SEPARATION OF SPEECH SIGNALS UNDER REVERBERANT CONDITIONS

Christine Servière

Laboratoire des Images et des Signaux, ENSIEG, BP 46 38402 Saint-Martin d'Hères France
Email :Christine.Serviere@lis.inpg.fr

ABSTRACT

BSS performance is not still enough for speech signals and long acoustic responses. An original frequency model, strictly equivalent to a time linear convolution, is used for speech signals under highly reverberant conditions. If the responses are virtually sectioned in K blocks of N samples, the time linear convolutions are strictly transformed in frequency domain at frequency ν, into FIR filtering of K taps where the K taps are the complex gains of the K sectioned blocks at the same frequency ν. Short values of the DFT, N, can be employed, although the length of the responses remains long enough (K.N samples) to suit with acoustic responses. Finally, the separation is achieved with a natural gradient algorithm based on a maximum-entropy cost function. The proposed method is then tested on speech signals.

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. It aims at the retrieval of independent sources and tests the statistical independence of the separated signals according to different measures as higher-order cumulants or mutual information.

Most research is done for BSS with instantaneous mixtures. However, in the context of speech signal separation, BSS must be necessary achieved for convolutive mixtures as the acoustic impulse responses contain typically several echoes and reverberation. The separation can be applied in time or in frequency-domain. However, working in frequency domain is now commonly admitted in speech applications. Indeed, the iterative learning rule is complicated for large-tap FIR filter in time domain and the convergence strongly degrades [1,2]. In frequency-domain, convolutive mixtures are usually reduced to simultaneous mixtures and BSS for instantenous mixtures can be used with great performances. In this paper we also achieve BSS in frequency domain for the sake of simplicity and stability. It performs usually good results when no reverberation is present. However, under reverberant conditions, the separation performance remains not still enough. The frame size of discrete Fourier transforms (DFT), N, must be discussed in detail, versus the length of a room impulse response L. N must verify N>>L in order to estimate an unmixing filter [3,4]. 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 one [3,4]. Unfortunately, the separation performance is not increasing with N and is saturated before reaching a sufficient performance because few data are available in frequency domain for a constant duration of the observations lengths [1,2]. BSS methods then fail to test the independence of the estimated sources.

Besides, 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 only correct when the real impulse response lengths are short in comparison to N. In order to resolve the problem, we propose in section 2 to use a complete model in frequency domain, exactly equivalent to a time linear convolution. The idea is derived from the overlap-add method [5]. The responses are virtually sectioned in K blocks of N samples. Data are then transformed in frequency domain at frequency 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 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 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 enough (K.N taps) according to the 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.

2. SOUND MIXING MODEL

2.1 BSS mixing mdel

We consider a M input, M output convolutive problem. Each microphone $x_j(t)$ receives a linear convolution (noted *) of each sound source $s_i(t)$:

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

where h_{ij} is the impulse response from source i to output j. In frequency domain, it is usually reduced to:

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

where $X(n,\nu)$ (respectively $S(n,\nu)$) is the N-points DFT of the nth data vector $X(n)$ (respectively $S(n)$).
This frequency model is simple but equation (2) is only an approximation as it implies a circular convolution in time-domain and is justified only for N>>L. But if N>>L, not enough data $X(n,\nu)$ are available to estimate the sources for a usual duration of the observations lengths (3 or 4 seconds). So, we compute a complete frequency expression, strictly equivalent to a time linear convolution. This model is exposed in two parts. Consider a long impulse response **h** of length L : $\boldsymbol{h}=[h_0, ..., h_{L-1}]^T$. It can be virtually sectioned in K segments of length N and elementary impulse responses $\boldsymbol{h_i}$:

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

Due to the principle of superposition, the resulting time signal $x(n)$ of the linear convolution between **h** and a signal $s(n)$ is the addition of the linear convolution of all the elementary filters $\boldsymbol{h_i}$. Each elementary output is calculated in section 2.2. Then the complete model is exposed for long responses in section 2.3. Uppercase symbols will denote frequency variables, lowercase symbols stand for time variables.

2.2 Frequency model

Consider $x(n)$, the linear convolution between $s(n)$ and a FIR filter H. Let $\boldsymbol{h}=[h_0, ..., h_{N-1}]^T$ be its response of N taps. $\boldsymbol{x(n)}$ denotes the data block $[x(n-N), ..., x(n), ..., x(n+N-1)]^T$ and is given by: (3)

$$\underbrace{\begin{bmatrix} x(n) \\ x(n+1) \\ . \\ . \\ x(n+N-1) \end{bmatrix}}_{\boldsymbol{x(n)}} = \begin{bmatrix} s(n) & s(n-1) & ... & s(n-N+1) \\ s(n+1) & s(n) & ... & s(n-N+2) \\ . & . & ... & . \\ s(n+N-1) & s(n+N-2) & ... & s(n) \end{bmatrix} . \underbrace{\begin{bmatrix} h_0 \\ h_1 \\ . \\ h_{N-1} \end{bmatrix}}_{\boldsymbol{h}}$$

Extend the Toeplitz matrix (NxN) in (3) to a circulant matrix (2Nx2N), noted $\boldsymbol{\chi}$, built with samples of signal $s(n)$.
The resulting vector is: $\boldsymbol{x'(n)}=[x'(n-N), ...,x'(n+N-1)]^T$: (4)

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

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

It can be seen from equations (3) and (4) that the N last components of vector $\boldsymbol{x'(n)}$ are equal to $\boldsymbol{x(n)}$. So, vectors $\boldsymbol{x(n)}$ and $\boldsymbol{x'(n)}$, verify :

$$\boldsymbol{f}.x(n)=f.x'(n) \qquad (5)$$

if $\boldsymbol{f}$ is a (2Nx2N) diagonal matrix: $\boldsymbol{f}=diag(0,...,0, f_0, ..., f_{N-1})$ and $[f_0, ..., f_{N-1}]^T$ a window of length N. Consequently it comes from equations (4) and (5) that :

$$f.x(n)=f.x'(n)=f.\chi.h' \qquad (6)$$

W denotes the symmetric matrix (2Nx2N) whose kth, lth element is $\boldsymbol{W_{kl}}=\exp(-j2\pi kl/2N)$. Multiplying equation (6) by the DFT matrix $\boldsymbol{W}$ leads to :

$$W.f.x(n) = W.f.x.h' = \underbrace{Wf\,W^{-1}}_{F} . \underbrace{WxW^{-1}}_{S(n)} . \underbrace{Wh'}_{H} \qquad (7)$$

As $\boldsymbol{\chi}$ is a circulant matrix, it owns the DFT matrix $\boldsymbol{W}$ as eigenvectors. 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 $\boldsymbol{\chi}$, i.e. the 2N-points DFT of : $[s(n-N), ..., s(n), ..., s(n+N-1)]^T$.
They are denoted : $\boldsymbol{S(n)}=diag(S(n,\nu_0), ..., S(n,\nu_{2N-1}))$

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

$$\boldsymbol{H} = W.[h_0, ..., h_{N-1}, 0, ..., 0]^T=[H_0, ..., H_{2N-1}]^T \qquad (8)$$

As a conclusion, the model (7) becomes :

$$W.f.x(n) = F.S(n).H \qquad (9)$$

where $\boldsymbol{W.f.x(n)}$ represents the 2N-points DFT of the block $[x(n-N), ..., x(n), ..., x(n+N-1)]^T$, multiplied by the window $[0, ..,0., f_0, \ldots, f_{N-1}]^T$.Consequently, eq. (9) relies the DFT of the signal $s(n)$ and the DFT of the impulse response $\boldsymbol{H}$ to the time linear convolution $x(n)$. It is the complete frequency model equivalent to the time linear convolution.

2.3 Case of a long impulse response

Consider the long impulse response **h** of length L. It can be virtually sectioned in K segments of length N of elementary impulse responses $\boldsymbol{h_i}$, with (L=K.N) :

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

Due to the principle of superposition, the resulting time signal $x(n)$ of the linear convolution between H and $s(n)$ is the addition of the linear convolution of all the elementary filters $\boldsymbol{h_i}$. Each elementary output is given by (9) 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.

$$\boldsymbol{H_i} = W.[h_{iN}, ..., h_{iN+N-1}, 0, ..., 0]^T \qquad i=0,...,K-1 \qquad (10)$$
$$\boldsymbol{H_i} = [H_i(\nu_0), ..., H_i(\nu_{2N-1})]^T$$

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

$$W.f.x(n) = F.\sum_{i=0}^{K-1} S(n-iN).H_i \qquad (11)$$

where $\boldsymbol{S(n-iN)}$ is the 2Nx2N diagonal matrix whose elements are the DFT of the data block: $[s(n-iN-N), ..., s(n-iN), ..., s(n-iN+N-1)]^T$. Eq.(11) can also be written with the following expression, similar to equation (9) :

$$W.f.x(n) = \qquad (12)$$
$$F.\underbrace{[S(n)\ S(n-N)\ ...\ S(n-(K-1)N)]}_{S'(n)}.\underbrace{[H_0\ H_1\ ...H_{(K-1)}\ ]^T}_{H'}$$

where ***S'(n)*** is a (2Nx2KN) row-block matrix obtained by stacking the K diagonal matrices ***S(n-iN)*** (for i=0,…, K-1). Let ***H'*** be the 2KN vector obtained by stacking vectors $\boldsymbol{H_i}$ (i=0,…,K-1). The 2N points DFT of the windowed signal *x(n)*, noted $\boldsymbol{X_f(n,\nu)}$, is given by: (13)

$$\boldsymbol{X_f(n,\nu) = W.f.x(n) = F.} \underbrace{\boldsymbol{S'(n).H'}}_{\boldsymbol{S(n,\nu).H_0(\nu)+...+S(n-KN,\nu).H_{K-1}(\nu)}} \qquad \nu=\nu_0,...,\nu_{2N-1}$$

As ***F*** is not a diagonal matrix, $\boldsymbol{X_f(n,\nu)}$ at frequency bin ν is a linear combination of the terms :

$$S(n,\nu_i).H_0(\nu_i)+...+S(n-KN,\nu_i).H_{K-1}(\nu_i)$$

at all frequency bins ν_i and eq. (13) is not easy to handle since all frequencies are mixed. Besides, the system (13) cannot be obviously inversed as ***F*** is of rank N : recall that $\boldsymbol{F = W.f.W^{-1}}$ and $\boldsymbol{f}=diag(0,...,0,f_0,\ ...,f_{N-1})$.
However, the useful informations issued of the DFT are restricted on the N first frequencies for real-valued signals and the N first components of vector (***S'(n). H')*** can be recovered under some conditions. If the window is a hamming function, **F** is a banded matrix (see the 10th row of F on figure 1). Consequently, system (13) of (2N) equations can be separated into two systems of N equations. The partionned square matrix **F**(k,l) (restricted to k=1….N, l=1… N) of length NxN can be numerically inversed by F_N^{-1} [6].

After inversion of the first system of (13) by F_N^{-1}, the model becomes:

$$\boldsymbol{Z}(n,\nu) = F_N^{-1}.\boldsymbol{X_f(n,\nu)} = \boldsymbol{S'(n).H'} \qquad \nu=\nu_0,\ ...,\ \nu_{N-1} \quad (14)$$

And the ith equation of (14) is equal to :

$$\boldsymbol{Z}(n,\nu_i) = S(n,\nu_i).H_0(\nu_i)+...+S(n-KN,\nu_i).H_K(\nu_i) \qquad (15)$$

Suppose now that the DFT are computed and data blocks of ***x(n)***, delayed of N samples. $Z(kN,\nu_i)$ can be seen, at each frequency bin ν_i , as the filtering between the FIR filter of K taps ($H_i(\nu_i)$ $i=0,...,K-1$) and the K DFT of the sectioned signal $S(kN-iN,\nu_i)$ $i=0,...,K-1$.
For each frequency bin ν_i, the model is equal to a linear convolution, versus the time index *k*:

$$Z(kN,\nu_i) = \boldsymbol{H}^* S(kN,\nu_i) \qquad (16)$$

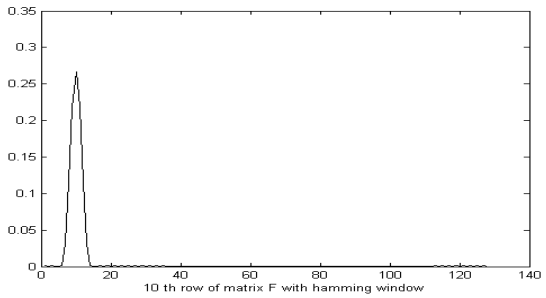

figure 1: 10th row of the module of matrix F with a hamming function for N=64

3. APPLICATION TO BSS

For more simplicity, consider here a 2 inputs, 2 outputs convolutive problem. Two sensors receive mixtures of sources *s1(n)* and *s2(n)*, mixed with filters H^{11}, H^{12}, H^{21} and H^{22}. From section 2.3, for each frequency bin ν_j, the mixing model is equal to :

$$Z1(kN,\nu_i) = H^{11} * S1(kN,\nu_i) + H^{12} * S2(kN,\nu_i) \qquad (17)$$
$$Z2(kN,\nu_i) = H^{21} * S1(kN,\nu_i) + H^{22} * S2(kN,\nu_i)$$

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 are 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 mixtures $Z(kN,\nu_i) = [\,Z1(kN,\nu_i)\ Z2(kN,\nu_i)\,]^T$, but information maximization methods best suited with acoustically-mixed sounds. For each frequency bin ν_i, we search the convolutive separating system **w** will yields outputs $y(kN,\nu_i)$ that do not contain any mutual information:

$$y(kN,\nu_i) = \sum_{p=0}^{K'-1} \mathbf{w}_p(kN,\nu_i) Z(kN-pN,\nu_i) \qquad (18)$$

where $\mathbf{w}_p(kN,\nu_i)$ is a sequence of K' (2x2) matrices.

We perform the separation with a natural gradient algorithm based on a maximum-entropy cost function [7]. It 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 [8].

The complex-valued matrices $\mathbf{w}_p(kN,\nu_i)$ are updated according to [7]:

$$\mathbf{w}_p((k+1)N,\nu_i) = \boldsymbol{w}_p(kN,\nu_i) +$$
$$\mu(kN)\left[\boldsymbol{w}_p(kN,\nu_i) - \phi(y((k-P)N,\nu_i)u((k-p)N,\nu_i)^H\right] \qquad (19)$$
$$u(kN,\nu_i) = \sum_{q=0}^{K'-1} \boldsymbol{w}^T{}_{P-q}(kN,\nu_i) Z(kN-pN,\nu_i)$$

The optimum choice for each function $\phi(y_i)$ depends on the statistics of each extracted source (y_i) at convergence. The optimal choice $\phi(y_i) = -d\log(p_i(y_i))/dy_i$ 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 $S(n, \nu_i)$. From [9], we assume Laplacian priors for the sources and we use the following activation function: $\phi(u) = u/|u|, u \neq 0$

As remarked in [10], 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. 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 (18), the extracted sources $y(kN, \nu_i)$ are estimations of the DFT of the sources computed on successive data blocks of N samples. Under that condition, we can then assume that the dependencies between time-frequency samples are weak.

4. EXPERIMENTS IN REVERBERANT ROOM

The performance of the proposed algorithm is tested on real data available from [11] of two people speaking simultaneously in a room. Two mixtures are constructed with real measured impulse responses and generate highly reverberant mixtures (figure 2). The reverberation time is around 250ms.

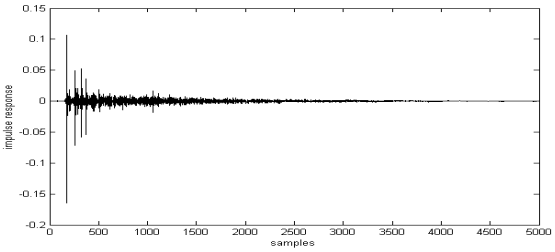

Figure 2 : example of a room impulse response

The source signals are sampled at 22.05kHz and we used 4 seconds for learning. The performance is evaluated with 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 overcome as proposed in [12]. The FFT length was set to 128 to 512 and the segments number is varying from 1 to 10 (figure 3).

We remark that NRR is increasing with K'. Very good performances can be obtained with relative short values of the DFT. 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 performance is saturated and best results were obtained for N=1024 (NRR=9.3 dB). Indeed the performances decrease for larger values of N since we have too few of frequency data.

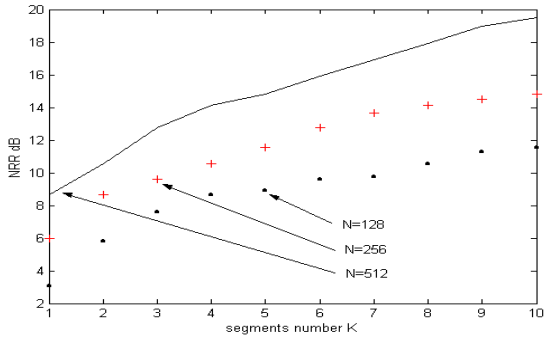

Figure 3 : NRR in function of the segments number K'

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

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

[3] 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

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

[5] O.Ait Amrane, E.Moulines, Y.Grenier, " Structure and convergence analysis of the generalized multi-delay adaptive filter", EUSIPCO'92, pp115-118

[6] C.Servière, 'Separation of speech signals with segmentation of the impulses responses under reverberant conditions', ICA'03, pp 511-516, Nara April 2003.

[7] 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

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

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

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

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

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