SEPARATION OF SPEECH SIGNALS WITH SEGMENTATION OF THE IMPULSE RESPONSES UNDER REVERBERANT CONDITIONS

Christine Servière

LIS ENSIEG
BP 46 38402 Saint-Martin d'Hères France
Christine.Serviere@lis.inpg.fr

ABSTRACT

BSS performance is not still enough for realistic acoustic signals, particularly when the lengths of the impulse responses are long. We propose to use a complete model expressed in frequency domain but which is exactly equivalent to a linear convolution in time-domain. This model is applied to acoustic responses by segmenting the overall impulse responses in K short segments. The data are then transformed for each frequency bin into convolutive mixtures of K taps. Finally, the separation is achieved with a natural gradient algorithm based on a maximum-entropy cost function. The interest of this complete model consists in combining both short Fourier Transforms (with N samples) and convolutive mixtures with few taps K. The resulting impulse responses lengths in the time-domain will be of (K.N) samples and can best suit with long real acoustic responses.

1. INTRODUCTION

Blind Source Separation (BSS) consists in recovering signals of different physical sources $s_i(t)$ from a finite set of observations $x_i(t)$ recorded by sensors. Under the only hypothesis of mutually independent sources, BSS extracts the contributions of the sources independently of the propagation medium. These methods have been successfully used in many fields as medicine, telecommunications, audio processing or rotating machine diagnosis.

There have been a lot of proposals to achieve BSS for the separation of speech signals [1-3], introduced for separating convolved mixtures. BSS methods aim at the retrieval of the inverse of the propagation medium between sources and sensors in time or frequency-domains. They usually test the statistical independence of the separated signals with higher-order moments, negentropy or the mutual information [4-5]. However, the most successful require working in the frequency domain [3] [5-7], even if the separation performance declines significantly in a reverberant environment [8].

When the inverse of the mixing FIR-filter is calculated in time domain, the iterative learning rule is complicated for large-tap FIR filter and the convergence degrades when the reverberation increases, as noted in [1]. In practice, it only works with short-tap filter (less than 100 taps) [1].

In frequency-domain, the convolutive mixtures are simplified into simultaneous mixtures and the complex-valued inverse of the mixing matrix is calculated at each frequency bin. Then it is easy to estimate the separation filters and for example to converge them in an iterative ICA learning with a high stability [1]. Nevertheless, it also presents some disadvantages, degrading the performances under heavily reverberant conditions [1]. In [1], the authors showed that the two approaches were mutually complementary and proposed a multi-stage ICA. A first frequency-domain method separates sources, which are good initialisations for a time domain stage. The latter performs the separation by removing the residual cross-components.

In this paper we achieve BSS in frequency domain for the sake of simplicity and stability. However, two disadvantages exist : permutation among source signals and choice of the discrete Fourier transform (DFT) length. Various solutions have been proposed to solve the permutation problem [2] [6] [9] but usually require many computations.

The frame size of FFT, N, used in frequency domain must be discussed in detail, in particular versus the length of a room impulse response T. It is commonly believed that N must verify T<<N in order to estimate an unmixing filter [2] [6]. Indeed, firstly, if N is too short versus the inverse filter lengths, the impulse reponses are truncated. It often occurs with room acoustics, as the inverse system generally contains more parameters than the mixing system [10]. Secondly, in frequency domain, the

convolutive mixture is usually reduced to an instantaneous complex mixture for each frequency bin. It is only an approximation as it implies a circular convolution (and not a linear one) in time domain. This approximation is all the more correct since the real impulse response lengths are short in comparison with N.

However, for large value of N, N>>T, it is proved in [1] that the separation performance is saturated before reaching a sufficient performance because few data are available in frequency domain for a constant duration of the observations lengths. BSS methods then fail to test the independence of the estimated sources. The poor performance of BSS in a long reverberation environment with N>>T can be explained so. Besides, the computation cost required for solving the permutation problem is increasing with N.

In order to resolve the above problems, we propose in section 2 to use a complete model in frequency domain which is exactly equivalent to a linear convolution in time-domain. The idea is derived from the overlap-add method. The impulse responses are sectioned in K blocks of N samples. After pre-processing, the data are transformed in frequency domain at frequency bin ν, into a FIR filtering of K taps where the K taps are the complex gains of the K sectioned blocks at the same frequency bin ν.

Consequently, we have replaced the problem of the inversion of filters of K.N taps with the inversion of N filters of K taps. The interest of this complete model consists in combining both short Fourier Transforms (with N samples) and convolutive mixtures with few taps K, although the length of the inverse impulse responses remains long (K.N taps) according to the real acoustic responses.

Finally, the separation is achieved in section 3 with a natural gradient algorithm based on a maximum-entropy cost function. The proposed method is then tested in section 4 on speech signals mixed with real measured impulse responses.

2. SOUND MIXING MODEL

2.1. BSS mixing model

In this paper, we consider a M input, M output convolutive problem. Each microphone j receives a direct copy of each sound source $s_i(t)$ (at different propagation delays between each source and microphone) as well as several reflected and modified copies of each source. These acoustic effects can be modelled in a linear system:

$$x_j(t) = \sum_{i=1}^{N} h_{ij} * s_i(t) \qquad (1)$$

where h_{ij} is the impulse response from source i to microphone j and the operator * denotes linear convolution. In frequency domain, the convolutive mixture is usually reduced to:

$$X(\nu) = A(\nu)S(\nu) \qquad (2)$$

where $X(\nu)$ (respectively $S(\nu)$) is the N-points discrete Fourier transform of the nth data vector $X(n)$ (respectively $S(n)$).

This frequency model is simple but generates two types of errors versus real-world mixtures for short length N (N<<T). Indeed, $A(\nu)$ represents here the DTF of the impulse responses of length N, and not the total impulse responses. Besides equation (2) is only an approximation as it implies a circular convolution in time-domain and is justified only for N>>T. Consequently it should be interesting to work with N>>T for a T-points room impulse response as underlined in [2-3]. However in that case, it has been discussed in [1] that we usually have not enough available data $X(\nu)$ available to estimate the sources for a constant (or short) duration of the observations lengths.

We propose in the next section to model the mixtures with a complete expression in frequency domain but strictly equivalent to a linear convolution in time domain. In the following, uppercase symbols will denote frequency-domain variables, lowercase symbols will stand for time-domain variables and boldface will denote vectors and matrices.

2.1. Frequency model

Consider $x(n)$, the linear convolution of the signal $s(n)$ and a FIR filter H. Let $\boldsymbol{h}=[h_0, \ldots, h_{N-1}]^T$ be the impulse response of filter H with N taps.

The signal $x(n)$ is given by a linear convolution:

$$\begin{bmatrix} x(n) \\ x(n+1) \\ . \\ . \\ x(n+N-1) \end{bmatrix} = \begin{bmatrix} s(n) & s(n-1) & \ldots & s(n-N+1) \\ s(n+1) & s(n) & \ldots & s(n-N+2) \\ . & . & \ldots & . \\ s(n+N-1) & s(n+N-2) & \ldots & s(n) \end{bmatrix} . \begin{bmatrix} h_0 \\ h_1 \\ . \\ h_{N-1} \end{bmatrix} \qquad (3)$$

Extend the right term in equation (3) with a Toeplitz matrix (NxN) to a new formulation of length 2N with a Toeplitz circulant matrix built with samples of signal $s(n)$. We designate the resulting vector as: $\boldsymbol{x'(n)}=[x'(n-N), \ldots, x'(n+N-1)]^T$. $\boldsymbol{x'(n)}$ is given as :

$$\begin{bmatrix} x'(n-N) \\ . \\ x'(n) \\ . \\ x'(n+N-1) \end{bmatrix} = \underbrace{\begin{bmatrix} s(n-N) & \dots & s(n+1) & \dots & s(n-N+1) \\ . & \dots & . & \dots & . \\ s(n) & \dots & s(n-N+1) & \dots & s(n+1) \\ . & \dots & . & \dots & . \\ s(n+N-1) & \dots & s(n) & \dots & s(n-N) \end{bmatrix}}_{\boldsymbol{\chi}} \cdot \begin{bmatrix} h_0 \\ . \\ h_{N-1} \\ 0 \\ 0 \end{bmatrix}$$

(4)

where matrix $\boldsymbol{\chi}$ is a (2Nx2N) circulant matrix. (4) can also be written as : $\boldsymbol{x'(n)=\chi h'}$ (5)

where vector $\boldsymbol{h'}$ contains the impulse response $\boldsymbol{h}$ padded with N zeros.

Let $\boldsymbol{x(n)}$ denote the block of samples $[x(n-N), ..., x(n), ..., x(n+N-1)]^T$, we see from (4) that the N last components of vector $\boldsymbol{x'(n)}$ $[x'(n), ..., x'(n+N-1)]^T$ are equal to the N last ones of vector $\boldsymbol{x(n)}$. Define $[f_0, ..., f_{N-1}]^T$ to be a window of length N and let $\boldsymbol{f}$ be the PxP (with P=2N) diagonal matrix: $\boldsymbol{f}=diag(0,...,0, f_0, ..., f_{N-1})$.

From the previous remark on vectors $\boldsymbol{x(n)}$ and $\boldsymbol{x'(n)}$, it is obvious that they verify :

$$\boldsymbol{f.x'(n)=f.x(n)} \quad (6)$$

Consequently we obtain from equations (5) and (6) that :

$$\boldsymbol{f.x(n)=f.\chi.h'} \quad (7)$$

W will denote the symmetric matrix (2Nx2N) whose *k*th, *l*th element is $\boldsymbol{W_{kl}}=\exp(-j2\pi kl/2N)$. Multiplying equation (7) by the DFT matrix $\boldsymbol{W}$ leads to :

$$\begin{aligned} \boldsymbol{W.f.x(n)} &= \boldsymbol{W.f.\chi.h'} \\ &= \boldsymbol{Wf\,W^{-1}.\,W\chi W^{-1}.Wh'} \end{aligned} \quad (8)$$

As $\boldsymbol{\chi}$ is a circulant matrix, it owns the DFT matrix $\boldsymbol{W}$ as eigenvectors [11]. It can be so deduced that $\boldsymbol{W\chi W^{-1}}$ is equal to a diagonal matrix $\boldsymbol{S(n)}$ (2Nx2N), whose elements are the DFT coefficients of the first column of matrix $\boldsymbol{\chi}$: $[s(n-N), ..., s(n), ..., s(n+N-1)]^T$.

Let $\boldsymbol{F= W.f.W^{-1}}$ be the 2Nx2N circulant matrix whose elements are the DFT coefficients of the window *f*. Let $\boldsymbol{H}$ denote the 2Nx1 vector of the DFT coefficients of the impulse response vector $\boldsymbol{h}$, padded with N zeros :

$$\boldsymbol{H} = [H_0, ..., H_{2N-1}]^T = \boldsymbol{W}[h_0, ..., h_{N-1}, 0, ..., 0]^T \quad (9)$$

With the previous notations introduced in frequency domain, equation (8) becomes :

$$\boldsymbol{W.f.x(n) = F.\,S(n).\,H} \quad (10)$$

where $\boldsymbol{W.f.x(n)}$ represents the 2N-points DFT of the block of samples $[x(n-N), ..., x(n), ..., x(n+N-1)]^T$, multiplied by the window $[0, .., 0., f_0, \dots, f_{N-1}]^T$.

Equation (10) relies the DFT coefficients of signal $[s(n-N), ..., s(n), ..., s(n+N-1)]^T$ and the DFT coefficients of the impulse response $\boldsymbol{H}$ to the time-domain linear convolution $[x(n-N), ..., x(n), ..., x(n+N-1)]^T$.

$\boldsymbol{S(n)}$ is the diagonal matrix (2Nx2N), whose elements are the 2N-points DFT coefficients of : $[s(n-N), ..., s(n), ..., s(n+N-1)]^T$. They are denoted $S(n, \nu_0), ..., S(n, \nu_{2N-1})$.

$$\boldsymbol{S(n)}=diag(S(n,\nu_0), ..., S(n,\nu_{2N-1})) \quad (11)$$

This formulation (12) is no more an approximation but the complete model in frequency domain equivalent to the time domain linear convolution. We can remark that the DFT of the windowed signal $x(n)$ at one frequency bin ν_j is a linear combination of terms $(S(n,\nu_j)H(\nu_j))$ at all frequency bins as the matrix F is usually not a diagonal matrix. In the case of long impulse responses as acoustic ones, equation (10) is not easy to handle. However, it can be extended by segmentating the overall impulse response in data blocks.

2.2. Segmentation of the response in case of long impulse response

Let L, N, K denote respectively the length of the impulse response, the block size and the number of segments (L=K.N). Consider the long impulse response **h** of length L. It can be sectioned in K segments of length N of elementary impulse responses $\boldsymbol{h_i}$:

$\boldsymbol{h_i} = [h_{iN}, ..., h_{iN+N-1}]^T$ i=0, ..., K-1

Due to the principle of superposition, the resulting time signal $x(n)$ of the linear convolution between the filter H and the signal $s(n)$ is the addition of the linear convolution of all the elementary filters $\boldsymbol{h_i}$. Each elementary output is given by (10) where vector $\boldsymbol{H}$ is replaced with the elementary vector $\boldsymbol{H_i}$, which is the vector of the DFT coefficients of $\boldsymbol{h_i}$ padded with N zeros.

$$\begin{aligned} &\boldsymbol{H_i = W}.\,[h_{iN}, ..., h_{iN+N-1}, 0, ..., 0]^T \quad i=0,...,K-1 \\ &\boldsymbol{H_i} = [H_i(\nu_0), ..., H_i(\nu_{2N-1})]^T \end{aligned} \quad (12)$$

Consequently, the signal $x(n)$ is of the form :

$$\boldsymbol{W.f.x(n)} = \boldsymbol{F.} \sum_{i=0}^{K-1} \boldsymbol{S(n-iN).\,H_i} \quad (13)$$

It can also be written with the following expression, similar to equation (10) :

W.f.x(n) =

$$F.\underbrace{[S(n)\ S(n-N)\ \ldots\ S(n-(K-1)N)]}_{S'(n)}.[H_0\ H_1\ \ldots H_{(K-1)}\ ]^T$$

(14)

where ***S'(n)*** is a (2Nx2KN) row-block matrix obtained by stacking the K diagonal matrices ***S(n-iN)*** (for i=0, K-1).
S(n-iN) is the 2Nx2N diagonal matrix whose elements are the DFT coefficients of the block of the input samples $[s(n-iN-N), \ldots, s(n-iN), \ldots, s(n-iN+N-1)]^T$.
Finally let ***H'*** be the 2KNx1 vector obtained by stacking the vectors H_i (i=0,…,K-1).
The relation between the DFT of the windowed signal *x(n)* and DFT of the input signal *s(n)* is the following :

W.f.x(n) =***F. S'(n). H'*** (15)

The interest of this complete model is that the DFT are processed on data blocks of 2N points on the observations *x(n)* whereas the length of the impulse response is KN. K is a parameter, varying according to the reverberation length. If we develop equation (15), we remark that ***W.f.x(n)*** is a 2Nx1 vector which contains the 2N-points DFT coefficients of : $[0, \ldots 0, f_0\ x(n), \ldots, f_{N-1}\ x(n+N-1)]^T$. Denote them : $[X_f(\nu_0), \ldots, X_f(\nu_{2N-1})]^T$
The term ***F. S'(n). H'*** is the product between the 2Nx2N matrix ***F= W.f.W*** $^{-1}$ which only depends on the chosen window and the 2Nx1 vector ***S'(n).H'***. The ith component of ***(S'(n). H')*** is equal to :

$$S(n, \nu_i).H_0(\nu_i)+\ldots+ S(n-KN, \nu_i).H_K(\nu_i) \quad (16)$$

where $S(n-iN, \nu_i)$ is the DFT of the block of the input samples $[s(n-iN-N), \ldots, s(n-iN), \ldots, s(n-iN+N-1)]^T$ at frequency bin ν_i. As matrix ***F*** is generally not diagonal, $X_f(\nu_i)$ is a linear combination of 2N terms such as $S(n, \nu_i).H_0(\nu_i)+\ldots+ S(n-KN, \nu_i).H_K(\nu_i)$ for $\nu=\nu_0, \ldots, \nu_{2N-1}$.

3. APPLICATION TO BSS

3.1. Complete mixture model in frequency domain

For more simplicity, consider here a 2 inputs, 2 outputs convolutive problem. Let be *x1(n)* the first sensor. It receives a mixture of source *S1(n)* and *S2(n)*.

W.f.x1(n)=F.R1 (17)

Where vector R1 is equal to:

$$\begin{bmatrix} \sum_{i=0}^{K-1} S1(n-iN, \nu_0)H_i^{11}(\nu_0) + S2(n-iN, \nu_0)H_i^{12}(\nu_0) \\ . \\ \sum_{i=0}^{K-1} S1(n-iN, \nu_{2N-1})H_i^{11}(\nu_{2N-1}) + S2(n-iN, \nu_{2N-1})H_i^{12}(\nu_{2N-1}) \end{bmatrix}$$

where $H_i^{11}(\nu_0)$ and $H_i^{12}(\nu_0)$ are respectively the ith elements of the DFT of the sectioned propagation filters between the source *S1(n)* (respectively *S2(n)*) and sensor *x1(n)*. The complete model (17) is not easy to handle, as all frequency bins are mixed with matrix ***F***. A similar system of equations can be written between the second sensor *x2(n)* and the two sources, mixed with the same matrix ***F***.

W.f.x2(n)=F.R2 (18)

The aim of the next section is then to inverse each of the two systems (17)(18) to provide equations such as :

$$Z(n, \nu_j) = \sum_{i=0}^{K-1} S1(n-iN, \nu_j)H_i^{11}(\nu_j) + S2(n-iN, \nu_j)H_i^{12}(\nu_j)$$

(19)

where quantities $Z(n, \nu_j)$ can be computed only from data *x(n)*. Due to the DFT properties on real-valued data, we need only to recover the N first equations in (19), corresponding to the N first frequencies for *j=0, …, N-1*.

3.2. Analysis of matrix *F*

Recall that ***F= W.f.W*** $^{-1}$. ***F*** is a 2Nx2N circulant matrix. As ***f****=diag(0,…,0, f*$_0$*, …, f*$_{N-1}$*),* it is clear that ***F*** is of rank N. Consider the shape of a row of the module of matrix *F(k,l)* for two different types of windows: the rectangular one and a hamming window. The ith row of ***F*** is drawn in figure 1 for the rectangular window and in figure 2 for the hamming function. We see that in the first case the (2N-1) elements of R1 (or R2) are totally mixed whereas F is a banded matrix in the second case.

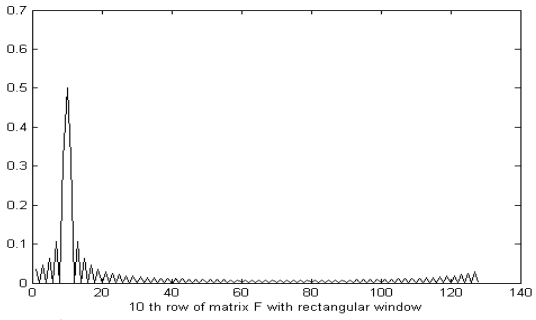

Fig 1 : 10th row of the module of matrix F with a rectangular window

In that latter case, the system of (2N-1) equations can be separated into two systems of N equations with good approximation, by neglecting several bins with very weak power around the Nth frequency. The partionned square matrix F(k,l) (restricted to k=1….N, l=1… N) of length NxN can be so numerically inversed.

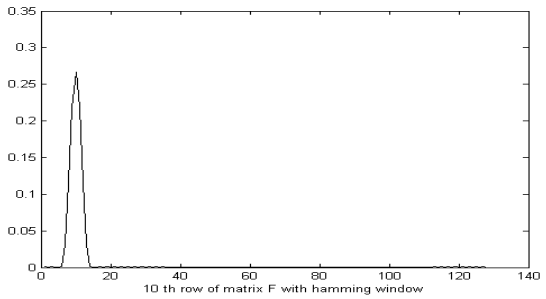

Fig 2 : 10^{th} row of the module of matrix F with a hamming window

3.3. BSS with a natural gradient algorithm

After inversion of the first system (17), the problem becomes : (20)

$$Z1(n,\nu_j) = \sum_{i=0}^{K-1} S1(n-iN,\nu_j)H_i^{11}(\nu_j) + S2(n-iN,\nu_j)H_i^{12}(\nu_j)$$

Where $Z1(n,\nu_j)$ is computed from data *x1(n)*. Suppose that the DFT are computed on data blocks of *x1(n)*, delayed of N samples. The inversion of system (17) provides a time-frequency signal denoted $Z1(kN,\nu_j)$ at frequency ν_j. Under this condition, at each frequency bin ν_j, $Z1(kN,\nu_j)$ can be seen as the filtering between FIR filters with K taps ($H_i^{11}(\nu_j)$ and $H_i^{12}(\nu_j)$ *i=0,...,K-1*) and the K DFT of the sectioned sources S*(kN-iN, ν_j)* *i=0,..., K-1*. For each frequency bin ν_j, the BSS model is equal to:

$$Z1(kN,\nu_j) = H^{11} * S1(kN,\nu_i) + H^{12} * S2(kN,\nu_i) \quad (21)$$
$$Z2(kN,\nu_j) = H^{21} * S1(kN,\nu_i) + H^{22} * S2(kN,\nu_i)$$

So, we have replaced the problem of inversion of filters of K.N taps with the inversion of N filters of K taps. Each inverse filter is estimated independently and modelled with a FIR filter of K' taps. The parameters N and K' can be set to much smaller values than in time-domain or classical frequency-domain. The first interest is that relative short values of N can be used for the DFT, even in the case of long responses and highly reverberant conditions. The permutation indeterminacy and the choice of N are so strongly simplified. For short duration of signals, enough data can be available to achieve the separation. The second interest is that the parameter K' can be chosen small enough to assure a good convergence for the separation filters.

BSS methods developed in time domain for convolved complex sources can be applied to the mixture $Z(kN,\nu_j) = [\, Z1(kN,\nu_j)\ Z2(kN,\nu_j)\,]^T$. After reviewing several BSS algorithms, it can be seen in [7] that information maximization methods best suited with acoustically-mixed sounds. For each frequency bin ν_j, we search the convolutive separating system $\mathbf{w}_j$ will yields outputs $y_j(kN)$ that do not contain any mutual information:

$$y(kN,\nu_j) = \sum_{p=0}^{K'-1} \mathbf{w}_{j,p}(kN) Z_j(kN-pN,\nu_j) \quad (22)$$

where $\mathbf{w}_{j,p}$ is a sequence of K' (2x2) matrices.

We perform the separation with a natural gradient algorithm based on a maximum-entropy cost function [12]. The natural gradient search method [13] is a modified gradient search whereby the standard gradient search direction is altered according to the local Riemannian structure of the parameter space. The resulting search direction is then guaranteed to be invariant to the statistical relationships between the parameters of the model, thus providing statistically efficient learning performance [13].

The complex-valued matrices $\mathbf{w}_{j,p}$ are updated according to [12]:

$$\boldsymbol{w}_{j,p}((k+1)N) = \boldsymbol{w}_{j,p}(kN) + \mu(kN)\left[\, \boldsymbol{w}_{j,p}(kN) - \phi(y((k-P)N,\nu_j)u((k-p)N,\nu_j)^H \right] \quad (23)$$
$$u(kN,\nu_j) = \sum_{q=0}^{K'-1} \boldsymbol{w}^T{}_{j,P-q}(kN) Z(kN-pN,\nu_j)$$

As noted in [13], the optimum choice for each function $\phi(y_i)$ depends on the statistics of each extracted source (y_i) at convergence. The optimal choice of nonlinear activation functions $\phi(y_i) = -d\log(p_i(y_i))/dy_i$ yields the fastest convergence behavior and best steady-state performance if $p_i(y_i)$ is the true p.d.f. of the ith extracted source. Suboptimal choices for these nonlinearities still allow the algorithm to perform separation of the sources, although for a large mismatch there is no guarantee of convergence to the desired solution. Here, $p_i(y_i)$ must suit the p.d.f. of the sources expressed in frequency-domain S*(n, ν_i)*. From [14], we assume Laplacian priors for the sources and we can use the following activation function: $\phi(u) = u/|u|, u \neq 0$

As remarked in [15], maximization of the entropy at the output of the network leads to separation and deconvolution since redundant delayed versions of the same signal result in less entropy overall. Due to this principle, one major drawback that the feedforward architecture suffers in time domain is that it introduces temporal whitening on the recovered sources. Yet, speech signals have short-term dependencies (up to some 5-6msecs, translating to 40-50 samples for a 8kHz-sampled signal). Using the model (21), the extracted sources y*(kN, ν_j)* are estimations of the DFT of the sources computed on successive data blocks of N samples. Under

that condituon, we can then assume that the dependencies between time-frequency samples are weak.

4. EXPERIMENTS IN REVERBERANT ROOM

In order to record the performance of the proposed algorithm, we tested it on some real data available from [16] of two people speaking simultaneously in a room. The two mixtures are constructed with the westner's matlab routine roomix.m. It uses real measured impulse responses and generates highly reverberant mixtures, usually difficult to separate (figure 3). The reverberation time is around 250ms.

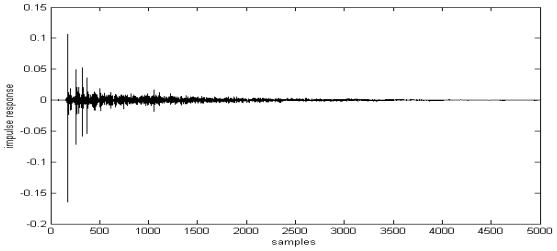

Figure 3 : example of a room impulse response

The source signals are sampled at 22.05kHz and we used 4 seconds for learning. In order to evaluate the performances, we computed the noise reduction rate (NNR in dB), defined as the output signal-to-noise ratio (SNR) in dB in the first estimated source minus input SNR in dB in one sensor. The second source acts as the noise in the SNR. The well-known permutation problem is overcomed as proposed in [9] using the properties of the DFT. The FFT length was set to 128 to 512 and the segments number is varying from 1 to 10 (figure 4).

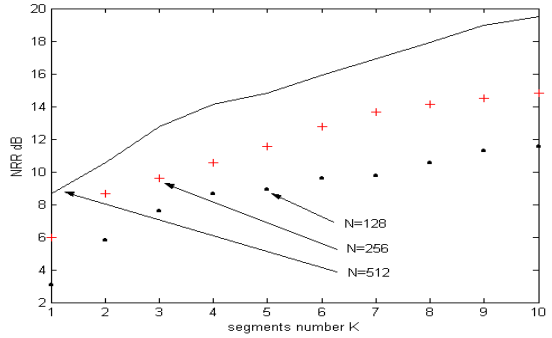

Figure 4 : NRR in function of the segments number K'

We remark that the NRR is increasing with K'. Very good performances can be obtained thus relative short values of the DFT are employed. For example, the best NRR is around 19dB for N=512 and K=10, which is equivalent to a length of 5120 points for the inverse filter. We also tried a classical frequency domain algorithm on the same data with N set to 256 to 8192. The best results were obtained for N=1024 (NRR=9.3 dB). Indeed the performances decrease for larger N since we have too few of frequency data.

5. CONCLUSION

An original frequency model, strictly equivalent to a time linear convolution, is used for BSS of speech signals under highly reverberant conditions. It includes a segmentation of the responses into K segments. Exploiting this model, data are transformed for each frequency bin into convolutive mixtures of K taps. Finally, the separation is achieved with a natural gradient algorithm based on a maximum-entropy cost function. Short values of the DFT N can be employed, although the length of the inverse responses remains long enough (K.N samples), according to real-world responses.

6. REFERENCES

[1] T. Nishikawa, H. Saruwatari and K. Shikano,"Blind source separation based on multi-stage ICA combining frequency-domain ICA and time-domain ICA," *ICASSP'02*, Publisher, Orlando, pp. 917-920, May.

[2] L Parra and C. Spence,"Convolutive blind separation of non-stationary sources", *IEEE Trans. Speech Audio Processing*, vol 8, n°3, pp 320-327, May 2000

[3] P. Smaragdis; " Blind separation of convolved mixtures in the frequency domain", *Neurocomputing*, vol 22, pp 21-34,1998

[4] K. Torkkola, "Blind separation of convolved sources based on information maximisation", *IEEE Workshop on neural networks for signal processing*, 1996, pp 423-432

[5] A. Bell, T. Sejnowski, "An information-maximisation approach to blind separation and blind deconvolution', *Neural Computation*, 7,n06, pp 1129-1159

[6] M.Z. Ikram and D.R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment", *ICASSP'00*, pp 1041-1044,

[7] A. Westner, "Object-based audio capture : separating acoustically-mixed sounds", M.S. Thesis, Massachusetts Institute of Technology", 1998

[8] S. Araki, S. Makino, R. Mukai, T. Nishikawa and H. Saruwatari, "Fundamental limitation of frequency domain blind source separation for convolved mixture of speech', *ICA'01*, pp 132-137, Dec 2001

[9] V. Capdevielle, C. Servière, JL. Lacoume, "Blind separation of wide band sources in frequency domain", *ICASSP 95*, Detroit, Mai 1995, pp 2080-2083

[10] K. Torkkola, "Blind separation for audio signals. Are we there yet?", *ICA 99*, pp 239-2444

[11] RM. Gray, "On the asymptotic eigenvalue distributions of Toeplitz matrices", *IEEE Trans IT*, vol 18, nov 1972

[12] S. Amari, S. Douglas, A. Cichoki and H. Yang,' Novel on line adaptive learning algorithms for blind deconvolution using the natural gradient approach', *11th IFAC Symposium on System Identification*, SYSID 97, Kitakyushu, Japan, 8-11 July 1997, pp 1057-1062

[13] S. Amari, A. Cichoki and H. Yang,' A new learning algorithm for blind signal separation', in *Avances in Neural Information Processing Systems 8,* pp 752-763, MIT Press, Cambridge, MA, 1996

[14] M. Davies, 'Audio source separation', *Mathematics in Signal Processing V*, 2000

[15] X. Sun and S. Douglas,"A natural gradient convolutive blind source separation algorithm for speech mixtures", ICA'01, pp59-64, Dan Diego

[16] http://sound.media.mit.edu/ica-bench