A NEW METHOD FOR SEPARATION OF SPEECH SIGNALS IN CONVOLUTIVE MIXTURES

Mahmood Ferdosizadeh , Massoud Babaie-Zadeh, and Farrokh A.Marvasti

Advanced Communications Research Center(ACRI), Sharif University of Technology
Tehran, Iran
mferdosi@mehr.sharif.edu, mbzadeh@yahoo.com, marvasti@sharif.edu

ABSTRACT

In this paper, the performance of the gradient method based on Score Function Difference (SFD) in the separation of i.i.d. and periodic signals will be investigated. We will see that this algorithm will separate periodic signals better than i.i.d. ones. By using this experimental result and the fact that voiced frames of speech signals are approximately periodic, a modified algorithm named VDGaradient has been proposed for separation of speech signals in synthetic convolutive mixtures. In this method, voiced frames of speech signal will be used as the input to the gradient method, then the resulting separating system will be applied to separate sources completely.

1. INTRODUCTION

The concept of Blind Source Separation (BSS) or Independent Component Analysis (ICA) is to recover the source signals from their mixtures, assuming that there is no information about the sources and the mixing system. Suppose that there are N convolutive mixtures $x_1(n),x_2(n),\ldots,x_N(n)$ from N independent sources $s_1(n),s_2(n),\ldots,s_N(n)$ and $\mathbf{A} = [A_{ij}(z)]_{N\times N}$ is the mixing matrix that each of its entries is a filter. Then we can write:

$$x_i(n) = \sum_{j=1}^{N}\sum_{k=1}^{K} a_{ij}(k)s_j(n-k) \qquad (1)$$

where $a_{ij}(k)$ is the k-th coefficient of the mixing filter $A_{ij}(z)$ and K is the maximum length of the mixing filters. In matrix form (1) can be written as:

$$\mathbf{x}(n) = [\mathbf{A}(\mathbf{z})]\mathbf{s}(\mathbf{n}) \qquad (2)$$

where $\mathbf{x}(n) = (x_1(n),x_2(n),\ldots,x_N(n))^T$ and $\mathbf{s}(n) = (s_1(n),s_2(n),\ldots,s_N(n))^T$. The purpose of BSS is to find the separating system $\mathbf{B}$ such that each element of output vector $\mathbf{y}(n) = [\mathbf{B}(z)]\mathbf{x}(n)$ be a filtered version of one and only one of the sources.

In convolutive mixtures, it has been shown [1] that independence of output signals of the separating system is sufficient for separation. To produce independent outputs from mixtures, several methods have been proposed. Some of them are based on cancellation of cross-spectra [1], cancellation of second [2] or higher order cross-moment [3], and higher order cross-cumulants [3]. In some of the proposed methods, the mutual information is used as the independence criterion [4, 5, 6].

The output random variables $y_1,y_2,\ldots,y_N$ with probability density functions $p_{y_1}(y_1),p_{y_2}(y_2),\ldots,p_{y_N}(y_N)$ are independent if and only if $p_{\mathbf{y}}(\mathbf{y}) = \prod_{i=1}^{N} p_{y_i}(y_i)$, where $p_{\mathbf{y}}(\mathbf{y})$ is the joint probability density function of the random variables. Kullback-Leibler divergence can be used for the measurement of the distance between $p_{\mathbf{y}}(\mathbf{y})$ and $\prod_{i=1}^{N} p_{y_i}(y_i)$, which is called mutual information[7]:

$$I(\mathbf{y}) = \int p_{\mathbf{y}}(\mathbf{y}) \ln \frac{p_{\mathbf{y}}(\mathbf{y})}{\prod_{i=1}^{N} p_{y_i}(y_i)} d\mathbf{y} \qquad (3)$$

$I(\mathbf{y})$ is a non-negative function which is zero if and only if the random variables y_is are independent. Thus the BSS method can operate based on the minimization of the mutual information of the outputs. In the gradient method that was proposed in [4] the steepest descent algorithm has been used to minimize the mutual information of the outputs. In this method the coefficients of the separating filters are updated in the opposite direction of the gradient of mutual information. The gradient of the mutual information with respect to the parameters of the separating system has been derived using Score Function Difference (SFD) of the outputs[8]. SFD, which is the difference between Marginal Score Function (MSF) and Joint Score Function (JSF),is in fact a non-parametric gradient for mutual information[8]. The main advantage of the SFD based methods is that they can be extended for separation of more complicated mixtures [8], but they have low performance in separation of speech signals.

In this paper we will investigate the performance of the gradient method in the separation of stationary sources and non-stationary signals such as speech signals in synthetic convolutive mixtures. In section 2, we will see that the performance of the gradient method depends on the nature of the sources. For example, two types of the sources, i.i.d. and periodic, will be tested. In section 3, a modified gradient method will be proposed for separation of speech signals in synthetic convolutive mixtures. Finally, in section 4 experimental results will be explained and the performance of the traditional gradient method and its modified version named VD-gradient will be compared.

2. GRADIENT ALGORITHM IN SEPARATION OF STATIONARY SOURCES

In this section, we will focus on the separation of two types of stationary sources using the gradient method: i.i.d. and periodic sources. Figures 1 and 2 show 1500 samples of two i.i.d. and two periodic sources, respectively which have been used in simulations. In all of simulations of the paper,

This work has been partially funded by Sharif University of Technology, by French Embassy in Tehran, and by Center for International Research and Collaboration (ISMO).

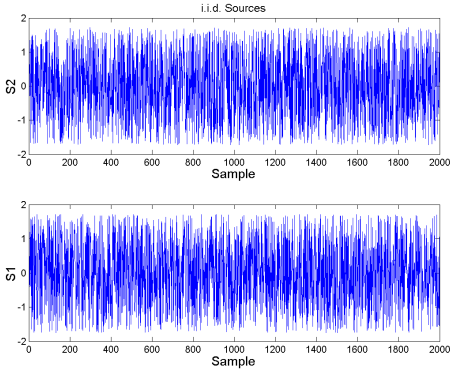

Figure 1: 1500 sample from two i.i.d. sources

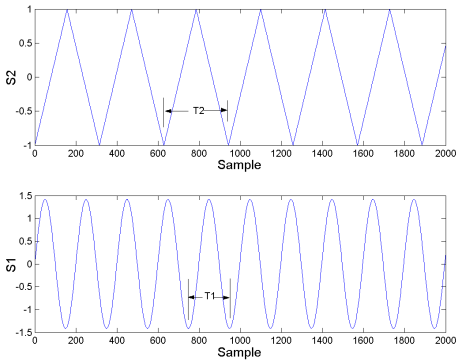

Figure 2: 1500 sample from two periodic sources, a sinuous and a sawtooth with periods T_1 and T_2

the mixing matrix is assumed to be:

$$A(z) = \begin{pmatrix} 1+0.2z^{-1}+0.1z^{-1} & 0.5+0.3z^{-1}+0.1z^{-1} \\ 0.5+0.3z^{-1}+0.1z^{-1} & 1+0.2z^{-1}+0.1z^{-1} \end{pmatrix} \quad (4)$$

and the learning rate is $\mu = 0.01$. To measure the separation performance, SNR of each output and the average SNR have been defined as:

$$SNR_1 = 10\log\frac{E\{[C_{11}(z)]s_1(n)+[C_{12}(z)]s_2(n)\}^2}{E\{[C_{12}(z)]s_2(n)\}^2}$$

$$SNR_2 = 10\log\frac{E\{[C_{21}(z)]s_1(n)+[C_{22}(z)]s_2(n)\}^2}{E\{[C_{12}(z)]s_2(n)\}^2} \quad (5)$$

$$SNR = \frac{SNR_1+SNR_2}{2} \quad (6)$$

where $\mathbf{C}(\omega) = \mathbf{A}(\omega)\mathbf{B}(\omega)$ is the overall frequency response matrix of the mixing-separating system. Note that in these equations the term $[C_{ij}(z)]s_j$ shows the effect of the source j on the output i.

As shown in Figure 3, in separation of periodic sources, when the ratio of the main frequencies of the sources is an integer, gradient method will diverge (which was expected, because in this case the sources are not independent). Also

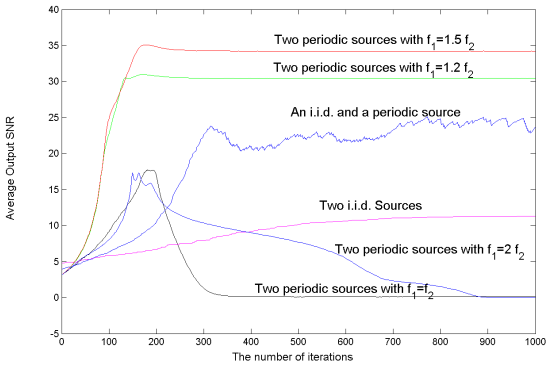

Figure 3: A comparison between the SNR of the gradient method in separation of two sources with 4 different cases of inputs: two i.i.d. sources, a periodic source and an i.i.d. one, two periodic sources with integer or non-integer ratio of the main frequencies f_1 and f_2

the gradient method in separation of two periodic sources with non-integer ratio of the main frequencies works better than that of two i.i.d. sources. In the next section we will use these results to propose a modified version of the gradient method for separation of speech signals.

3. VDGRADIENT ALGORITHM

The voiced frames of a speech signal are approximately periodic. From this fact and the results of the previous section we can say that if we choose a few number of samples from voiced frames of the mixture signals, then we hope that the gradient algorithm can use the information of these mixture samples to calculate the separating filters. Then we can use these filters to separate the sources completely.
Figure 4 shows a block diagram for this algorithm. In the Voiced detection(VD) block of this figure the instants in which both sources are voiced are detected using the information of the mixtures. Then these parts of the mixtures are buffered as the input to the gradient algorithm. After convergence of the gradient method, the resulting filter coefficients are copied and applied to the whole of mixture samples to separate the sources completely.
Choosing the voiced parts of the sources we get other benefits because the power of the sources in voiced frames is greater than that of unvoiced frames, and hence:

- We can hope that the effect of the sources in the mixture signals are not too different.
- Detection of these frames in the mixture signals is easy (it will be discussed in the next section).
- Low power noises can not make the voiced detection false alarm in the VD block.

3.1 VD block

The duty of the VD block is to detect the voiced-voiced situation of the sources from their mixtures. Note that the problem is different from the traditional voiced detection problem in speech signals. In this case the source signals are not available but we have two convolutive mixtures. To design this

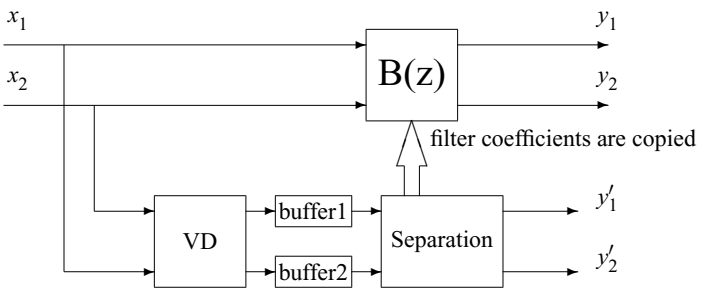

Figure 4: VD-Gradient

block, we can use the fact that in the frequency domain the energy of the periodic signals are concentrated about some frequencies. Thus we can detect voiced frames of the source signals from the peaks in the spectral density of the mixtures. As shown in Figure 5, the process of the proposed voiced frame detection is as follow:

- A predefined number of samples from mixtures are buffered.
- A threshold equal to a fixed ratio of the buffered mixture power is set.
- The Power Spectrum Density (PSD) of the mixtures is calculated.
- Frequency components of the mixture PSD where their power is less than the threshold are ignored.
- The power of the remaining PSD is calculated.
- If the average of the calculated power of the mixtures (θ) is greater than a decision threshold, the buffered frames are detected as voiced frames.

In fact when the sources are unvoiced then the mixture signals have approximately flat distribution in the frequency domain. Thus if we choose a threshold several times greater than the power of the buffered samples then the major part of the energy of the mixtures will be removed. But when the sources are voiced then the major part of the energy is concentrated in the harmonic frequencies, thus thresholding can not remove the energy of these harmonics.

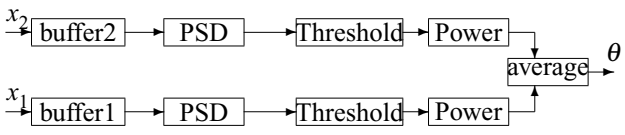

Figure 5: VD-block

Figure 6 shows the value of the parameter θ for different cases of the source signals. For each case, four window (indexed from 1 to 4) from different parts of the mixtures are selected and the horizontal axis shows the index of these windows. As we can see in the case where both sources are voiced, the value of the parameter θ is the greatest. But when one of the sources is voiced and another unvoiced, the VD-block may have a false alarm. This is due to the energy of unvoiced frames, which is very low and the major power of the mixtures is from the voiced source. To overcome this problem, we need an algorithm to address the following question: Have both sources nearly the same power contribution in the buffered mixtures? This question has been answered in [9]. In that paper a parameter (ω) has been introduced as follow:

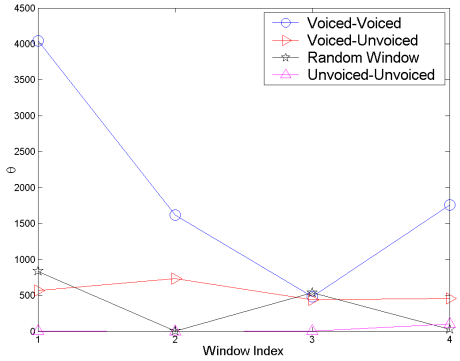

Figure 6: The value of the parameter θ for different cases of the buffered mixture samples. The threshold is chosen 20 times greater than the average power of the buffered samples

$$(\omega) = \frac{|S_{x_1x_2}(\omega)|^2}{S_{x_1x_1}(\omega).S_{x_2x_2}(\omega)} \tag{7}$$

where $S_{x_1x_1}(\omega)$ and $S_{x_2x_2}(\omega)$ are the spectral densities of the mixtures x_1 and x_2 and $S_{x_1x_2}(\omega)$ is their cross spectral density. In [9] it has been shown that the parameter (ω) is near 1 if one of the sources is in the silent or low power mode. Thus we define a new parameter named ζ as follow:

$$\zeta = \sum_{\omega \le \omega_0} 1 - (\omega) \tag{8}$$

ω_0 is a frequency band that contains most of the power of the voiced frames. Figure 7 shows the value of the parameter ζ for two cases of the sources, voiced-voiced and voiced-unvoiced. We see that in the case of voiced-voiced, the value of this parameter is much higher than the other cases. From Figures 6 and 7 it can be seen that $\theta > \theta_T$ and $\zeta > \zeta_T$ is a good criterion for the VD decision where the threshold levels $\theta > \theta_T$ and $\zeta > \zeta_T$ are chosen appropriately.

4. SIMULATION RESULTS

To show the performance of the proposed method we have selected 1500 samples (the sampling frequency is 16KHz) from different parts of the mixture signals and then the gradient algorithm was applied to this buffered samples. Three different cases were examined:

- Case 1: A section of the mixture signals in which the both sources are voiced.
- Case 2: A section of the mixture signals in which the both sources are unvoiced.
- Case 3: A section of the mixture signals that was chosen randomly.

Note that in case 1 the appropriate samples were selected automatically (by VD block), while in the other cases they were

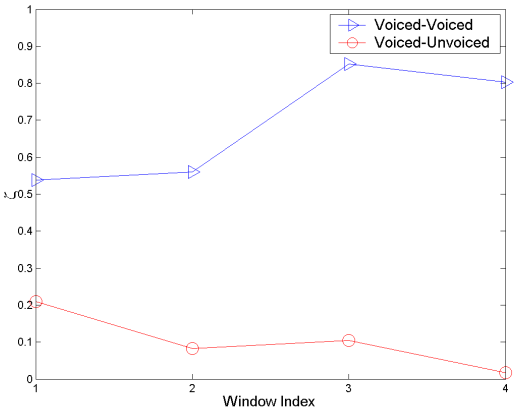

Figure 7: The value of the parameter ζ in two cases of the sources: Voiced-Voiced and Voiced-Unvoiced

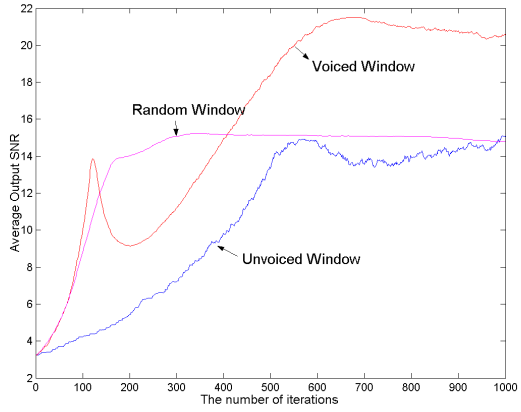

Figure 8: performance of the gradient method on different cases of the two speech sources: Voiced-Voiced, Voiced-Unvoiced, Unvoiced-Unvoiced and random selection of buffered samples (average of 20 random selection).

selected manually. Figure 8 shows the simulation results for these cases. For case 3, the sketched curve is the SNR averaged over 10 randomly selected windows and it can be interpreted as the performance of the ordinary gradient method for separation of speech signals, when there is no strategy in selection of mixture samples. Note that although in the block named as "separation" in Figure 4 we use only 1500 samples of the mixtures, the SNR is calculated using y_1 and y_2 instead of y_1' and y_2'. We see that the voiced-voiced case has the best performance, thus we see that using VDGradient instead of the ordinary gradient method improves the SNR by approximately 7dB.

5. CONCLUSION

In this paper we investigated the performance of the gradient algorithm for separation of speech signals. The result shows that this algorithm converges when we apply mixtures from stationary parts of the sources and specially when both sources are voiced. Then we proposed an algorithm named VDGradient. In this algorithm, a VD block selects a section of mixtures in which the sources are voiced. Then this part of the mixtures is applied for the calculation of separation filters. Also a method based on the periodic nature of the voiced frames was proposed for design of the VD block.

REFERENCES

[1] D.Yellin and E. Weinstein,"Criteria for multichannel signal separation," *IEEE Trans. Signal Processing*, pp.2158–2168, August 1994.

[2] U.A.Lindgren and H. Broman, "Source separation using a criterion based on second-order statistics", *IEEE Trans. Signal Processing*, pp.1837–1850, July 1998.

[3] H.L. Nguyen Thi and C. Jutten, "Blind Sources Separation For Convolutive mixtures", *Signal Processing*, Vol.45, pp.209–229, 1995.

[4] M.Babaie-Zadeh, C.Jutten, and K. Nayebi, "Separating convolutive mixtures by mutual information minimization", in *Proc. IWANN2001*, Granada, Spain,Juin 2001,834–842.

[5] A.Taleb and C.Jutten, "Entropy optimization, application to blind source separation",*ICANN*,Lausanne, Switzeland,October 1997,529-534.

[6] M.Babaie-Zadeh and C.Jutten and K.Nayebi, Using Joint Score Functions in Separating Post Non-Linear Mixtures,in *Scientia-Iranica 9(4)*,pp.409–418,2002.

[7] T. M. Cover and J. A. Thomas, "Elements of Information Theory",Wiely Series in Telecommunications,1991.

[8] M. Babaie-Zadeh and C. Jutten,"A general approach for mutual information minimization and its application to Blind Source Separation",*Signal Processing*, Vol.85,No.5, pp.975–995, 2005.

[9] Y.Deville and B.Albouy, "Alternative Structures and Power Spectrum Criteria for Blind Segmentation and Separation of Convolutive Speech Mixtures", *fourth International Conference on Independent Component Analysis and Blind Source Separation (ICA2003)*, Nara, Japan,April 2003,361-366.