Blind Source Separation Using Time-Frequency Masking

Abbas MOHAMMED[1], *Tarig BALLAL*[2], *Nedelko GRBIC*[1]

[1] Dept. of Signal Processing, Blekinge Institute of Technology, Box 520, 37225 Ronneby, Sweden
[2] School of Computer Science, College of Engineering, University College Dublin, Ireland

amo@bth.se, tarigb@yahoo.com, ngr@bth.se

Abstract. *In blind source separation (BSS), multiple mixtures acquired by an array of sensors are processed in order to recover the initial multiple source signals. While a variety of Independent Component Analysis (ICA)-based techniques are being used, in this paper we used a newly proposed method: The Degenerate Unmixing and Estimation Technique (DUET). The method applies when sources are W-disjoint orthogonal; that is, when the time-frequency representations, of any two signals in the mixtures are disjoint sets. The method uses an online algorithm to perform gradient search for the mixing parameters, and simultaneously construct binary time-frequency masks that are used to partition one of the mixtures to recover the original source signals. Previous studies have demonstrated the robustness of the method. However, the investigation in this paper reveals significant drawbacks associated with the technique which should be addressed in the future.*

Keywords

Blind source separation, DUET, Time-frequency masking.

1. Introduction

The goal of blind source separation (BSS) is to recover a set of *unobserved* signals or "sources" from a set of *observed* mixtures. Typically the observations are obtained at the output of a set of sensors, where each sensor receives a different combination of the source signals. The adjective "blind" stresses two facts [1]:

1) The source signals are *not* observed.

2) No information is available about the mixing system.

The lack of prior knowledge about the source signals and the mixing system is always compensated by assumptions that should be met by the unknown sources [1]. Some common assumptions are that the sources are statistically independent [2], are statistically orthogonal [3], are non-stationary [4], or can be generated by finite dimensional model spaces [5].

This paper investigates the Degenerate Unmixing and Estimation Technique (DUET), a method that applies when sources are W-disjoint orthogonal; that is, when the time-frequency representations, of any two signals in the mixtures are disjoint sets. The method uses an online algorithm to perform gradient search for the mixing parameters, and simultaneously construct binary time-frequency masks that are used to partition one of the mixtures to recover the original source signals. Exploiting the W-disjoint orthogonality property, the method requires only two mixtures to separate an arbitrary number of sources.

Previous publications (e.g. [6], [7]) have demonstrated the robustness of the method even when tested with real data; up to 19 dB SIR (signal to interference ratio) gain has been achieved with instantaneous mixtures, up to 5 dB with echoic real mixtures, and separation of 3 sources using only 2 mixtures was realized emphasizing the main advantage of the method.

In this paper we investigate the method in more details. Our investigation reveals significant drawbacks, consideration of which is important for improving the DUET method in the future. The slowness of convergence and presence of artifacts constitute two of these drawbacks. Additionally, the presence of white noise was seen to violate the basic assumption which leads to failure when dealing with noisy mixtures.

The organization of this paper is as follows. Section 2 defines the source assumption. Section 3 shows how the source assumption is used to define the signal model. In Section 4, we present a method for estimating the mixing parameters. A summary of the algorithm is given in Section 5. In Section 6, we describe the demixing process. Section 7 presents and discusses the results. Finally, Section 8 concludes the paper.

2. Source Assumptions

The main assumption for DUET is that, the time-frequency representations of the source signals contained in a mixture should be *disjoint* (or non-overlapping). This condition generated a concept which is referred to as the *W-disjoint orthogonality* [3], [2]. Given a windowing function $W(t)$, two signals $s_i(t)$ and $s_i(t)$ are said to be W-Disjoint

Orthogonal (W-DO) if the supports of the short-time Fourier transforms (STFTs) of $s_i(t)$ and $s_i(t)$ are disjoint [6], [8].

The STFT of $s_i(t)$ is defined as [9]

$$S_j(\omega,\tau)=\int_{-\infty}^{\infty} s_j(t)\,w(t-\tau)\,e^{-i\omega t}\,dt\,. \tag{1}$$

The support of $S_j(\omega,\tau)$ is denoted as the set of the (ω,τ) pairs for which $S_j(\omega,\tau)\neq 0$.

Since the W-disjoint orthogonality assumption is not exactly satisfied for many categories of signals, the concept of approximate W-disjoint orthogonality introduced in [8] provides a practical version for the basic assumption. Approximate W-disjoint orthogonality assumes that at each point of the time-frequency representation of a mixture, the power of, at most, one source signal will be *dominant*. In other words, the assumption that sources other than the active (or the dominant) source has *zero* energy is replaced by the assumption that these sources have *relatively low* energy compared to the dominant source.

With such an assumption, and if the set of time-frequency points where one source dominates all the other sources is sufficient to represent the dominant source, masking the remaining time-frequency points (points where the dominant source has relatively low energy) suggests a good method for extracting the dominant source one of the mixtures.

3. Signal Model

Let's assume that we have N sources, s_j, $j\in\{1,2,\ldots,N\}$, that arrive at two sensors to compose two mixtures, $x_1(t)$ and $x_2(t)$ defined as

$$x_1(t)=\sum_{j=1}^{N} a_{j1}\,s_j(t-\delta_{j1}) \tag{2}$$

$$x_2(t)=\sum_{j=1}^{N} a_{j2}\,s_j(t-\delta_{j2}) \tag{3}$$

where a_{j1} and a_{j2} are attenuation factors corresponding to paths between source j and sensor 1 and 2, δ_{j1} and δ_{j2} are delays corresponding to paths between source j and sensor 1 and 2. In this case we can absorb the attenuation and delay parameters of the first mixture $x_1(t)$ into the definition of the sources, (2) and (3) can then be expressed as

$$x_1(t)=\sum_{j=1}^{N} s_j(t), \tag{4}$$

$$x_2(t)=\sum_{j=1}^{N} a_j\,s_j(t-\delta_j) \tag{5}$$

where the relative parameters $a_j = a_{j2}/a_{j1}$ and $\delta_j = \delta_{j2} - \delta_{j1}$.

We will refer to (a_j,δ_j) as *attenuation-delay parameters*, or simply, *mixing parameters*.

In time-frequency domain, (4) and (5) can be written as

$$\begin{bmatrix} X_1(\omega,\tau)\\ X_2(\omega,\tau)\end{bmatrix}=\begin{bmatrix} 1 & \cdots & 1\\ a_1e^{-i\omega\delta_1} & \cdots & a_Ne^{-i\omega\delta_N}\end{bmatrix}\begin{bmatrix} S_1(\omega,\tau)\\ \vdots\\ S_N(\omega,\tau)\end{bmatrix}. \tag{6}$$

4. Attenuation-Delay Estimation

For W-disjoint orthogonal sources, it is noticed that at most one of the N sources will be non-zero for a given time-frequency point. Therefore, (6) can be written as

$$\begin{bmatrix} X_1(\omega,\tau)\\ X_2(\omega,\tau)\end{bmatrix}=\begin{bmatrix} 1\\ a_je^{-i\omega\delta_j}\end{bmatrix}S_j(\omega,\tau) \tag{7}$$

where j corresponds to the active source. Solving (7) gives

$$a_j e^{-i\omega\delta_j}X_1(\omega,\tau)-X_2(\omega,\tau)=0\,. \tag{8}$$

Defining

$$\rho_j=\rho(a_j,\delta_j,\omega,\tau)=\frac{1}{1+a_j^2}\left|a_je^{-i\omega\delta_j}X_1(\omega,\tau)-X_2(\omega,\tau)\right|^2 \tag{9}$$

ρ_j should be zero if $s_j(t)$ is the active source at point (ω,τ).

It is noticed that at least one ρ must be zero, which implies that the minimum value of the set $\{\rho_1,\ldots,\rho_N\}$ is always zero. We define

$$J(\tau)=\sum_{\omega}\min(\rho_1,\ldots,\rho_N)\,. \tag{10}$$

As proved in [8], minimizing $J(\tau)$ is equivalent to maximizing the log-likehood of the mixing parameters estimates. Again as in [8], (10) can be approximated as

$$J(\tau)=\sum_{\omega}-\frac{1}{\lambda}\ln(e^{-\lambda\rho_1}+\ .....\ +e^{-\lambda\rho_N}) \tag{11}$$

where λ is a smoothing parameter, which has partial derivatives,

$$\frac{\partial J(\tau)}{\partial\delta_j}=\sum_{\omega}\frac{e^{-\lambda\rho_j}}{\sum_{k=1}^{N}e^{-\lambda\rho_k}}\frac{-2\omega a_j}{1+a_j^2}\,\mathrm{Im}(X_1(\omega,\tau)\,\overline{X_2(\omega,\tau)}\,e^{-i\omega\delta_j}) \tag{12}$$

$$\frac{\partial J(\tau)}{\partial a_j}=\sum_{\omega}\frac{e^{-\lambda\rho_j}}{\sum_{k=1}^{N}e^{-\lambda\rho_k}}\frac{2}{(1+a_j^2)^2}((a_j^2-1)\,\mathrm{Re}(X_1(\omega,\tau)\,\overline{X_2(\omega,\tau)}\,e^{-i\omega\delta_j} + a_j(\,|X_1(\omega,\tau)|^2-|X_2(\omega,\tau)|^2)) \tag{13}$$

where Im( ) and Re( ) are the imaginary and real parts of a complex value, and | | is the complex magnitude.

5. Algorithm

The complete algorithm that is used for learning the mixing parameters is summarized as follows:

- Initialize the amplitude-delay estimates $(\hat{a}_j(k), \hat{\delta}_j(k))$ to random values, where k is a time index, $j \in \{1,2,\ldots,N\}$, N is the number of sources which is assumed to be known.
- Calculate ρ_j, $\forall j$ using $(\hat{a}_j(k), \hat{\delta}_j(k))$ and (9).
- Calculate the gradients from (12) and (13).
- Update the mixing parameters estimates according to

$$\hat{a}_j(k) = \hat{a}_j(k-1) - \mu \frac{\partial J(\tau)}{\partial a_j} \tag{14}$$

$$\hat{\delta}_j(k) = \hat{\delta}_j(k-1) - \mu \frac{\partial J(\tau)}{\partial \delta_j} \tag{15}$$

where μ is a learning constant.

6. Demixing

The ρ estimates can be used to construct binary time-frequency masks. We use the following equation to calculate the elements of a mask:

$$\Omega_j(\omega,\tau) = \begin{cases} 1 & \rho(\hat{a}_j, \hat{\delta}_j, \omega, \tau) \le \rho(\hat{a}_m, \hat{\delta}_m, \omega, \tau), \forall m \neq j \\ 0 & otherwise \end{cases} \tag{16}$$

As illustrated in Fig. 1, the time-frequency representation of one source can be obtained using

$$S_j(\omega,\tau) = \Omega_j(\omega,\tau)\, X_1(\omega,\tau). \tag{17}$$

Finally, using the inverse transform the original sources can be recovered.

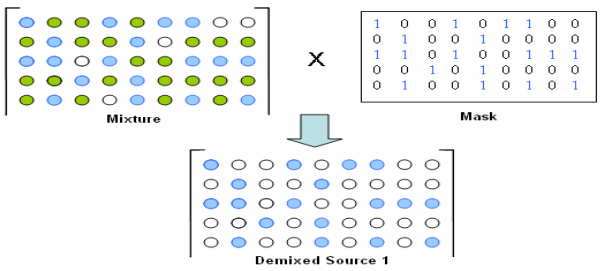

Fig. 1. Using binary time-frequency masking to recover the time-frequency representation of one source from the time- frequency representation of one mixture. At blue points source 1 is active, at green points source 2 is active, at white points no source is active. Note that the green points are completely masked.

7. Results

First we tested the *approximate* W-disjoint orthogonality of the source (speech) signals. For measuring the approximate W-disjoint orthogonality of speech signals, we used the measure introduced in [8]. Fig. 2 shows the values of the approximate W-disjoint orthogonality for one speech source for different threshold levels. For threshold x, a source is assumed to be the dominant (or approximately the only active) source if it is x dBs above the other (assumed non-active) source. The percentage value of the W-Disjoint Orthogonality (W-DO %) is the percentage of energy of the source that are contribution of the time-frequency points where it dominates the other source by x dBs. Fig. 1 clearly reflects the fact that speech sources are sufficiently W-DO for large range of thresholds.

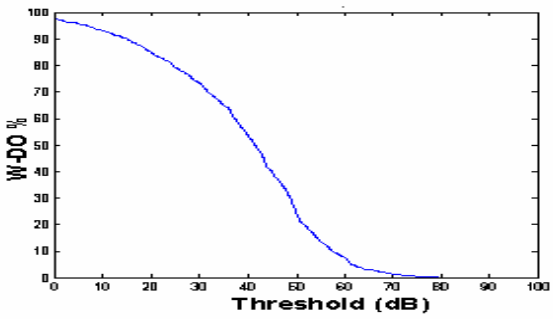

Fig. 2. Measuring the approximate W-Disjoint Orthogonality (W-DO) of a sample speech source. On the horizontal axis there are the values of different thresholds used in the measurement. The W-DO % values are the percentage of energy of the source that are contribution of the time-frequency points where it dominates the other source by a number of dBs equal to the value of the threshold. The figure shows that speech signals are sufficiently W-DO at different thresholds.

The algorithm was tested using both instantaneous mixtures and echoic mixtures. Up to 19 dB SIR (signal to interference ratio) gain was achieved with instantaneous mixtures, up to 5 dB with echoic real mixtures, and separation of 3 sources using only 2 mixtures was realized emphasizing the main advantage of the method. For the two-from-two case, we have found that normally more than 90% of the energy of each source is recoverable. Tab. 1 shows the (original) source recovery ratios and also the contribution of the original sources in the each output from a sample test.

	Input Source 1	Input Source 2
Output Source 1	92.1%	6.1%
Output Source2	7.9%	93.9%
	100.0%	100.0%

Tab. 1. The source recovery ratios and source contribution ratios in BSS of 2 sources from 2 mixtures using DUET. The first column shows that the energy from input source 1 is divided among two different output sources. The first row shows the percentage contributions of energy from input sources that are recovered in output source 1. The two less values are considered as interference. The table also indicates that the DUET performs BSS by portioning of mixture energy.

We measured the time required by the algorithm to reach the average value of the SIR gain for different inputs, and we found that the average value for this time is approximately 1.4 seconds. From a convergence point of view, we think this is quite slow. Fig. 3 shows the improvement (gain) in the SIRs during the convergence process from a sample test. The slowness of convergence may be due to the approximation introduced in eq. (11). We noticed that this approximation with a specific value of λ, is accurate only for a limited range of ρ values and will not be accurate any more if the range changes. Further, analyzing the real ρ values that are produced by the algorithm has shown that the approximation introduces a large error (see Fig. 4). We, therefore, suggest introducing a variable or adaptive amplification factor (λ).

We also noticed that, with a relatively high level of white noise the algorithm normally fails. White noise destroys the approximate W-disjoint orthogonality assumption. This is due to the fact that white noise occupies the entire time-frequency domain. Improving the performance of the algorithm in noisy environments and studying the effect of different noise levels on the W-disjoint orthogonality property are important for real usability of the DUET method.

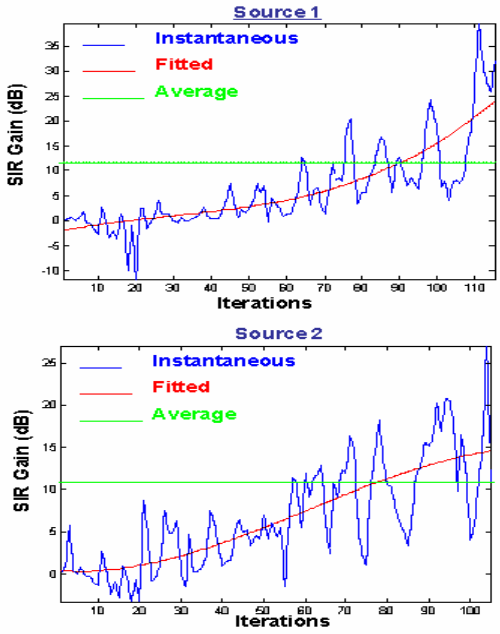

Fig. 3. The evolution of the SIR gains for two sources separated from two instantaneous mixtures. The overall average SIR gains are around 10 dB which is less due the convergence process in the first iterations of the algorithms. The average convergence time from different tests is 1.4 sec.

An important drawback that should also be addressed by future research is the presence of artifacts in the form of distortions especially when dealing with echoic mixtures. Using continuous masks instead of binary masks is supposed to solve this problem. Thus, new criteria for calculating the masks elements are needed for this purpose. Araki et al. [7] were able to reduce the artifacts by combing the method with ICA. In their approach, they used directivity pattern based *continuous* masks instead of *binary* mask. Still, the effectiveness of introducing ICA is questionable.

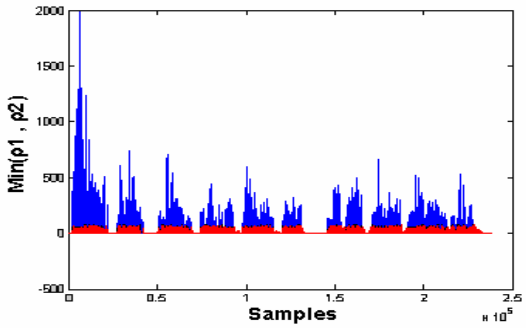

Fig. 4. Min(ρ1 , ρ2) for ρ values produced by the DUET for two mixtures of two sources; blue: the actual min. values from Matlab function min(.,.); red: min. value from approximation Min(ρ_1 , ρ_2) = - ln{exp(- λ ρ_1) + exp(- λ ρ_2) }/λ introduced in the algorithm for λ=10 which gives the best algorithm performance. The blue graph shows that Min(ρ1 , ρ2) is truly being minimized by the maximum likehood learning method despite the inaccuracy of the approximation. The accuracy of the approximation can be increased by having λ a function of ρ_1 and ρ_2.

8. Conclusions

The focus of this paper is on blind source separation applied to speech signals. The Degenerate Unmixing and Estimation Technique is used for this purpose. The approach utilizes binary time-frequency masks as tools for source separations. This paper demonstrates the powerfulness of the *basic* DUET approach that uses a simple intuitive idea to estimate the mixing parameters, and the powerfulness of time-frequency masks as an efficient tool for signal separation. However, the paper has also revealed significant drawbacks associated with the technique and that should be addressed in the future.

References

[1] CARDOSO, J.-F. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 1998, vol. 86, no. 10, pp. 2009–2025.

[2] BELL, A. J., SEJNOWSKI, T. J. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 1995, pp. 1129-1159.

[3] WEINSTEIN, E., FEDER, M., OPPENHEIM, A. Multichannel signal separation by decorrelation. *IEEE Transaction on Speech and Audio Processing*, 1993, vol. 1, pp. 405-413.

[4] PARRA, L., SPENCE, C. Convolutive blind separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, 2000, vol 8, no. 3, pp. 320-327.

[5] BROMAN, H., LINDGREN, U., SAHLIN, H., STOICA, P. Source separation: A TITO system identification approach. *Signal Processing*, 1999, vol. 73, pp. 169-183.

[6] JOURJINE, A., RIKARD, S., YILMAZ, O. Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures. In *IEEE ICASSP'2000*. Istanbul (Turkey), 2000, vol. 5, pp. 2985–2988.

[7] ARAKI, S., MAKINO, S., SAWADA, H., MUKAI, R. Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ICA. In *ICA2004 Fifth International Conference on Independent Component Analysis and Blind Signal Separation*. 2004, pp. 898-905.

[8] RICKARD, S., BALAN, R., ROSCA, J. Real-time time-frequency based blind source separation. In *Proc. Int. Workshop Independent Component Analysis and Blind Source Separation*. San Diego, CA (USA), 2001, pp. 651–656.

[9] ALLEN, J. B. Short term spectral analysis, synthesis and modification by discrete Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1977, vol. 25, no. 3, pp. 235-238.

About Authors...

Abbas MOHAMMED is an Associate Professor at the Department of Signal Processing, Blekinge Institute of Technology, Sweden. He was awarded the "PhD degree" from the Liverpool University, UK, in 1992 and the Swedish "Docent degree" in Radio Communications and Navigation from the Blekinge Institute of Technology in 2001. He was the recipient of the Blekinge Research Foundation Award "Researcher of the Year Award and Prize" for 2006. He is a Fellow of the Institution of Engineering and Technology (IET) and the UK's Royal Institute of Navigation (RIN). He is an Associate Editor to the International Journal of Navigation and Observation, a Board Member of the IEEE Signal Processing Swedish Chapter and an Editorial Board Member of the Radio Engineering Journal. He has also been a Guest Editor for several special issues of international journals. He is an author of over 120 papers in international journals and conference proceeding in the fields of signal processing, telecommunications and navigation systems. He has also developed techniques for measuring skywave delays in Loran-C receivers and received the 1994 Best Paper Award from the International Loran Association, USA, in connection to this work. His research interests are in space-time signal processing and MIMO systems, channel modeling, antennas and propagation, satellite and High Altitude Platforms, and radio navigation systems.

Tarig BALLAL was born in 1978 in Shendi, Sudan. He received his BSc honours in electrical engineering in 2001 from the University of Khartoum, Sudan, his MSc in electrical engineering with emphasize on telecommunications in 2005 from the Blekinge Institute of Technology, Sweden. Since June 2007, he is a PhD student at the School of Computer Science, College of Engineering, Mathematical & Physical Sciences, University College Dublin, Ireland. His major research interests include signal processing, telecommunication systems, and localization and positioning systems.

Nedelko GRBIC (IEEE M'97) was born in Sweden in 1971. He received his B.S degree at the University/College of Falun/Borlänge in 1993 and his MSc degree at the Blekinge Institute of Technology 1997, in Sweden. He received his Ph.D in Applied Signal Processing in 2001 and he was appointed an associate professor in 2006 at the Blekinge Institute of Technology. He was the recipient of the Blekinge Research Foundation Award "Researcher of the Year Award and Prize" for 2007. His research interests include array techniques in the field of speech enhancement, adaptive beamforming, blind equalization, blind signal separation and blind speech extraction in various applications such as binaural hearing aids, handsfree speech communication, conference telephony and underwater acoustics.