source signals. We could have; however, note that the proposed dual update algorithm will estimate ***A*** up to permutations which requires finding the proper order for the source signal separation in each wavelet sub-band. In order to overcome this problem we apply the "dual update" algorithm twice – once in the sparse Fourier domain and once in the sparse wavelet domain. We use the ***A*** matrix estimated in the Fourier domain while restoring the source signals in the wavelet domain and minimize the log likelihood function $\min_{W(\mathbf{S})} \left(L\left(W(\mathbf{S}) \middle| \mathbf{A}, W(\mathbf{X}) \right) \right)$. For this minimization also we apply the "dual update" algorithm. Note that even though we apply the same "dual update" algorithm here, we only update the ***S*** matrix and stop the iterative procedure when the stopping criterion is reached.

To verify the performance of our algorithm we conducted experiments using speech signals and noise signals as source signals, and two sensors (microphones). Experimental details and the simulation results are provided in the next section.

3 SIMULATIONS

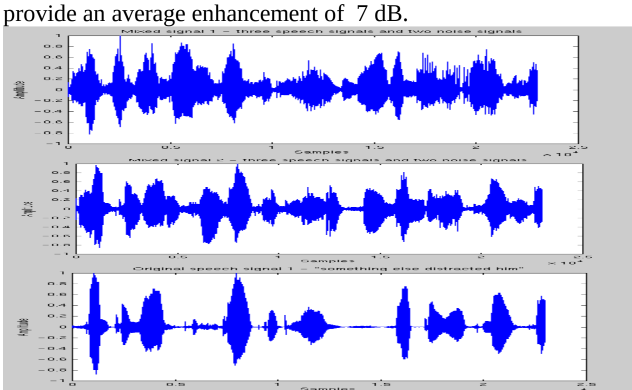
First three speech signals from three female speakers are considered. These correspond to three sentences from the TIMIT database. A babble noise that was selected from the NOISEX0 database is considered as the noise source. This was randomly split into two parts and thus we have two noise source signals. Three speech signals and two noise signals were mixed by generating a 2x3 **A** matrix randomly. For the generation of this mixing matrix, first random numbers were generated using a uniform random number generator, these random numbers were then used in calculating the angles as mentioned above and these angles were next used in forming a **A** matrix. This mixing matrix was then used to mix three speech signals and two noise signals. As a result of this mixing we would get two mixtures that corresponds to received signals at two sensors (microphones). From the two mixed signals, three speech signals were separated using our algorithm. As mentioned above, our algorithm first estimates the **A** matrix in the Fourier domain by applying the proposed dual update algorithm. This estimated **A** matrix is then used to estimate the source signals in the wavelet domain as mentioned above. For the application of the wavelet transform, wavelet packet approach was used. The wavelet packet approach generates a library of bases for a given orthogonal wavelet function. Each of these bases offers a particular way of decomposing or representing (expanding) a signal. The most suitable decomposition of a given signal is then selected by applying the minimum entropy constraint. For the wavelet packet decomposition biorthogonal wavelet of order 6 was used. The dual update algorithm was applied in the wavelet sparse domain to estimate the source signals as described in section 2.2.3. The estimated source signals in the wavelet domain were transformed to the time domain by applying the inverse wavelet transform. In Figure 2, two mixed signals, three original and separated speech signals are plotted. For full details of this study refer to [6]. From this figure it can be seen that all the speech signals are well separated from the mixed signal. In order to quantify how well the signals were separated, we computed the (a) SNR of mixed signals as:

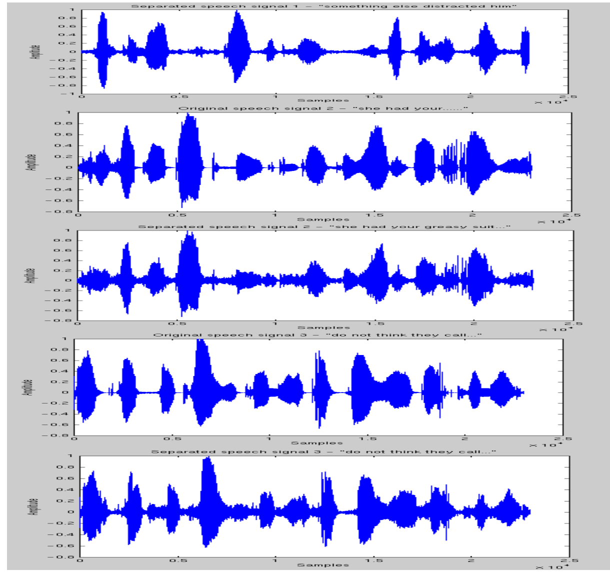
$$SNR_{\mathbf{s}_{i_{mix_k}}} = 10\log_{10}\left(\frac{\operatorname{var}(\mathbf{s}_i)}{\operatorname{var}\left(mix_k - \mathbf{s}_i\right)}\right) i = 1,2,3 \,\&\, k = 1,2$$

and (b) the SNR of estimated source signal as:

$SNR_{\mathbf{s}_i} = 10\log_{10}\left(\frac{\operatorname{var}(\mathbf{s}_i)}{\operatorname{var}\left(\hat{\mathbf{s}}_i - \mathbf{s}_i\right)}\right) i = 1,2,3$. By comparing these two SNRs, it was observed that an average enhancement of 11 dB was obtained. Whereas the state of the art technique [3] provide an average enhancement of 7 dB.

(a)
(b)
(c)





(d)
(e)
(f)
(g)
(h)

Figure 2: (a) & (b) Two mixed signals received at two sensors, (c) original signal 1 and (d) separated signal 1, (e) original signal 2 and (f) separated signal 2, and (g) original signal 3 and (h) separated signal 3, respectively, from top to bottom.

For other speech signals of different languages and different speakers also an average of 11dB SNR enhancement was obtained [6]. Our results indicate this algorithm works well both for different types of noises and different speech signals (languages and speakers). Note that even though for mathematical simplicity we assumed noise as Gaussian noise, this algorithm works well for non-Gaussian noise as indicated by the previous example.

4 CONCLUSIONS

Our results show that our proposed approach can separate the source signals very well when the number of signal sources are more than the number of sensors. We are extending this algorithm to handle complex signals. We are also exploring applications of this algorithm in robust speech recognition and in urban cellular communication.

5 References:

1. M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition," University of New Mexico technical report No. CS99-1, 1999.
2. L. Q. Zhang, S. Amari and A. Cichocki, "Natural gradient approach to blind separation of over and under complete mixtures," Proceedings of ICA'99, Aussois, France 1999.
3. Te-Won Lee, M. S. Lewicki, M. Girolami and S. J. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," IEEE Signal processing letters, Vol. 6, No. 4, pp. 87-90, April 1999.
4. H-C. Wu, J. C. Principe and D. Xu, "Exploring the time-frequency microstructure of speech for blind source separation,", ICASSP'98, pp. 1145-1148.
5. P. E. Gill, W. Murray and M. H. Wright, Practical optimization
6. S.Kadambe and A. Ossadtchi, "Over-complete BSS by applying sparse decomposition and information theoretic based probabilistic approach and its application to robust speech recognition", to be submitted for IEEE trans. On speech & audio processing.