BLIND SEPARATION OF CONVOLUTIVE MIXTURE OF SPEECH SIGNALS

Hakim Boumaraf[†]*, Dinh-Tuan Pham*[†]*, and Christine Servière*[‡]

[†]Laboratory of Modeling and Computation, INPG/UJF/CNRS
B.P. 53X, 38041 Grenoble cedex 9, France
phone: +33 4 76 51 44 23, fax: +33 4 76 63 12 63, email: {Hakim.Boumaraf, Dinh-Tuan.Pham}@imag.fr
web: www-lmc.imag.fr/lmc-sms/{Hakim.Boumaraf, Dinh-Tuan.Pham}
[‡] Laboratoire des Images et des Signaux
BP 46, 38402 St Martin d'Hère Cedex, France
phone: +33 4 76 51 44 23, fax: +33 4 76 63 12 63, email: Christine.Serviere@inpg.fr

ABSTRACT

We present in this paper an improvement for our previous blind source separation of speech signals based on the joint diagonalization of the time varying spectral matrices of the observations and the use of energy profiles to handle the problem of permutation ambiguity in the frequency domain. Two new techniques are proposed to improve the estimation of profiles which are used for permutation corrections. Simulations using real impulse response of acoustic room show that these novel profiles estimation methods improve the efficiency of our algorithm, which performs well even in the difficult reverberation environment characterized by long response .

1. INTRODUCTION

Blind separation of convolutive mixture of speech signals has been the subject of many researches [4, 9, 11], but the performance of the proposed algorithms in realistic setting is still not quite satisfactory [3], due mainly to the long impulse response of the mixing filter. Time domain approach would be too computational costly and suffers from the difficulty of convergence since it requires the adjustment of too many parameters. Frequency domain approach has the advantage that it reduces the problem to a set of independent problems of separation of instantaneous mixtures in each frequency bin, but it creates the additional difficult problem of permutation ambiguity. Further, since the finite Fourier transform tends to produce nearly Gaussian variables [2], higher (than second) order statistics in the frequency domain contain little useful information for these separation problems. Fortunately, speech signals are highly non stationary so one can exploit this nonstationarity to separate their mixture based only their second order statistics [6], which leads to a joint diagonalization problem. This idea has been introduced by Para and Spence [4], but these authors used an ad-hoc criterion, while in our two earlier papers [7, 8], we use a criterion based on the Gaussian mutual information and related to the maximum likelihood. Such criterion has in fact been considered in [11], but without using the nonstationarity idea.

The main problem in a frequency domain approach is to resolve the permutation ambiguity. In [7, 8] two methods are proposed, based on the continuity of the demixing filter and on the use of source energy profiles (similar to an idea in [1]), respectively. However, the use of profiles is based on a very crude model for the time varying spectrum of the source. In this paper we improve this model, leading to a better permutation ambiguity resolution.

2. MODEL AND METHODS

Consider an acoustic situation in which K sensors receive signals from K sources. The observed sequences $\{x_1(t)\},\ldots,\{x_K(t)\}$ are assumed to be linear mixtures of sources sequences $\{s_1(t)\},\ldots,\{s_K(t)\}$ with delay:

$$x_k(t) = \sum_{n=-\infty}^{\infty}\sum_{i=1}^{K} H_{ki}(n)s_i(t-n), \qquad (1)$$

where $H_{kj}(n)$ are elements of the impulse response matrix $\mathbf{H}(n)$ of the mixing filter. The goal is to recover the sources through a demixing filter: the reconstructed source sequences $\{y_1(t)\},\ldots,\{y_K(t)\}$ are the components of $\{\mathbf{y}(t) = \sum_{n=-\infty}^{\infty}\mathbf{G}(n)\mathbf{x}(t-n)\}$, where $\mathbf{G}(n)$ is the impulse response matrix of the filter and $\mathbf{x}(t) = [x_1(t) \ \cdots \ \mathbf{x}_K(t)]^{\mathbf{T}}$, $^{\mathbf{T}}$ denoting the transpose. In the blind context, the idea is to adjust the filter such that $\{y_k(t)\}$ are as mutually independent as it is possible. In a second order approach, only the inter-spectra between the reconstructed sources at every frequency are needed to express dependence, but since we are dealing with nonstationary signals, we shall consider the time varying spectra, that is the localized spectra around each given time point. It is precisely the time evolution of these spectra which helps us to separate the sources. From (1), the time varying spectrum of the vector observation sequence is $S_{\mathbf{x}}(t,f) = \mathbf{H}(f)S_{\mathbf{s}}(t,f)\mathbf{H}^*(f)$ where $\mathbf{H}(f) = \sum_{n=-\infty}^{\infty} e^{j2\pi nf}\mathbf{H}(n)$ denotes the frequency response of the mixing filter, $S_{\mathbf{s}}(t,f)$ is the diagonal matrix with diagonal elements being the time varying spectra of the sources and * denotes the transpose conjugated. As in [7, 8], we aim to make the spectrum of the reconstructed source vector $\mathbf{G}(f)S_{\mathbf{x}}(t,f)\mathbf{G}^*(f)$ to be as close to diagonal as it is possible, according to the following diagonalization criterion:

$$\sum_t \left\{ \frac{1}{2}\log\det \mathrm{diag}[\mathbf{G}(f)\hat{S}_{\mathbf{x}}(t,f)\mathbf{G}^*(f)] - \log\det|\mathbf{G}(f)| \right\}$$

where $\mathrm{diag}(\cdot)$ denotes the operator which builds a diagonal matrix from its argument and the summation is over the time points of interest. A simple and very fast algorithm to minimize this criterion has been already developed [5].

In practice, the spectrum $\hat{S}_{\mathbf{x}}(t,f)$ is estimated over a (high resolution) grid of frequencies. It is important to have a good estimator, since the final separation would depend on it. We chose a new variant of the estimator used in [7, 8] which has better performance (reducing side-lobe level). We form

the sliding short term periodogram using a *Hanning taper window*

$$P_{\mathbf{x}}(\tau,f)=\Big[\sum_t H_N(t-\tau)\mathbf{x}(t)e^{2\pi ift}\Big]\Big[\sum_t H_N(t-\tau)\mathbf{x}(t)e^{2\pi ift}\Big]^*.$$

where $H_N(t)=\sqrt{2/(3N)}[1-\cos(2\pi t/N+\pi/N)]$ for $0\le t<N$, 0 otherwise. The above periodogram will be averaged over m consecutive equispaced points $\tau_1,\ldots,\tau_m$ yielding the estimated spectrum at time $(\tau_1+\tau_m+N-1)/2$:

$$\hat{S}_{\mathbf{x}}\Big(\frac{\tau_1+\tau_m+N-1}{2},f\Big)=\frac{1}{m}\sum_{k=1}^{m}P_x(\tau_k,f)$$

The frequencies are taken to be of the form $f=n/N$, $n=0,\ldots,N-1$, with N being chosen to be a power of 2 to take advantage of the Fast Fourier Transform. The frequency resolution is determined by the taper window length N and the time resolution by $m\delta$ where $\delta=\tau_i-\tau_{i-1}$. Using $\delta\gg 1$ helps to reduce the computational cost but slightly degrades the estimator: actually δ can be a small fraction of N without a significant degradation. Of course a compromise between time and frequency resolution has to be made. Our method is more flexible for adjusting these resolutions than that of [7,8] and further helps to reduce the bias.

3. THE PERMUTATION AMBIGUITY PROBLEM

The advantage of the frequency domain approach, as explained in the introduction, comes however with a price: the ambiguity of its solution. The joint diagonalization only provides the matrices $\mathbf{G}(f)$ up to a scale change and a permutation: if $\mathbf{G}(f)$ is a solution then so is $\mathbf{\Pi}(f)\mathbf{D}(f)\mathbf{G}(f)$ for any diagonal matrix $\mathbf{D}(f)$ and any permutation matrix $\mathbf{\Pi}(f)$. Thus, one only gets a separation filter of frequency response matrix $\mathbf{G}(f)$ of the form $\mathbf{\Pi}(f)\mathbf{D}(f)\hat{\mathbf{H}}^{-1}(f)$ where $\hat{\mathbf{H}}(f)$ is a consistent estimator of $\mathbf{H}(f)$ but $\mathbf{\Pi}(f)$ and $\mathbf{D}(f)$ are *arbitrary* permutation and diagonal matrices. The scale ambiguity is intrinsic to the blind separation of convolutive mixtures and cannot be lifted, but the permutation ambiguity must be reduced to a global ambiguity *independent of the frequency*.

Permutation ambiguity in convolutive separation is a difficult problem, especially in audio applications [3]. In the literature several methods have been proposed to solve this problem. One method constrains the separation filter to have FIR support [11], others introduce some coupling between solutions in frequency domain [1, 9] or otherwise using the continuity of the frequency response [9, 11].

This paper extends and improves the method introduced in [8], that exploits direct intrinsic properties of sounds by constructing energy distribution profiles (in logarithmic scale) and using them to lift the permutation ambiguity. The main idea is that, for a speech signal at least, the energy over different frequency bins appears to vary in time in a similar way, up to a gain factor. For example, if a time block contains a long period of pause, one would expect that it energy would be nearly zero in all frequency bins. The method in [8] exploits this idea by implicitly assuming a model of time varying spectrum of the k-th source of the form $S_k(t,f)=\exp[E_k(t)]S'_k(f)$, where $E_k(t)$ denotes the "profile". For uniqueness of the above representation, $E_k(t)$ shall be constrained to satisfy $\sum_{j=1}^{L}E_k(t_j)=0$, where $t_1,\ldots,t_L$ denote the considered time points. One can then estimate $E_k(t_j)$ by

$$\hat{E}_k(t_j)=\frac{1}{N}\sum_{n=0}^{N-1}\tilde{E}_{\pi(k;n/N)}\Big(t_l,\frac{n}{N}\Big) \tag{2}$$

$$\tilde{E}_k(t_j,f)=\log\hat{S}_k(t_j,f)-\frac{1}{L}\sum_{l=1}^{L}\log\hat{S}_k(t_l,f) \tag{3}$$

where $\hat{S}_k(t_l,f)$ is the k-th diagonal element of the matrix $G(f)S_{\mathbf{x}}(t_l,f)\mathbf{G}(f)$ and $\pi(1;f),\ldots,\pi(K;f)$ is a permutation of $1,\ldots,K$, which corrects the permutation errors in the output of the diagonalization algorithm so that $\hat{S}_{\pi(k;f)}(t_l,f)$ is indeed an estimate of the spectral density of the k-th source at frequency f. Note that the scale ambiguity is automatically eliminated, since this ambiguity amounts to adding to $\log\hat{S}_k(t_j,f)$ a term depending only on f, which leaves $\hat{E}_k(t_j,f)$ unchanged. However the frequency permutation corrections $\pi(1;n/N),\ldots,\pi(K;n/N)$ are unknown. Therefore, we proceed iteratively. We start with some initial permutation corrections ($\pi(i;n/N)=i$ for ex.), compute the profiles as in (2), then update the permutation correction for each frequency $f=n/N$ by minimizing the criterion

$$\sum_{k=1}^{K}\sum_{l=1}^{L}[\tilde{E}_{\pi(k)}(t_l,f)-\hat{E}_k(t_l)]^2$$

over all possible permutations $\pi(1),\ldots,\pi(K)$ of $1,\ldots,K$ and set the new permutation correction $\pi(1;f),\ldots,\pi(K;f)$ to be the one realizing the minimum. The profile is the re-estimated and so on until convergence.

The model "$S_k(t,f)=\exp[E_k(t)]S'_k(f)$" is however very crude, as can be seen by examining the time-frequency spectrum of speech samples. In spite of this, the above method works reasonably well, since this model only serves to make permutation corrections. But the method can be improved by adopting a more realistic model, as in this paper where we allow the profile to depend on the frequency but only *mildly*. Specifically, we now model $S_k(t,f)$ as $\exp[E_k(t,f)]S'_k(f)$, where $E_k(t,f)$ is a slowly varying function of f for each t and satisfies $\sum_{l=1}^{L}E_k(t_l,f)=0$. Note that if only the last condition is required, the natural estimate of $E_k(t,f)$ would be $\hat{E}_{\pi(k;f)}(t,f)$ where $\hat{E}_k(t,f)$ is given by (3) and $\pi(1;f),\ldots,\pi(K;f)$ is the permutation correction. By adding the information of *slow variation* with respect to the frequency variable, one is led to a second estimate $\hat{\hat{E}}_k(t,f)$ for $E_k(t,f)$, which will be described below. The algorithm then proceeds as before except that the permutation correction for each frequency $f=n/N$ is now updated by minimizing

$$\sum_{k=1}^{K}\sum_{l=1}^{L}[\tilde{E}_{\pi(k)}(t_l,f)-\hat{\hat{E}}_k(t_l,f)]^2$$

over all possible permutations $\pi(1),\ldots,\pi(K)$ of $1,\ldots,K$ and set $\pi(1;f),\ldots,\pi(K;f)$ as the one realizing the minimum.

3.1 Estimation by moving average

One estimates the profile for the k-th source by taking the moving average of the natural profile estimator after permutation correction $\tilde{E}_{\pi(k,f)}(t_l,n/N)$

$$\hat{\hat{E}}_k\Big(t_l,\frac{n}{N}\Big)=\frac{1}{\tau}\sum_{p=n-(\tau-1)/2}^{p=n+(\tau-1)/2}\tilde{E}_{\pi(k,p/N)}\Big(t_l,\frac{p}{N}\Big)$$

where τ denotes the window width, assumed to be odd, which controls the degree of smoothness of the estimator.

3.2 Estimation using Discrete Fourier Sequence (DFS)

Since $\hat{E}_{\pi(k,f)}(t_l, n/N)$ as a function of n (for fixed t_l) is periodic with period N, one can express it as a Fourier sum:

$$\hat{E}_{\pi(k,f)}\left(t_l, \frac{n}{N}\right) = \sum_{r=0}^{N-1} C_r e^{j2\pi rn/N} \tag{4}$$

where $\{C_0, \ldots, C_{N-1}\}$ is the discrete Fourier transform of $\{\hat{E}_{\pi(k,f)}(t_l, n/N), n = 0, \ldots, N-1\}$:

$$C_r = \frac{1}{N} \sum_{n=0}^{N-1} \hat{E}_{\pi(k,f)}(t_l, n/N) e^{-j2\pi rn/N}. \tag{5}$$

In the representation (4) the terms correspond to indexes r far from 0 and N induce rapid oscillations of the function. Thus to estimate $E_k(t_l, n/N)$, one may simply suppress in this representation the terms of index r for which $\min(r, N-r) > L$, L being a parameter controlling the degree of smoothness of the estimator. Specifically, the estimator is given by

$$\hat{\hat{E}}_k\left(t_l, \frac{n}{N}\right) = \sum_{r=0,\ldots,L,N-L,\ldots,N-1} C_r e^{j2\pi rn/N}.$$

Note that $C_r = C_{N-r}$ so that the above right hand side is actually a sum of a constant and cosine functions.

4. DESIGN AND SIMULATION RESULTS

To validate our algorithm, we experiment with real acoustic response impulse measured by McMaster University [10] to hearing aid in BLISS project. The used response impulse was measured in a 3.4m $\times$ 3.4m $\times$ 2.6m room. We chose in this simulation one that corresponds to the combination: angle 315° & 45°, height 18cm, distance 0.9m (measured from KEMAR [10]). The length of the impulse respose is truncated to 1024 lags after downsampling the speech source signals to frequency 11025Hz. The signals have a duration of about 2.98s. Figure 1 shows the impulse responses of the mixing filter with all it's echos.

We take as block length $N = 2048$ with an overlap of $1-(\delta-1)/N = 75\%$ (yielding 57 time blocks) and estimate the spectral matrices by averaging over 5 blocks ($m = 5$). As in [7, 8], we consider the performance index

$$r(f) = |(\mathbf{GH})_{12}(f)(\mathbf{GH})_{21}(f)/[(\mathbf{GH})_{11}(f)(\mathbf{GH})_{22}(f)]|^{1/2}$$

where $(\mathbf{GH})_{ij}(f)$ is the ij element of the matrix $\mathbf{G}(f)\mathbf{H}(f)$. For a good separation, this index should be close to 0 or infinity (in this case the estimated sources are permuted). When r crosses the value 1, this means that a permutation has occurred.

We will present, in follow, the results of DFS method. The results for the "moving average" method are similar. The DFS method has the advantage that it is easier to tune since the parameter L should be quite small. Figure 2 plots $\min(r, 1)$ and $\min(1/r, 1)$ versus frequency (in Hz), before (with old profiles estimation [8] applied in) and after applying the new method of profiles estimation (DFS with $L = 3$).

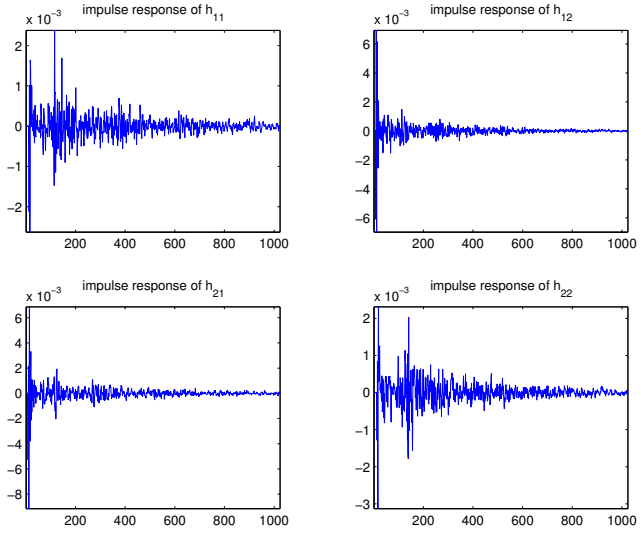

Figure 1: Impulse responses of the considered real acoustic filter

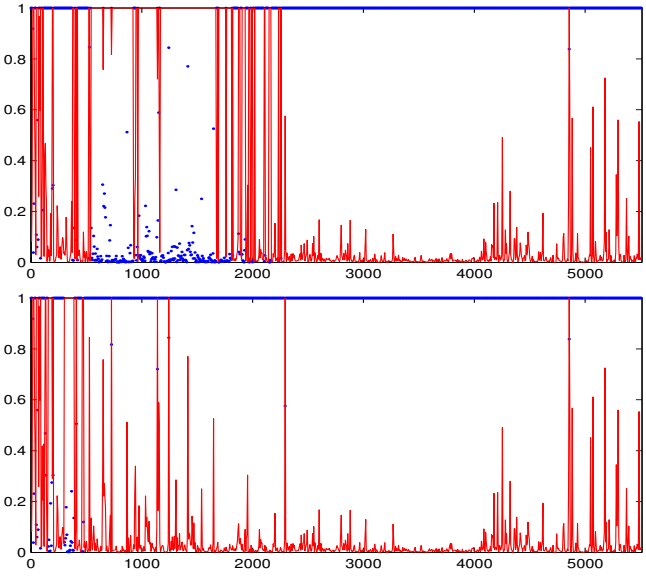

Figure 2: Separation index (solid red) and its inverse (blue dots) truncated at 1, before (upper panel) and after (lower panel) applying the new permutation correction, frequency in Hz

One can see that the new method eliminates many permutation errors (relative to a global permutation) which can not be eliminated by the old method.

Figure 3 plots the distance $\|\hat{\hat{E}}_1(\cdot, f) - \hat{\hat{E}}_2(\cdot, f)\|$ between the two reference profiles (solid blue) together with the distances from a raw source profiles to it reference profiles $\|\tilde{E}_k(\cdot, f) - \hat{\hat{E}}_k(\cdot, f)\|, k = 1$ (point red), 2 (dotted black). It appears that our method should work well at higher frequency where the two reference profiles are more separated.

The impulse response of the global filter $(\mathbf{G} * \mathbf{H})(n)$ is shown in figure 4. One can see that $(\mathbf{G} * \mathbf{H})_{11}(n)$ is much smaller than $(\mathbf{G} * \mathbf{H})_{12}(n)$ and $(\mathbf{G} * \mathbf{H})_{22}(n)$ is somewhat smaller than $(\mathbf{G} * \mathbf{H})_{21}(n)$, meaning that the sources are well

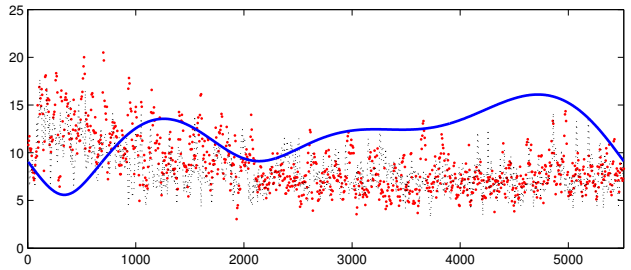

Figure 3: Distances between reference profiles (solid blue) and between a source profile and its reference profile (point red and dotted black), frequency in Hz

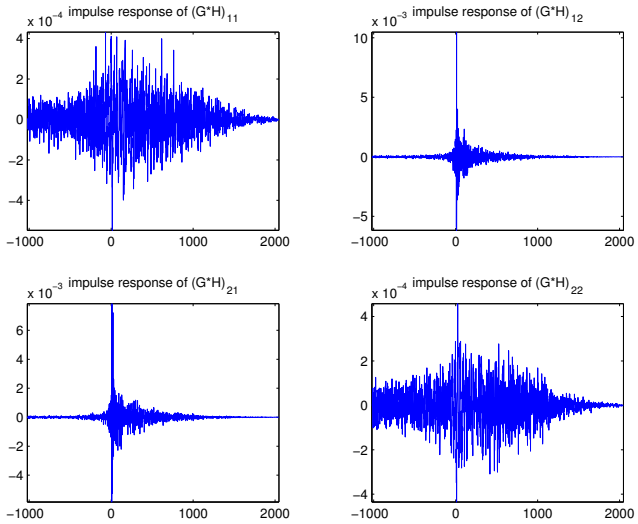

Figure 4: Impulse response of the global filter $(\mathbf{G} * \mathbf{H})(n)$

separated (and permuted). This can be confirmed by looking at the original sources, the mixtures and the separated sources, displayed in figure 5 (noting that there is a global permutation).

5. CONCLUSION

We have improved the blind separation of speech signals algorithm in [8] by introducing two new profiles estimation methods. The proposed algorithm is able to separate convolutive mixtures with fairly long impulse responses arising from real acoustic environments that contain strong echos.

REFERENCES

[1] J. Anemüler and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proceeding of ICA 2000 Conference*, Helsinki, Finland, June 2000, pp. 215–220.

[2] D. R. Brillinger, *Time series: Data Analysis and Theory*. Holt, Rinehart and Winston: New-York, 1975.

[3] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Evaluation of separation and dereverberation performance in frequency domain blind source separation," *Acoustical Science and Technology*, vol. 25, No.2, pp. 119–126, Mar. 2004.

[4] L. Parra and C. Spence, "Convolutive blind source separation of non-stationary sources," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[5] D.-T. Pham, "Joint approximate diagonalization of positive definite matrices," *SIAM J. on Matrix Anal. and Appl.*, vol. 22, no. 4, pp. 1136–1152, 2001.

[6] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of non stationary sources," *IEEE Trans. Signal Processing*, vol. 49, no. 9, pp. 1837–1848, 2001.

[7] D.-T. Pham, C. Servière, and H. Boumaraf, "Blind separation of convolutive audio mixtures using nonstationarity," in *Proceeding of ICA 2003 Conference*, Nara, Japan, Apr. 2003.

[8] ——, "Blind separation of speech mixtures based on nonstationarity," in *Proceeding of ISSPA 2003 Conference*, Paris, France, July 2003.

[9] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," in *International Workshop on Independence & Artificial Neural Networks*, University of La Laguna, Tenerife, Spain, Feb. 1998.

[10] L. Trainor, R. Sonnadara, K. Wiklund, J. Bondy, S. Gupta, S. Becker, I. C. Bruce, and S. Haykin, "Development of a flexible, realistic hearing in noise test environment (R-HINT-E)," *Signal Processing*, vol. 84, pp. 299–309, 2004.

[11] H.-C. Wu and J. C. Principe, "Simultaneous diagonalization in the frequency domain (SDIF) for source separation," in *Proceeding of ICA 1999 Conference*, Aussois, France, Jan. 1999, pp. 245–250.

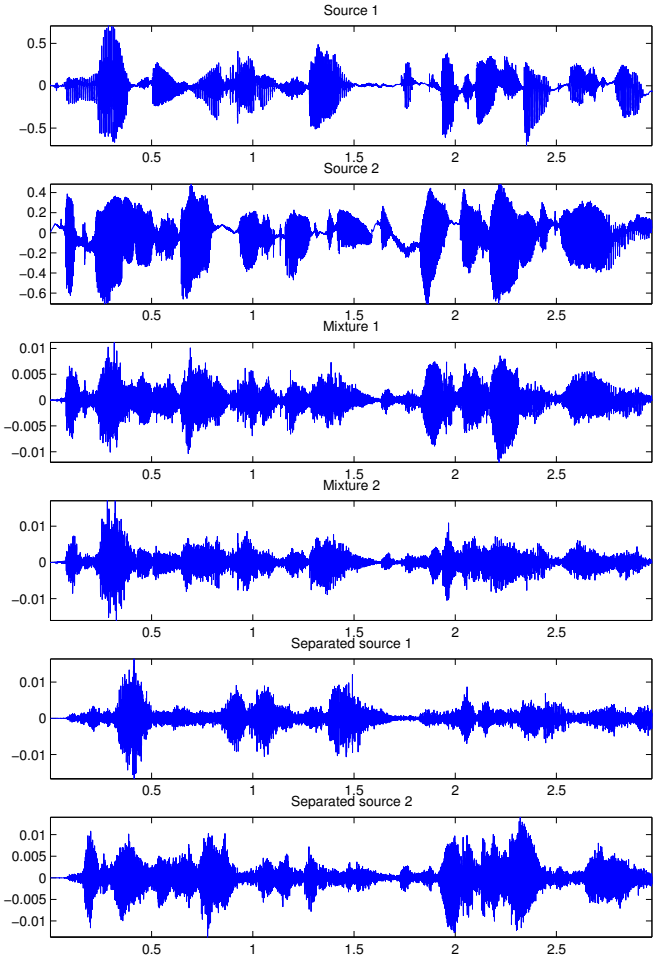

Figure 5: Sources, mixtures and estimated sources