SETIT 2007
4th International Conference: **S**ciences of **E**lectronic, **T**echnologies of **I**nformation and **T**elecommunications
March 25-29, 2007 – TUNISIA

Blind Source Separation of Speech Mixtures using a Simple and Computationally Efficient Time-Frequency Approach

Tarig Ballal, Nedelko Grbic and Abbas Mohammed

School of Engineering, Blekine Institute of Technology, 372 25 Ronneby, Sweden
tbkh03@student.bth.se, ngr@bth.se, amo@bth.se

Abstract: A very simple and extremely computationally efficient algorithm for blind separation of two speech sources from two mixtures is presented in this paper. The algorithm exploits the approximate W-disjoint orthogonality of speech signals and assumes specific sensors (microphones) setting that allows the sources to possess a feature we call *cross high-low diversity*. Two sources are said to be *cross high-low diverse (CH-LD)* if the two sources are not both *close* to the same sensor. A source is said to be close to a sensor, if its energy at that sensor is *higher* than its energy at the other sensor. With this assumption and the W-disjoint orthogonality, it was found that a speech source can easily be extracted from any of the two mixtures with good SIRs (signal-to-interference ratios) based on simple algorithm that compares the ratios of the magnitudes of the time-frequency representations of the two mixtures. The proposed algorithm was tested using different mixtures and has proved to be efficient with both instantaneous and echoic real mixtures. Finally, performance optimization and future expendability to non-CH-LD sources was found possible.

Key words: Blind source separation, Speech processing, Cross high-low diversity, W-disjoint orthogonality.

INTRODUCTION

The approximate W-disjoint orthogonality is an interesting property of speech signals that has been exploited by a few algorithms for blind separation of speech signals (e.g. [Jourjine 2000, Rickard 2001]). The approximate W-disjoint orthogonality requires only one source to be *dominant* and possess a large percentage of the energy at each point of the time-frequency representation of a mixture. Generally, it was found that speech signals can be well represented by the set of time-frequency points where they possess a high percentage of their energy. Except the artifacts caused by sudden falls of energy to zero (discussed and healed in [Araki 2004]), removing points with just little energy was found to be unharmful to speech signals. Therefore, a source can be separated simply by extracting the *high-point* from a mixture.

In a mixture of several sources each time-frequency point contains energy that is a contribution from different sources and we need to make just simple decisions as which source is dominant at each point. Having successfully made such decisions, sources can be separated with a small amount of interferences. In [Rickard 2001] maximum likelihood parameters estimation is used in order to decide which source is dominant at which point.

In this paper, we exploit the sensor (microphones) settings to achieve the above-mentioned decision making in a much simpler way. In case of to two speech sources, it is possible to adopt a microphone setting such that each of two microphones is (physically) *closer* and/or *more directed* to one of the sources. With such setting, each of the two sources will have two different (one relatively *high* and one relatively *low*) amounts of energy in the two mixtures recorded via the microphones. Additionally, the two sources will *not both* possess high (or low) energy in the same mixture. We refer to such sources as *cross high-low diverse (CH-LD).* For CH-LD speech sources, we have found that, deciding which source is dominant at a time-frequency point becomes very simple. The values of the ratios of the magnitudes of the time-frequency representations of the two mixtures can effectively and *directly* be used for this purpose as utilized by our proposed algorithm. Two other algorithms that have utilized the *mixture ratios* are DUET [Jourjine 2000, Rickard 2001] and TIFROM [Abrard 2001a, Abrard 2001b]. However both methods are much more computationally complex as compared to our proposed method.

This paper is organised as follows. In section 2 we define the source assumptions and accordingly derive a simple model. In section 3 we present our proposed

algorithm. Section 4 discusses the results obtained from different tests. Section 5 is a summary for the paper.

1 Signal Model

The algorithm we propose here is based on two basic assumptions; the first is related to the amplitude diversity provided by two sensors (microphones); the second is related to a property of the source signals.

First Assumption: Sources should be *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 *close* to the same sensor. A source is said to be close to a sensor, if its energy at that sensor is *higher* than its energy at the other sensor. This assumption can be satisfied exploiting the sensors settings/directions.

For simple instantaneous mixing, two mixtures can be described as

$$x_1 = a_{11}s_1 + a_{21}s_2 \quad (1)$$

$$x_2 = a_{12}s_1 + a_{22}s_2 \quad (2)$$

In this paper we use the short-time Fourier transform (STFT) as the time-frequency representation of a signal. The STFT of $s_j(t)$ is defined as [Allen 1997]

$$S_j(\omega,\tau) = \int_{-\infty}^{\infty} s_j(t)\, w(t-\tau) e^{-i\omega t}\, dt \quad (3)$$

Taking the STFT for both (1) and (2) yields

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

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

We define

$$a_1 = a_{11}/a_{12},\ and \quad a_2 = a_{21}/a_{22} \quad (6)$$

We refer to a_1 and a_2 as the relative attenuation parameters.

The CH-LD assumption is fully satisfied when either

$$a_1 > 1 \ \ and \ \ a_2 < 1 \quad (7)$$

or

$$a_2 > 1 \ \ and \ \ a_1 < 1 \quad (8)$$

Second Assumption: Sources should be "at least" *approximately W-disjoint orthogonal (W-DO).* 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. The support of $S_j(\omega,\tau)$ is denoted as the set of the (ω,τ) pairs for which $S_j(\omega,\tau) \neq 0$, [Jourjine 2000, Rickard 2001].

According to this property, and in case of a mixture of a number of sources, only one source should be active in each time-frequency point of the time-frequency representation of the mixture. Practically, the time frequency representations of real signals (e.g. speech signals) are found to overlap and normally more than one source can be active at any time-frequency point. However, it was found that in most cases, one source is dominant and contributes most of the signal energy at that point. This leads to the concept of *approximate W-disjoint orthogonality* [Rickard 2001]. Two source signals are said to be *approximately W-DO*, if at *most* of the points (ω,τ) either of the following conditions is satisfied:

$$10\log_{10}[|S_1(\omega,\tau)|^2 / |S_2(\omega,\tau)|^2] \geq x \quad (9)$$

$$10\log_{10}[|S_2(\omega,\tau)|^2 / |S_1(\omega,\tau)|^2] \geq x \quad (10)$$

where x is the number of decibels (dBs) the energy of the dominant source exceeds the energy of the non-dominant source at a certain time-frequency point. x should be sufficiently large so as to be close from the original W-disjoint orthogonality. For sufficiently large value of x, (9) and (10) can be written as

$$|S_1(\omega,\tau)| >> |S_2(\omega,\tau)| \quad (11)$$

and

$$|S_2(\omega,\tau)| >> |S_1(\omega,\tau)| \quad (12)$$

Applying (11) and (12) to (4) and (5) yields

$$|X_1(\omega,\tau)| \approx a_{k1}|S_k(\omega,\tau)| \quad (13)$$

$$|X_2(\omega,\tau)| \approx a_{k2}|S_k(\omega,\tau)| \quad (14)$$

where $S_k(\omega,\tau)$, $k \in \{1,2\}$, is the dominant source at (ω,τ).

From (13 and (14) we can obtain

$$r(\omega,\tau) = |X_1(\omega,\tau)| / |X_2(\omega,\tau)| \approx a_{k1}/a_{k2} = a_k \quad (15)$$

At a point (ω,τ), an element of the matrix $r(\omega,\tau)$ represents an estimate for either a_1 or a_2.

Further, we realised that $r(\omega,\tau)$ can be decomposed into three matrices as follows:

$$r(\omega,\tau) = r_1(\omega,\tau) + r_2(\omega,\tau) + r_3(\omega,\tau) \quad (16)$$

where

$r_1(\omega,\tau)$ is all zeros except for (ω,τ) satisfying (9), $r_2(\omega,\tau)$ is all zeros except for (ω,τ) satisfying (10), and $r_3(\omega,\tau)$ is all zeros except for (ω,τ) satisfying neither (9) nor (10).

According to the above definitions, $r_1(\omega,\tau)$ is supposed to contain estimates of a_1 that correspond to s_1, and $r_2(\omega,\tau)$ is supposed to contain estimates of a_2 that correspond to s_2. Practically, we have found that $r_3(\omega,\tau)$ contain random values that spread in a wide area in the (ω,τ,r) space and that cluster around the unity (see Fig. 1). From the definition also, s_1 and s_2 are separable with good SIRs if the following conditions hold

1) $r_1(\omega,\tau)$ and $r_2(\omega,\tau)$ are *sufficiently* disjoint.

2) The amount of mixture energy from points (ω,τ) that satisfies $r_3(\omega,\tau) \neq 0$ is *sufficiently* small as compared to the total mixture energy.

2 Proposed Algorithm

Our proposed algorithm uses (6), (7), (8) and (15) to determine which source (1 or 2) is dominant, and accordingly uses a mixture value as an estimate for one of the sources. For simplicity, let's assume that the sources satisfy (7). In this case the algorithm can be summarized as follows:

1) Define $R = 1$.

2) Calculate $r(\omega,\tau), \forall\omega, \forall\tau$ according to (15).

3) If $r(\omega,\tau) > R$
Then $\hat{S}_1(\omega,\tau) = X_1(\omega,\tau), \forall\omega, \forall\tau$

4) If $r(\omega,\tau) < R$
Then $\hat{S}_2(\omega,\tau) = X_2(\omega,\tau), \forall\omega, \forall\tau$

In step 3 and 4, we used two different mixtures instead of partitioning a single mixture. Since we are dealing with CH-LD sources, $\hat{S}_i(\omega,\tau)$ is supposed to have better SIR when it is extracted from $X_i(\omega,\tau)$ where it initially possesses a relatively high input SIR.

Obviously, the algorithm requires just a few computations to generate the estimates of the separated sources and that is extremely fast and more suitable for real-time operation.

3 Tests and Results

First we tested the relative attenuation parameters estimates defined by (15) and (16). Practical tests has shown that the estimates corresponding to each of the two sources constitutes to separate sets. Fig. 1 shows the time-frequency plot of $r(\omega,\tau)$ and emphasizes the decomposition we introduced in (16). For visualizing $r(\omega,\tau)$, we have found it more illustrative to combine both time and frequency in one axis. This is done by simply ploting columns of $r(\omega,\tau)$ in consecutive order starting from column that corresponds to time=0.

Tests with different instantaneous mixtures have proved the efficiency of the method; up to 22 dB SIR gain has been achieved. The algorithm has also been tested with real echoic mixtures and has achieved up to 5 dB SIR gains.

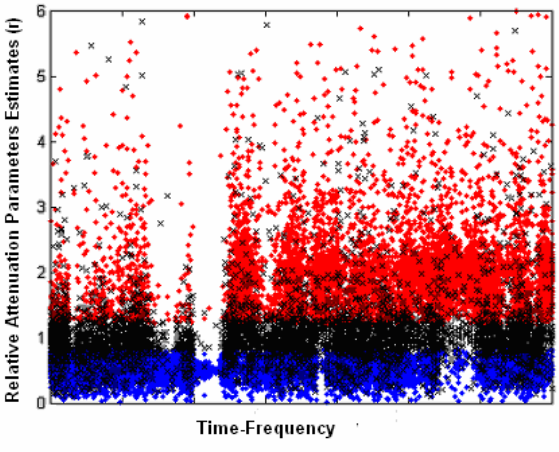

Fig. 1 *The estimates of the relative attenuation mixing parameters estimated from two instantaneous mixtures of two speech source. The two mixtures have a_1=2 and a_2 = 0.5. 5 dB dominance is assumed. The red and blue points represent r_1 and r_2 respectively. The Black x's represent r_3. As shown, r_1, r_2 and r_3 cluster around a_1, a_2 and unity, respectively. r_1 and r_2 are separate and do not cross the unity, and r_3 is found to contribute approximately 4% and 2% of the total energy of X_1 and X_2 respectively proving the separability of the sources using R=1 and with reasonable SIRs.*

The presence of time-frequency points that do not satisfy the approximate W-disjoit othogonality condition represented by $r_3(\omega,\tau)$, and the possibility of introducing large errors due to the approximation of the W-disjoint orthogonality, has motivated us to seek a way to improve the algorithm. We carried out several tests to check the performance of the algorithm with different values of R. Surprisingly, for each of the several pairs of mixtures we have tested, we have found that the SIR gains can be optimized at a certain

value of R we refer to as R_{opt}. For the mixtures we have tested, we have found that R_{opt} is typically greater than unity. Fig. 2 underscores the existence of such value.

Another important future extension is to generalize the algorithm to non-CH-LD sources. Our tests have proved the separability of such sources using different values of R, with the possibility of optimizing performance at some R_{opt} (Fig. 3). In both cases of CH-LD and non-CH-LD sources, Tests has revealed the existence of some relationship between the value of R_{opt} from one side and the input SIRs and the difference $(a_1 - a_2)$ from the other side. Studying this relationship is an important future research proposal. Additionally devising an algorithm that searches for the optimal value of R constitutes an important future extension for the proposed algorithm.

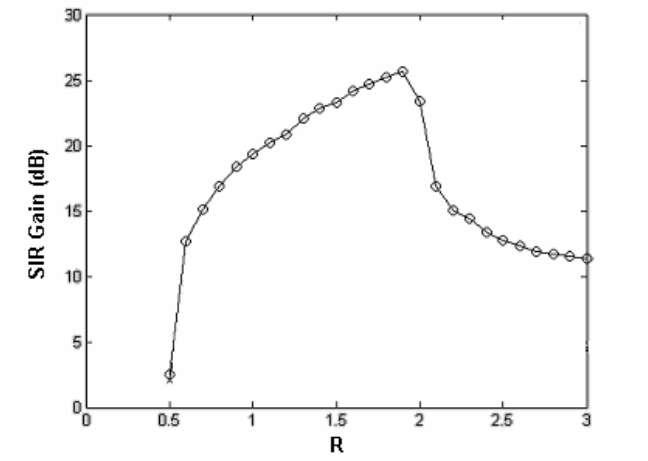

Fig. 2: *The SIR gains of one separated source extracted by the proposed algorithm. Different values of R are used with the same mixtures. Sources are CH-LD speech sources. The two mixtures have a_1=2 and a_2 = 0.5. The peak of the graph corresponds to the optimal performance of the algorithm.*

Conclusions

In this paper we presented a simple and extremely computationally efficient algorithm for blind source separation of speech mixtures. The algorithm exploits the approximate W-disjoint orthogonality property of speech signals. Additionally, the concept of cross high-low diversity, another basic assumption required by the algorithm was introduced. The tests have proved the efficiency of the method, and demonstrated the possibility to optimize the performance of the algorithm even when applied to non-cross high-low diversity (non-CH-LD) sources.

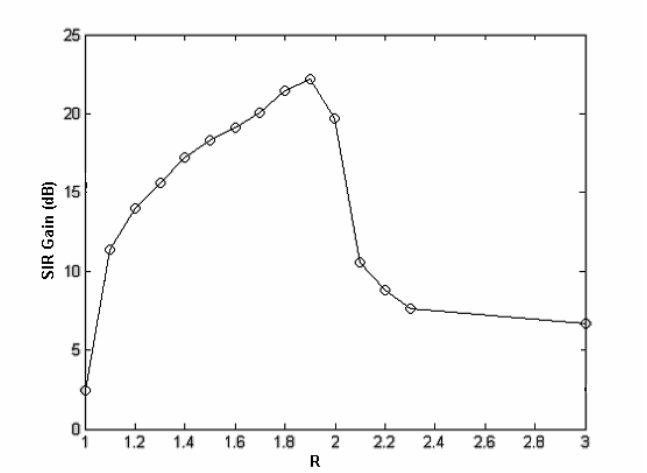

Fig. 3: *The SIR gains of one separated source extracted by the proposed algorithm. Different values of R are used with the same mixtures. Sources are non-CH-LD speech sources. The two mixtures have a_1=3 and a_2 = 2. The peak of the graph corresponds to the optimal performance of the algorithm.*

REFERENCES

– (Abrard & al. 2001a) F. Abrard, Y. Deville, and P. R. White, "A new Source Separation Approach for Instantaneous Mixtures Based on Time-Frequency Analysis", *Proceedings of ECM^{2}S*, Toulouse, France, May 2001.

– (Abrard & al. 2001b) F. Abrard, Y. Deville, and P. R. White, " From Blind Source Separation to Blind Source Cancellation in The Underdetermined Case: A new Approach Based on Time-Frequency Analysis", *Proceedings of ICA 2001*, San Diego, CA, Dec. 2001.

– (Allen 1997) Jont B. Allen, " Short Term Spectral Analysis, Synthesis and Modification by Discrete Fourier Transform", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 235-238, June 1977.

– (Aracki & al. 2004) S. Araki, S. Makino, H. Sawada and R. Mukai, "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, Sept. 2004.

– (Jourjine & al. 2000) A. Jourjine, S. Rickard, and O. Yilmaz, " 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, pp. 2985–88, June 2000.

– (Rickard & al. 2001) S. Rickard, R. Balan, and J. Rosca, "Real-Time Time-Frequency Based Blind Source Separation ", *Proc. Int. Workshop Independent Component Anal. Blind Source Separation*, San Diego, CA, pp. 651–656, Dec 2001.