FREQUENCY DOMAIN BLIND SOURCE SEPARATION OF A REDUCED AMOUNT OF DATA USING FREQUENCY NORMALIZATION

Enrique Robledo-Arnuncio[†‡], *Hiroshi Sawada*[†], *Shoji Makino*[†]

[‡]Center for Signal and Image Processing
Georgia Institute of Technology
Atlanta, GA 30332-0250, USA
era@ece.gatech.edu

[†]NTT Communication Science Laboratories,
2-4 Hikaridai, Seika-cho, Kyoto 619-0237, Japan
sawada@cslab.kecl.ntt.co.jp

ABSTRACT

The problem of blind source separation (BSS) from convolutive mixtures is often addressed using independent component analysis in the frequency domain. The separation performance with this approach degrades significantly when only a short amount of data is available, since the estimation of the separation system becomes inaccurate. In this paper we present a novel approach to the frequency domain BSS using frequency normalization. Under the conditions of almost sparse sources and of dominant direct path in the mixing systems, we show that the new approach provides better performance than the conventional one when the amount of available data is small.

1. INTRODUCTION

There exist several methods to separate an instantaneous linear mixture of multiple sources. If the sources are statistically independent, independent component analysis (ICA) algorithms are commonly used [1]. These can find a separation system up to two ambiguities: an arbitrary permutation in the ordering of the source estimates and an arbitrary gain applied to each source.

In many practical applications, such as the separation of acoustic mixtures, it is necessary to deal with convolutive mixtures. A possible approach to these is to move the problem to frequency domain [2], using tools like the short time Fourier transform (STFT), making the convolutive mixture approximately become a collection of instantaneous mixtures, one for each frequency bin. For each of these, a separation matrix needs to be estimated.

Sometimes the amount of available data is small. For example, at the initialization of real time systems, or when it is necessary to track a changing mixing system. With the frequency domain BSS approach the performance in these situations can be very poor, due to the large amount of separation matrices to estimate.

An obvious approach to improve the accuracy in the estimation of the separation parameters is to reduce the number of such parameters. If the mixing system has some structure it may be possible to capture its most significant aspects using a reduced number of parameters. A naive way to attempt such reduction is to use a smaller frame size in the STFT analysis. This can only provide a limited success, since with a small frame size the instantaneous mixture model at each frequency becomes less accurate.

In this paper we present a novel approach to perform the BSS of convolutive mixtures using the STFT. The new method allows to obtain a reduced number of separation parameters while using a long STFT frame. It relies on two assumptions often found in speech applications: the dominance of the direct path and the sparseness of the sources. Under these assumptions, we propose a normalization method which can remove from the measurement vectors the frequency dependence introduced by the mixing system, allowing one to perform ICA over multiple frequency bins simultaneously. Finally, we present experimental results which illustrate how BSS using this approach outperforms conventional frequency domain BSS (FD-BSS) in the separation of speech sources when the amount of available data is small.

The normalization scheme presented here derives from one used in the past to remove the frequency dependence from the inverse of the separation system, in the context of permutation alignment [3]. This previous normalization mechanism has also been used to normalize the measurement vectors in [4], but there the purpose was again to allow clustering, and not to perform ICA on the normalized data. In this paper we introduce a simple modification to this normalization scheme to make it preserve source information.

The general approach in this paper has some similarity in aim to the one proposed in [5] where, also under the assumption of free-field propagation, a frequency invariant transformation is used to transform the convolutive problem into an instantaneous one. The main difference is that the frequency normalization approach proposed here does not require a big number of sensors. Taking into account the sparseness of the sources, only two sensors are required, although more can be used. Also, our approach does not require knowledge about the location of the sensors, other than a maximum bound to the distance between each sensor and a reference sensor.

2. CONVENTIONAL FREQUENCY-DOMAIN BSS

An acoustic mixture is often modelled as the linear combination of several statistically independent sources, each of which has been transformed through a different linear time-invariant acoustic response,

$$\mathrm{x}_j(t) = \sum_{k=1}^{N} \sum_{r=0}^{R-1} \mathrm{h}_{jk}(r)\mathrm{s}_k(t-r) \qquad \text{for } j = 1 \dots M, \quad (1)$$

where t is the discrete time index, and N and M are the number of sources and microphones. In the blind separation problem, the mixing filters h_{jk} and the sources s_k are unknown, and the goal is to find an estimate of s_k from the mixture x_j, knowing that the sources are statistically independent from each other. In this paper we will focus in the case $N = M$.

If the length of the mixing filters is big, as it is often the case with acoustic mixtures, the computational cost can be reduced by

The first author performed the work while at NTT Communication Science Laboratories.

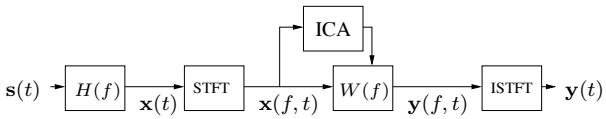

Fig. 1. Frequency domain blind source separation (FD-BSS).

moving the problem to frequency domain. Figure 1 shows the different stages of this approach.

The first step is to perform the L-point STFT of the measured signals, $\mathrm{x}_j(t)$:

$$x_j(f,t) = \sum_{r=-L/2}^{L/2-1} \mathrm{x}_j(t+r)\mathrm{win}(r)e^{-i2\pi fr}, \qquad (2)$$

where f is the frequency index, and win is the analysis window. If this window is long enough, the mixing system becomes approximately instantaneous. In vector notation,

$$\mathbf{x}(f,t) = H(f)\mathbf{s}(f,t), \qquad (3)$$

where $\mathbf{x}(f,t) = [x_1(f,t)\dots x_M(f,t)]^T$ is the vector of measurements at each time-frequency point, $\mathbf{s}(f,t)$ is the corresponding source signal vector, and $H(f)$ is a square scalar (complex) mixing matrix.

The next step, as shown in Figure 1 is to use an ICA algorithm to estimate a separation system $W(f)$ from the vectors $\mathbf{x}(f,t)$. The conventional approach to this is to perform ICA separately for each frequency, and to use additional knowledge about the problem to choose the permutation for each separation matrix consistently.

3. PROPOSED APPROACH

The approach proposed in this paper replaces the ICA block in Figure 1 by a new procedure based on the frequency normalization of the measurement spectrograms.

The mixing matrix in (3), $H(f)$, is frequency dependent. The goal of the frequency normalization is to obtain an alternative representation of the measurement vectors that makes them relate to the sources through a frequency independent mixing matrix. With this representation, it becomes possible to rearrange the spectrograms by grouping multiple frequency lines together, and to apply ICA on each of such groups of frequency bins as one would do on a time sequence of measurements. This reduces the number of separation matrices to estimate, while preserving the FFT frame size.

Figure 2 illustrates this process. First, the original time-frequency measurement vectors are normalized, as explained in Section 3.2 below. Then, frequency bins are grouped and ICA is applied on each group (Section 3.3). Finally, the resulting separation matrices are denormalized as explained in Section 3.4. The next section presents the two assumptions required for the justification of this new method.

3.1. Assumptions

The following two assumptions will be used to validate the normalization equation. The experimental results presented later show how the approach gives useful results even if the assumptions are only approximately satisfied.

1. The mixing system consists of a *direct propagation path* for each source, with some frequency-dependent attenuation $q(f)$:

$$h_{jk}(f) = \frac{q(f)}{d_{jk}}\exp\left[-i2\pi fc^{-1}d_{jk}\right], \qquad (4)$$

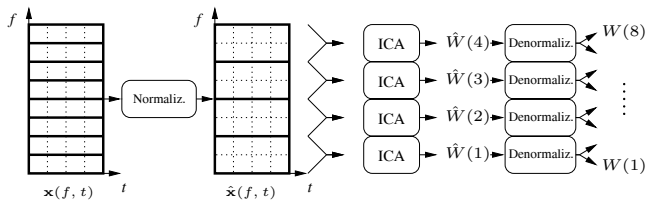

Fig. 2. Main steps for the proposed approach. The measured time-frequency vectors, shown as a spectrogram on the left, are normalized, and the frequency bins are grouped forming several separation bands (four in the figure). ICA is then performed on each of them.

where d_{jk} is the distance between the k'th source and the j'th sensor, and c is the speed of sound.

2. The time-frequency representation of the sources is *sparse*. This means that at any time-frequency point at most one source is active. This allows to express the measurements as:

$$x_j(f,t) = h_{jk}(f)s_k(f,t), \qquad \text{for some } k \qquad (5)$$

3.2. Normalization equation

The normalized measurement values, $\hat{x}_j(f,t)$, are constructed by modifying the phase of the original measurement values as follows:

$$\hat{x}_j(f,t) = |x_j(f,t)|\exp\left(i\frac{\arg\left(\frac{x_j(f,t)}{x_J(f,t)}\right)}{4fc^{-1}d_{max}} + i\arg\left(x_J(f,t)\right)\right), \qquad (6)$$

where J is the index for one of the sensors chosen as reference, and d_{max} is the maximum distance between this reference sensor and any other. Indexes f and t cover the complete time-frequency range.

Equation (6) involves a non-linear transformation of the phase, but direct substitution of equations (4) and (5) into (6) gives the following relationship between the source and the normalized vectors:

$$\hat{\mathbf{x}}(f,t) = \hat{H}C_1(f)\mathbf{s}(f,t), \qquad (7)$$

where $\hat{H}$ is an unknown *frequency independent* mixing matrix, whose elements relate to the unknown time delay differences:

$$\hat{h}_{jk} = \frac{1}{d_{jk}}\exp\left(-i\frac{\pi}{2}\frac{(d_{jk}-d_{Jk})}{d_{max}}\right), \qquad (8)$$

and $C_1(f)$ is a *diagonal* matrix. We will assume that $\hat{H}$ is invertible.

Thus, the assumptions lead to a normalized spectrogram that relates to the source spectrogram through a unique mixing matrix and through a frequency dependent scaling. It is thus possible to perform ICA once on the whole spectrogram or on a subset of the frequency bins, as discussed next. The frequency-dependent scaling can be aggregated with the arbitrary scaling introduced by ICA.

3.3. Separation of the normalized data

As shown in Figure 2, the f axis of the normalized spectrogram is divided in several bands (groups of adjacent frequency bins), each of them with equal width. ICA is then applied on the $\hat{\mathbf{x}}(f,t)$ vectors within each band. We use the term *separation band* to refer to each of these groups of frequency bins on which a single execution of ICA is applied. In the case of conventional FD-BSS, each separation band consist of one single frequency bin.

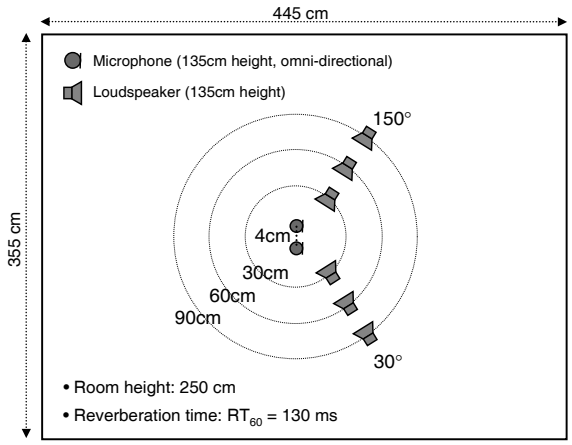

Fig. 3. Configuration of the experiment room.

In the experiments described below, ICA was performed on each separation band using a complex version of FastICA [1], and the solution was improved by using InfoMax [6] and the natural gradient using a non-linear function based on the polar coordinate [7].

These techniques assume that there is some underlying random process, whose statistics can be estimated from the available "sequence" of measurements. This corresponds to some kind of ergodicity requirement. As it is often done in speech signal processing, we will assume that this requirement is satisfied to some extent, and verify it indirectly through the experimental results.

In the following discussion we will assume that ICA has been applied successfully on a group of frequency bins, so that a separation matrix $\hat{W}$ has been obtained. This matrix will thus relate to the unknown matrix $\hat{H}$ in equation (7) through the following equation:

$$\hat{W} = PD\hat{H}^{-1}, \tag{9}$$

where D is a diagonal matrix which accounts for the gain ambiguity, and P is a permutation matrix which accounts for the permutation ambiguity. The desired (unnormalized) separation filter should relate in a similar way to the inverse of the original mixing matrix H. In the next section the computation of such a filter is described.

3.4. De-normalization of the separation system

To compute the de-normalized separation system, it is convenient to work with the inverse of $\hat{W}$, which is in turn related to the matrix $\hat{H}$:

$$\hat{A} = \hat{W}^{-1} = \hat{H}D^{-1}P^{-1}. \tag{10}$$

Again, D^{-1} is a diagonal matrix and P^{-1} is a permutation matrix. Note that although $\hat{H}$ is unknown, its relation to the parameters of the mixing system is known (8). This knowledge lets one define the following as the de-normalization equation for the elements of $\hat{A}$:

$$a_{jk}(f) = |\hat{a}_{jk}|\exp\left(i \arg\left(\frac{\hat{a}_{jk}}{\hat{a}_{Jk}}\right) 4fc^{-1}d_{max}\right) \tag{11}$$

where $\hat{a}_{jk}$ are the elements of $\hat{A}$, and $a_{jk}(f)$ are the elements of the de-normalized matrix $A(f)$. Substituting (4), (8) and (10) into (11),

$$W(f) = A(f)^{-1} = PC_2(f)H^{-1}(f), \tag{12}$$

where $C_2(f)$ is a diagonal matrix. This equation means that $W(f)$ is a separation system. The arbitrary scaling caused by $C_2(f)$ can be removed using the minimum distortion principle [8]. If multiple bands are used, permutation inconsistencies between their $A(f)$ matrices are avoided by normalizing and clustering their columns [3].

4. EXPERIMENTS SETUP

We have carried out several experiments to analyze the performance of the proposed approach in real acoustical conditions. The mixture segments were prepared from a set of anechoic speech recordings mixed through several different room response recordings, using two sources and two microphones. The different mixture segment lengths used were 0.5, 0.25 and 0.16 seconds. For each of these lengths, the experiments were repeated for several different source segments, and performance results were averaged.

The source signals for each experiment were constructed from recordings of Japanese and English speakers, both male and female, sampled at 8 kHz. To simplify the analysis of the results, the source segments were selected so that the difference of energy between both sources in each segment was not greater than 10dB.

The setup for the room recordings is shown in Figure 3. Three different recordings were made, each with the two loudspeakers at a certain distance from the center between the two microphones. Also, synthetic anechoic room responses for the same inter-microphone distance and source directions (far field) were generated.

Estimates of the sources were computed using both the proposed approach and conventional FD-BSS. For the proposed approach a STFT frame of 512 samples was used, and experiments were performed using 1, 4, 16 and 256 separation bands. For the conventional FD-BSS, the STFT frame size was twice the number of bands, except for the case of one band, which corresponds to time domain instantaneous BSS. In all cases the frame shift was one fourth of the frame length, and a Hanning window was used. The resulting number of available data points (samples) for each ICA execution is sumarized in Table 1.

	Proposed approach			
Separation bands	256	16	4	1
Samples per band, 0.5sec sources	28	448	1792	7168
Samples per band, 0.25sec sources	12	192	768	3072
Samples per band, 0.16sec sources	7	112	448	1792
	Conventional FD-BSS			
Frame size	512	32	8	1
Samples per bin, 0.5sec sources	28	497	1997	4000
Samples per bin, 0.25sec sources	12	247	997	2000
Samples per bin, 0.16sec sources	7	157	637	1280

Table 1. Available number of samples per separation band

Separation performance was evaluated with the signal to interference ratio (SIR) gain (SIR at the outputs minus SIR at the sensors), averaged across all the sources for each experiment condition.

5. RESULTS

The separation performance results clearly illustrate the benefits due to the reduction in the number of estimated separation matrices. Figure 4 shows how, when using the proposed approach, SIR results improve as the number of separation matrices to estimate (abscissa axis) is reduced. In contrast, the performance of conventional FD-BSS drops as the number of separation matrices is reduced, due to the corresponding reduction in the frame size.

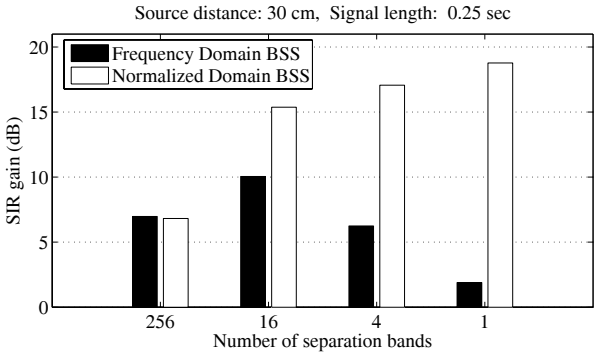

Fig. 4. Average SIR gain for different number of separation bands.

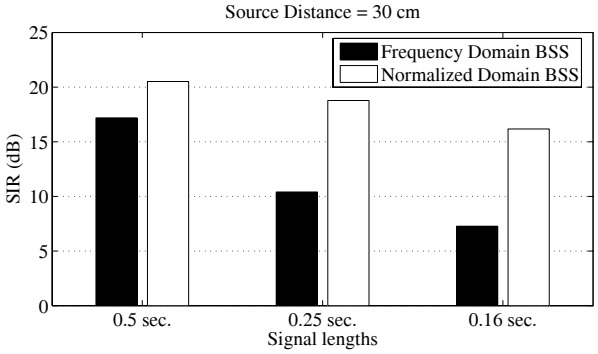

Fig. 5. Average SIR gain for different signal lengths. For the normalized BSS one separation band was used. For the FD-BSS the experiments were performed with several different frame sizes, and the sizes which yielded best separation performance were chosen.

The results in Figure 4 show how conventional FD-BSS performs best for some intermediate frame size. The degradation of the separation performance for big frame sizes (left side of the figure) can be attributed to the reduced number of frequency samples used to estimate each separation matrix (see Table 1). The degradation for small frame sizes (right side of the figure), on the other hand, can be attributed to the fact that a smaller frame size makes the assumption of instantaneous mixture in frequency domain less accurate and constrains the separation filter length too much. The proposed approach avoids using a small frame size.

Figure 5 illustrates the variations in performance when different amounts of data are available. The advantage of using frequency normalization is specially clear for very short signal lengths.

Finally, Figure 6 shows the separation performance for different mixing systems with different direct to reverberant ratios. As this ratio increases, the free-field assumption becomes more accurate, and the separation performance increases. Using frequency normalization, this increment becomes clearly higher.

6. CONCLUSION

We have presented a novel approach to perform blind separation of convolutive mixtures that outperforms conventional FD-BSS for input signals of small length. The approach is based on a special normalization of the spectrogram of the measured signals.

We have shown that, under the assumptions of free-field mixture and sparse sources, the normalized data relates to the sources through a matrix of frequency independent mixing parameters. This property allows one to combine multiple frequency bins for the estimation of a single separation matrix, reducing the estimation error.

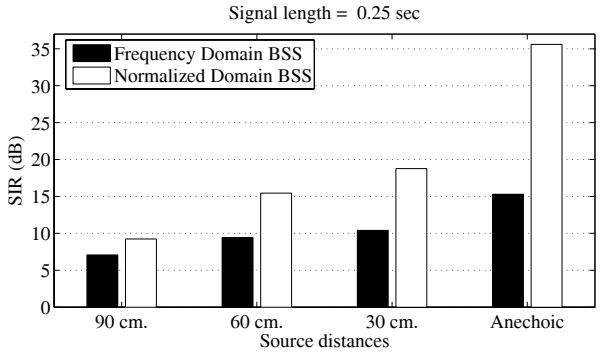

Fig. 6. Average SIR gain for sources at different distances. The direct-to-reverberant ratio increases for closer sources. The anechoic situation has infinite direct-to-reverberant ratio. The STFT frame size and number of separation bands were chosen as for Figure 5.

Finally, we have shown through experimental evaluation how this approach improves the separation performance even when the free-field and the sparseness assumptions are not strictly satisfied. This makes the new approach a useful tool for BSS applications where only a small amount of data is available, for example when fast initialization or tracking are required.

7. REFERENCES

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley, 2001.

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21 – 34, Nov. 1998.

[3] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind Extraction of a Dominant Source Signal from Mixtures of Many Sources," in *Proc. ICASSP*, Philadelphia, PA, USA, Mar. 2005, pp. 61–64.

[4] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Blind source separation by observation vector clustering," in *Proc. IWAENC*, Eindhoven, The Netherlands, Sept. 2005.

[5] W. Liu and D. P. Mandic, "Semi-blind Source Separation for Convolutive Mixtures Based on Frequency Invariant Transformation," in *Proc. ICASSP*, Philadelphia, PA, USA, Mar. 2005, pp. 285–288.

[6] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[7] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation," *IEICE Transactions Fundamentals*, vol. E86-A, no. 3, pp. 590–596, Mar. 2003.

[8] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. 4th Int. Symposium on Independent Component Analysis and Blind Signal Separation*, San Diego, California, USA, Dec. 2001, pp. 722–727.