잡음 환경하에서의 음성 분리

장인선, 최승진
포항공과 대학교 컴퓨터 공학과

Convolutive source separation in noisy environments

Inseon Jang, Seungjin Choi
Department of Computer Science and Engineering, POSTECH

jinsn@postech.ac.kr, seungjin@postech.ac.kr

Abstract

This paper addresses a method of convolutive source separation that based on SEONS (Second Order Nonstationary Source Separation) [1] that was originally developed for blind separation of instantaneous mixtures using nonstationarity. In order to tackle this problem, we transform the convolutive BSS problem into multiple short-term instantaneous problems in the frequency domain and separated the instantaneous mixtures in every frequency bin. Moreover, we also employ a H infinity filtering technique in order to reduce the sensor noise effect. Numerical experiments are provided to demonstrate the effectiveness of the proposed approach and compare its performances with existing methods.

I . Introduction

Convolutive source separation is a fundamental problem which plays a critical role in cocktail party speech recognition. The goal of convolutive source separation is to restore original unknown sources from their mixtures where sources are convolved with an unknown multivariate linear time invariant FIR filter which reflects the characteristics of propagation media. In this paper, we trasform the convolutive BSS problem into multiple short-term instantaneous problems in the frequency domain and separating the instantaneous mixtures in every frequency bin in order to tackle this problem [2], [3]. We propose a method based on SEONS (Second Order Nonstationary Source Separation) [1] that was originally developed for blind separation of instantaneous mixtures using nonstationarity. This method is insensitive to the temporally white noise since it is based on only time-delayed cross-spectral density matrix.

We also employ a H infinity filtering technique in order to reduce the sensor noise effect. Since the design criterion of the H infinity filtering technique is based on the worst case disturbances (modeling errors of state space model and additive noises), the model is less sensitive to uncertainty in the exogenous signal statistics and system model dynamics. And H infinity filtering technique requires no a priori knowledge of the noise statistics (the only assumption is that the noise signals have a finite energy) and is known that it performs better than traditional methods (for example, Wiener filter and Kalman filter) [4].

II . Problem Formulation in Frequency

Let $\boldsymbol{x}(n)=[x_1(n),\cdots,x_M(n)]^T$ be a vector whose elements $\{x_i(n)\}$ are the signals measured at an array of microphones. Denotes the speech source vector by $\boldsymbol{s}(n)=[s_1(n),\cdots,s_N(n)]^T$ (where $M\geq N$). Taking the delay and multipath effect into account, the observation data $\boldsymbol{x}(n)$ is modeled by

$$x_i(n) = \sum_{j=1}^{N} \sum_{\tau=0}^{P} h_{ij}(\tau) s_j(n-\tau) + v_i(n), \quad (1)$$

where $\{h_{ij}(n)\}$ is the room impulse response between the jth speech source and the ith microphone and $v_i(n)$ is the additional sensor noise.

The problem is to reconstruct the unknown sources from the sensor data by assuming statistically independent sources without any other prior knowledge about the sources and the mixing process. The frequency domain approach to blind source separation of convolutive mixtures is to transform the problem into multiple short-term instantaneous BSS problems in the frequency domain and the independent component analysis (ICA) is applied to the instantaneous mixtures in every frequency bin. Using a T-point windowed discrete Fourier transformation (DFT), time-domain signals x_j can be converted into frequency-domain time-series signals $X_j(\omega, n)$

$$X_i(\omega, n) = \sum_{\tau=0}^{T-1} x_i(n+\tau) w(\tau) e^{-j2\pi\omega\tau}, \quad (2)$$

where $w(\tau)$ denotes a window function and T is the frame size of the DFT. We apply corresponding expressings to $H_{ij}(\omega)$, $S_i(\omega, n)$ and $V_i(\omega, n)$, in which $H_{ij}(\omega)$ does not depend on the discrete time index n due to the assumption that the mixing system is time invariant and the same assumption will be applied to the separation system as below. As shown in [2], a linear convolution can be approximated by a circular convolution if $P \ll T$, that is

$$\boldsymbol{X}(\omega, n) = \boldsymbol{H}(\omega)\, \boldsymbol{S}(\omega, n) + \boldsymbol{V}(\omega, n), \quad (3)$$

where $\boldsymbol{S}(\omega, n) = [S_1(\omega, n), \cdots, S_N(\omega, n)]^T$ and $\boldsymbol{X}(\omega, n) = [X_1(\omega, n), \cdots, X_M(\omega, n)]^T$ are the time-frequency representation of the source signals and the observed signals respectively.

Similar to some time domain methods for instantaneous mixtures, an alternative method for convolutive blind separation in the frequency domain is to estimate a backward model in every frequency bin ω,

$$\boldsymbol{Y}(\omega, n) = \boldsymbol{W}(\omega)\, \boldsymbol{X}(\omega, n), \quad (4)$$

where $\boldsymbol{Y}(\omega, n) = [Y_1(\omega, n), \cdots, Y_N(\omega, n)]^T$ is the time-frequency representations of the estimated source signals. The parameters in $\boldsymbol{W}(\omega)$ are determined so that the elements $Y_1(\omega, n), \cdots, Y_N(\omega, n)$ become mutually independent. Uisng a inverse T-point windowed discrete Fourier transformation (IDFT), frequency-domain demixing filter $\boldsymbol{W}(\omega)$ can be converted into time-domain demixing filter $\boldsymbol{W}(n)$. When we try to combine the results for the individual frequency bins in the time domain, the permutation problem occurs becasue fo the inherent permutation ambiguity in the rows of $\boldsymbol{W}(\omega)$. Hence To solve this problem, we use existing method which is constraints on the filter models in the frequency domain.

1) Apply the Robust Whitening method to (5):

(i) We partition the observed data into several nonoverlapping blocks,

(ii) Compute

$$\boldsymbol{M}_x^p(\omega, k) = \frac{1}{2}[\boldsymbol{R}_x^p(\omega, k) + \{\boldsymbol{R}_x^p(\omega, k)\}^H],$$

for some time lag τ_p in every bolck.

(iii) Calculate $\boldsymbol{C}(\omega) = \sum_{i=1}^{J} \alpha_i \boldsymbol{M}_x^p(\omega, k)$ by the FSGC (Finite Step Global Convergence) method and perform an eigenvalue decomposition of $\boldsymbol{C}(\omega)$

(iv) The robust whitening matrix is

$$\boldsymbol{Q}(\omega) = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\omega)\, \boldsymbol{U}^H(\omega)$$

where $\boldsymbol{U}(\omega)$ contains the eigenvectors associated with N principal singular values $\boldsymbol{\Sigma}(\omega)$.

2) Calculate

$$\boldsymbol{M}_z^p(\omega, k) = \boldsymbol{Q}(\omega)\, \boldsymbol{M}_x^p(\omega, k)\, \boldsymbol{Q}^H(\omega)$$

for $k = 1, \cdots, K$ and $p = 1, \cdots, J$.

3) Find a unitary joint diagonalizer $\boldsymbol{V}(\omega)$ of $\{\boldsymbol{M}_z^p(\omega, k)\}$ using the joint approximate diagonalization method, which satisfied

$$\boldsymbol{V}(\omega)\, \boldsymbol{M}_z^p(\omega, k)\, \boldsymbol{V}(\omega) = \boldsymbol{\Lambda}_{k,p}$$

where $\boldsymbol{\Lambda}_{k,p}$ is a set of diagonal matrices.

4) The demixing filter is computed as $\boldsymbol{W}(\omega) = \boldsymbol{V}^H(\omega)\, \boldsymbol{Q}(\omega)$.

Table 1 SEONS for convolved mixtures

Ⅲ. SEONS in Frequency domain

It has been shown in [2], [3] that there is a set of second-order conditions that can be specified to perform blind separation for nonstationary signals. Hence we propose a method based on SEONS [1] that was originally developed for blind separation of instantaneous mixtures using nonstationarity.

The separation criteria is based on minimizing the off-diagonal elements of cross-spectral density matrices of the observed signal over K epochs with some time delay. The signal is assumed to be stationary over each epoch.

To estimate the cross-spectral density matrices of the observed signal, we first divide the input sequence into K epochs and we use the following formula to estimate the time-delayed cross-spectral density matrix at the kth epoch

$$\boldsymbol{R}_x^p(\omega,k)=\frac{1}{N_s}\sum_{i=0}^{N_s-1}\boldsymbol{X}(\omega,k,i)\{\boldsymbol{X}^p(\omega,k,i)\}^H \tag{5}$$

where

$$\boldsymbol{X}(\omega,k,i)=\sum_{n=-\infty}^{\infty}\boldsymbol{x}(n)\,\boldsymbol{w}(n-iT_s-kT_b)e^{-j2\pi n},$$

$$\boldsymbol{X}^p(\omega,k,i)=\sum_{n=-\infty}^{\infty}\boldsymbol{x}(n)\,\boldsymbol{w}(n-iT_s-kT_b-\tau_p)e^{-j2\pi n}$$

for $p=1,\cdots,J$. And k is the epoch index, N_s is the number of overlapping windows inside each epoch, T_s is the time shift between two overlapping windows, T_b is the size of each epoch, $w(n)$ is the windowing sequence and τ_p is time delay and H represents the conjugate transpose operation.

Also the time-delayed cross-spectral density matrix of the observed signals $\boldsymbol{x}(n)$ at kth epoch is given by

$$\boldsymbol{R}_x^p(\omega,k)=\boldsymbol{H}(\omega)\,\boldsymbol{R}_s^p(\omega,k)\,\boldsymbol{H}^H(\omega) \tag{6}$$

where $\boldsymbol{R}_x^p(\omega,k)$ is the cross-spectral density matrix ofthe sources, which by assumption is diagonal for all $\omega\in[0,2\pi)$ and $k=1,\cdots,K$.

We can estimate whitening matrix $\boldsymbol{Q}(\omega)$ by Robust Whitening method. Applying $\boldsymbol{Q}(\omega)$ to $\boldsymbol{R}_x^p(\omega,k)$, we can write

$$\boldsymbol{R}_z^p(\omega,k)=\boldsymbol{Q}(\omega)\,\boldsymbol{R}_x^p(\omega,k)\,\boldsymbol{Q}^H(\omega) \tag{7}$$

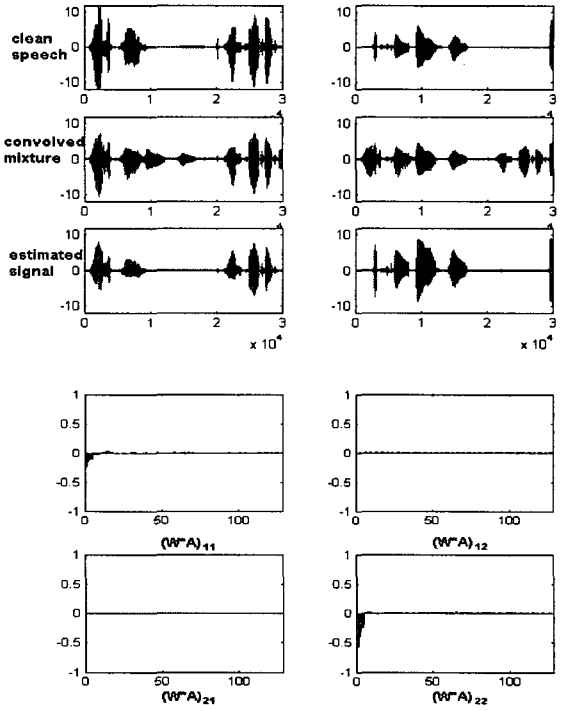

Figure 1 Results of separation (in no noisy case) (top) The convolved mixing and demixing system (in 10 dB AWGN case) (bottom). The delta-like response indicates successful blind demixing.

The unitary diagonalizer $\boldsymbol{V}(\omega)$ can be estimated from the orthogonal joint diagonalization of $\boldsymbol{R}_z^p(\omega,k)$ for all $k=1,\cdots,K,\ p=1,\cdots,J$.

In Table 1, there is the SEONS algorithm for convolved mixtures.

Ⅳ. Experiments

In this section, we demonstrate the performance of the proposed method by simulation, when applied to arificially mixed signals.

A system with two inputs and two outputs is considered, that is $N=M=2$. The source signals are the clean speech samples in the ST-NB 95 database. Both of the signals are sampled at 8kHz with a duration of 3.75s seconds. We artificially mix the two sources by the system to be order of $P=4$ as follows

$$H_{11}(z^{-1})=1+0.6703z^{-1}+0.5488z^{-2}$$
$$-0.4493^{-2}+0.3679z^{-4},$$

$$H_{12}(z^{-1}) = 0.6 - 0.4022z^{-1} + 0.3293z^{-2} - 0.2696z^{-3} + 0.2207z^{-4},$$

$$H_{21}(z^{-1}) = 0.5 + 0.3352z^{-1} + 0.2744z^{-2} - 0.2247z^{-3} + 0.1839z^{-4},$$

$$H_{22}(z^{-1}) = 1 - 0.5742z^{-1} + 0.4863z^{-2} - 0.3965z^{-3} + 0.2649z^{-4}.$$

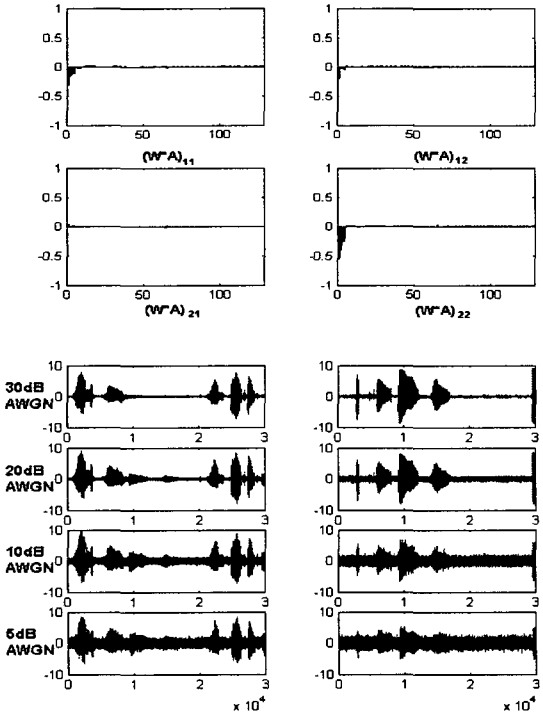

Figure 2 Results of separation (in a varity of noisy case) (top), the convolved mixing and demixing system (in 10 dB AWGN case) (bottom)

Through the simulation, the parameters are set to be $T=512$, $K=5$, $J=3$, $T_s=4$. Firstly, we estimate source signal from convolved mixture in no noisy case. In Fig. 1, we can see that the proposed method show quite a good separation performance.

Secondly, we estimate source signal form convolved mixture with a varity of AWGNs. In Fig.2, we can also see that the proposed method shows a good separation performance although there is an additive white Gaussian noise (AWGN). But there is still noise effect and reverberant effect. In order to reduce the noise effect, we employ a H infinity filtering technique in order to reduce the sensor noise effect. In Fig.3, we can see that the noise effect is reduced as compared to Fig. 2.

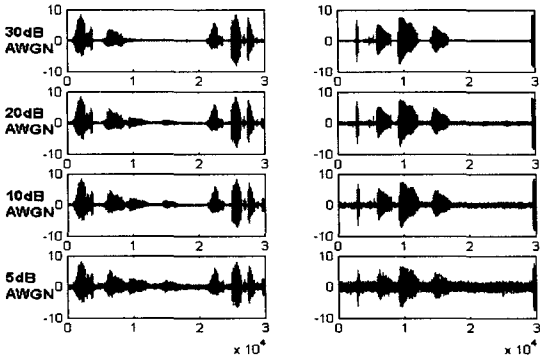

Figure 3 Results of separation and denosing (a varity of noisy case with H infinity filter)

V. Conclusion

In this paper, we proposed a method based on SEONS in order to restore original unknown sources from their convolved mixtures. SEONS is originally developed for blind separation of instantaneous mixtures using nonstationarity. We trasformed the convolutive BSS problem into multiple short-term instantaneous problems in the frequency domain and separated the instantaneous mixtures in every frequency bin by using SEONS. Moreover, in experiments, we employed a H infinity filtering technique in order to reduce the sensor noise effect. Simulations verified the high performance of proposed method.

References

[1] S. Choi, A. Cichocki and A. Belouchrani, "Second order nonstationary source separation," *Journal of VLSI Signal Processing*, vol. 32, no.1-2, pp. 93~104, Aug. 2002.

[2] 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.

[3] K. Rahbar and J. P. Reilly, "Blind source separation algorithm for MIMO convolutive mixtures", in *Proc. ICA2001*, Dec. 9-13, 2001.

[4] X. Shen and L. Deng, "A dynamic system approach to speech enhancement using the H infinity filtering algorithm", *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 4, pp. 391~399, July 1999.