Semi-blind Source Separation Based on ICA and Overlapped Speech Detection*

Jiří Málek[1], Zbyněk Koldovský[1,2], and Petr Tichavský[2]

[1] Faculty of Mechatronic and Interdisciplinary Studies
Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic
jiri.malek@tul.cz,
[2] Institute of Information Theory and Automation, Pod vodárenskou věží 4,
P.O. Box 18, 182 08 Praha 8, Czech Republic

Abstract. We propose a semi-blind method for separation of stereo recordings of several sources. The method begins with computation of a set of cancellation filters for potential fixed positions of the sources. These filters are computed from one-source-only intervals selected upon cross-talk detection. Each source in some of the fixed positions is canceled by the corresponding filter, by which the other sources are separated. The former source can be then separated by adaptive suppression of the separated sources. To select the appropriate cancellation filter, we use Independent Component Analysis. The performance of the proposed method is verified on real-world SiSEC data with two fixed and/or moving sources.

Keywords: Semi-blind Separation, Audio Source Separation, Cancellation Filter, Independent Component Analysis.

1 Introduction

Separation of multiple audio signals recorded in a natural environment is a discipline comprising several situations. These mainly differ in mutual positions of microphones and sources, room reverberation and variability of the environment. The SiSEC 2012 evaluation campaign[1] defines several tasks. In this paper, we consider the task "Determined convolutive mixtures under dynamic conditions". The goal is here to separate utterances of several speakers where at most two of them speak simultaneously from random fixed positions or moving positions (one source). The scenario is practical as it simulates a meeting situation. Signals recorded by four microphones are available, but we focus on using only two microphones, which are more accessible in practice.

The problem can be solved in a blind way, that is, by using only general assumptions such as the sparsity or independence. The latter assumption enables the use of Independent Component Analysis (ICA) either in the frequency

* This work was supported by Grant Agency of the Czech Republic through the project P103/11/1947.

[1] http://sisec.wiki.irisa.fr/

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 462–469, 2012.

domain or in the time domain. A drawback of blind methods consists in their limited efficiency due to the generality of the conception. The recent effort is therefore to take advantage of blind approaches together with incorporated a priori knowledge. These approaches have the common label *semi-blind*.

The known features of the SiSEC scenario (a priori knowledge) are as follows.

F1 Maximum two sources are active at the same time instant.
F2 At least one of the active sources is located at a fixed position.
F3 There is a finite number of potential fixed positions, and for each such position there exists an interval (even just one second short) in which a source sounds from this position but other sources are silent.
F4 Different sources are mutually independent.

In this paper, we propose a separation method that takes advantage of the above features as much as possible. The method utilizes two basic tools: cancellation filters and the ICA. The use of cancellation filters for separation of audio sources has been already proved to be useful even in difficult environments [2]. The approach is however restricted to sources having fixed known position. In this paper, we go one step further by applying ICA to find the filter assuming that the source is in one of possible (but unknown) positions.

A cancellation filter (CF) is a filter that cancels a targeted signal and passes the other signal through. Its output thus gives, on one hand, a separated (non-target) signal, which, on the other hand, can be suppressed from the original recording by an adaptive filter to separate the targeted signal. The CF is a time-invariant filter, so it cannot cancel a moving source.

In some situations such as a meeting, CFs can be computed for potential positions of sources in advance. Then, when an active speaker is detected at a given position and its speech overlaps with another speaker, the speeches can be separated using the corresponding CF(s). The SiSEC scenario considered here can be seen as one such situation. The CFs can be found based on F3, and the separation is possible thanks to F1 and F2. For easy reference, let the set of the computed CFs be called the cancellation filter-bank (CFB).

The only problem to cope with is the fact that the positions of active sources are not known at a given time. Based on F4, we propose a sophisticated method that uses ICA to separate the signals without knowing their positions. Following the idea of [3] and [4], ICA is applied to a data matrix that is defined using the a priori known CFB. In this sense, the method is "semi-blind". The details are given in Section 4. The following section describes the mixing model and the way to derive a CF. Section 3 describes how the CFB for the SiSEC data was derived. Results of the separation of the SiSEC data are presented in Section 5.

2 Problem Statement

Let s denote a targeted signal whose position is fixed. A stereo mixture of this signal with a noise is, in general, described by

$$\begin{aligned} x_{\mathrm{L}}(n) &= \{h_{\mathrm{L}} * s\}(n) + y_{\mathrm{L}}(n), \\ x_{\mathrm{R}}(n) &= \{h_{\mathrm{R}} * s\}(n) + y_{\mathrm{R}}(n) \end{aligned} \tag{1}$$

where n is the time index, $*$ denotes the convolution, $x_{\mathrm{L}}(n)$ and $x_{\mathrm{R}}(n)$ are, respectively, the signals from the left and right microphone, and $h_{\mathrm{L}}(n)$ and $h_{\mathrm{R}}(n)$ denote the microphone-source impulse responses. The noise signals on respective microphones are denoted by y_{L} and y_{R}. The signals are independent of s and, in our case, they correspond to responses (images) of the other speaker (or may be equal to zero). When the position of the "noise" speaker is fixed, the roles of the target and "noise" are interchangeable.

2.1 Cancellation Filter

To cancel the target s, we can seek a filter g that satisfies

$$\{g * h_{\mathrm{L}}\}(n) = h_{\mathrm{R}}(n), \tag{2}$$

because then the signal

$$\begin{aligned} v(n) &= \{g * x_{\mathrm{L}}\}(n) - x_{\mathrm{R}}(n) \\ &= \{g * h_{\mathrm{L}} * s\}(n) + \{g * y_{\mathrm{L}}\}(n) - \{h_{\mathrm{R}} * s\}(n) - y_{\mathrm{R}}(n) \\ &= \{g * y_{\mathrm{L}}\}(n) - y_{\mathrm{R}}(n) \end{aligned} \tag{3}$$

does not contain any contribution of $s(n)$, while y_{L} and y_{R} are passed through.

The filter g can be found using a noise-free interval $n = N_1, \ldots, N_2$, i.e. when $y_{\mathrm{L}}(n) = y_{\mathrm{R}}(n) = 0$, as a solution to the least square problem

$$g = \arg\min_{g} \sum_{n=N_1}^{N_2} \Big| \{g * x_{\mathrm{L}} - x_{\mathrm{R}}\}(n) \Big|^2. \tag{4}$$

We will call g the cancellation filter, although the true CF is the MISO filter on the right-hand side of (3), comprising of g and $-\delta$ (the unit impulse).

3 Building the CFB

According to F3, it is possible to compute the CF for each potential (fixed) position of a source. Our strategy is therefore to find one-source-only intervals and compute the CF according to (4), for each interval. This can be done manually, that is, in a supervised way, which we take into consideration. On the other hand, an automatic selection may be needed in real-time applications. Therefore, we propose two approaches to find the one-source-only intervals automatically.

The need is to distinguish three possible situations: silence, one speaker active, and two speakers active. The silence is easily detected by thresholding the energy of signals on microphones. It is more challenging to distinguish one speaker talk from a cross-talk.

Our first approach uses single (left) microphone only and is based on the linear predictive coding (LPC) of the observed signal. LPC models the signal as an autoregressive process of a selected order and measures the energy of

the residual signal (the prediction error). In the literature, see e.g. [7], it was observed that the prediction error of single speech signal is lower than that of an overlapped speech signal. The first approach therefore does the detection by thresholding the linear prediction error.

The second approach utilizes both microphones and measures the coherence of the signals [5]. The coherence is equal to one when the signal from one microphone is a delayed version of the signal from the other microphone, which ideally happens when the signal comes from a single direction without any reverberation. The reverberation must be taken into account, so the detection is based on thresholding the coherence.

The automatic selection proceeds as follows, examples of selected intervals are shown in Figure 1.

1. The detection criterion is computed throughout available data and smoothed by the moving-average filter (the length is 250 ms).
2. The intervals where the smoothed criterion is lower (higher) than a threshold are selected.
3. For each block of a sufficient length ($\geq$ 1 s), compute the CF according to (4) a store it into the CFB.

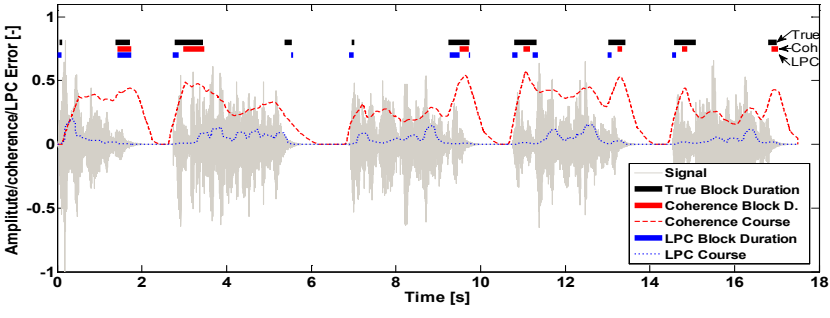

Fig. 1. An example of detected one-source-only blocks by thresholding LPC error and coherence. "True" blocks denote manually selected blocks.

The automatic procedure (but also the manual one) has the potential problem of computing several CFs for the same position. The duplicated CFs can be recognized by using a similarity measure (e.g. the mean square distance). However, there still may be CFs that differ quite much due to estimation errors but correspond to the same position. Fortunately, our method is robust in this respect thanks to the applied ICA, as it is explained in the following section.

4 Source Separation Using ICA and CFB

The SiSEC data can be divided into intervals in which two overlapping sources sound from unknown positions. In this section, we focus on processing one such interval $n = N_1, \ldots, N_2$ and propose a method that separates the signals from

the mixtures $x_{\mathrm{L}}(n)$ and $x_{\mathrm{R}}(n)$. The features F1-F4 are taken into account, so it is assumed that a CFB containing CF for each potential position of stationary sources is available.

Let g_i, $i = 1, \ldots, P$ denote CFs in the CFB. We define a data matrix as

$$\mathbf{X} = \begin{bmatrix} \{g_1 \star x_{\mathrm{L}}\}(N_1) & \ldots & \{g_1 \star x_{\mathrm{L}}\}(N_2) \\ \vdots & \vdots & \vdots \\ \{g_P \star x_{\mathrm{L}}\}(N_1) & \ldots & \{g_P \star x_{\mathrm{L}}\}(N_2) \\ x_{\mathrm{R}}(N_1) & \ldots & x_{\mathrm{R}}(N_2) \end{bmatrix} \tag{5}$$

and search for its independent components (ICs) by an ICA algorithm[2]. The ICA yields the de-mixing $(P+1) \times (P+1)$ matrix $\mathbf{W}$ and independent components $\mathbf{C} = \mathbf{W}\mathbf{X}$ which are linear combinations of rows of $\mathbf{X}$. It is highly expectable that at least one such combination (independent component) corresponds to the signal in which one source having fixed position is canceled. There are two key reasons for this claim.

1. The output of the kth CF can be expressed by

$$[\underbrace{0, \ldots, 0, 1}_{k}, 0, \ldots, -1] \cdot \mathbf{X} = \{g_k \star x_{\mathrm{L}}\} - x_{\mathrm{R}}, \tag{6}$$

   which means that the subspace spanned by rows of $\mathbf{X}$ contains the outputs of all CFs in the CFB. Since one source is in one of the potential positions (although unknown), there exists a linear combination of rows of $\mathbf{X}$ that cancels the source.
2. Such linear combination is an independent signal since it contains the contribution of one source only.

Since the order of the independent components (ICs) is random, the one that corresponds to the signal with canceled source must be found. This problem is easily resolved by finding the largest element (in absolute value) of the last column of $\mathbf{W}$. To explain, the ℓth element of the last column of $\mathbf{W}$ determines how much the last row of $\mathbf{X}$ contributes to the ℓth IC. Since only the last row of $\mathbf{X}$ contains samples of x_{R} (the other rows contain x_{L}), its contribution must be significant so that a source in the IC be canceled. Similarly, when there are two stationary sources in the mixture, we select two components corresponding to the two largest elements.

4.1 Separation by Adaptive Post-filtering

Once an independent signal is obtained, it can be considered as a separated one thanks to F1; let it be denoted by $v(n)$. The other source can be obtained by an adaptive Wiener-like filter that suppresses $v(n)$ from x_{L} and x_{R}.

[2] An arbitrary ICA algorithm can be used. We utilize the BGSEP algorithm from [6] for its speed and accuracy.

Let $X(k,\ell)$ and $V(k,\ell)$ be the short-time Fourier transform of $x_{\mathrm{L}}(n)$ (or x_{R}) and $v(n)$, respectively, where k is the frequency index and ℓ is the time-frame index. The adaptive filter, which is sometimes called a soft mask or the frequency-domain Wiener filter [8], is defined in the time-frequency domain by

$$W(k,\ell) = \frac{|X(k,\ell)|^2}{|X(k,\ell)|^2 + \tau|V(k,\ell)|^2}. \tag{7}$$

The time-frequency representation of the final output signal is

$$\widehat{S}(k,\ell) = W(k,\ell)X(k,\ell). \tag{8}$$

The free positive parameter τ allows control of the trade-off between the Signal-to-Interference ratio (SIR) and Signal-to-Distortion ratio (SDR) of the output signal.

5 Experiments

The SiSEC datasets "Determined convolutive mixtures under dynamic conditions" were recorded in a room with reverberation time about 700 ms. The sampling rate of signals is 16 kHz. From the four channel recordings in development dataset, we use signals from microphone 2 and 3, whose distance is 2 cm. The distances of the sources from microphones are about 1 m.[3]

The datasets are divided into intervals in which two sources are active. Each interval is processed separately and the separated signals are evaluated using the BSS_EVAL toolbox [9]. We use the criteria SIR, SDR and SAR (Signal-to-Artifact ratio) and SIR improvement (the difference between the SIR of mixed and separated signals). The resulting criteria are averaged over all intervals.

In our experiments, we distinguish the three ways of obtaining the CFB needed for our method (Section 3). *MAN* means the manual selection of one-source-only intervals. The automatic selections are denoted by *LPC* (LPC with the AR order 18) and *COH* (coherences with the length of the FFT window 128 samples and zero overlap).

5.1 Random Sources Activity in Unknown Static Positions

In this situation, two active speakers are located at unknown fixed positions on a semi-circle with radius 1 m. In Setup 1, the competing sources are always located on different angular sides with respect to the center of the array, that is one speaker is in $(-90°;0°)$ while the other one is in $(0°;90°)$. In the Setup 2, the two competing sources can be located in the whole angular space $(-90°;90°)$, but never in the same position. We consider two ways the separated signals

[3] The results for other microphone/source distances achieved on the SiSEC datasets can be found on the SiSEC results web page `http://www.irisa.fr/metiss/SiSEC11/dynamic/main.html`.

could be obtained. They can either be both obtained as ICs (Section 4) in which one source is canceled (denoted by *ica*) or both as the outputs of the adaptive Wiener-like filter (Section 4.1) that suppresses the obtained component from original recordings (denoted by *wf*); the parameter τ in (7) was put equal to 10. The results averaged over both separated sources are summarized in Table 1.

Table 1. Separation of fixed sources with random location

	method	SIR[dB]	SIR impr.[dB]	SDR[dB]	SAR[dB]
Setup 1	MAN (ica)	17.18	15.65	4.02	4.88
	MAN (wf)	12.16	10.62	1.17	3.24
	LPC (ica)	12.76	11.22	2.95	4.56
	LPC (wf)	9.64	8.10	-0.33	2.60
	COH (ica)	14.34	12.80	2.96	4.26
	COH (wf)	10.33	8.80	0.13	2.66
Setup 2	MAN (ica)	14.67	12.79	1.36	2.88
	MAN (wf)	11.42	9.54	0.71	3.33
	LPC (ica)	12.57	10.69	1.61	3.25
	LPC (wf)	8.62	6.74	-0.58	2.89
	COH (ica)	11.99	10.11	0.79	2.89
	COH (wf)	8.20	6.32	-1.15	3.06

The manually selected CFB leads to a better performance in terms of all criteria. The unsupervised approaches give comparable results, which points to their efficiency. The separation is better when signals are taken as the ICs than when they are obtained by the adaptive filter, especially in terms of SDR and SAR. This is explained by the fact that the sources are in fixed positions, so invariant filters (ica), which generate less distortions, are sufficient for the separation.

5.2 A Moving Source

In this scenario, one source is moving within the angular space (0°;90°) and its distance from microphones is varying between 0.5 m and 1.2 m. The position of the second source is fixed within the angular space (−90°;0°) either at one position during the whole dataset (Setup 1) or random position (Setup 2). Here, the moving source can be separated as the IC only, while the stationary source must be separated by the adaptive filter. Table 2 shows the results. In the case of Setup 1 (fixed source at one position), the label *single* denotes the case, when the CFB contains one filter only. This filter is able to supress the fixed source in the whole recording, i.e. the ICA utilization is not necesary.

The performance achieved with the single manually selected filter in Setup 1 confirms the suitability of the CF utilization for this type of separation scenario. In case of automatically constructed CFBs, the performance is lower, because the CFB contain CFs for positions where the moving source appeared for a moment. These CFs cause random confusion of the separated sources and deteriorate the performance.

Table 2. Separation of mixtures of one fixed and one moving source

	method	src. position	SIR[dB]	SIR impr.[dB]	SDR[dB]	SAR[dB]
Setup 1	MAN (single)	moving	13.00	12.62	6.62	8.23
	MAN (wf)	fixed	19.48	16.33	3.74	3.98
	LPC (ica)	moving	7.04	6.66	1.31	4.25
	LPC (wf)	fixed	16.72	13.57	0.17	0.50
	COH (ica)	moving	7.99	7.61	1.33	3.65
	COH (wf)	fixed	16.73	13.57	0.92	1.24
Setup 2	MAN (ica)	moving	10.28	10.26	-1.47	1.81
	MAN (wf)	fixed	14.82	11.29	1.70	2.24
	LPC (ica)	moving	9.10	9.08	1.17	3.73
	LPC (wf)	fixed	15.88	12.35	0.03	0.44
	COH (ica)	moving	10.03	10.01	1.23	3.65
	COH (wf)	fixed	15.29	11.75	-0.60	0.01

6 Conclusion

We presented a solution for the task presented in SISEC evaluation campaign using ICA and cancellation filters. The proposed method can be easily extended to situations where there are more than two sources [10].

References

1. Benesty, J., Makino, S., Chen, J. (eds.): Speech Enhancement, 1st edn. Springer, Heidelberg (2005)
2. Li, J., Sakamoto, S., Hongo, S., Akagi, M., Suzuki, Y.: Two-stage binaural speech enhancement with Wiener filter based on equalization-cancellation model. In: Proc. of WASPAA 2009, New Paltz, New York, pp. 133–136 (October 2009)
3. Koldovský, Z., Tichavský, P.: Time-Domain Blind Separation of Audio Sources on the basis of a Complete ICA Decomposition of an Observation Space. IEEE Trans. on Speech, Audio and Language Processing 19(2), 406–416 (2011)
4. Koldovský, Z., Tichavský, P., Málek, J.: A Semi-Blind Noise Extraction Using Partially Known Position of the Target Source. Submitted to a Conference (2011)
5. Albouy, B., Deville, Y.: Alternative structures and power spectrum criteria for blind segmentation and separation of convolutive speech mixtures. In: Proc. of ICA 2003, Nara, Japan, April 1-4, pp. 361–366 (2003)
6. Tichavský, P., Yeredor, A.: Fast Approximate Joint Diagonalization Incorporating Weight Matrices. IEEE Trans. on Signal Processing 57(3), 878–891 (2009)
7. Sundaram, N., et al.: Usable speech detection using linear predictive analysis - a model based approach. In: Proc. of ISPACS, Awaji Island, Japan, pp. 231–235 (2003)
8. Koldovský, Z., Nouza, J., Kolorenč, J.: Continuous Time-Frequency Masking Method for Blind Speech Separation with Adaptive Choice of Threshold Parameter Using ICA. In: Proc. of Interspeech 2006, Pittsburgh PA, USA, September 17-21, pp. 2578–2581 (2006)
9. Févotte, C., Gribonval, R., Vincent, E.: BSS EVAL toolbox user guide. IRISA, Rennes, France, Tech. Rep. 1706 (2005), `http://www.irisa.fr/metiss/bss_eval`
10. Rutkowski, T.M., et al.: Identification and tracking of active speaker's position in noisy environments. In: Proc. of IWAENC 2003, Kyoto, Japan, pp. 283–286 (2003)