COMBINED BEAMFORMING AND FREQUENCY DOMAIN ICA FOR SOURCE SEPARATION

Nilesh Madhu, André Gückel and Rainer Martin

{firstname}.{lastname}@rub.de
Institute of Communication Acoustics (IKA), Ruhr-Universität Bochum
44780 Bochum, Germany.

ABSTRACT

In this paper we propose a combined approach using beamforming and frequency domain ICA for the source separation problem in reverberant environments. The proposed method gains, and makes use of, the location of (any one) target source for permutation resolution. This approach is computationally less intensive and yields results comparable to current, state-of-the-art approaches. Additionally, we propose and justify the use of a non-linear post-processor to further improve crosstalk suppression in the recovered signals.

1. INTRODUCTION

Linear independent component analysis (ICA) has proven itself as a statistical tool for the blind demixing of (primarily) instantaneous mixtures of statistically independent random variables. The advantage of ICA lies in the fact that, given nothing more than the observations of the mixtures, it is nevertheless possible to obtain the underlying independent components, even obtaining an estimate of the mixing system in the process. However, the price to pay for this absence of *a priori* knowledge is the ambiguity associated with scaling and permutation. While these are less critical for time-domain separation algorithms, the resolution of these ambiguities are fundamental for proper recovery when using frequency domain approaches on broadband signals, where source separation is performed independently in each frequency bin.

ICA approaches utilizing beamformers in some way or another have been proposed before [1, 2, 3, 4]. In [1], the 'beam-pattern' of the demixing matrix is used to generate direction of arrival (DOA) information, subsequently used for permutation resolution; in [2], the demixing matrices select, for each frequency bin, the best option from a null-beam solution and an ICA solution, depending upon a quality criterion that is based upon the coherence function; [3] uses an anechoic approximation to the mixing model and forms a cumulant based cost function for optimizing the null-beamformers for the considered model, whereas [4] argues that proper initialization of the demixing matrices – using geometric constraints based on beamformers – obviates the need for permutation correction in the proposed frequency-domain, second-order statistics (SOS) algorithm. The DOA-based approach of [5] or the clustering approach of [6] for permutation solution are similar to the beam-pattern approach of [1] in that each cluster in the pattern is allocated to a source.

All the approaches above, however, do not utilize the beamformed signals themselves, and while a proper initialization of the demixing matrices as in [4] decreases the reconstruction error, the problem is not completely solved. The approach proposed in this paper makes use of the *null* and *direct path* characteristics of beamformers to generate *reference sources* for permutation resolution. This approach is computationally less expensive as compared to the method proposed in [7], while, at the same time, yielding comparable results. Another advantage of the proposed approach is that it is tolerant to position estimation errors and, further, the position of only one source need be estimated (for a 2×2 system).

The paper is organized as follows: first, the system model is introduced along with the ICA approach. The permutation inconsistency, inherent to ICA, is then briefly discussed. This is followed by an overview of our proposed approach. Finally, the results obtained using the proposed approach are compared with those using the approach of [7]. Additionally, a rather simple post-processor, based on binary masks, is introduced, and its application justified.

2. SYSTEM OVERVIEW

Consider a two-speaker, two-microphone setup in a room. When the speakers talk simultaneously, the (discrete-time) received mixture $x_l(n)$ at the microphones can be modelled as:

$$\begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} = \begin{bmatrix} a_{11}(n) & a_{12}(n) \\ a_{21}(n) & a_{22}(n) \end{bmatrix} * \begin{bmatrix} s_1(n) \\ s_2(n) \end{bmatrix}, \qquad (1)$$

where the $a_{lm}(n)$ represent the (discrete-time) room impulse responses (RIR) from source m to microphone l, and $*$ represents the convolution operator. As the mixing is not instantaneous, the time domain formulation does not lead to a practical application of the standardized ICA [8]. However, taking the short-time discrete Fourier transform, we obtain the following approximate representation at *each* frequency bin k:

$$\begin{bmatrix} X_1(k) \\ X_2(k) \end{bmatrix} = \begin{bmatrix} A_{11}(k) & A_{12}(k) \\ A_{21}(k) & A_{22}(k) \end{bmatrix} \begin{bmatrix} S_1(k) \\ S_2(k) \end{bmatrix}. \qquad (2)$$

Thus, the temporal, convolutive problem of (1) is now represented by N instantaneous mixtures, where N is the length of the Fourier transform. This transformation now allows us to apply ICA to each instantaneous mixture, a process that shall subsequently be referred to as frequency domain ICA (FDICA) [9]. Thus, FDICA optimizes, for each bin k, a demixing matrix $\mathbf{W}(k)$ such that:

$$\mathbf{Y}(k) \quad = \mathbf{W}(k)\mathbf{X}(k) \quad \approx \begin{bmatrix} S_1(k) \\ S_2(k) \end{bmatrix}. \qquad (3)$$

2.1. Caveats of FDICA

The solution presented in (3) is ideal. However, in the absence of any *a priori* knowledge of the mixing system, ICA solutions usually have the form:

$$\mathbf{Y}(k) \approx \mathbf{P}(k)\mathbf{D}(k)\mathbf{S}(k), \tag{4}$$

where $\mathbf{P}(k)$ is a 2×2 permutation matrix, and $\mathbf{D}(k)$ is a diagonal, scaling matrix.
Note that $\mathbf{P}$ and $\mathbf{D}$ are, in general, different in different frequency bins. We shall term a permutation and scaling as *local*, if it is for a particular frequency bin, and as *global*, if it is common to all bins. It is then obvious that source reconstruction from the spectral domain is only successful if the permutation and scaling matrices are global. In other words, $\mathbf{P}(k_a) = \mathbf{P}(k_b)$, and $\mathbf{D}(k_a) = \mathbf{D}(k_b)\ \forall\ a, b$.
While the scaling problem is rather reliably solved by the minimal distortion principle [10], permutation remains a rather formidable issue, especially due to its combinatorial nature. Recent literature contains various proposals to mitigate this problem using various properties of the speech signals, namely:

- inter-frequency amplitude envelope correlation (AmDeCor approach of [11])
- direction of arrival information (DOA) (DOA based approach of [1, 2, 5])

A recent approach that iteratively applies the above methods was proposed by [7]. Our experience was that while this approach is, indeed, rather robust, it is also computationally very expensive. In the following we propose our alternative which, instead of an iterative combination, demonstrates a 'one-step' approach. We shall assume that the azimuthal location of (at least one of) the sources is known. This information may be gained by, e.g., SRP-PHAT [12], clustering [13], Generalized Cross Correlation (GCC), or similar approaches.

3. THE COMBINED BEAMFORMING AND ICA APPROACH

3.1. Preprocessing

As in most ICA approaches, the search space is considerably reduced when standard preprocessing such as whitening and centering are performed. Indeed, it can be proven that such preprocessing constrains the search for the demixing matrix to a space of orthogonal (rotation) matrices [8]. However, before this is done, we introduce a spatial filtering of the signals.

3.1.1. Beamforming

Assume that, of the two sources, we know the approximate azimuthal location θ_t of one source – the 'target' source[1]. Then, the steering vector corresponding to the *direct* path from this position to the array would be:

$$\boldsymbol{\kappa}_t = [1\ \exp(-\jmath\,\omega_k d\cos(\theta_t)/c)]^T$$

where ω_k represents the k^{th} discrete frequency; d, the inter-microphone distance and c, the speed of sound in air. Correspondingly, we define two signals:

- the null-beam signal:

$$\hat{X}_{0,t}(k) = [1\ -\exp(\jmath\,\omega_k d\cos(\theta_t)/c)]\,\mathbf{X}(k)$$

- the direct-path signal: $\hat{X}_{d,t}(k) = \boldsymbol{\kappa}_t^H \mathbf{X}(k)$

Note that, except in anechoic environments, both signals still contain a mixture of the sources. However, the interferer would be predominant in $\hat{X}_{0,t}(k)$ (in the absence of grating lobes). This null-beam signal generates our 'reference-source' for permutation resolution, as will be shown in Section 3.3. Further, since the subsequent processing is done separately for each frequency bin, we shall drop the bin index k in the following.

3.1.2. Whitening

The input to this preprocessing stage is the composite vector $\hat{\mathbf{X}} = [\hat{X}_{0,\text{t}}\ \hat{X}_{d,\text{t}}]^T$ of the previous step. We use the standard PCA to de-correlate elements of the vector $\hat{\mathbf{X}}$ to obtain the whitened vector $\mathbf{U}$:

$$\mathbf{U} = \mathbf{R}_{\mathbf{X}}^{-1/2}\hat{\mathbf{X}}. \tag{5}$$

3.2. ICA

The vector $\mathbf{U}$ from (5) is input to the ICA stage. Due to the preprocessing, we now search for an orthogonal matrix $\mathbf{V}$ that decomposes the vector $\mathbf{U}$ into mutually independent components $\mathbf{Y} = \mathbf{V}\mathbf{U}$. Choosing the Kullback-Leibler distance as the cost function measure, and using the polar co-ordinate non-linearity of [14] to approximate the derivative of the probability density functions, we arrive at the following fast update rule:

$$\begin{aligned} \Delta\mathbf{V} &= \left(\mathbf{I} - E\left\{\boldsymbol{\varphi}(\mathbf{Y})\mathbf{Y}^H\right\}\right)\mathbf{V}, \\ \mathbf{V} &\leftarrow \Delta\mathbf{V}, \end{aligned} \tag{6}$$

where $\varphi(x) = \tanh(|x|)e^{\jmath\arg(x)}$, and $E\{.\}$ stands for the expectation operator. We found this particular choice of cost function and non-linearity to have the fastest convergence among the non-linearities proposed in [8, 9]

3.3. Permutation resolution

The result of the whitening and ICA steps is a scaled and permuted estimate of the underlying source components:

$$\mathbf{Y} \approx \mathbf{P}\mathbf{D}\mathbf{S}.$$

To resolve the permutation, we shall consistently assign the signal from the known location to channel 1 and the interferer to channel 2. For this, we use the null-beam signal as the reference signal. The required permutation is then such, that:

$$\mathbf{P}_\Pi = \underset{\Pi(i)}{\text{argmin}}\ \text{cor}\left(|(\mathbf{P}_{\Pi(i)}\mathbf{Y})_1|, |\hat{X}_{0,t}|\right), \tag{7}$$

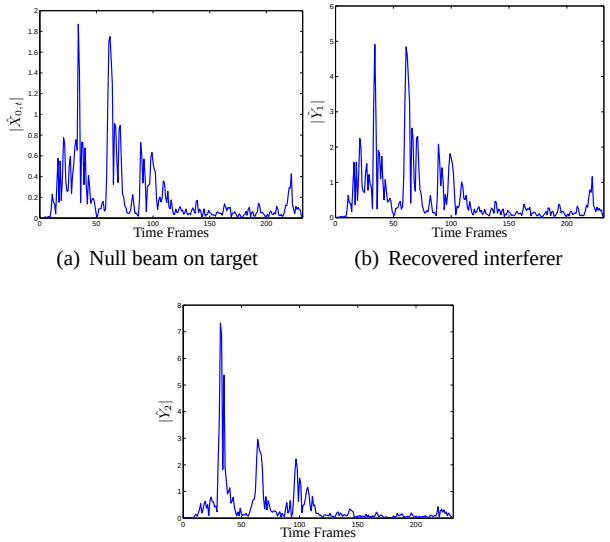
where $(\mathbf{P}_{\Pi(i)}\mathbf{Y})_1$ is the first element of the permuted output signal vector, corresponding to the permutation $\Pi(i)$. Thus (7) seeks that permutation that gives the *minimum* correlation of the amplitude envelopes of the null-beam signal $\hat{X}_{0,t}$ and the first channel of the aligned output signal $\mathbf{Y}_p$. This consideration arises from the fact that a null-beam should remove much of the direct path energy from the signal, S_t, resulting in amplitude

[1] the angle is measured with respect to the array axis

spectra that are least correlated with the corresponding recovered versions (Figure 1). The correlation is a normalized value and is defined for two variables x and y as:

$$\text{cor}(x, y) = \frac{E\{xy\} - \mu_x \mu_y}{\sigma_x \sigma_y}, \quad (8)$$

where μ represents the mean, and σ, the standard deviation of the variables.



(a) Null beam on target (b) Recovered interferer

(c) Recovered target

Figure 1: Amplitude envelopes at 1.0 kHz. Note the similarity between the envelopes of the interferer and the null beam.

The system schematic is presented in Figure 2. Once the permutation has been resolved, the minimum distortion principle [10] is used to solve the scaling. The time-domain signal can then be reconstructed, yielding the separated sources.

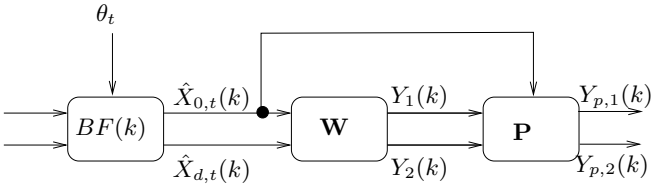


Figure 2: System Schematic. Illustrated for a particular frequency bin

4. NONLINEAR POSTPROCESSING

Time-frequency masking algorithms (e.g., [13, 15]), based on the disjoint nature of the supports of the short-time Fourier transform of speech signals, may be used to mask interferers at each time-frequency point. However, such masking requires some information to correctly allocate the time-frequency points in a mixture to any one source. One approach, using direction of arrival clustering, has been described in [13]. However, the disadvantage of this method is its rather poor performance in reverberant environments, as multipath propagation smears the time-frequency representations of the sources, invalidating the disjointness assumption.

Separating signals using ICA deconvolves the signals, in addition to separating them. However, the separated signals still contain crosstalk. This stems from the general inability of frequency domain ICA approaches to ideally demix sources in reverberant environments. But, at this stage, where the disjointness assumption is fulfilled to a better extent, it is possible to use masking approaches to further increase the interference suppression.

The simplest implementation of such a method would be, considering the time-frequency points (m, k) of the demixed, permutation and scaling aligned, signal vector $\mathbf{Y}_p(m, k)$ (where m represents the time-frame index, dropped for convenience in the previous discussion):

$$M_i(m, k) = \begin{cases} 1 & |Y_{p,i}(m, k)| > |Y_{p,3-i}(m, k)| \\ 0 & \text{Otherwise} \end{cases} \quad (9)$$

where $i \in \{1, 2\}$. We term this stage as the *nonlinear crosstalk canceller* (NCC).

5. EXPERIMENTAL SETUP AND RESULTS

The proposed algorithm, which we shall term Beamforming-ICA (BFICA), was compared vis-à-vis the robust and precise method (ICARP) of [7], and an 'ideal' method (ICAID), where the permutation ambiguity is resolved using the amplitude envelopes of the original sources. The ICAID approach represents the upper bound, in terms of performance, for the ICARP and BFICA algorithms. To keep the comparison fair, the core *ICA* algorithm of each approach is set to the one described in this paper. For the proposed approach, and the robust and precise method, the DOAs were given to the algorithms, and correspond to the values from the setup.

The experiments were conducted on data measured in a reverberant room ($T_{60} = 0.5$ s). The DFT size was taken as 1024 bins. The frame shift was 50%, and the sampling rate was 8000 kHz. Further, the data was tapered by a Hann window, before taking the Fourier transform. The microphones were omni-directional, and were placed 1 m from the sources, with an inter-microphone distance of 3 cm. The sources were selected from the TIMIT database and consisted of both male and female speakers. The algorithms were run for different combinations of speakers, over different azimuthal spacings (from 30° to 120°) between the sources, and over different positions of these *pairs* in the azimuthal plane. The results presented have been averaged over all the experiments.

The two instrumental measurement criteria selected for evaluating the algorithms were the Signal to Interference Ratio (SIR) *improvement* (SIRi) and the Signal to Distortion ratio (SDR). SIRi is computed as the difference between the Signal to Interference Ratio (SIR) *after* separation and the average SIR before separation (i.e., in the input signals). In each case the SIR and the SDR are computed in the time domain as proposed in [16]. The results are as shown in Figure 3.

Both, the proposed algorithm (ICABF), and the reference approach (ICARP), perform comparably. The ideal approach (ICAID) has the best performance (as expected), however, this approach still has some cross-talk (confirmed by listening tests and by the improvement in SIR due to post-processing). This is an indication of the degradation of ICA algorithms in reverberant and

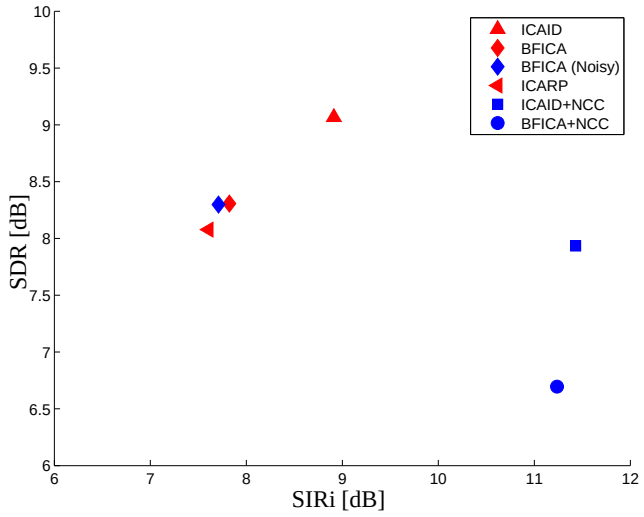


Figure 3: Experimental results in the SIR-SDR plane.

noisy environments. Results using the NCC stage, with binary masks, clearly indicate not only the reduction in cross-talk (increase in SIRi), but also the resultant degradation in the signal (SDR reduces). This highlights the inevitable trade-off between SIR and SDR – one cannot be improved without subsequently degrading the other.

Another interesting aspect of the proposed approach is its relative tolerance to DOA estimation errors in the localization stage. This was tested on the measured data, where the beamforming was done with corrupted DOA values (obtained by adding random Gaussian noise (of variance 5) to the actual DOA). The results (BFICA (Noisy)) clearly corroborate the tolerance of our approach.

6. CONCLUSIONS

This paper has introduced a combined beamforming and frequency domain ICA approach for source separation. The advantage of this method lies in its lower computational cost (as compared to current state-of-the-art approaches), its adroit utilization of beamformed signals for permutation resolution, and its tolerance to azimuth estimation errors. The results show that this approach is comparable to the state-of-the-art approaches. Further, the proximity of the SIR/SDR values for these approaches to the ideal approach show that, while there is still room for minor improvements in the area of ambiguity resolution, stronger focus should be on the ICA algorithm itself. Further, a rather simple post-processor using binary masks (NCC) was also proposed. Experiments show that NCC can improve the interference suppression by up to 4 dB, in individual cases. However, like all non-linear methods, this introduces an additional distortion in the recovered signals. The use of 'soft' masks, based on relative signal energy in the time-frequency plane reduces distortion and the signals sound more natural. This is an important aspect, as informal listening tests indicate that people prefer undistorted speech with more cross-talk, to distorted speech with less interference.

7. REFERENCES

[1] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind source separation," in *Proceedings of the ICASSP*, 2002, vol. I, pp. 881–884,

[2] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 666–678, Mar. 2006.

[3] W. Baumann, D. Kolossa, and R. Orglmeister, "Beamforming-based convolutive blind source separation," in *Proceedings of the ICASSP*, 2003, pp. 357–360,

[4] L. Parra and C. V. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, pp. 352–362, Sept. 2002.

[5] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Real-time blind source separation and DOA estimation using a small 3-D microphone array," in *Proceedings of the IWAENC*, Sept. 2005.

[6] S. Araki, H. Sawada, R. Mukai, and S. Makino, "A novel blind source separation method with observation vector clustering," in *Proceedings of the IWAENC*, Sept. 2005.

[7] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 530–538, Sept. 2004.

[8] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, New York, 2001.

[9] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, pp. 21–34, 1998.

[10] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proceedings of the International Conference on ICA*, Dec. 2001, pp. 722–727.

[11] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proceedings of the International Conference on ICA*, June 2000, pp. 215–220.

[12] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer Verlag, 2001.

[13] Ö. Yilmaz, A. Jourjine, and S. Rickard, "Blind separation of disjoint orthogonal signals: Demixing n sources from two mixtures," in *Proceedings of the ICASSP*, 2000.

[14] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A polar-coordinate based activation function for frequency domain blind source separation," in *Proceedings of the International Conference on ICA*, Dec. 2001, pp. 663–668.

[15] S. Rickard and Ö. Yilmaz, "On the W-Disjoint orthogonality of speech," in *Proceedings of the ICASSP*, May 2002.

[16] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *Proceedings of the ICASSP*, 2005, pp. 81–84.