A SIMPLE AND COMUTATIONALLY EFFICIENT ALGORITHM FOR REAL-TIME BLIND SOURCE SEPARATION OF SPEECH MIXTURES

Tarig Ballal, Nedelko Grbic and Abbas Mohammed
Department of Signal Processing, Blekinge Institute of Technology, 372 25 Ronneby, Sweden

Keywords: BSS, blind source separation, speech enhancement, speech analysis, speech synthesis.

Abstract: In this paper we exploit the amplitude diversity provided by two sensors to achieve blind separation of two speech sources. We propose a simple and highly computationally efficient method for separating sources that are W-disjoint orthogonal (W-DO), that are sources whose time-frequency representations are disjoint sets. The Degenerate Unmixing and Estimation Technique (DUET), a powerful and efficient method that exploits the W-disjoint orthogonality property, requires extensive computations for maximum likehood parameter learning. Our proposed method avoids all the computations required for parameters estimation by assuming that the sources are "cross high-low diverse (CH-LD)", an assumption that is explained later and that can be satisfied exploiting the sensors settings/directions. With this assumption and the W-disjoint orthogonality property, two binary time-frequency masks that can extract the original sources from one of the two mixtures, can be constructed directly from the amplitude ratios of the time-frequency points of the two mixtures. The method works very well when tested with both artificial and real mixtures. Its performance is comparable to DUET, and it requires only 2% of the computations required by the DUET method. Moreover, it is free of convergence problems that lead to poor SIR ratios in the first parts of the signals. As with all binary masking approaches, the method suffers from artifacts that appear in the output signals.

1 GENERAL INFORMATION

Blind source separation (BSS) consists of recovering unobserved signals or “sources” from several observed mixtures (Cardoso, 1998). Several algorithms based on different source assumptions have been proposed. Some common assumptions are that the sources are statistically independent (Bell et al., 1995), are statistically orthogonal (Weinstein et al., 1993), are non-stationary (Parra et al., 2000), or can be generated by finite dimensional model spaces (Broman et al., 1999).

The Degenerate Unmixing and Estimation Technique (DUET) algorithm (Jourjine et al., 2000) (Rickard et al., 2001) (Yilmaz, et al., 2004) and other proposed methods (Bofill et al., 2000) exploit the approximate W-disjoint orthogonality property of speech signals to perform source separation. Two signals are said to be W-disjoint orthogonal (W-DO) when their time-frequency representations, are disjoint sets (Jourjine et al., 2000) (Rickard et al., 2001). DUET 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. DUET was proved to be powerful in speech source separation. Additionally, it is proved to be more computationally efficient as compared to other existing methods (Rickard et al., 2001). However, the idea that separation of W-DO sources requires only classifying the time-frequency points of mixtures has motivated us to look for a simpler approach. In other words, for W-DO sources the source separation problem is as simple as classifying the time-frequency points of a mixtures as belonging to one source or another.

In this paper we propose a simple and highly computationally efficient approach to achieve the above-mentioned classification. For this purpose we have introduced an additional assumption, that is the sources are "cross high-low diverse (CH-LD)". In a system with two sensors, two sources are said to be CH-LD, if the two sources are not both *clos*e to the same sensor. A source is close to a sensor, if its energy

Ballal T., Grbic N. and Mohammed A. (2006).
A SIMPLE AND COMUTATIONALLY EFFICIENT ALGORITHM FOR REAL-TIME BLIND SOURCE SEPARATION OF SPEECH MIXTURES.
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pages 105-109
DOI: 10.5220/0001571901050109
Copyright © SciTePress

at that sensor is higher than its energy at the other sensor.

Obviously, such diversity can be obtained from the spatial domain. To provide such diversity, sensors settings/directions can be exploited. A real case that supports our assumption is the case of two microphones with two speakers each associated with one of the microphones. If the distance between the microphones is relatively large as compared to that between the speakers and the microphones, we will get two mixtures of two sources (near speaker and interference from far speaker) that exactly satisfy the required assumption.

With this new assumption and the W-disjoint orthogonality property, two binary time-frequency masks that can extract the original sources from any of the two mixtures, can be constructed directly from the amplitude ratios of the time-frequency points of the two mixtures. The organization of this paper is as follows. In section 2 we define the source assumptions. In section 3 we derive a simple signal model based on the source assumptions. In section 4 the proposed algorithm is presented. In section 5, we discuss the results obtained from practical tests. Finally, a summary of this paper is given in section 6.

2 SOURCE ASSUMPTIONS

There are two basic assumptions required by our proposed approach:

- First: the sources should be W-disjoint orthogonal (W-DO).
- Second: the sources should be cross high-low diverse (CH-LD).

The first assumption requires 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* (Jourjine et al., 2000) (Rickard et al., 2001). For W-disjoint orthogonal (W-DO) sources, only one source should be active in each time-frequency point of the time-frequency representation of the sources.

Given a windowing function $W(t)$, two signals $s_i(t)$ and $s_j(t)$ are said to be W-disjoint orthogonal (W-DO) if the supports of the short-time Fourier transforms (STFTs) of $s_i(t)$ and $s_j(t)$ are disjoint (Jourjine et al., 2000) (Rickard et al., 2001).

The STFT of $s_j(t)$ is defined as (Allen et al., 1977)

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

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 (Rickard et al., 2001) 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 a non-active source contributes *zero* energy is replaced by assuming that it contributes *relatively low* energy as compared to the dominant source. Additionally, if the source signals have sparse representations in the time-frequency domain, the W-disjoint orthogonality can be sufficiently satisfied as for speech signals (Araki et al., 2004).

The second assumption requires that at least one of the two sources has two different (one *high* and one relatively *low*) amplitudes in the two mixtures, and the two sources are not both high (or both low) in the same mixture. To illustrate this assumption, let us assume a simple instantaneous mixing model with two mixtures of two sources:

$$x_1 = a_{11}s_1 + a_{21}s_2 \qquad (2)$$

$$x_2 = a_{12}s_1 + a_{22}s_2 \qquad (3)$$

Taking the STFT for both (2) and (3) yields

$$X_1(\omega,\tau) = a_{11}S_1(\omega,\tau) + a_{21}S_2(\omega,\tau) \qquad (4)$$

$$X_2(\omega,\tau) = a_{12}S_1(\omega,\tau) + a_{22}S_2(\omega,\tau) \qquad (5)$$

The CH-LD is fully satisfied when one of the two following statements is fulfilled:

$$\frac{|a_{11}S_1(\omega,\tau)|}{|a_{12}S_1(\omega,\tau)|} > 1 \;\; and \;\; \frac{|a_{21}S_2(\omega,\tau)|}{|a_{22}S_2(\omega,\tau)|} \le 1 \qquad (6)$$

or

$$\frac{|a_{21}S_2(\omega,\tau)|}{|a_{22}S_2(\omega,\tau)|} > 1 \;\; and \;\; \frac{|a_{11}S_1(\omega,\tau)|}{|a_{12}S_1(\omega,\tau)|} \le 1 \qquad (7)$$

Simplifying (6) and (7), the CH-LD can be fully satisfied by satisfying either

$$\frac{|a_{11}|}{|a_{12}|} > 1 \quad and \quad \frac{|a_{21}|}{|a_{22}|} \leq 1 \tag{8}$$

or

$$\frac{|a_{21}|}{|a_{22}|} > 1 \quad and \quad \frac{|a_{11}|}{|a_{12}|} \leq 1 \tag{9}$$

From (8) and (9), we deduce that the CH-LD depends mainly on the sensor setting and not on the source signals. For a typical interference cancellation problem (8) and (9) are normally satisfied. For general source separation problems a variety of sensor settings that satisfies (8) and (9) do exist.

3 SIGNAL MODEL

Let's start with the mixing model described by (2), (3), (4) and (5), respectively. First we try to introduce the W-disjoint orthogonality assumption into the model. Assuming that source s_k, $k \in \{1,2\}$, is the active source at time-frequency point (ω, τ), (4) and (5) become

$$X_1(\omega,\tau) = a_{k1} S_k(\omega,\tau) \tag{10}$$

$$X_2(\omega,\tau) = a_{k2} S_k(\omega,\tau) \tag{11}$$

From (10), (11) and in order to separate the sources, we need to determine which source is active at each time-frequency point. In other words, we need to determine the values of k that satisfy (10) and (11) at each time-frequency point.

4 PROPOSED ALGORITHM

Our proposed algorithm constructs two binary time-frequency masks, $\Phi_i(\omega,\tau), i = \{1,2\}$, by testing the ratio $\frac{|X_1(\omega,\tau)|}{|X_2(\omega,\tau)|}$ for all time-frequency points. The masks are constructed simply using

$$\Phi_i(\omega,\tau) = \begin{cases} 1 & if \ \frac{|X_1(\omega,\tau)|}{|X_2(\omega,\tau)|} > 1 \\ 0 & otherwise \end{cases}, i \in \{1,2\} \tag{12}$$

$$\Phi_j(\omega,\tau) = 1 - \Phi_i(\omega,\tau), \ j \in \{1,2\}, \ j \neq i \tag{13}$$

(12) and (13) stem directly from (6), (7), (10) and (11).

The time-frequency representation of the original sources can be obtained using

$$S_j(\omega,\tau) = \Phi_j(\omega,\tau)\, X_1(\omega,\tau), \ j \in \{1,2\} \tag{14}$$

$X_2(\omega,\tau)$ can be used instead of $X_1(\omega,\tau)$ in (14) and will yield a scaled and phase shifted version of $S_j(\omega,\tau)$ providing a spatial diversity the can further be exploited to improve the outputs. Finally, the inverse transform is used to obtain the original sources.

It is noticed that for instantaneous mixing, a similar algorithm can be derived without the CH-LD assumption be satisfied. For an instantaneous mixing model, the attenuations parameters are fixed leading to ratios of absolute values in (6), (7), (8) and (9) that are equal to constants (e.g. c_j, $j = \{1,2\}$). In this case a mask can be constructed according to

$$\Phi_i(\omega,\tau) = \begin{cases} 1 & if \ \frac{|X_1(\omega,\tau)|}{|X_2(\omega,\tau)|} = c_i \\ 0 & otherwise \end{cases}, i \in \{1,2\} \tag{15}$$

For real mixing models with reverberations, the ratios of absolute values in (6), (7), (8) and (9) will take random values that cluster around some two values constituting two clusters corresponding to the two sources. To be able to demix the sources, these clusters should be separate and no intersection should occur between them. If no inter-cluster intersection takes place, two clusters corresponding to the two CH-LD sources will be separated by a surface (imagine a 3-D plot of the ratios over the (ω,τ) plane) for which the ratio equals unity. If the sources are not CH-LD demixing the sources is still possible if we find the separating surface, which is beyond the scope of this paper. Finally, if reverberations cause inter-cluster intersection, separating the sources using the proposed method will not be possible in this case. But generally, practical tests with real echoic mixtures have proved the separation of sources even when reverberation is present.

5 RESULTS

The algorithm was implemented and tested using both artificial mixtures and real mixtures. Up to 22 dB SIR (signal to interference ratio) gain has been achieved with instantaneous artificial mixtures, up to 5 dB with echoic (i.e., containing reverberations in addition to the main signals) real mixtures. The interference here is the energy contribution from the other (undesired) source that should ideally be completely masked. For

the same mixtures the DUET showed approximately the same performance. Block size and block overlap were respectively 512 and 384 for the STFT. Fig. 1 shows two speech sources Separated from two echoic real mixtures using our proposed method, and using DUET. The mixtures are from an office room recording done by Te-Won Lee (http://inc2.ucsd.edu/~tewon/). Two Speakers have been recorded speaking simultaneously. Speaker 1 says the digits from one to ten in English and speaker 2 counts at the same time the digits in Spanish (uno dos, etc.). The recording was done in a normal office room. The distance between the speakers and the microphones was about 60cm in a square ordering. The figure illustrates the efficiency of our proposed method.

Since the proposed algorithm does not require any parameter learning, convergence problems are avoided. This explains the improved SIR in the first few milliseconds as compared to that of our implementation of the DUET method and as reflected in Fig. 1.

We noticed that, when white noise is present, the DUET normally fails. Our proposed algorithm was seen to withstand white noise. The algorithm has been verified to work even with low input signal to noise ratios, but still the noise remains in the outputs. Addressing this problem and generalizing the method for cases with more than two sources are two important future research goals.

As with all binary masking approaches, an important drawback that should also be addressed by future research is the presence of artifacts in the form of distortions. Using continuous masks instead of binary masks is supposed to solve the problem. Araki et al. (Araki et al., 2004) has addressed the artifacts problem associated with the DUET method and were able to reduce the artifacts by combing DUET with ICA (Independent Component Analysis). Combining our method with ICA in a similar way can be proposed as a solution to the associated artifacts problem.

6 CONCLUSIONS

In this paper we proposed a new method for blind source separation of W-disjoint orthogonal sources using time-frequency masks. Our focus was on separating two sources from two mixtures. We also introduced the cross high-low diversity assumption, an assumption that can be satisfied exploiting the sensors setting/directions. The method uses the amplitude ratios of the time-frequency representations of two mixtures to directly construct binary time-frequency masks to separate the sources. The method has shown performance that is comparable to that of the DUET method despite using only 2% of the computations required by DUET. Moreover, it is free of convergence problems. As with all binary masking approaches, the method suffers from artifacts that appear in the output signals.

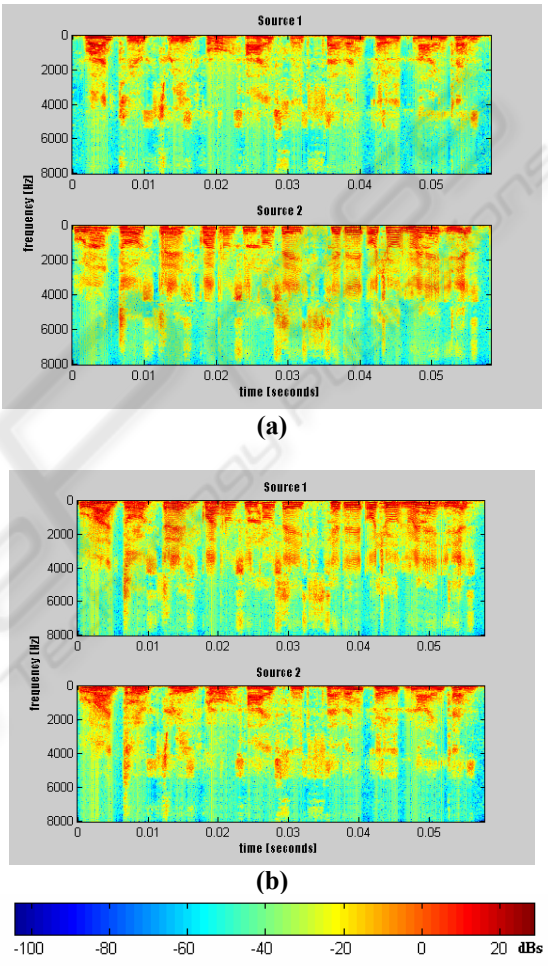

Figure 1: Separation of two speech sources from two echoic real mixtures. The figure shows the spectrograms of the separated sources: a) using our proposed method and b) using DUET. Sources have different permutation in each case. While DUET does not have a specific assumption about source permutation, our proposed method assumes that source 1 is the one that has its higher energy at sensor 1, and source 2 is the one that has its higher energy at sensor 2. As appears, there are no significant differences between the separated sources in fig. (a) and those in fig. (b). This illustrates the major advantage of our proposed method; that is despite using only 2% of the computations required by DUET, it can achieve results that are comparable to those achieved by DUET.

REFERENCES

Cardoso, J.-F., 1998. Blind Signal Separation: Statistical Principles. In *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025.

Bell, A.J. and Sejnowski, T.J., 1995. An Information Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, pp. 1129-1 159.

Weinstein, E., Feder, M. and Oppenheim, A., 1993 Multichannel Signal Separation by Decorrelation. *IEEE Transaction on Speech and Audio Processing*, vol. 1, no. 4, pp. 405-413.

Parra, L. and Spence, C., 2000. Convolutive Blind Source Separation Based On Multiple Decorrelations. *IEEE Transactions on Speech and Audio Processing*, March 2000.

Broman, H., Lindgren, U., Sahlin, H and Stoica, P, 1999. Source Separation: A TITO System Identification Approach. *Signal Processing*, vol. 73, pp. 169-183.

Jourjine, A., Rickard, S. and Yilmaz, O., 2000. Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures. *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey.

Rickard, S., Balan, R. and Rosca, J., 2001. Real- Time Time-Frequency Based Blind Source Separation. *Proceedings of the International Workshop of Independent Component Analysis and Blind Source Separation*, San Diego, CA.

Yilmaz, O Rickard, S, July 2004. Blind Separation of Speech Mixtures via Time-Frequency Masking. *IEEE Transactions on Signal Processing*, Vol. 52.

Bofill, P. and Zibulevsky, M., 2000. Blind Separation of More Sources than Mixtures Using Sparsity of Their Short-time Fourier Transform. *International Workshop on Independent Component Analysis and Blind Signal Separation*, Helsinki, Finland.

Allen, Jont B., June1977. Short Term Spectral Analysis, Synthesis and Modification by Discrete Fourier Transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25.

Araki, S., Makino, S., Sawada H. and Mukai, R., Sept. 2004. Underdetermined Blind Separation of Convolutive Mixtures of Speech with Directivity Pattern Based Mask and ICA. *ICA2004 (Fifth International Conference on Independent Component Analysis and Blind Signal Separation)*, pp. 898-905.

http://inc2.ucsd.edu/~tewon/