BEAMFORMING-BASED CONVOLUTIVE SOURCE SEPARATION

Wolf Baumann, Dorothea Kolossa and Reinhold Orglmeister

Berlin University of Technology
Electronics and Medical Signal Processing Group, Einsteinufer 17, 10587 Berlin
{w.baumann, d.kolossa, orglmeister}@ee.tu-berlin.de

ABSTRACT

A robust independent component analysis (ICA) algorithm for blind separation of convolved mixtures of speech signals is introduced. It is based on two parallel frequency dependent beamforming stages, each of which cancels the signal from one interfering source by frequency dependent null-beamforming. The zero-directions of the beamforming stages are optimized to yield maximally independent outputs, which is achieved via second and higher order statistics. Optimization is carried out in the frequency domain for each frequency band separately, so that phase distortions caused by the room impulse responses are compensated. In contrast to other frequency domain source separation algorithms, this structure does not suffer from permutation of frequency bands, while retaining the major advantage of blind methods, that do not require an external estimate of the direction of arrival (DOA).

1. INTRODUCTION

Independent component analysis is of great importance in the field of blind source separation, where the task is to recover original source signals from a set of mixed signales without knowledge about the mixing process. If the source signals are statistically independent, separation can be performed with the help of ICA methods, which maximize independence of the output signals.

Successfull applications include separation of biomedical data (like ECG or FMRI), sonar or seismographic data. Depending on the mixing process, e.g. linear/nonlinear or instantaneous/convolved the task of source separation can become very difficult. In practice, algorithms for the separation of convolutive mixtures are computationally expensive and often restricted to certain room conditions.

To make separation of convolved sources more robust, recent approaches apply geometrical constraints in ICA algorithms to solve permutations between frequency bands, e.g. [1]. Other approaches iterate between beamforming and ICA stages, like [2].

In contrast, our approach is actually a combination of beamforming and ICA because it consists of a beamforming structure with frequency dependent null-steering, where the null-directions are adjusted to make output signals as independent as possible. Through the use of a combination of second and higher order statistics, and in constraining the direction of arrival of the source signals, our algorithm is robust even under reverberant conditions and needs no additional permutation correction.

2. MODEL AND MATHEMATICAL PRELIMINARIES

We restrict the following consideration to a 2×2 mixing system, i.e. two simultaneously talking speakers are recorded by two microphones. The convolutive mixing process can be expressed as

$$\mathbf{x}(t) = \mathbf{A} * \mathbf{s}(t), \tag{1}$$

with $\mathbf{x}$, $\mathbf{A}$ and $\mathbf{s}$ representing the recorded signals, the mixing matrix and the source signals, respectively. Here, the mixed signals $\mathbf{x}$ are superpositions of the filtered source signals $\mathbf{s}$, and the mixing matrix $\mathbf{A}$ contains the room impulse responses between each source and microphone.

From a beamforming viewpoint, some simplifications are possible. Assuming farfield conditions and a standard delay and sum beamforming model, the mixing matrix reduces to a matrix that contains simple time shifts according to the different time delays in dependence on the look direction. In the frequency domain the mixing matrix then contains frequency dependent phase shifts, and the mixing process is expressed as

$$\mathbf{X}(j\omega) = \mathbf{A}_{ph}(j\omega) \cdot \mathbf{S}(j\omega) \tag{2}$$

with the phase shift mixing matrix

$$\mathbf{A}_{ph}(j\omega) = \begin{bmatrix} 1 & 1 \\ e^{-j\omega\frac{d}{c}\sin\left(\varphi_1(\omega)\right)} & e^{-j\omega\frac{d}{c}\sin\left(\varphi_2(\omega)\right)} \end{bmatrix} \tag{3}$$

which depends on the frequency ω, the speed of sound c, the distance d between microphones and the angles of the impinging sources $\varphi_1(\omega)$ and $\varphi_2(\omega)$, as shown in Figure 1. The phase shifts are given relative to the phase on microphone one, so the first row of $\mathbf{A}_{ph}(j\omega)$ is normalized to one.

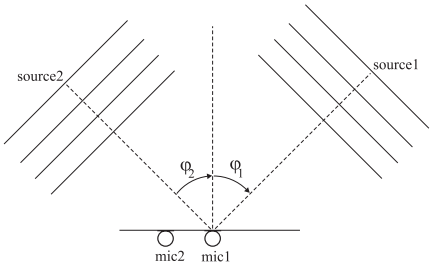

Fig. 1. Farfield beamforming model

Note that the matrix $\mathbf{A}_{ph}(j\omega)$ differs from a standard delay and sum beamforming model because the angles φ_1 and φ_2 are not

restricted to be the same for all frequencies. This key difference is absolutely necessary for the compensation of phase distortions, caused by the reverberations inherent in room impulse responses.

It has to be mentioned that the model in (2) does not consider attenuations caused by the room impulse responses but describes the mixing process as a superposition of delayed source signals. Though this is not an exact reproduction of the real world conditions, it has proven to be sufficient for close microphone arrangements and furthermore reduces the search space.

Legitimation for this model is given by analyzing the separation filters of some ICA algorithms for convolutive mixtures. If treated as two filter and sum beamformers, the ICA filters can be examined with respect to their spatial response. Figure 2 gives a good example of such a beampattern.

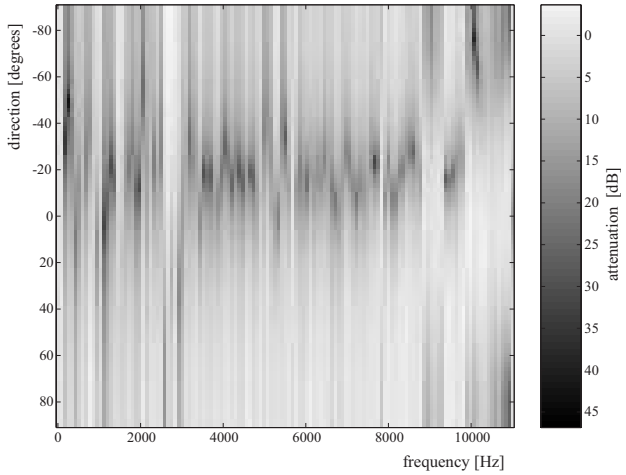

Fig. 2. Pattern of an ICA filter

It is clearly to be seen that separation is based on forming a spatial null in a certain direction. This fact can be used to build a separation matrix which performs frequency dependent null-beamforming. The separation matrix $\mathbf{W}(j\omega)$ is obtained by taking the inverse of the phase shift mixing matrix $\mathbf{A}_{ph}(j\omega)$ multiplied by a scaling factor that improves attenuation for low frequencies

$$\mathbf{W}(j\omega) = \frac{|e_1 - e_2|}{e_1 - e_2} \begin{bmatrix} -e_2 & 1 \\ e_1 & -1 \end{bmatrix} \tag{4}$$

with

$$e_1 = e^{-j\omega \frac{d}{c} \sin\left(\varphi_1(\omega)\right)} \quad \text{and} \quad e_2 = e^{-j\omega \frac{d}{c} \sin\left(\varphi_2(\omega)\right)}. \tag{5}$$

This ensures an unattenuated transfer function in one look direction while forcing the other one to be zero. Note that the two beamformers (see Figure 3), implemented by the separation matrix $\mathbf{W}$, need to be optimized jointly, because the constant direction of the first one is the zero direction of the second one and vice versa.

The joint adjustment of $\varphi_1(\omega)$ and $\varphi_2(\omega)$ is simplified when the sources are restricted to lie in different quadrants. In this case, no permutations between frequency bands are possible, because φ_1 and φ_2 can only take positive or negative values, respectively. While restricting the source directions, one additionally avoids non-invertible constellations, e.g. if the sources impinge from the same direction. However, to our expirience, those constellations are not separable in general.

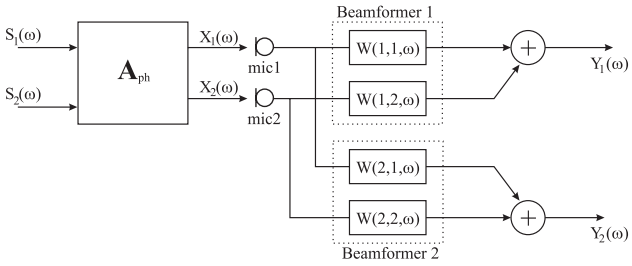

Fig. 3. Model in the frequency domain

3. ALGORITHM

Independent component analysis aims to find that demixing matrix $\mathbf{W}(\omega)$, which makes the outputs as statistically independent as possible. In constrast to second order methods like PCA, independent component analysis not only utilizes the second order information contained in the covariance matrix $\mathbf{C_{xx}}$ of the data, but additionally incorporates higher order statistical information. Whereas gaussian distributions are characterized completely by the first and second moments of the variables – i.e. by the variables means and variances $E(x)$ and $E(x^2)$ – for non-gaussian variables, the higher order statistics, formed by expectations of higher order polynomials of the data, contain additional information about the data distributions. This additional information can be utilized to obtain unmixing matrices which not only decorrelate the data but also minimize correlations between arbitrary functions of the outputs y_i, thus leading to results which are independent in the original sense of factorizable probability desity functions: $p(y_1, y_2) = p(y_1) \cdot (y_2)$.

3.1. Cost function

Within the ICA framework, different measures of independence have been proposed. One natural criterion, taken from information theory, is the mutual information I of the outputs y_i defined as

$$I(y_1 \ldots y_N) = H(\mathbf{y}) - \sum_{i=1}^{N} H(y_i) \tag{6}$$

where H stands for the entropy of a variable. An equivalent criterion measures and optimizes the negentropy of the demixed components (e.g. [3]), which can be considered as maximizing the variables deviation from gaussianity. However, as higher order statistics can be computationally expensive to calculate and large amounts of data are needed for reliable estimation, idependent component analysis is often carried out by minimizing the data's cross-statistics up to the fourth order. A useful measure of the data's statistics are cross-cumulants, which are defined as the Taylor series expansion coefficients of the data's second characteristic function $\Psi(\omega) = ln(E(e^{j\omega x}))$. In contrast to the moments of the data (the Taylor series expansion coefficients of the first characteristic function $\Phi(\omega) = E(e^{j\omega x})$), cross-cumulants have two very desireable properties which motivate their use in independent component analysis:

- Cumulants are additive for independent variables - i.e., if $z = x + y$ and x and y are independent, then $cum(z) = cum(x) + cum(y)$.

- For gaussian random variables, all cumulants above order two are zero, thus cumulants also provide a measure of a variables degree of deviation from gaussianity, which is of use especially in information theoretic methods.

If and only if the data is statistically independent, all of their cross-cumulants are zero, therefore, the set of cross-cumulants up to order four constitutes a cost-function that can be estimated from the demixed sources $\mathbf{Y}$ in order to obtain independent data. Thus, cross cumulants have been used in extracting independent components from mixtures, notably in [4]. In the proposed algorithm, the cost function J, which is optimized to obtain the null-direction of the beamformer, consists of a sum of the second and fourth order cross-cumulants. Cross-cumulants of order three are neglected, since they are zero for symmetrical distributions like those of $\mathbf{Y}(j\omega)$:

$$J(Y_1', Y_2') = E(abs(Y_1' \cdot Y_2')) + Cum(Y_1', Y_2'), \quad (7)$$

where $Cum(Y_1', Y_2')$ refers to the cross-cumulant of Y_1' and Y_2' defined by:

$$\begin{aligned} Cum(Y_1', Y_2') &= E[|Y_1'|^2 \cdot |Y_2'|^2] - E[|Y_1'|^2] \cdot E[|Y_2'|^2] - \\ &\quad |E[Y_1' \cdot Y_2^{*'}]|^2 - |E[Y_1' \cdot Y_2']|^2. \end{aligned} \quad (8)$$

Before calculating the cost function using (7) and (8), the variables Y must be centered and normalized to unit variance:

$$Y' = \frac{Y - E(Y)}{\sqrt{E((Y - E(Y))^2)}}. \quad (9)$$

3.2. Optimization

The cost function to be optimized,

$$J(\mathbf{Y}) = J(\mathbf{W}\,\mathbf{X}) = J(\mathbf{W}(\varphi 1, \varphi 2)\,\mathbf{X}), \quad (10)$$

depends on the tuning variables only indirectly via the composition of the mixing matrix, with $\mathbf{W}(\varphi 1, \varphi 2)$ defined by (4) and (5).

Therefore, a calculation of the cost function gradient leads to

$$\nabla_\varphi J(\mathbf{W}\,\mathbf{X}) = \mathbf{X}\,\nabla_W J(\mathbf{W}(\varphi 1, \varphi 2)\mathbf{X})\nabla_\varphi \mathbf{W}(\varphi 1, \varphi 2) \quad (11)$$

with

$$\nabla_\varphi \mathbf{W}(\varphi 1, \varphi 2) = \frac{\delta \mathbf{W}}{\delta \mathbf{e}} \cdot \frac{\delta \mathbf{e}}{\delta \varphi}. \quad (12)$$

An explicit calculation of the gradient therefore is computationally very expensive, so that an empirical gradient descent was employed. For this, the procedure adopted was as follows:

- At each frequency band do:
- Set initial angles φ_1 and φ_1 to mean of previously found directions:

$$\varphi_{1/2}(k) = \frac{1}{k-1} \sum_{m=1}^{k-1} \varphi_{1/2}(k). \quad (13)$$

- Set initial stepsize ν.
- Calculate empirical gradient by

$$\begin{aligned} \frac{\delta J}{\delta \varphi_1} &\approx \frac{J(\varphi_1 + \delta, \varphi_2)}{\delta} \quad \text{and} \\ \frac{\delta J}{\delta \varphi_2} &\approx \frac{J(\varphi_1, \varphi_2 + \delta)}{\delta}. \end{aligned} \quad (14)$$

Table 1. Configurations

config	θ_1	θ_2	recordings
A	45°	25°	speaker 1, speaker 2, both speakers
B	10°	25°	speaker 1, speaker 2, both speakers

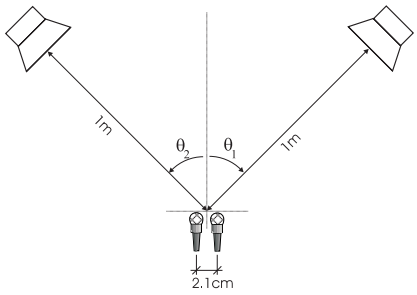

Fig. 4. Experimental setup

- Normalize the empirical gradient vector ΔJ to unit length and multiply with stepsize to obtain search steps. At this stepsize, conduct linesearch until no further improvement is obtained at given direction and stepsize.
- While ν remains larger than minimum stepsize ν_{min}, determine new stepsize by $\nu_{i+1} = \nu_i / m$ and calculate new empirical gradient.

4. RESULTS

Recordings were made in an office room with dimensions of about 10×15m. The distance between the loudspeakers and the two microphones (Behringer ECM 8000) was set to one meter (see Figure 4). We used speech signals from the TI20 database with two different male speakers, which were played back and recorded, once simultaneously and once separately, in two different setups of loudspeakers. The following table gives an overview of the configurations.

4.1. SNR calculation

Working with real room recordings, the calculation of correct SNR values is difficult due to convolutions with the transfer function between source and microphone. To avoid the additional expenditure of correlating and normalizing signals, we calculated the SNR improvement with the help of the separately recorded source signals (only one speaker present). The separation filters were applied to these recordings to determine the relative attenuation and amplification, respectively. The resulting ouputs $\mathbf{Y}(j\omega) = \mathbf{W}(j\omega) \cdot \mathbf{X}(j\omega)$ were compared with respect to each other:

$$\begin{aligned} SNR_1 &= 10 \cdot log_{10} \frac{Var(Y_{11})}{Var(Y_{12})} \\ SNR_2 &= 10 \cdot log_{10} \frac{Var(Y_{22})}{Var(Y_{21})}, \end{aligned} \quad (15)$$

where Y_{ij} stands for the signal on output i if the signal of speaker j is filtered with $\mathbf{W}(j\omega)$.

4.2. Experimental evaluation

The algorithm was tested on both recordings, which were first transformed to the frequency domain at a resolution of $N_{FFT} = 512$. For calculating of the spectrogram, the signals were divided into overlapping frames with a Hanning window and an overlap of $3/4 \cdot N_{FFT}$ and the STFT was then calculated. Since the recordings sample rate was 22kHz, this corresponds to a frame length of 23 ms with a 17 ms overlap. Parameters of the algorithm were set to an intitial stepsize $\nu = 25$, stepsize-division $m = 5$ and minimum stepsize $\nu_{min} = 0.5°$. With these settings, the number of necessary cost function evaluations was 2216 for one and 2188 for the other configuration. One resulting beampattern for configuration B, with the speaker signals impinging from $-45°$ and $25°$ are shown in Figure 5. As can be seen from the pattern, the zeros of

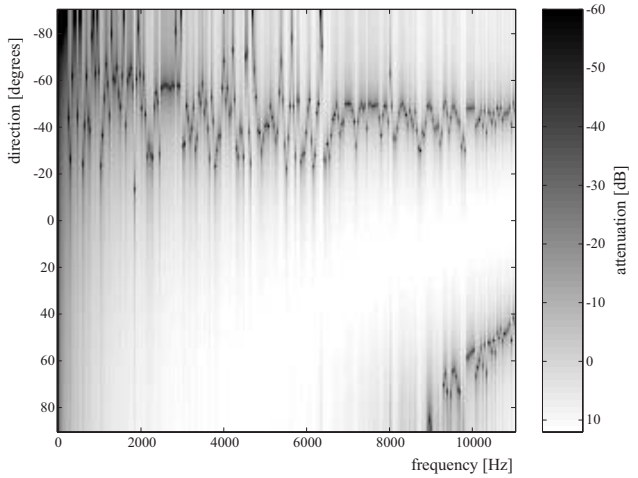

Fig. 5. First pattern for configuration B

the beamformer tend to an incoming direction of $-45°$, which is the main direction of the incoming interferer.

4.3. Comparison to other ICA algorithms

Two other methods for frequency domain convolved source separation were compared to the proposed algorithm. One was a frequency domain implementation of the JADE algorithm [4], the other was the convolutive source separation algorithm described in [5]. JADE was also applied in the frequency domain, with the spectrogram calculated also for $N_{FFT} = 512$ with an overlap of $3/4 \cdot N_{FFT}$. Permutations were corrected by minimizing the difference of null directions in adjacent frequency bins. For Parras algorithm, the time domain filter length was set to $Q = 128$. The resulting demixing-filters W_{JADE} and W_{Parra} were also applied to the single-source recordings to obtain an equivalent measure of SNR improvement for the three compared methods. Tables 2 and 3 show the comparison.

5. CONCLUSION

A new algorithm for blind separation of convolutive mixtures has been presented. It is based on frequency dependent null-steering, where the optimal angles are found using a combination of second and higher order statistics. This offers the advantage over conventional beamforming techniques that there is no need to know the exact direction of arrival.

The algorithm has been tested on real room recordings and has been compared to other standard algorithms. Its computational effort is, though no explicit gradient is available, in the same order of magnitude.

The results were evaluated in terms of SNR improvement. The model has proven to be sufficient for separating real room recordings, leading to an SNR improvement of up to 8 dB for two signals of male speakers. Thus the algorithm compares favorably to other blind methods for separation of convolved sources.

Table 2. SNR improvements for configuration A.

	JADE	Parra	Proposed Algorithm
SNR_1	3.6dB	4.7dB	7.8dB
SNR_2	5.4dB	-0.1dB	8.0dB

Table 3. SNR improvements for configuration B.

	JADE	Parra	Proposed Algorithm
SNR_1	5.4dB	3.9dB	6.9dB
SNR_2	5.6dB	1.4dB	5.0dB

Table 4. Average computation time.

	JADE	Parra	Proposed Algorithm
Config. A	57.6sec	79.4sec	96.0sec
Config. B	58.7sec	82.2sec	93.0sec

6. REFERENCES

[1] Lucas Parra and Christopher Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," in *IEEE Transaction on Speech and Audio Processing*, September 2002, vol. 10, pp. 352–362.

[2] Toshiya Kawamura Hiroshi Saruwatari and Kiyohiro Shikano, "Blind source separation based on fast-convergence algorithm using ica and array signal processing," in *ICA 2001*, 2001, pp. 412–417.

[3] A. Hyvärinen, "Fast and robust fixed-point algorithm for independent component analysis," *IEEE Trans. on Neural Networks*, vol. 10, pp. 626–634, 1999.

[4] J.-F. Cardoso, "High order contrasts for independent component analysis," *Neural Computation*, vol. 11, pp. 157–192, 1999.

[5] L. Parra and C. Spence, "Convolutive blind source separation of non-stationary sources," in *IEEE Trans. on Speech and Audio Processing*, May 2000, pp. 320–327.