Separating Underdetermined Convolutive Speech Mixtures

Michael Syskind Pedersen[1,2], DeLiang Wang[3], Jan Larsen[1], and Ulrik Kjems[2]

[1] Informatics and Mathematical Modelling, Technical University of Denmark, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark
[2] Oticon A/S, Kongebakken 9, DK-2765 Smørum, Denmark
[3] Department of Computer Science and Engineering & Center for Cognitive Science, The Ohio State University, Columbus, OH 43210-1277, USA
{msp, jl}@imm.dtu.dk, uk@oticon.dk, dwang@cse.ohio-state.edu

Abstract. A limitation in many source separation tasks is that the number of source signals has to be known in advance. Further, in order to achieve good performance, the number of sources cannot exceed the number of sensors. In many real-world applications these limitations are too restrictive. We propose a method for underdetermined blind source separation of convolutive mixtures. The proposed framework is applicable for separation of instantaneous as well as convolutive speech mixtures. It is possible to iteratively extract each speech signal from the mixture by combining *blind source separation* techniques with *binary time-frequency masking*. In the proposed method, the number of source signals is not assumed to be known in advance and the number of sources is not limited to the number of microphones. Our approach needs only two microphones and the separated sounds are maintained as stereo signals.

1 Introduction

Blind source separation (BSS) addresses the problem of recovering N unknown source signals $\mathbf{s}(n) = [s_1(n), \ldots, s_N(n)]^T$ from M recorded mixtures $\mathbf{x}(n) = [x_1(n), \ldots, x_M(n)]^T$ of the source signals. The term 'blind' refers to that only the recorded mixtures are known. An important application for BSS is separation of speech signals. The recorded mixtures are assumed to be linear superpositions of the source signals. Such a linear mixture can either be instantaneous or convolutive. The instantaneous mixture is given as

$$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) + \boldsymbol{\nu}(n), \tag{1}$$

where $\mathbf{A}$ is an $M \times N$ mixing matrix and n denotes the discrete time index. $\boldsymbol{\nu}(n)$ is additional noise. A method to retrieve the original signals up to an arbitrary permutation and scaling is independent component analysis (ICA) [1]. In ICA, the main assumption is that the source signals are independent. By applying ICA, an estimate $\mathbf{y}(n)$ of the source signals can be obtained by finding a (pseudo)inverse $\mathbf{W}$ of the mixing matrix so that

$$\mathbf{y}(n) = \mathbf{W}\mathbf{x}(n). \tag{2}$$

J. Rosca et al. (Eds.): ICA 2006, LNCS 3889, pp. 674–681, 2006.

Notice, this inversion is not exact when noise is included in the mixing model. When noise is included as in (1), $\mathbf{x}(n)$ is a nonlinear function of $\mathbf{s}(n)$. Still, the inverse system is assumed to be approximated by a linear system.

The convolutive mixture is given as

$$\mathbf{x}(n) = \sum_{k=0}^{K-1} \mathbf{A}_k \mathbf{s}(n-k) + \boldsymbol{\nu}(n) \tag{3}$$

Here, the source signals are mixtures of filtered versions of the original source signals. The filters are assumed to be causal and of finite length K. The convolutive mixture is more applicable for separation of speech signals because the convolutive model takes reverberations into account. The separation of convolutive mixtures can either be performed in the time or in the frequency domain. The separation system for each discrete frequency ω is given by

$$\mathbf{Y}(\omega, t) = \mathbf{W}(\omega)\mathbf{X}(\omega, t), \tag{4}$$

where t is the time frame index. Most methods, both instantaneous and convolutive, require that the number of source signals is known in advance. Another drawback of most of these methods is that the number of source signals is assumed not to exceed the number of microphones, i.e. $M \geq N$.

If $N > M$, even if the mixing process is known, it may not be invertible, and the independent components cannot be recovered exactly [1]. In the case of more sources than sensors, the *underdetermined/overcomplete* case, successful separation often relies on the assumption that the source signals are sparsely distributed in the time-frequency domain [2], [3]. If the source signals do not overlap in the time-frequency domain, high-quality reconstruction could be obtained [3].

However, there is overlap between the source signals. In this case, good separation can still be obtained by applying a binary time-frequency (T-F) mask to the mixture [2], [3]. In *computational auditory scene analysis*, the technique of T-F masking has been commonly used for years (see e.g. [4]). Here, source separation is based on organizational cues from auditory scene analysis [5]. More recently the technique has also become popular in blind source separation, where separation is based on non-overlapping sources in the T-F domain [6]. T-F masking is applicable to source separation/ segregation using one microphone [4],[7],[8] or more than one microphone [2], [3]. T-F masking is typically applied as a binary mask. For a binary mask, each T-F unit is either weighted by one or zero. An advantage of using a binary mask is that only a binary decision has to be made [9]. Such a decision can be based on, e.g., clustering [2], [3], [6], or direction-of-arrival [10]. ICA has been used in different combinations with the binary mask. In [10], separation is performed by first removing $N-M$ signals via masking and afterwards applying ICA in order to separate the remaining M signals. ICA has also been used in the other way around. In [11], it has been applied to separate two signals by using two microphones. Based on the ICA outputs, T-F masks are estimated and a mask is applied to each of the ICA outputs in order to improve the signal to noise ratio (SNR).

In this paper, we propose a method to segregate an arbitrary number of speech signals in a reverberant environment. We extend a previously proposed method for separation of instantaneous mixtures [12] to separation of convolutive mixtures. Based on the output of a square (2×2) blind source separation algorithm and binary T-F masks, our method segregates speech signals iteratively from the mixtures until an estimate of each signal is obtained.

2 Blind Extraction by Combining BSS and Binary Masking

With only two microphones, it is not possible to separate more than two signals from each other because only one null direction can be placed for each output. This fact does not mean that the blind source separation solution is useless in the case of $N > M$. In [12] we examined what happened if an ICA algorithm was applied to an underdetermined 2-by-N mixture. When the two outputs were considered, we found that the ICA algorithm separates the mixtures into subspaces, which are as independent as possible. Some of the source signals are mainly in one output while other sources mainly are present in the other output.

A flowchart for the algorithm is given in Fig. 1. As described in the previous section, a two-input-two-output blind source separation algorithm has been applied to the input mixtures, regardless the number of source signals that actually exist in the mixture. The two output signals are arbitrarily scaled. Different methods have been proposed in order to solve the scaling ambiguity. Here, we assume that all source signals have the same variance as proposed in [1] and the outputs are therefore scaled to have the same variance.

The two re-scaled output signals, $\hat{y}_1(n)$ and $\hat{y}_2(n)$, are transformed into the frequency domain e.g. using the Short-Time Fourier Transform STFT so that two spectrograms are obtained:

$$\hat{y}_1 \rightarrow Y_1(\omega, t) \tag{5}$$

$$\hat{y}_2 \rightarrow Y_2(\omega, t), \tag{6}$$

where ω denotes the frequency and t is the time frame index. The binary masks are then determined for each T-F unit by comparing the amplitudes of the two spectrograms:

$$\mathrm{BM1}(\omega, t) = \tau |Y_1(\omega, t)| > |Y_2(\omega, t)| \tag{7}$$

$$\mathrm{BM2}(\omega, t) = \tau |Y_2(\omega, t)| > |Y_1(\omega, t)|, \tag{8}$$

where τ is a threshold. Next, each of the two binary masks is applied to the original mixtures in the T-F domain, and by this non-linear processing, some of the speech signals are *removed* by one of the masks while other speakers are removed by the other mask. After the masks have been applied to the signals, they are reconstructed in the time domain by the inverse STFT. If there is only a single signal left in the masked output, defined by the selection criteria

in Section 2.3, i.e. all but one speech signal have been masked, this signal is considered extracted from the mixture and it is saved. If there are more than one signal left in the masked outputs, the procedure is applied to the two masked signals again and a new set of masks are created based on (7), (8) and the previous masks. The use of the previous mask ensures that T-F units that have been removed from the mixture are not reintroduced by the next mask. This is done by an element-wise multiplication between the previous mask and the new mask. This iterative procedure is followed until all masked outputs consist of only a single speech signal. When the procedure stops, the correlation between the segregated sources are found in order to determine whether a source signal has been segregated more than once. If so, the source is re-estimated by merging the two correlated masks. It is important to notice that the iteratively updated mask always is applied to the original mixtures and not to the previously masked signal. Hereby a deterioration of the signal due to multiple iterations is avoided.

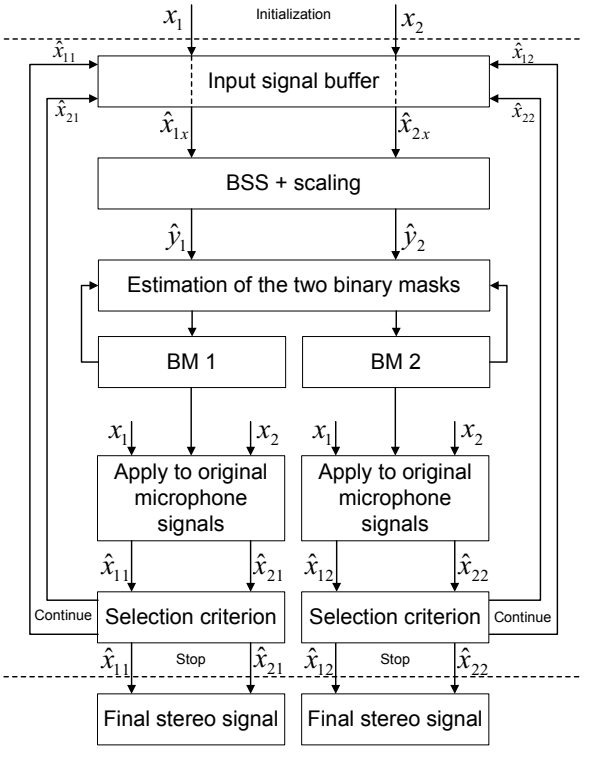

Fig. 1. Flowchart showing the main steps of the proposed algorithm. From the output of the BSS algorithm, binary masks are estimated. The binary masks are applied to the original signals which again are processed through the BSS step. Every time the output from one of the binary masks is detected as a single signal, the signal is stored. The iterative procedure stops when all outputs only consist of a single signal. The flowchart has been adopted from [12].

2.1 Finding the Background Signals

Since some signals may have been removed by both masks, all T-F units that have not been assigned the value '1' are used to create a *background mask*, and the procedure is applied to the mixture signal after the remaining mask is applied, to ensure that all signals are estimated. Notice that this step has been omitted from Fig. 1.

2.2 Extension to Convolutive Mixtures

Each convolutive mixture is given by a linear superposition of filtered versions of each of the source signals. The filters are given by the impulse responses from each of the sources to each of the microphones. An algorithm capable of separating convolutive mixtures is used in the BSS step. Separation still relies on the fact that the source signals can be grouped such that one output mainly contains one part of the source signals and the other output mainly contains the other part of the signals. In order to avoid arbitrary filtering, only the cross channels of the separation filters have been estimated. The direct channel is constrained to be an impulse. Specifically, we employ the frequency domain convolutive BSS algorithm by Parra and Spence [13][1].

2.3 Selection Criterion

In order to decide if all but one signal have been removed, we consider the envelope statistics of the signal. By considering the envelope histogram, it can be determined whether one or more than one signal is present in the mixture. If only one speech signal is present, many of the amplitude values are close to zero. If more speech signals are present, less amplitude values are close to zero. In order to discriminate between one and more than one speech signals in the mixture, we measure the width of the histogram as proposed in [14] as the distance between the 90% and the 10% percentile normalized to the 50% percentile, i.e.

$$\text{width} = \frac{P_{90} - P_{10}}{P_{50}}. \tag{9}$$

Further processing on a pair of masked signals should be avoided if there is one or zero speech signals in the mixture. If the calculated width is smaller than two, we assume that the masked signal consists of more than one speech signal. We discriminate between zero and one signal by considering the energy of the segregated signal. This selection criterion is more robust to reverberations than the correlation-based criterion used in [12].

3 Evaluation

The algorithm described above has been implemented and evaluated with instantaneous and convolutive mixtures. For the STFT, an FFT length of 2048

[1] Matlab code is available from `http://ida.first.gmd.de/~harmeli/download/download_convbss.html`

has been used. A Hanning window with a length of 512 samples has been applied to the FFT signal and the frame shift is 256 samples. A high frequency resolution is found to be necessary in order to obtain good performance. The sampling frequency of the speech signals is 10 kHz, and the duration of each signal is 5 s. The thresholds have been found from initial experiments. In the ICA step, the separation matrix is initialized by the identity matrix, i.e. $\mathbf{W} = \mathbf{I}$. When using a binary mask, it is not possible to reconstruct the speech signal as if it was recorded in the absence of the interfering signals, because the signals partly overlap. Therefore, as a computational goal for source separation, we employ the *ideal binary mask*[9]. The ideal binary mask for a signal is found for each T-F unit by comparing the energy of the desired signal to the energy of all the interfering signals. Whenever the signal energy is higher, the T-F unit is assigned the value '1' and whenever the interfering signals have more energy, the T-F unit is assigned the value '0'. As in [8], for each of the separated signals, the percentage of energy loss P_{EL} and the percentage of noise residue P_{NR} are calculated as well as the signal to noise ratio (SNR) using the resynthesized speech from the ideal binary mask as the ground truth:

$$P_{EL} = \frac{\sum_n e_1^2(n)}{\sum_n I^2(n)}, \quad P_{NR} = \frac{\sum_n e_2^2(n)}{\sum_n O^2(n)}, \quad \text{SNR} = 10 \log_{10} \left[\frac{\sum_n I^2(n)}{\sum_n (I(n) - O(n))^2} \right],$$

where $O(n)$ is the estimated signal, and $I(n)$ is the recorded mixture resynthesized after applying the ideal binary mask. $e_1(n)$ denotes the signal present in $I(n)$ but absent in $O(n)$ and $e_2(n)$ denotes the signal present in $O(n)$ but absent in $I(n)$. The input signal to noise ratio, SNR_i, is found too, which is the ratio between the desired signal and the noise in the recorded mixtures.

Convolutive mixtures consisting of four speech signals have also been separated. The signals are uniformly distributed in the interval $0° \leq \theta \leq 180°$. The mixtures have been obtained with room impulse responses synthesized using the image model [15]. The estimated room reverberation time is $T_{60} \approx 160$ ms. The distance between the microphones is 20 cm. The method has been evaluated with and without the proposed selection criterion described in Section 2.3. When the selection criterion was not used, it has been decided when a source signal has been separated by listening to the signals. The separation results are shown in Table 1 and Table 2. The average input SNR is -4.91 dB. When the selection criterion was applied manually, the average SNR after separation is 1.91 dB

Table 1. Separation results for four convolutively mixed speech mixtures. A manual selection criterion was used.

Signal No.	P_{EL}(%)	P_{NR}(%)	SNR_i (dB)	SNR (dB)
1	66.78	20.41	-4.50	1.35
2	32.29	41.20	-4.50	1.24
3	52.86	19.08	-3.97	2.12
4	15.78	30.39	-6.67	2.91
Average	41.93	27.77	-4.91	1.91

Table 2. Separation results for four convolutively mixed speech mixtures. The selection criterion as proposed in Section 2.3 was used.

Signal No.	P_{EL}(%)	P_{NR}(%)	SNR_i (dB)	SNR (dB)
1	39.12	46.70	-4.50	0.63
2	64.18	18.62	-4.50	1.45
3	26.88	33.73	-3.97	2.40
4	45.27	32.49	-6.67	1.69
Average	43.86	32.88	-4.91	1.54

with an average SNR gain of 6.8 dB. When selection criterion was applied as proposed, the average SNR after separation is 1.45 dB with an average SNR gain of 6.4 dB, which is about half a dB worse than selecting the segregated signals manually. It is not always that all the sources are extracted from the mixture. Therefore the selection criterion could be further improved. For separation of instantaneous mixtures an SNR gain of 14 dB can be obtained, which is significantly higher than that for the reverberant case. This may be explained by several factors. Errors such as misaligned permutations are introduced from the BSS algorithm. Also, convolutive mixtures are not as sparse in the T-F domain as instantaneous mixtures. Further, the assumption that the same signals group into the same groups for all frequencies may not hold. Some artifacts (musical noise) exist in the segregated signals. Especially in the cases, where the values of P_{EL} and P_{NR} are high. Separation results are available for listening at `www.imm.dtu.dk/~msp`.

As mentioned earlier, several approaches have been recently proposed to separate more than two sources using two microphones by employing binary T-F masking [2], [3], [10]. These methods use clustering of amplitude and time differences between the microphones. In contrast, our method separates speech mixtures by iteratively extracting individual source signals. Our results are quite competitive although rigorous statements about comparison are difficult because the test conditions are different.

4 Concluding Remarks

A novel method of blind source separation of underdetermined mixtures has been described. Based on sparseness and independence, the method iteratively extracts all the speech signals. The linear processing from BSS methods alone cannot separate more sources than the number of recordings, but with the additional nonlinear processing introduced by the binary mask, it is possible to separate more sources than the number of sensors. Our method is applicable to separation of instantaneous as well as convolutive mixtures and the output signals are maintained as stereo signals. An important part of the method is the detection of when a single signal exists at the output. Future work will include better selection criteria to detect a single speech signal, especially in a reverberant environment. More systematic evaluation and comparison will also

be given in the future. The assumption of two microphones may be relaxed and the method may also be applicable to other signals than speech which also have significant redundancy.

Acknowledgements

The work was performed while M.S.P. was a visiting scholar at The Ohio State University Department of Computer Science and Engineering. M.S.P was supported by the Oticon Foundation. M.S.P and J.L are partly also supported by the European Commission through the sixth framework IST Network of Excellence: PASCAL. D.L.W was supported in part by an AFOSR grant and an AFRL grant.

References

1. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley (2001)
2. Roman, N., Wang, D.L., Brown, G.J.: Speech segregation based on sound localization. J. Acoust. Soc. Amer. **114** (2003) 2236–2252
3. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Processing **52** (2004) 1830–1847
4. Wang, D.L., Brown, G.J.: Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. Neural Networks **10** (1999) 684–697
5. Bregman, A.S.: Auditory Scene Analysis. 2 edn. MIT Press (1990)
6. Jourjine, A., Rickard, S., Yilmaz, O.: Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures. In: Proc. ICASSP. (2000) 2985–2988
7. Roweis, S.: One microphone source separation. In: NIPS'00. (2000) 793–799
8. Hu, G., Wang, D.L.: Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Trans. Neural Networks **15** (2004) 1135–1150
9. Wang, D.L.: On ideal binary mask as the computational goal of auditory scene analysis. In Divenyi, P., ed.: Speech Separation by Humans and Machines. Kluwer, Norwell, MA (2005) 181–197
10. Araki, S., Makino, S., Sawada, H., Mukai, R.: Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ICA. In: Proc. ICA'2004. (2004) 898–905
11. Kolossa, D., Orglmeister, R.: Nonlinear postprocessing for blind speech separation. In: Proc. ICA'2004, Granada, Spain (2004) 832–839
12. Pedersen, M.S., Wang, D.L., Larsen, J., Kjems, U.: Overcomplete blind source separation by combining ICA and binary time-frequency masking. In: Proceedings of the MLSP workshop, Mystic, CT, USA (2005)
13. Parra, L., Spence, C.: Convolutive blind separation of non-stationary sources. IEEE Trans. Speech and Audio Processing **8** (2000) 320–327
14. Büchler, M.C.: Algorithms for Sound Classification in Hearing Instruments. PhD thesis, Swiss Federal Institute of Technology, Zurich (2002)
15. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Amer. **65** (1979) 943–950