BLISS

IST-1999-14190

Blind Source Separation and Applications

Technical report on implementation of linear methods and validation on acoustic sources

Deliverable D18

Report Version: Final
Report Preparation Date: 28. September 2003
Report Contributors: FhG, INPG, McMaster University
Classification: Public
Contract Start Date: 1 June 2000 Duration: 41 months
Project Co-ordinator: INPG
Partners: HUT, INPG, FhG, INESC, McMaster University

Project funded by the European Community under the "Information Society Technologies" Programme (1998-2002)

Technical report on implementation of linear methods and validation on acoustic sources

Stefan Harmeling, Paul v. Bünau, Andreas Ziehe, Dinh-Tuan Pham

Contents

1 Introduction

This report presents a comparison study of the methods, that have been implemented for the BLISS project, and the new method of Pham. For this purpose, we utilise the real-room recordings that McMaster University created as a contribution to the BLISS project.

After introducing the different algorithms (in Sec. 2), the database, a new performance measure for convolutive mixtures (proposed recently by FhG FIRST) and the experimental setup is explained (Secs. 3, 4, and 5). A detailed discussion of our findings follows (Sec. 6), and finally, a conclusion is given.

2 Methods for convolutive BSS

We compare INPG's new method to some of the state-of-the-art methods for convolutive BSS which are explained in the following. The algorithm of Murata, Ikeda, and Ziehe is available on Ikeda's website [5]. Futhermore, FhG FIRST implemented the algorithm of Parra and Spence and the method of Anemüller and made the code publicly available on the BLISS project website.

We consider the convolutive blind source separation problem: the source signals at time point t are given as a column-vector $s(t)$. The convolutive mixture is represented as a filter matrix A, i.e. each entry a_{ij} is a linear time-invariant filter. The mixed signals $x(t)$ can then be written as:

$$\begin{aligned} x_1(t) &= (a_{11} \star s_1)(t) + (a_{12} \star s_2)(t) \\ x_2(t) &= (a_{21} \star s_1)(t) + (a_{22} \star s_2)(t) \end{aligned}$$

or more compact

$$x(t) = (A \star s)(t). \tag{1}$$

The demixing system is also modeled as a filter matrix:

$$y(t) = (B \star x)(t) = ((B \star A) \star s)(t) \tag{2}$$

We call $(B \star A)$ the *global system*. The global system will be used to evaluate the performance of the different demixing algorithms.

2.1 Algorithm of Murata, Ikeda, and Ziehe

In a pioneering work [8], Murata, Ikeda, and Ziehe proposed to apply temporal decorrelation [7, 4, 13] for each subband in the frequency domain (see Fig. 1) in order to solve the convolutive blind source separation problem by exploiting the rich time-frequency structure of speech signals.

Recall from Eq. (1), that we model the n input signals $x(t)$ from m unknown sources $s(t)$ by the equation

$$x(t) = (A \star s)(t)$$

where A is an $n \times m$ matrix of filters and $\star$ denotes convolution. Hence, $x(t)$ is regarded as a linear combination of filtered sources. This translates to

$$\hat{x}(\alpha, t) \approx A(\alpha)\hat{s}(\alpha, t)$$

in the frequency domain, where we have $\hat{s}(\alpha, t)$, short-time Fourier transforms (abbr. STFT) of unknown sources using window length T starting at time t, mixed by the $n \times m$ matrix

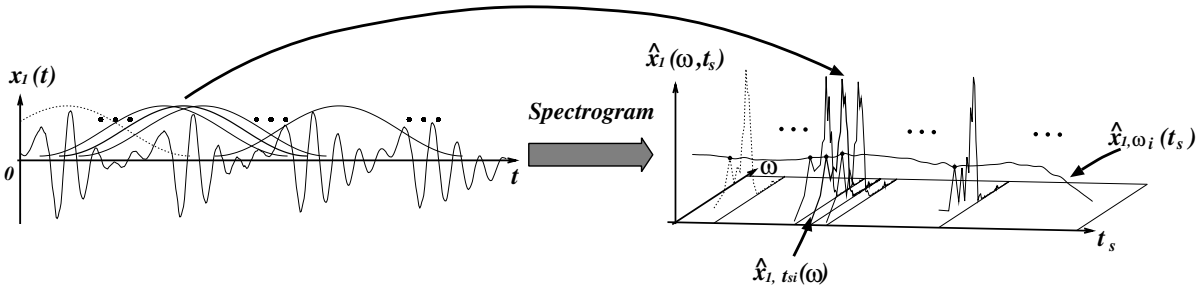

Figure 1: The specgram is calculated by applying the short-time Fourier transfrom to time-shifted windows.

$A(\alpha)$ of transformed filters to obtain the inputs $\hat{x}(\alpha, t)$. This approximation is good if the time frame length T is large enough to properly reflect the convolutive mixture A in time domain.

Because each frequency band is handled separately, the scaling and permutation indeterminacy (which is not a problem for instantaneous mixtures) has to be taken care of to correctly reconstruct the demixed signals. Murata eta al. [8] solve the scaling indeterminacy by the assumption that the energy of the signal does not change by the mixing process, i.e. by back-projecting the demixed signals onto the sensor space. The permutation ambiguity is resolved using the large-scale temporal structure of speech signals. Fig. 2 illustrates the structure of this approach. More details can be found in [8, 9]. For the comparison in this paper we use the MATLAB code from Ikeda's website [5].

2.2 Algorithm of Parra and Spence

The algorithm by Parra and Spence (as described in [10]) also tries to find demixing matrices in the frequency domain similar to Murata, Ikeda, Ziehe's approach. In particular, Parra and Spence's method utilizes a joint diagonalization of time-shifted cross-power spectra between the input signals, which is carried out by standard gradient-based optimization. The permutation problem is solved by restricting the number of non-zero filter taps in the time domain, which leads to smooth results in frequency domain.

The goal is to find a set of matrices $\{W(\alpha)\}$ that simultaneously diagonalize the cross-power spectra

$$\bar{R}_x(\alpha, t) = \frac{1}{N} \sum_{n=0}^{N-1} \hat{x}(\alpha, t + nT) \hat{x}(\alpha, t + nT)^H.$$

For each frequency, we consider K spectra estimated by N non-overlapping short-time Fourier transforms, so that the k-th cross-power spectrum to diagonalize for frequency channel α is $\bar{R}_x(\alpha, kTN)$.

With $\|x\|$ being the L_2-norm of x, we measure the error by

$$E(\alpha, k) = \left\| W(\alpha)\ \bar{R}_x(\alpha, kTN)\ W^H(\alpha) \right\|^2.$$

The optimal diagonalizers $\left\{\hat{W}_\alpha\right\}$ are then found by minimizing the error:

$$\hat{W}_\alpha = \arg\min_{W(\alpha)} \sum_{k=0}^{K-1} E(\alpha, k)$$

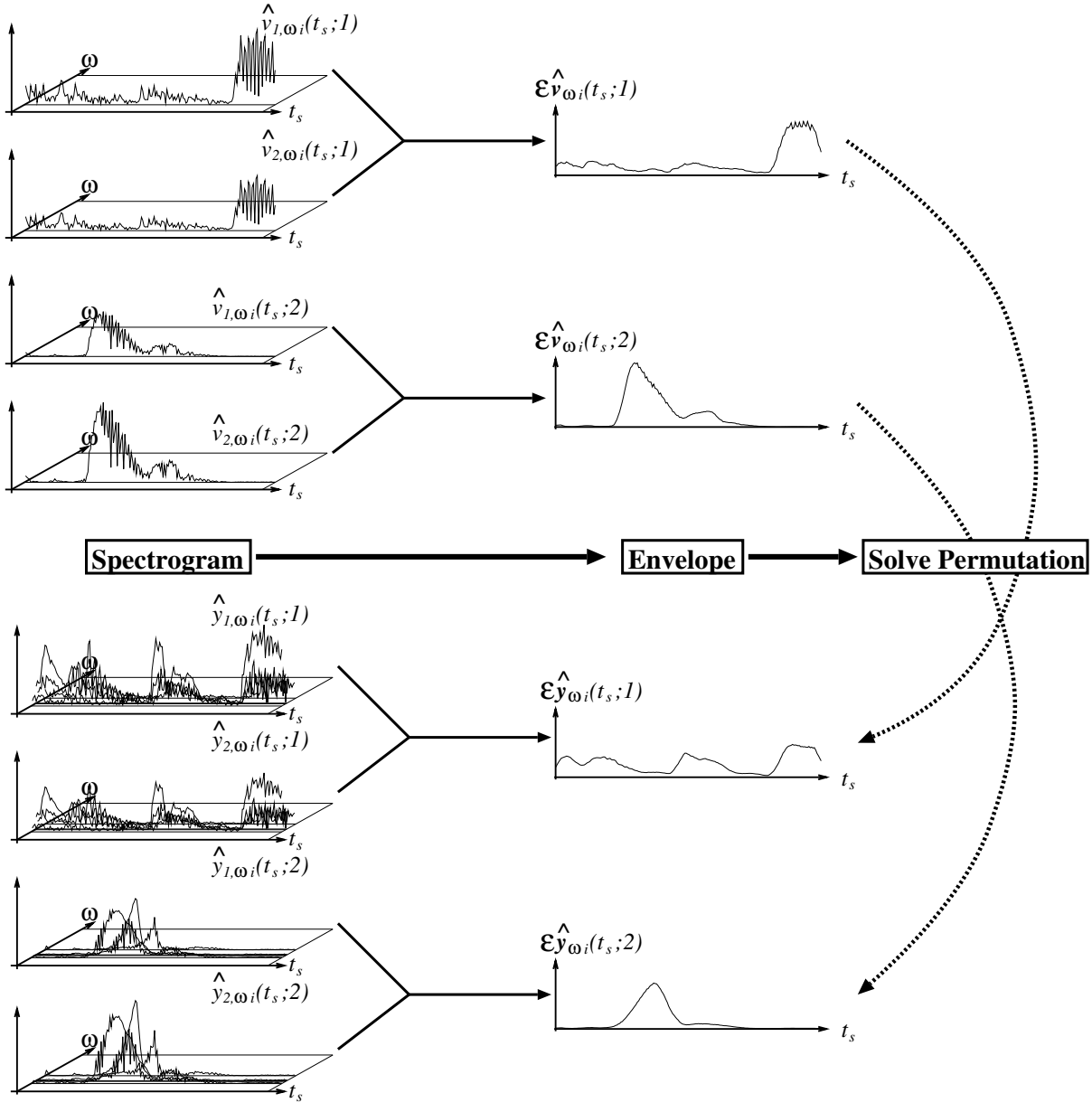

Figure 2: Exploiting the correlations between the envelopes of the signals in each frequency band allows to solve the permutation problem.

The permutation problem is solved by posing an additional constraint to this optimization problem: the taps of each filter in time-domain should be zero for indices $i > Q \ll T$. This ensures smooth impulse responses in the frequency domain which penalizes permutations. In practice, this condition is enforced by projecting the preliminary solution onto the subspace of legal diagonalizers after each optimization step (for further details see [10]).

2.3 Algorithm of Anemüller

The algorithm of Anemüller [3, 2] is based on the observation, that for speech signals amplitude changes in neighboring frequency channels are correlated. Hence, a useful criteria for unmixing speech signals is as follows: find source signals whose frequency channels are correlated, but not inter-correlated across different sources. Using a cost function that quantifies inter-correlation, the algorithms finds a set of demixing matrices (one for each frequency channel) by standard numerical optimization tools. In the following, we briefly describe Anemüller's method, formulated in a concise matrix notation which allows straightforward implementation.

First of all, we cast the demixing problem in the frequency domain (as the algorithm of Murata, Ikeda, Ziehe). From the n input signals (or microphones) $x_i(t)$ we compute STFTs $\hat{x}_{i,\alpha}(\hat{t})$ for each frequency channel α (using short-time Fourier transforms with an appropriate window function). The goal is to find a set of $m \times n$ matrices $\left\{\hat{W}_\alpha\right\}$ for computing the unmixed STFTs

$$\left[\hat{u}_{1,\alpha}(\hat{t}) \cdots \hat{u}_{m,\alpha}(\hat{t})\right]^\top = \hat{W}_\alpha \left[\hat{x}_{1,\alpha}(\hat{t}) \cdots \hat{x}_{n,\alpha}(\hat{t})\right]^\top.$$

The unknown sources are then easily computed from $\hat{u}_{i,\alpha}(\hat{t})$ by using the overlap-add method.

For measuring the amplitude correlation between frequency channels (abbr. AMCor), we use the covariance of their magnitude spectra $\left|\hat{u}_{i,\alpha}(\hat{t})\right|$. With $\operatorname{cov}\{\cdot\}$ denoting the covariance, AMCor between frequency channels α, β and sources i, j is expressed as

$$c_{\alpha,\beta}^{i,j} = \operatorname{cov}\left\{\left|\hat{u}_{i,\alpha}(\hat{t})\right|, \left|\hat{u}_{j,\beta}(\hat{t})\right|\right\}.$$

By computing AMCor for all frequencies, we obtain correlation matrices $C^{i,j}$ for each pair (i, j) of sources. The cost function H is the sum of squared correlations between different sources for all frequency channels, which can be written as

$$H\left(\left\{\hat{W}_\alpha\right\}\right) = \sum_{\alpha,\beta,i,j\neq i} \left(c_{\alpha,\beta}^{i,j}\right)^2$$

or

$$\tilde{H}\left(\left\{\hat{W}_\alpha\right\}\right) = \sum_{i,j\neq i} \operatorname{tr}\left(\left[C^{i,j}\right]^\top C^{i,j}\right)$$

where we discarded the normalization of the covariance estimate.

H penalizes $\left\{\hat{W}_\alpha\right\}$ for amplitude correlations across different frequencies. Hereby, H also prevents local permutations: imagine there are channels α, β incorrectly assigned to sources i, j. Then we have $c_{\alpha,\beta}^{i,j} > 0$ (with $i \neq j$) which will add to the costs.

Experiments show that optimizing over all $\hat{W}_\alpha$ at once is both expensive and prone to get stuck in local minima. Considering only a single $\hat{W}_\alpha$ while fixing all $\hat{W}_\beta$ ($\beta \neq \alpha$) proved to be a good heuristic.

An iteration of the algorithm consists of a full optimization sweep over all demixing matrices $\hat{W}_\alpha$. The frequency channels $\alpha_1 \dots \alpha_l$ are processed in the following order:

1. Find the channel α_k that has the maximum signal energy. Optimize $\hat{W}_{\alpha_k}$.
2. Consider frequency channels in ascending order ($\alpha_{k+1} \dots \alpha_l$) until the highest frequency α_l is reached.
3. Optimize $\hat{W}_{\alpha_i}$ for the lower half ($\alpha_{k-1} \dots \alpha_1$) of frequency channels in descending order.

As we fix all demixing matrices while optimizing a single $\hat{W}_\alpha$, the computational effort for evaluating the cost-function H can be signifcantly reduced. More precisely, when considering frequency channel α, all terms $c^{i,j}_{\gamma,\delta}$ of AMCor with $\gamma \neq \alpha$ and $\delta \neq \alpha$ remain constant and are thus neglectable. In the following, we show how this is implemented in practice. From now on, we use matrix notation and limit the presentation to the $m = n = 2$ case. The extension to an arbitrary number of sources and microphones is straightforward.

We extend the cost-function by adding a set of m matrices $\{U_1 \dots U_m\}$ as arguments. These matrices represent centered magnitude spectra for each source, that are not subject to optimization but necessary for computing the covariances. Frequency channels go along the rows and time increases column-wise. Formally speaking

$$U_k = \Xi \left[\left[\hat{u}_{k,\alpha_1}(\hat{t}) \cdots \hat{u}_{k,\alpha_l}(\hat{t}) \right]^\top \right]$$

where we introduced Ξ as an abbreviation for centering the rows that can be written as

$$\Xi [A] = A - A\frac{1}{c}1^{c\times c}$$

with $1^{c\times c}$ as a $c \times c$ matrix of ones and c being the number of rows of A.

Before formulating the cost-function H, we introduce some further notations: by e_i we denote the i-th standard unit column vector, E_n is the $n \times n$ identity matrix and E^i_n is E_n with the (i,i)th element set to zero. We use x^2 for $xx^\top$.

Another abbreviation makes the formulas easier to read: we use ω_k for the centered absolute value of the demixed STFTs in the frequency channel α_k. That is

$$\omega_k = \Xi \left[\left| \hat{W}_{\alpha_k} \left[\hat{x}_{1,\alpha_k}(\hat{t}) \cdots \hat{x}_{n,\alpha_k}(\hat{t}) \right]^\top \right| \right].$$

In order to make the cost-function more concise, we break it into pieces using the following definition: $\xi^k_{i,j}$ is the sum of unormalized squared AMCor between frequency α_k of source i and all channels α_z of source j where $z \neq k$, that is

$$\begin{aligned} \left(\xi^k_{i,j}\right)^2 &= n_{\hat{t}}^2 \sum_{z\neq k} \left(c^{i,j}_{\alpha_k,\alpha_z}\right)^2 \\ &= \left(e_i{}^\top \omega_k U_j{}^\top E^k_l\right)^2 \end{aligned}$$

where $n_{\hat{t}}$ denotes the number of time-slices obtained by the STFTs.

Using these notations, the cost-function for the $m = n = 2$ can be written as:

$$\tilde{H}_{U_1,U_2,\hat{x}_{1,\alpha_k}(\hat{t}),\hat{x}_{2,\alpha_k}(\hat{t}),k}\left(\hat{W}_{\alpha_k}\right) = \left(\xi_{1,2}^k\right)^2 + \left(\xi_{2,1}^k\right)^2 + \left(e_1{}^\top \left(\omega_k\right)^2 e_2\right)^2.$$

The last term adds the AMCor $\left(c_{\alpha_k,\alpha_k}^{1,2}\right)^2$ which is not included in $\xi_{1,2}^k$ and $\xi_{2,1}^k$.

Though the latter formulation of $\tilde{H}$ seems a litte bit obfuscated, it allows an simple gradient calculation and helps for a straightforward implementation. Let us first introduce the gradient of

$$\psi_{Q,x}\left(W\right) = Q\left|Wx\right|$$

which is

$$\begin{aligned}
\mathrm{Re}\left[\nabla\psi_{Q,x}\right]\left(W\right) &= \mathrm{Re}\left[x\right]\left(\left(Q \odot \frac{1}{\left|Wx\right|^\top}\right)\mathrm{Re}\left[Wx\right]\right) + \mathrm{Im}\left[x\right]\left(\left(Q \odot \frac{1}{\left|Wx\right|^\top}\right)\mathrm{Im}\left[Wx\right]\right) \\
\mathrm{Im}\left[\nabla\psi_{Q,x}\right]\left(W\right) &= -\mathrm{Im}\left[x\right]\left(\left(Q \odot \frac{1}{\left|Wx\right|^\top}\right)\mathrm{Re}\left[Wx\right]\right) + \mathrm{Re}\left[x\right]\left(\left(Q \odot \frac{1}{\left|Wx\right|^\top}\right)\mathrm{Im}\left[Wx\right]\right).
\end{aligned}$$

We get

$$\begin{aligned}
\mathrm{Re}\left[\nabla\tilde{H}_{U_1,U_2,\hat{x}_{1,\alpha_k}(\hat{t}),\hat{x}_{2,\alpha_k}(\hat{t}),k}\left(\hat{W}_{\alpha_k}\right)\right] &= \mathrm{Re}\left[\nabla\psi_{Q,x}\right]\left(\hat{W}_{\alpha_k}\right) \\
\mathrm{Im}\left[\nabla\tilde{H}_{U_1,U_2,\hat{x}_{1,\alpha_k}(\hat{t}),\hat{x}_{2,\alpha_k}(\hat{t}),k}\left(\hat{W}_{\alpha_k}\right)\right] &= \mathrm{Im}\left[\nabla\psi_{Q,x}\right]\left(\hat{W}_{\alpha_k}\right)
\end{aligned}$$

with

$$\begin{aligned}
Q &= 2\left(I_{n_{\hat{t}}} - 1^{n_{\hat{t}}\times n_{\hat{t}}}\right)\left[U_2{}^\top {\xi_{1,2}^k}^\top e_2{}^\top + U_1{}^\top {\xi_{2,1}^k}^\top e_1{}^\top + \left[e_1{}^\top\left(\omega_k\right)^2 e_2\right]\omega_k{}^\top\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\right] \\
x &= \left[\hat{x}_{1,\alpha_k}(\hat{t})\ \hat{x}_{2,\alpha_k}(\hat{t})\right]^\top
\end{aligned}$$

as the gradient of our cost-function (more details can be found Anemüller's PhD thesis [2]).

2.4 Algorithm of Pham

D. T. Pham (INPG) considers the blind separation of convolutive mixtures based on a new highly-efficient joint diagonalization algorithm (developed in the BLISS project workpackage 1) of time varying spectral matrices of the recorded signals, hereby optimally exploiting both: non-stationarity and non-whiteness. In order to eliminate the permutation ambiguity, the novel method relies on the continuity of the frequency response of the filter. The method achieves the separation of audio mixtures in which the mixing filter has quite long impulse responses. The details of the algorithm of Pham from INPG can be found in the two attached papers [11, 12].

3 Database of real-room recordings

McMaster University created in the BLISS project a comprehensive database of real-room recordings [6]: live-capture audio mixtures and a realistic hearing in noise test environment (R-HINT-E). A human head and torso model called KEMAR was placed in the centre of three different rooms:

pc323dc: a semi-reverberent 10'x10' room with carpeted floor and acoustically treated ceiling panels and with 2 x 24 oz velour drapes independently hung around the perimeter of the room to minimise reflections

pc323do: the same pc323dc 10'x10' room without the 2 x 24 oz velour drapes independently hung around the perimeter of the room

pc335: a semi-reverberent classroom located in the Psychology Department at McMaster University

KEMAR has in each ear a small microphone array with three microphones. A single loudspeaker was moved to different locations around KEMAR (with different angles, heights and distances). For each location the room impulse response was measured. Using these room responses and some source signals, many different mixtures can be produced by convolving the sources (for more details see [6]).

One might object that convolving source signals with the recorded room responses lead to different signals, than if we recorded the source signals *directly* in the room. To study this, six sentences were played from all different locations, really recorded (in addition to measuring the room responses) and compared to the signals obtained by convolution with the corresponding room responses. In [6]), Trainor et al. show that the measured and the convolved sentences are in very good agreement.

4 Validation methods

In order to enable an informative comparison, a meaningful performance measure is needed. In the following, we introduce a new performance measure for convolutive mixtures that has been recently developed at FhG FIRST. This performance index analyses the global system $(B \star A)$ (see Eq. (2)), which is available to our experimental setup, since our comparison is based on McMaster's database that provides us with the true mixing matrix A.

4.1 Indeterminacies of convolutive BSS

A useful performance measure should be invariant to the indeterminacies of the blind source separation problem. For the convolutive case these are:

1. Arbitrary permutation, i.e. exchanging the rows of B does not change the quality of the solution.

2. Arbitrary scaling, i.e. multiplying some row of B by some factors does not change the quality of the solution either.

3. Arbitrary filtering, i.e. applying some filter to some extracted signal does not change the quality of the separation. However, the sound quality might be influenced. But since a real-world mixing system (the room impulse function) is often not invertible, the demixed signals are usually filtered version of the source signals.

Obviously, the second point—scaling—can be seen as a special case of the filtering indeterminacy. We distinguish those two, because they are dealt with differently: the scaling indeterminacy is evened out, but the filtering indeterminacy is only circumvented (see next paragraph).

4.2 Amari index for convolutive mixtures

The global system, the filter matrix $C = B \star A$, is the basis of our performance measure which comprises of the following steps:

1. In order to overcome the scaling indeterminacy we normalize the rows of C: stack all filters of a row of C into one long vector and normalize its norm to one. Then distribute the parts of that vector to the original row entries. Doing this for all rows of C evens out the scaling indeterminacy.

2. Calculate for each filter in the filter matrix C, the norm (viewing each filter as a vector and assuming they have the same length to insure comparability). These norms can be arranged in a matrix $\widetilde{C}$ which has the same size as C, but the entries of which are numbers instead of filters. $\widetilde{C}$ summarizes the proportions that each true source s_j contributes to the estimated sources y_i. This circumvents the filtering indeterminacy, since $\widetilde{C}$ can not distinguish whether the contribution of a source s_j to the estimated source y_i originates from a filter in C_{ij} that is just a delta peak or from a more complicated filter. In *realistic* real-room mixtures, obtaining a filtered version of the true sources is most likely all we can hope for, because the room response function might not be invertible. However, if an entry in $\widetilde{C}$ is small or close to zero, we can be sure that there was no contribution from the corresponding source. Therefore, to assess the degree of separation, we measure in the next step how close $\widetilde{C}$ is to a permutation matrix.

3. Calculate the Amari index of $\widetilde{C}$, i.e. normalize the rows of $\widetilde{C}$ (if necessary) and calculate (as suggested by Amari [1]):

$$\sum_i \left(\sum_j \frac{|\widetilde{C}_{ij}|}{\max_k |\widetilde{C}_{ik}|} - 1 \right) + \sum_j \left(\sum_i \frac{|\widetilde{C}_{ij}|}{\max_k |\widetilde{C}_{kj}|} - 1 \right).$$

   The Amari index is small when each row and each column is dominated by one large element, i.e. if the evaluated matrix is close to a permutation matrix.

We call this performance index defined by the above steps *Amari index for convolutive mixtures*. Note, that the Amari index for convolutive mixtures is zero if and only if there are for an estimated source no contributions from the other sources, which coincides with the notion that all sources are separated from each other.

5 Experimental setup

For the comparison study in this paper, we employ McMaster University's hearing database (R-HINT-E) to create mixtures with two sources originating from different directions (all

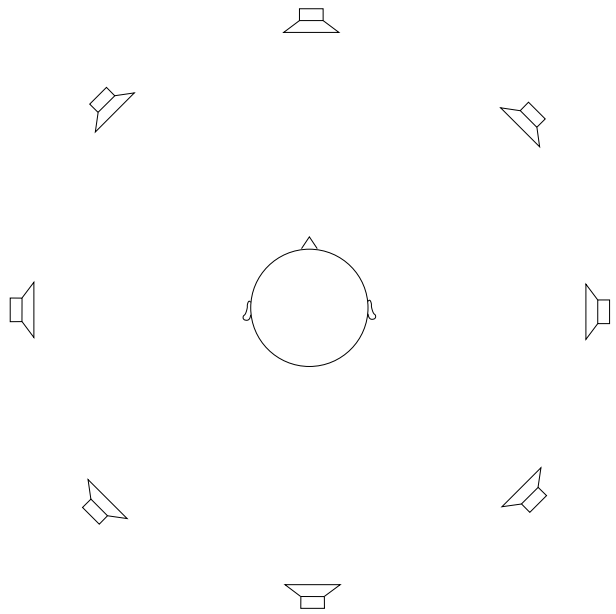

Figure 3: The room impulse response function are measured for each angle and each microphone in the ears.

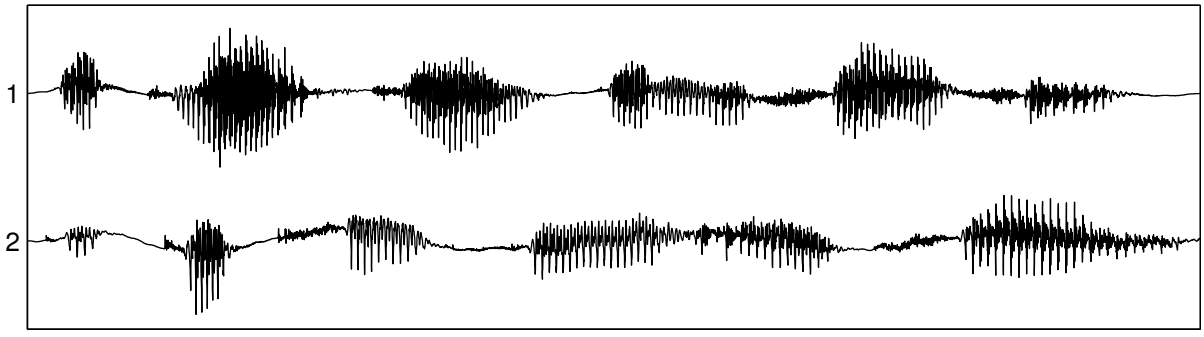

Figure 4: The speech signals used in the experiments.

combinations of the angles: 0, 45, 90, 135, 180, 225, 270, 315 degree, see Fig. 3) at a fixed distance (3 feet) and a fixed height (4 feet). This corresponds to 28 different mixtures (and therewith experiments) per room (without repetitions and without equal angles). Since some of the algorithms can only deal with two microphones we restrict ourselves to one microphone for each ear (in the database called left1 and right1). We use speech source 1 and 2 from the speech files supplied with the database (about 18,000 samples, see Fig. 4). All signals are downsampled to 11025 Hz. Note that all four algorithms use short-time Fourier transforms. The length of the Fourier transform has been fixed for all algorithm to 512 to allow a fair comparability.

6 Results

Fig. 5 is an example how we visualized our results: shown are the Amari indices of all different angle combinations for the room 323dc. The gray level of each of the 64 squares encodes the Amari index of the corresponding combination of angles. We observe:

- On the main diagonal the value is 4 since that is the maximal Amari index for a 2×2 mixing matrix. Two signals coming from exactly the same direction can be considered like one single channel, which is not separable with the ICA approaches presented in this study.
- Combining the angles 45, 90 and 135 with 45, 90, 135 degrees leads to large Amari indices (dark areas) as well, indicating that if both sources are on the same side of the head (in this case at the right side) the sources arrive very well mixed at the ears. This applies analogously to the left side for combinations of the angles 225, 270, 315 with 225, 270, 315.
- Combining a source from the left side (angles 45, 90, 135) with a source from the right side (angles 225, 270, 315) are not very much mixed (light areas). The reason for this is that the signals are quite well separated due to acoustic shielding by the KEMAR head.

The first row of Fig. 6 shows the Amari indices of the mixtures for all three rooms (the upper left panel is a small version of Fig. 5). These plots represent the initial situations. The panels of each further row show the results after applying one of the algorithms. Especially, for the hearing-aids-scenario it is important to know how much we can improve upon the initial situations (the panels in the first row of Fig. 6), i.e. how close we get to a plot in which every box in the off-diagonal is white (i.e. perfect separation):

parra: Comparing the panels in the second row (parra) with the panels in the first row, we observe that the squares changed towards a lighter gray. In other words, the algorithm of Parra and Spence is able to improve the separation in all three cases. Especially, if the sources are on different sides of KEMAR, the separation is almost perfect.

murata: The algorithm of Murata, Ikeda, and Ziehe fails on the dataset from McMaster University: the panels in the third row (murata) show very dark squares, even darker than the ones in the first row, indicating that after applying this algorithm

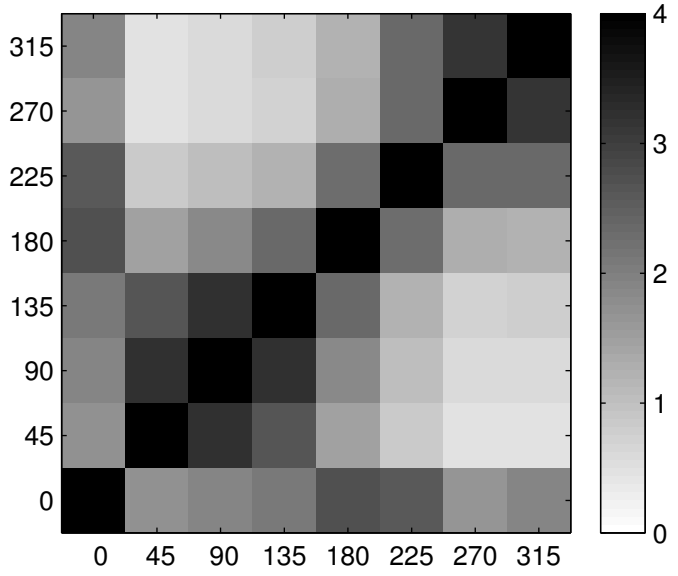

Figure 5: The Amari indices for all different angles in room PC323dc.

the data is even more mixed than before. The reason for such an behaviour might be that the method, which should solve the permutation, miscarries due to the short signal length or due to correlations among the envelopes of the two sources.

pham: The panels in the fourth row show the best results: the off-diagonal squares are almost white, implying that the algorithm of Pham is able to largely improve the degree of separation. Compared to the other rows (showing the results for the other algorithm) it is clear that Pham's algorithm performed best. However, comparing the three panels in the fourth row (Pham) shows that the different rooms (PC323dc, PC323do, PC335) pose problems of increasing difficulty. The left panel (PC323dc, low reverberation) shows almost a diagonal (the simplest mixture), and the right panel (PC335, class room, high reverberation) shows some darker squares which indicate that the demixing is not perfect.

anemueller: The results of the algorithm of Anemueller (fifth row) are similar to the results of Murata, Ikeda, and Ziehe's algorithm. The reason for the failure might be that the used speech signals are quite short (about 18,000 samples), so that there might not be enough statistics to estimate the cross-frequency correlations properly. Therefore, the algorithm probably failed to solve the permutation problem across the frequencies.

Obviously, letting the algorithms find longer filter would lead to better results. But to study and compare the different proposed algorithms (and to limit computing time), short filters (length of the Fourier transform 512) were chosen. Furthermore, the performance could be increased by exploiting the signals of all six microphones which were recorded in McMaster Unversity's database.

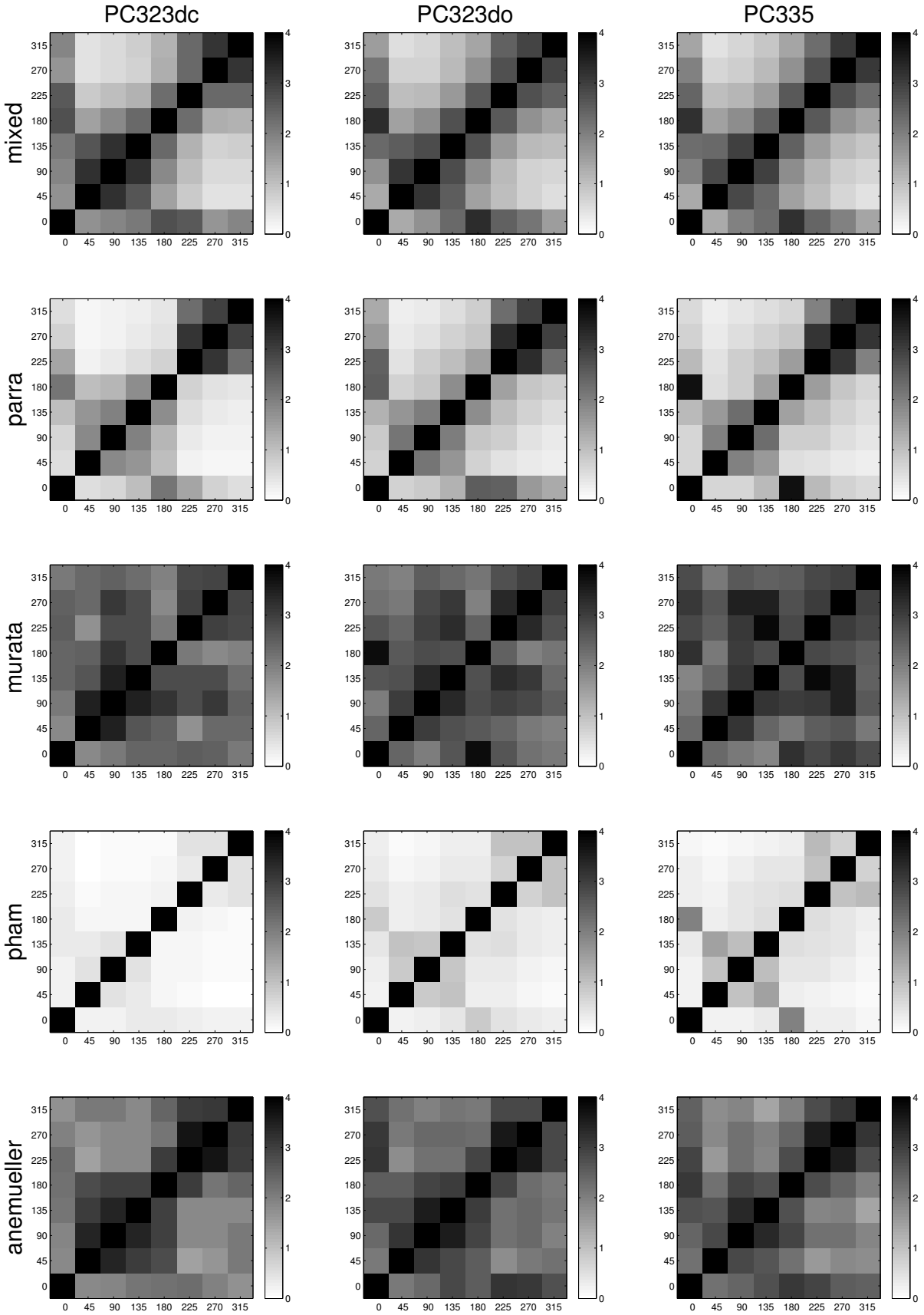

Figure 6: Each panel presents the results for one special algorithm applied to the data recorded in one of the three rooms. The plot in each panel visualizes the Amari indeces for the different angles.

7 Conclusion

The algorithm developed at INPG (Pham), combining a powerful diagonalization algorithm (from BLISS project workpackage 1) and a novel, efficient solution to the permutation problem, has the best performance among the four tested algorithms.

The database created by McMaster University is very useful for comparison studies of algorithms, because it allows systematic experiments with a broad variety of setups, hereby allowing in combination with the Amari index for convolutive mixtures (developed at FhG FIRST) to differentiate the algorithms' performances in a meaningful way. Many other comparison studies are conceivable using McMaster's database. It has the potential to become *the benchmark* for algorithms that tackle the Cocktail-party problem.

Future work can now be based on the achievements of the BLISS project: all implemented methods and McMaster's database are available online to the public.

8 Publications included in the deliverable

This deliverable includes two attached papers that describe the new method for convolutive blind source separation developed by INPG (Pham):

Publication 1 [**11**] D. T. Pham, Ch. Servière, and H. Boumaraf, "Blind separation of speech mixtures based on nonstationarity", in *Proceedings of ISSPA 2003 conference*, 2003.
Pham considers the blind separation of convolutive mixtures based on the joint diagonalization of time varying spectral matrices of the observation records. The goal is to separate audio mixtures in which the mixing filter has quite long impulse responses and the signals are highly non stationary. In order to eliminate the permutation ambiguity, the method relies on the continuity of the frequency response of the filter. Simulations show that the method works well when there are no strong echos in the mixing filter. But if it is not the case, the permutation ambiguity cannot be sufficiently removed.

Publication 2 [**12**] D. T. Pham, Ch. Servière, and H. Boumaraf, "Blind separation of convolutive audio mixtures using nonstationarity", in *Proc. of 4th Int. Symp. on Independent Component Analysis and Blind Source Separation (ICA2003)*, pages 257–262, Nara, Japan, April 2003.
This paper presents a method for blind separation of convolutive mixtures of speech signals, based on the joint diagonalization of the time varying spectral matrices of the observation records and a novel technique to handle the problem of permutation ambiguity in the frequency domain. Simulations show that the method works well even for rather realistic mixtures in which the mixing filter has a quite long impulse response and strong echos.

References

[1] S.-I. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind source separation. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 757–763. MIT Press, Cambridge, MA, 1996.

[2] J. Anemüller. *Across-frequency processing in convolutive blind source separation.* PhD thesis, Universität Oldenburg, 2001.

[3] J. Anemüller and B. Kollmeier. Amplitude modulation decorrelation for convolutive blind source separation. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 215–220, Helsinki, Finland, 2000.

[4] A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on Signal Processing*, 45(2):434–444, 1997.

[5] S. Ikeda, N. Murata, and A. Ziehe. MATLAB code for convolutive blind source separation. *http://www.ism.ac.jp/˜shiro/research/blindsep.html*, 2001.

[6] K. Wiklund J. Bondy S. Gupta S. Becker I. C. Bruce S. Haykin L. Trainor, R. Sonnadara. Development of a flexible, realistic hearing in noise test environment (r-hinte). *Signal Processing*, to appear.

[7] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72:3634–3636, 1994.

[8] N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. Technical Report 98-2, RIKEN BSIS, 1998.

[9] N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1-4):1–24, August 2001.

[10] L. Parra and C. Spence. Convolutive blind source separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, pages 320–327, May, 2000.

[11] D. T. Pham, Ch. Servière, and H. Boumaraf. Blind separation of speech mixtures based on nonstationarity. In *Proceedings of ISSPA 2003 conference.*

[12] D. T. Pham, Ch. Servière, and H. Boumaraf. Blind separation of convolutive audio mixtures using nonstationarity. In A. Cichocki and N. Murata, editors, *Proc. of 4th Int. Symp. on Independent Component Analysis and Blind Source Separation (ICA2003)*, pages 257–262, Nara, Japan, April 2003.

[13] A. Ziehe and K.-R. Müller. TDSEP—an efficient algorithm for blind separation using time structure. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN'98)*, pages 675–680, Skvde, Sweden, 1998.