Single Channel Blind Source Separation using the Best Characteristic Basis

Bin Gao, W.L. Woo, and S.S. Dlay
School of Electrical, Electronic and Computer Engineering
University of Newcastle
Newcastle upon Tyne, NE1 7RU
UNITED KINDOM
Email: bin.gao@ncl.ac.uk, w.l.woo@ncl.ac.uk, s.s.dlay@ncl.ac.uk

Abstract: - This paper proposes a novel technique for separating single channel recording of speech mixture using a hybrid of maximum likelihood and maximum a posteriori estimators. In addition, the new algorithm proposes a new approach that accounts for the time structure of the source signals by encoding them into a set of basis filters that are characteristically the most significant. Real time testing of the new algorithm has been conducted and the obtained results are very encouraging.

Keywords: Blind Source Separation, Independent Component Analysis, ML and MAP Estimation

I. INTRODUCTION

Linear blind deconvolution has been well learnt so far and a plethora of methods has been proposed. Most methods are based on higher order statistics which require non-Gaussian source signals [1-3]. Extension of BSS to solving nonlinear mixtures has also been introduced [4-9]. In addition these methods yield good performances only if number of observations more or equal to number of independent components. However, in the practical applications e.g. speech and imaging processing, only one-channel recording is available and the methods based on one channel audio source separation are more practical. But this crux topic has not yet developed enough to make its way out of laboratories. The proposed solutions of [10] Single Channel Source Separation (SCSS) problem are categorized into three branches: firstly, model-based SCBSS, secondly, underdetermined blind source separation, and finally, computational auditory scene analysis (CASA). Model-based SCSS [11-13] techniques are similar to model-based single channel speech enhancement techniques. In other words SCSS can be considered as a speech enhancement in which both the target and interference. In underdetermined BSS techniques [14-17], the sources are projected onto a set of basis functions whose coefficients are as sparse as possible. By using independent component analysis (ICA) nonnegative matrix factorization, or sparse coding, In CASA-based techniques [18-20], the goal is to replicate the process of human auditory system by exploiting signal processing approaches. The main idea is based on an appropriate transform (such as the short-time Fourier transform (STFT)), the observation signal is segmented into time-frequency cells; then using some criteria to group one source.

Our research area is based on the underdetermined blind source separation. The key point is to exploit a priori knowledge of sources such as the basis functions to generate sparse coding. Then the sources are projected onto a set of basis functions whose coefficients are as sparse as possible. The proposed separation algorithm exploit hybrid of maximum likelihood and maximum a posteriori estimators to recover independent components. If the basis functions are not chosen correctly, this will deteriorate the performance of signal separation.

In this paper, the Single Channel mixing problem is considered for the first time and the objective is to provide optimal estimation of the source signals, The contribution of this paper is to provide a novelty method to extract the most significant characteristic features based on [21] general ICA algorithm. The generalized hybrid of maximum likelihood and maximum a posteriori algorithm is then derived where to estimate sources. In the proposed algorithm, the female and male speeches are exploited to test the separations results.

II. PROPOSED MODEL

We assume there are two independent male and female speech sources mixed together. Let \( y^t \) be the observed signal:

\[
y^t = \kappa_1 x_1^t + \kappa_2 x_2^t
\]

Where \( \kappa_i \) is the \( i^{th} \) mixing coefficient, \( x_i^t \) is the \( i^{th} \) source signal, while the individual sources are constructed by basis functions and their coefficients. The basis functions and coefficients learned by ICA constitute an efficient representation of the [15] given time-ordered sequences of a sound source by estimating the maximum likelihood densities. The theoretical basis of the approach is sparse coding. Here sparse means only a small number of coefficients in the representation differ significantly from zero. Then the individual sources can be expressed as:

\[
x_i^t = \sum_{k=1}^{M} a_{ik} s_{ik} = A_i s_i^t
\]

\[
s_i^t = W_i x_i^t
\]

Where \( A_i = [a_{i1}, a_{i2}, ..., a_{iM}] = W_i^{-1} \) is the basis matrix which contains basis functions of \( i^{th} \) source signal and \( s_i^t \) is the basis coefficient of the \( i^{th} \) source signal. \( s_i^t \) and \( x_i^t \) are both vectors. It satisfies:
\[
x'_{i1} = [a_{i1}, a_{i2}, \ldots, a_{iM}] \times [s_{i1}, s_{i2}, \ldots, s_{iM}]
\]

Here \( M = P \), then the basis functions \( A \), and inverse matrix \( W_i \) satisfy full rank. Therefore this equation satisfies generate general linear ICA model. The learning ICA algorithm can be exploited for capturing the source coefficient density. The generalized [19] exponential density model is expressed as:

\[
G(s; q, \beta, u) = \frac{q^q}{2\Gamma(\frac{q}{2})} \exp(-\beta |s-u|^q)
\]

where \( \Gamma(.) \) is the gamma function. The coefficient is determined by parameters mean \( u \) exponent \( q \) and variance \( \beta^{-1} \) By using Maximum log likelihood estimator we obtain the gradient ascent adaption of coefficient density.

\[
\varphi(s) = \frac{\partial \log p(s | q, u, \beta)}{\partial s} = -\text{sign}(s-u)q \beta |s-u|^{q-1}
\]

While the Gaussian, Laplacian and strong Laplacian—speech signal distribution are affected by exponent \( q \). In this case, for simplicity we set mean and variance are zero and unit respectively. For exponent \( q \), we prove the separation results of recovered speech signal are better when \( q < 1 \).

<table>
<thead>
<tr>
<th>Recovered male speech</th>
<th>MSE</th>
<th>( \kappa )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.78064</td>
<td>0.62653</td>
</tr>
<tr>
<td>1</td>
<td>1.4087</td>
<td>0.46589</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Recovered female speech</th>
<th>MSE</th>
<th>( \kappa )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.6979</td>
<td>0.37347</td>
</tr>
<tr>
<td>1</td>
<td>1.5085</td>
<td>0.53411</td>
</tr>
</tbody>
</table>

Table 1: Performance comparisons

In Table 1, the results of Mean Square Error (MSE) are shown and \( \lambda \) denotes the recovered gain. (We initialized \( \lambda = 0.5 \) respectively). The recovered male-female speech results prove that when exponent \( q \) is smaller; the estimation values of individual sources are more close to the original source. If the parameters set mean and variance are given, the optimal \( q \) can be determined from the maximum a posteriori estimator.

\[
\frac{\partial \log p(q, s)}{\partial q} = \frac{\partial \log p(s | q, u, \beta)}{\partial q} + \frac{\partial \log p(q)}{\partial q}
\]

\[
\frac{\partial \log p(s | q, u, \beta)}{\partial q} = \frac{1}{q} + \frac{\Gamma'(q^{-1})}{q^2 \Gamma(q^{-1})} - |s|^q \log q
\]

So the gradient-based expression (7) for updating \( q \) can be derived by evaluating the appropriate derivatives of exponential density model. The inverse of basis functions also can be updated by exploiting this model.

\[
\frac{\partial \log p(s | q, u, \beta)}{\partial W} = W^{-T} + \varphi(s)x^T \propto [I - \varphi(s)s^T]W
\]

Here the natural gradient is exploited to extension and update the inverse of basis functions. In (8) \( I \) is the identity matrix and we assume \( M = P \), then the \( W \) is square matrix.

III. LEARNING RULES

The maximum likelihood (ML) estimator [15] is chosen to maximize the likelihood function (joint distribution). According to the basic fundamental of Independent Component Analysis, statistical independent is defined in terms of probability density the sources are said to be independent if only if:

\[
p_{x_1^{1:T}, x_2^{1:T}}(x_1^{1:T}, x_2^{1:T}) = p_{x_1^{1:T}}(x_1^{1:T})p_{x_2^{1:T}}(x_2^{1:T})
\]

Where \( T \) denotes number of samples we observed and \( p(.) \) denotes the probability density function. Here only two sources will be considered. The source vectors are passed through the fixed basis filters \( W_i \) to generate set of basis coefficients:

\[
\Pr(x_1^{1:T}; W_i) = \prod_{t=1}^{T} \Pr(x_1^t; W_i)
\]

\[
= \prod_{t=1}^{T} \Pr(s_1^t | \text{det } W_i)
\]

\[
= \prod_{t=1}^{T} \prod_{k=1}^{M} p(s_1^t | u, \beta, q) \text{det } W_i
\]

where \( \Pr(.) \) denotes the probability of source. The likelihood function \( L \) can be expressed as:

\[
L = \log \left[ \prod_{t=1}^{T} \Pr(x_1^{1:T}) \Pr(x_2^{1:T}) \right]
\]

\[
\propto \sum_{k=1}^{M} \sum_{t=1}^{T} \log p(s_{1k}^t) + \sum_{k=1}^{M} \sum_{t=1}^{T} \log p(s_{2k}^t)
\]

(11)

For simplicity, we set \( z^t_i = \kappa_i x^t_i \) and exploit gradient-based algorithm evaluate the individual sources:

\[
\frac{\partial L}{\partial \kappa_i} \propto \sum_{n=1}^{M} \sum_{k=1}^{M} \left[ \lambda_2 \frac{\partial \log p(s_{1k}^n | W_{1kn})}{\partial \kappa_i} - \lambda_1 \frac{\partial \log p(s_{2k}^n | W_{2kn})}{\partial \kappa_i} \right]
\]

(12)

While \( m = t - n + 1 \), \( \varphi(s) \) can be obtained by equation 6 and \( w_{dn} = W_i(k, n) \). We force the sum of gain \( \kappa_i \) equal to constant: \( \kappa_1 + \kappa_2 = 1 \). The gain can be estimated by MAP estimator when given the current estimated individual sources.

\[
\Pr(\kappa | x_1^{1:T}, x_2^{1:T}) \propto \Pr(x_1^{1:T}) \Pr(x_2^{1:T}) p_\kappa(\kappa)
\]

\[
\log \Pr(x_1^{1:T}) \Pr(x_2^{1:T}) p_\kappa(\kappa) \propto L + \log p_\kappa(\kappa)
\]

(13)

We exploit gradient descent algorithm to solve \( \kappa_i \)
\[
\frac{\partial \log p(s'_{ik})}{\partial \kappa_i} = \frac{\partial \log p(s'_{ik})}{\partial s'_{ik}} \frac{\partial s'_{ik}}{\partial \kappa_i} = \frac{\partial}{\partial \kappa_i} \left( \frac{\kappa_i s'_{ik}}{\kappa_i^2} \right) = \frac{\kappa_i s'_{ik}}{\kappa_i^2} \cdot \kappa_i = 1 - \kappa_i^2 \quad (14)
\]

### IV. BEST CHARACTERISTIC BASIS

The basis functions and coefficients [23, 24] learned by ICA constitute an efficient representation. Given segments of predefined length out of a time-ordered sequence of scalar values, ICA infers time domain basis filters generating the most probable coefficients. For comparing the separation results between general basis functions and characteristic features, firstly we generate basis functions by exploiting traditional ICA methods (e.g. FastICA).

#### Table 2: Performance comparison \(n \times n\) basis

<table>
<thead>
<tr>
<th>Number of Basis function</th>
<th>MSE</th>
<th>(\kappa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.78064</td>
<td>0.62653</td>
</tr>
<tr>
<td>9</td>
<td>1.0548</td>
<td>0.40019</td>
</tr>
</tbody>
</table>

We may explain the results by analyzing the power spectral of the speech.

In Figure 3, the structure of male (dash line)-female (dot line) power spectral is very similar. There exists substantial overlap between the male and female speeches. So the problem stated here is one of finding sets of basis functions that explain one class of signal well. The bases functions are exactly extracted from speech signal themselves. Therefore, the male-female basis function also exhibits similar features. A single channel containing a mixture of different classes of signals can be projected on the combined bases sets, and sources are isolated by considering the projections on the subsets of bases. If more basis functions are chosen, the two sources are more likely projected on the combined basis sets and the separation results may lead to more errors.

The best characteristic basis functions requires that each basis be incomplete in the general space of all possible signals but complete in the subspace of a single class of signal, and that the classes are disjoint in signal space. Based on separation results, there is too much overlap in signal space between two speakers when decrease number of basis functions (e.g. the separation results in table 2 by exploiting 4 basis functions). One way around this obstacle would be to do the separation in some feature space where there is both better class separation, and the possibility of transformation back to signal space. This proposed method based on cross-correlation function to extract most similar features from
various male basis functions and female basis functions. Figures 4 and 5 show the process to extract best features from general basis functions.

In determining $W_1$, characteristically the best basis functions for the source signals will first be determined. In this paper, cross-correlation is used to identify the characteristically most similar features inherent in the speech signals. This is carried out by firstly normalizing a set of basis functions (as shown in Fig. 4 and 5) which are obtained using standard signal separation algorithm. The two general basis functions are cross-correlated and largest N values (represent two basis functions are most similar) from the cross-correlation matrix is used to obtain the most characteristically similar features (e.g. the red marked basis from both male-female speech basis function). The cross-correlated model can be expressed as:

$$C_{BA} = B^T A = \begin{bmatrix} B_1 & B_2 & \cdots & B_N \\ A_1 & A_2 & \cdots & A_N \end{bmatrix}$$

(15)

where $B$ and $A$ denotes the second and first speech basis functions respectively (e.g. $B$ represents the second male basis and $A$ represents the first male basis). The best characteristic features (largest values of $B^T A$) then can be extracted from the cross-correlation matrix and reconstructed as 4 out of 16 characteristic basis functions (less number of basis get better separation results) from male-female speech. Figure 6 shows the extracted basis functions for male and female speeches:

We update the new basis filter $W = A^{-1}$ to obtain the most probably estimate of the source coefficients density. According to Mean Square Error, the separation results especially in terms of recovered female source have been substantially improved by exploiting the most characteristic features.

V. EXPERIMENTS AND ANALYSIS

A real time recording of two speech signals (male and female) is used to test the performance of the proposed algorithm. The mixing environment is non-reverberant and the time delay has been estimated and removed from the speech mixture. The sampling frequency used is 8 kHz.
A new algorithm based on the hybrid of ML and MAP estimator combined with the encoding of the best characteristic features of the speech has been developed. Real time speech recording has been conducted and the obtained results show significant performance of male-female speech separation using the proposed algorithm. The separation results in both recovered male and female speech improve 0.355% and 50.62% respectively.

### References