Overcomplete BSS for Convolutive Mixtures Based on Hierarchical Clustering

Stefan Winter*, Hiroshi Sawada, Shoko Araki, and Shoji Makino

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
{wifan,sawada,shoko,maki}@cslab.kecl.ntt.co.jp

Abstract. In this paper we address the problem of overcomplete BSS for convolutive mixtures following a two-step approach. In the first step the mixing matrix is estimated, which is then used to separate the signals in the second step. For estimating the mixing matrix we propose an algorithm based on hierarchical clustering, assuming that the source signals are sufficiently sparse. It has the advantage of working directly on the complex valued sample data in the frequency-domain. It also shows better convergence than algorithms based on self-organizing maps. The results are improved by reducing the variance of direction of arrival. Experiments show accurate estimations of the mixing matrix and very low musical tone noise.

1 Introduction

High quality separation of speech sources is an important prerequisite for further processing like speech recognition. Often the underlying mixing process is unknown, which requires blind source separation (BSS). In general we can distinguish two cases depending on the number of sources N and the number of sensors M:

$$\begin{array}{l|l} N > M & \text{overcomplete BSS} \\ \hline N \leq M & \text{(under-) complete BSS} \end{array}$$

Since undercomplete BSS ($N < M$) can be reduced to complete BSS ($N = M$) [1] we refer to both by complete BSS. Most approaches assume complete mixtures [2, 3], but in reality often the contrary is true. While the area of overcomplete BSS has obtained more and more attention [4–12], it still remains a challenging task.

Several of the proposed algorithms are based on histograms and developed for only two sensors [4–6]. Some could, in principle, be enhanced for higher dimensions M. But since histograms are based on densities, the so called curse

* The author is on leave from the Chair of Multimedia Communications and Signal Processing, University Erlangen-Nuremberg.

C.G. Puntonet and A. Prieto (Eds.): ICA 2004, LNCS 3195, pp. 652–660, 2004.

of dimensionality [13] sets practical limits on the number of usable sensors. Another problem occurs with complex numbers, which cannot be handled straightforwardly by histograms, but are necessary if BSS is performed in the frequency-domain. Some methods approach complex numbers by applying real-valued algorithms to the real and imaginary part or amplitude and phase [7, 8], which is not always applicable. Some approaches extract features like the direction-of-arrival (DOA) or work on the amplitude relation between two sensor outputs [4, 5, 9, 10]. In both cases only two sensors can contribute, no matter how many sensors are available.

Other algorithms like GeoICA [12] or AICA [11] resemble self-organizing maps (SOM) and could more easily be applied to convolutive mixtures. However, their convergence depends on initial values.

In this paper we propose the use of hierarchical clustering embedded into a two-stage framework of overcomplete BSS to deal with convolutive mixtures in the frequency-domain. This method can work directly on the complex valued samples. While it does not limit the usable numbers of sensors, it also prevents the convergence problems which can occur with SOM based algorithms.

After estimating the mixing matrix in the first stage, a maximum a-posteriori (MAP) approach is applied to finally separate the mixtures, assuming statistical independence and Laplacian pdfs for the sources [14].

In Sec. 2 we first explain the general framework before we give details about the hierarchical clustering in Sec. 3 and the MAP based source separation in Sec. 4. After this, we present experimental results in Sec. 5 demonstrating the performance for convolutively mixed speech data in a real room with reverberation time $T_R = 130\text{ms}$.

2 General Framework

We will consider a convolutive mixing model with N sources $s_i(t)$ $(i = 1 \dots N)$ and M $(M < N)$ sensors that yield linearly mixed signals $x_j(t)$ $(j = 1 \dots M)$. The mixing can be described by $x_j(t) = \sum_{i=1}^{N} \sum_{l=1}^{\infty} h_{ji}(l) s_i(t-l)$, where $h_{ji}(t)$ denotes the impulse response from source i to sensor j.

Instead of solving the problem in the time-domain, we choose a narrowband approach in the frequency domain by applying a short-time discrete Fourier transform (STDFT). Thus time-domain signals $\mathbf{s}(t) = [s_1(t), \dots, s_N(t)]^T$ and $\mathbf{x}(t) = [x_1(t), \dots, x_M(t)]^T$ are converted into frequency-domain time-series $\mathbf{S}(f,\tau) = [S_1(f,\tau), \dots, S_N(f,\tau)]^T$ and $\mathbf{X}_\tau = \mathbf{X}(f,\tau) = [X_1(f,\tau), \dots, X_M(f,\tau)]^T$ by an L-point STDFT, respectively. Thereby $f = 0,\ f_s/L,\ \dots,\ f_s(L-1)/L$ (f_s: sampling frequency; τ: time dependence). Let us define $\mathbf{H}(f) \in \mathbb{C}^{M \times N}$ as a matrix whose elements are the transformed impulse responses. We call the column vectors $\mathbf{h}_i(f)$ $(i = 1, \dots, N)$ mixing vectors and approximate the mixing process by

$$\mathbf{X}(f,\tau) = \mathbf{H}(f)\mathbf{S}(f,\tau) \tag{1}$$

This reduces the problem from convolutive to instantaneous mixtures in each frequency bin f. For simplicity we will omit the dependence on frequency and

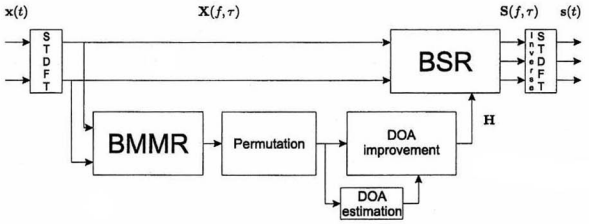

Fig. 1. Overall unmixing system.

time. Switching to the frequency domain has the additional advantage that the sparseness of the sources is increased [7]. This is very important, since the hierarchical clustering is based on the assumption of sparse sources.

The disadvantage of narrowband BSS in the frequency domain is the permutation problem, which results in wrong alignments of the frequency bins. In our framework we use a DOA based method to reduce the permutation problem [3]. We also apply the minimum-distortion-principle [2] to solve the scaling problem.

In complete BSS the mixing matrix $\mathbf{H}$ is square and (assuming full rank) invertible. Therefore the BSS problem can be solved by either inverting an estimate of the mixing matrix or directly estimating its inverse and solving (1) for $\mathbf{S}$.

However, this approach does not work in overcomplete BSS where the mixing matrix is not invertible. Therefore we follow a two-stage approach as proposed in [7] consisting of blind mixing model recovery (BMMR) and blind source recovery (BSR). To estimate the mixing matrix in the BMMR step, we propose the use of hierarchical clustering as described in detail in Sec. 3. To eventually separate the signals in the BSR step, we utilize a MAP based approach. Finally the inverse STDFT is applied to obtain time-domain signals. The overall system is depicted in Fig. 1.

3 Blind Mixing Model Recovery

Several algorithms have been proposed so far for BMMR. They usually have in common that they assume a certain degree of sparseness of the original signals. In this paper we consider signals that are sparse in the time-frequency domain. That means that different signals are rarely active at the same time-frequency instant (f, τ). This assumption leads to the conclusion that the samples in the mixed vector space $\mathbf{X}(f, \tau)$ cluster around the true mixing vectors $\mathbf{h}_i(f)$. This becomes clear when we consider the most sparse case when only a single source is active. Let us rewrite (1) as

$$\mathbf{X}(f, \tau) = \sum_{i=1}^{N} \mathbf{h}_i(f) S_i(f, \tau) \tag{2}$$

Assuming only one source active at (f, τ) means that the vector pointing to the resulting mixed sample $\mathbf{X}(f, \tau)$ is a scaled version of the corresponding mixing vector $\mathbf{h}_i(f)$. Depending on the actual sparseness of the source signals, the mixed signals will also have components of other signals and therefore be spread around the mixing vectors. In order to obtain a different cluster for each source signal S_i we assume a different mixing vector $\mathbf{h}_i(f)$ for each source signal.

3.1 Hierarchical Clustering

To avoid the problems discussed in Sec. 1, such as the curse of dimensionality or poor convergence, we propose the use of a hierarchical clustering algorithm following an agglomerative (bottom-up) strategy [13]. This means that at the beginning we consider each sample as a cluster that contains only one object. From there clusters are combined, so that the number of clusters decreases while the average number of objects per cluster increases. In the following we assume phase and amplitude normalized samples.

$$\mathbf{X} = \frac{\mathbf{X}}{|\mathbf{X}|_2} e^{-\varphi_{X_1}} \tag{3}$$

where φ_{X_1} denotes the phase of the first component of $\mathbf{X}$.

The combination of clusters into new clusters is an iterative process and based on the distance between the current clusters. Starting from the normalized samples, the distance between each pair of clusters is calculated, resulting in a distance matrix. The two clusters with the least distance are combined and form a new binary cluster. This process is called linking and repeated until the final number of clusters has decreased to a predetermined number c, $N \leq c \leq P$ (P: total number of samples).

For measuring the distance between clusters, we have to distinguish between two different problems. First we need a distance measure $d(\mathbf{X}_{\tau_1}, \mathbf{X}_{\tau_2})$ that is applicable to M-dimensional complex vector spaces. While there are several possibilities, we currently use the Euclidean distance defined by

$$d(\mathbf{X}_{\tau_1}, \mathbf{X}_{\tau_2}) = \sqrt{< (\mathbf{X}_{\tau_1} - \mathbf{X}_{\tau_2}), (\mathbf{X}_{\tau_1} - \mathbf{X}_{\tau_2})^* >} \tag{4}$$

where $< \cdot >$ stands for the inner product and * for complex conjugation.

When a new cluster is formed, we need to enhance this distance measure to relate the new cluster to the other clusters. The method we employ here is called the nearest-neighbor technique. Let C_1 and C_2 denote two clusters as illustrated in Fig. 2. Then the distance $d(C_1, C_2)$ between these clusters is defined as the minimum distance between its samples by

$$d(C_1, C_2) = \min_{\mathbf{X}_{\tau_1} \in C_1,\ \mathbf{X}_{\tau_2} \in C_2} d(\mathbf{X}_{\tau_1}, \mathbf{X}_{\tau_2}) \tag{5}$$

As mentioned earlier, most of the samples will cluster around the mixing vectors $\mathbf{h}_i$, depending on the sparseness of the original signals. Special attention must be paid to the remaining samples (outliers), which are randomly scattered

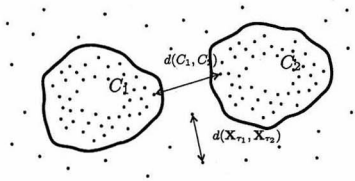

Fig. 2. Illustration of distances.

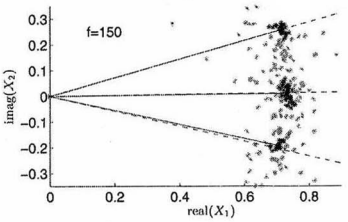

Fig. 3. Estimation of mixing vectors.

in the space between the mixing vectors. Usually they are far away from other samples and will be combined with other clusters only at higher levels of the clustering process (i.e. when only few clusters are left). This led us to the idea to set the final number of clusters at a high number

$$c \gg N \tag{6}$$

By doing so, we avoid linking these outliers with the clusters around the mixing vectors $\mathbf{h}_i$ and therefore distortions. This results in more robustness. More important, however, is the fact that we avoid combining desired clusters. Since the outliers are often far away from other clusters it might happen that desired clusters are closer to each other than to outliers. An example for the resulting clusters is shown in Fig. 3. Experimental details are given in Sec. 5.

3.2 Estimation of Mixing Matrix

Assuming that the clusters around the mixing vectors $\mathbf{h}_i$ have the highest densities and therefore the highest number of samples we finally choose the N largest clusters. Thereby the number of sources N must be known. To obtain the mixing vectors, we average over all samples of each cluster

$$\mathbf{h}_i = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} \mathbf{X}, \qquad 1 \leq i \leq N \tag{7}$$

where $|C_i|$ denotes the cardinality of cluster C_i. Thereby we assume that the influence of other sources has zero mean.

3.3 Advantages of Hierarchical Clustering

Among the most important advantages of the described hierarchical clustering algorithm is the fact that it works directly on the sample data in any vector space of arbitrary dimensions. The only requirement is the definition of a distance measure for the considered vector space. Therefore, it can easily be applied to complex valued data that occurs in frequency-domain convolutive BSS.

No initial values for the mixing vectors $\mathbf{h}_i$ are required. This means, in particular, that if the assumption of clusters with high densities around the mixing vectors is true, then the algorithm converges to those clusters.

Besides choosing a distance measure, there is only the single parameter c that determines the number of clusters. Experiments have shown that the choice for this parameter in the noiseless case is quite insensitive as long as it is above a certain limit that would combine desired clusters. Its choice is, in general, related to the sparseness of the sources. The sparser the signals are, the smaller the value of c can be chosen, because the number of outliers that must be avoided will be smaller.

While the considered signals must have some degree of sparseness, they do not have to be statistically independent at this point.

3.4 Reduction of DOA Variance

Experiments have shown that as long as there are clusters around the mixing vectors $\mathbf{h}_i$, the estimation results are of high quality. Even if the assumption of clear clusters is not true for all mixing vectors, the remaining ones are not influenced by poor estimation of others. In order to improve the wrongly estimated mixing vectors, we can utilize DOA information. While the mixing matrix is different for each frequency bin, the phase difference $\Delta\varphi_i$ between the components of a mixing vector $\mathbf{h}_i$ contains information about the relative physical position of its corresponding source. Assuming a linear sensor array in a far-field situation with plain wave fronts, the DOA θ_i is given by

$$\theta_i = cos^{-1}\left(\frac{\Delta\varphi_i v}{2\pi f d}\right) \tag{8}$$

where v denotes the sound velocity, d the distance between the corresponding sensors. Since θ_i is theoretically constant for all frequency bins, we can consider the DOA of the i-th signal as a random variable (RV) θ_i with mean μ_i and variance σ_i^2. While even the DOA of the original mixing matrix has a variance larger than 0, the results for the estimated mixing matrix can be improved if the variance of its DOAs is reduced.

For this purpose we define a new RV $\widehat{\theta}_i$ with reduced variance by

$$\widehat{\theta}_i = \sqrt{\varepsilon}\theta_i + (1 - \sqrt{\varepsilon})\mu_i, \qquad 0 \leq \varepsilon \leq 1 \tag{9}$$

While its mean is still μ_i, its variance $\widehat{\sigma_i}^2$ can be adjusted by ε and yields

$$\widehat{\sigma_i}^2 = \varepsilon\sigma_i^2 \tag{10}$$

We apply the new DOA by adjusting the phase of the mixing vectors $\mathbf{h}_i$. Since we do not need absolute DOA information, this improvement fully complies with the blind approach of BSS.

4 Blind Source Recovery

Since the mixing matrix cannot be inverted in overcomplete BSS, the unmixed signals cannot be directly obtained. Several approaches have been proposed to solve blind source recovery [14]. Among those we chose the shortest-path algorithm which is based on maximum a-posteriori (MAP) estimation, assuming statistical independence and Laplacian pdfs for the sources. Given the mixed signals $\mathbf{X}$ and the mixing matrix $\mathbf{H}$, the sources $\mathbf{S}$ are recovered by

$$\mathbf{S} = \arg \min_{\mathbf{X}=\mathbf{HS}} \sum_{i=1}^{N} |S_i| \tag{11}$$

This equation can be interpreted as finding the shortest-path decomposition, based on the mixing vectors $\mathbf{h}_i$ for each sample $\mathbf{X}_\tau$ separately. It means that each sample is assigned to exactly M signals. While (11) can, in general, be solved for real numbers by linear programming, we explicitly compute all $\binom{N}{M}$ possible decompositions and choose the one that minimizes $\sum_{i=1}^{N} |S_i|$. Taking a selection of M mixing vectors $\mathbf{h}_{i_1} \dots \mathbf{h}_{i_M}$, the decomposition is calculated by

$$\mathbf{S} = [\mathbf{h}_{i_1} \dots \mathbf{h}_{i_M}]^{-1} \mathbf{x} \qquad i_1, \dots, i_M \in \{1, \dots, N\} \tag{12}$$

5 Experimental Results

We performed experiments with the proposed algorithm using $N = 3$ speech signals and $M = 2$ sensors. The signals were taken from the Acoustical Society of Japan (ASJ) continuous speech corpus. The convolution was done with room impulse responses that were recorded at our laboratory. Further experimental conditions are given in Table 1. As performance measure, we used the signal-to-interference ratio $\mathrm{SIR}_i = 10 \log \left(\frac{\sum_t y_i^s(t)^2}{\sum_t y_i^{if}(t)^2} \right)$ where $y_i^s(t)$ is the portion of $y_i(t)$ that comes from $s_i(t)$ and $y_i^{if}(t) = y_i(t) - y_i^s(t)$. We also evaluated the signal-to-distortion ratio (SDR) as described in [15].

Table 1. Experimental conditions.

Direction of sources	50°, 90°, 120°
Distance of sensors	40 mm
Length of source signals	7.4 seconds
Reverberation time T_R	130ms
Sampling rate	8 kHz
Window type	von Hann
Filter length	1024 points
Shifting interval	256 points
Cluster threshold c (const $\forall f$)	100
Variance factor ε	0.8

As an upper limit for the performance of the whole system, scenario 1 in Table 2 shows the separation results when the original mixing matrix is used. This means that the permutation problem does not occur and the BSR part is given the best possible input.

Table 2. Performance of different parts of the separation system.

$N = 3$, $M = 2$, $T_R = 130$ms		Source 1	Source 2	Source 3	Average
Scenario 1	SIR (dB)	14.8	13.9	11.7	13.50
	SDR (dB)	13.39	6.83	10.55	10.26
Scenario 2	SIR (dB)	10.5	6.4	9.3	8.73
	SDR (dB)	7.47	2.82	5.99	5.43
Scenario 3	SIR (dB)	11.1	9.9	8.9	9.95
	SDR (dB)	9.65	4.05	5.95	6.55

Scenario 2 gives the results if we use the estimated mixing matrix without reduction of DOA variance. The last scenario shows the results if the estimated mixing matrix is used together with reduction of DOA variance. Figure 3 gives an example for the clustering for $f = 1164$Hz. To visualize, the real part of the first component X_1 versus the imaginary part of the second component X_2 is plotted. The N largest clusters (black) around the original mixing vectors $\mathbf{h}_i$ (dashed) can be clearly seen and result in precise estimations (solid).

Subjective evaluation of the separated sources showed very low musical tone noise.

6 Conclusion

We proposed the application of hierarchical clustering embedded into a two-stage framework of overcomplete BSS for convolutive speech mixtures. This method can work directly on the complex mixture samples. It also prevents the convergence problems which can occur with SOM based methods like GeoICA. Experimental results confirmed that the assumption of sparseness and, therefore, clusters around the mixing vectors is sufficiently fulfilled for convolutively mixed speech signals in the frequency domain.

References

1. Winter, S., Sawada, H., Makino, S.: Geometrical interpretation of the PCA subspace method for overdetermined blind source separation. In: Proc. ICA 2003. (2003) 775–780
2. Matsuoka, K.: Independent component analysis and its applications to sound signal separation. In: Proc. IWAENC 2003, Kyoto (2003) 15–18
3. Sawada, H., Mukai, R., Araki, S., Makino, S.: A robust and precise method for solving the permutation problem of frequency-domain blind source separation. In: Proc. ICA 2003. (2003) 505–510
4. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Transactions on Signal Processing (2004) (to appear).

5. Rickard, S., Yilmaz, O.: On the approximate W-disjoint orthogonality of speech. In: Proc. ICASSP 2002. (2002) 529–532
6. Vielva, L., Santamaria, I., Pantaleon, C., Ibanez, J., Erdogmus, D.: Estimation of the mixing matrix for underdetermined blind source separation using spectral estimation techniques. In: Proc. EUSIPCO 2002. Volume 1. (2002) 557–560
7. Bofill, P., Zibulevsky, M.: Blind separation of more sources than mixtures using sparsity of their short-time fourier transform. In: Proc. ICA 2000. (2000) 87–92
8. Bofill, P.: Underdetermined blind separation of delayed sound sources in the frequency domain. Neurocomputing **55** (2003) 627–641
9. Araki, S., Makino, S., Blin, A., Mukai, R., Sawada, H.: Blind separation of more speech than sensors with less distortion by combining sparseness and ica. In: Proc. IWAENC 2003. (2003) 271–274
10. Blin, A., Araki, S., Makino, S.: Blind source separation when speech signals outnumber sensors using a sparseness - mixing matrix estimation (SMME). In: Proc. IWAENC 2003. (2003) 211–214
11. Waheed, K., Salem, F.M.: Algebraic overcomplete independent component analysis. In: Proc. ICA 2003. (2003) 1077–1082
12. Theis, F.: Mathematics in independent component analysis. PhD thesis, University of Regensburg (2002)
13. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference, and prediction. Springer Series in Statistics. Springer-Verlag (2002)
14. Vielva, L., Erdogmus, D., Principe, J.C.: Underdetermined blind source separation using a probabilistic source sparsity model. In: Proc. ICA 2001. (2001) 675–679
15. Sawada, H., Mukai, R., de la Kethulle de Ryhove, S., Araki, S., Makino, S.: Spectral smoothing for frequency-domain blind source separation. In: Proc. IWAENC 2003. (2003) 311–314