BLIND SPEECH SEPARATION USING PARAFAC ANALYSIS AND INTEGER LEAST SQUARES

Kleanthis N. Mokios, Nicholas D. Sidiropoulos and Alexandros Potamianos

Dept. of Electronic and Computer Engineering Technical University of Crete, 73100 Chania, Greece

{ nikos,kleanthis,potam}@telecom.tuc.gr

ABSTRACT

We propose a new two-step frequency domain algorithm for blind speech separation (BSS) for unknown channel order. This new approach employs parallel factor analysis (PARAFAC) to separate the speech signals and a novel integer-least-squares-based method for matching the arbitrary permutations in the frequency domain. The proposed algorithm offers guaranteed convergence and good separation performance, measured both quantitatively and in subjective tests. Performance gains in signal-to-interference ratio of up to 10 db are achieved for certain source-sensor geometries.

1. INTRODUCTION

One of the key problems in speech processing for teleconferencing and mobile telephony applications is that of speaker separation from multichannel measurements. Such multichannel measurements provide the opportunity for speaker separation or speaker-background noise separation, which can be crucial for intelligibility. The objective of the "blind" speech separation problem is to separate the multiple source signals, using only the measured microphone signals, i.e., without exploring any additional information about the properties of the sources or the mixing channels.

In recent years, several methods have been developed to tackle the BSS problem, but only few of them exploit the inherent non-stationarity of the speech signals. The majority of the proposed methods that exploit non-stationarity, consider the simple case of instantaneous linear mixtures [3]. For the more general case of convolutive linear mixtures, the approaches can be divided into frequency domain [1], [2] and time domain methods [4]. In this paper we propose a new frequency domain technique that exploits non-stationarity of speech signals and deals with the more general case of convolutive linear mixtures. Similar approaches have been proposed in [1], [2]. The method in [1], however, employs a least-squares (LS) gradient descent procedure in order to separate signals, which is not guaranteed to converge. On the contrary, our technique is based on the formulation of the BSS problem as a conjugate symmetric PARAFAC model that is fitted optimally using a fast and monotonically convergent alternating LS algorithm. Furthermore, the conjugate symmetric PARAFAC model exhibits strong identifiability properties that allows the separation of a much higher number of signals than what is suggested in [1], [2]. The key problem in the BSS context is the resolution of the frequency-dependent permutation ambiguity. In [1] the frequency-dependent permutation ambiguity is dealt with by imposing a constraint on the length of the inverse-channel impulse response $\boldsymbol{W}(\tau)$, an approach that is loosely motivated. We propose an interpretable, novel approach based on integer-least-squares (ILS) to adjust the arbitrary permutations.

K. Mokios and A. Potamianos were partially supported by IST-FP6 HIWIRE Project.

To support our arguments we present numerical simulations using real speech data and compare the performance of our algorithm with the one proposed in [1].

2. PROBLEM STATEMENT

Assume I mutually uncorrelated (not necessarily statistically independent) sources $\boldsymbol{s}(t) = [s_1(t), \ldots, s_I(t)]^T$. These sources are convolved and mixed in a linear medium leading to J sensor signals $\boldsymbol{x}(t) = [x_1(t), \ldots, x_J(t)]^T$

$$\boldsymbol{x}(t) = \boldsymbol{A} * \boldsymbol{s}(t) + \boldsymbol{n}(t) = \sum_{\tau=0}^{L} \boldsymbol{A}(\tau)\boldsymbol{s}(t-\tau) + \boldsymbol{n}(t) \qquad (1)$$

for $t = 1, \ldots, N$, where $\boldsymbol{A}(\tau) = [\boldsymbol{\alpha}_1(\tau), \ldots, \boldsymbol{\alpha}_I(\tau)]^T \in \mathbb{R}^{J \times I}$ for $\tau = 0, \ldots, L$ is the mixing impulse response matrix, $\boldsymbol{\alpha}_i(\tau) = [\alpha_{1,i}(\tau), \ldots, \alpha_{J,i}(\tau)]^T \in \mathbb{R}^{J \times 1}$ is the spatial signature of the ith source for lag τ, $\boldsymbol{n}(t) = [n_1(t), \ldots, n_J(t)]^T$ is the additive noise vector, L is the maximum (unknown) channel length, N is the number of snapshots, $*$ denotes convolution and $(.)^T$ denotes the transpose. The objective of the blind separation problem is to estimate the inverse-channel impulse response matrix $\boldsymbol{W}(\tau)$ from the observed signals $\boldsymbol{x}(t)$, such that their convolution provides us with an estimate of the source signals $\widehat{\boldsymbol{s}}(t)$

$$\widehat{\boldsymbol{s}}(t) = \boldsymbol{W} * \boldsymbol{x}(t) \qquad (2)$$

Before proceeding further we make the following main assumptions: (i) The sources $\boldsymbol{s}(t)$ are zero mean, second-order quasi-stationary signals; i.e. the variances of the signals are slowly varying with time such that over short time intervals they can be assumed approximately stationary, and (ii) The noise $\boldsymbol{n}(t)$ is zero mean, uncorrelated at each sensor and independent of the sources.

Following, e.g. [1], one approach towards solving the problem is to transform it into the frequency domain and to solve the joint separation problem across frequency bins. Recall the well known property of the DFT that allows us to express circular convolutions as products. In (1), however, we assumed a linear convolution. A linear convolution can be approximated by a circular convolution if the size T of the DFT's frame is much larger than the length of the convolutional sum [1]. In such a case we can write approximately

$$\boldsymbol{x}(f,t) \approx \boldsymbol{A}(f)\boldsymbol{s}(f,t) + \boldsymbol{n}(f,t), for L \ll T \qquad (3)$$

where $\boldsymbol{x}(f,t) = \sum_{\tau=0}^{T-1} \boldsymbol{x}(t+\tau)e^{\frac{-i2\pi f\tau}{T}}$ is the DFT of the frame of size T starting at t, $[\boldsymbol{x}(t), \ldots, \boldsymbol{x}(t+T-1)$ and corresponding expressions apply for $\boldsymbol{s}(f,t)$ and $\boldsymbol{A}(f)$. The ith column of $\boldsymbol{A}(f)$ represents now the spatial signature of the ith source in the frequency domain, at frequency f. Let us focus on a time interval over which the measured signals can be assumed

stationary (see assumption (i)). The autocorrelation matrix of the vector of microphone outputs at frequency f is then

$$\begin{aligned} R_x(f,t) &= E[\boldsymbol{x}(f,t)\boldsymbol{x}^H(f,t)] \\ &= \boldsymbol{A}(f)E[\boldsymbol{s}(f,t)\boldsymbol{s}^H(f,t)]\boldsymbol{A}^H(f) + E[\boldsymbol{n}(f,t)\boldsymbol{n}^H(f,t)] \\ &= \boldsymbol{A}(f)D_s(f,t)\boldsymbol{A}^H(f) + D_n(f,t) \end{aligned} \quad (4)$$

where $E[.]$ denotes the expectation operator and $(.)^H$ denotes the Hermitian transpose. Since we assume mutually uncorrelated sources we postulate diagonal autocorrelation matrix of the sources $D_s(f,t)$, while by assumption the autocorrelation matrix of the additive noise $D_n(f,t)$ is also diagonal.

Now assume that the matrices $\boldsymbol{A}(f), f = 0,\ldots,T-1$ are available. Then, by taking the Moore-Penrose pseudo-inverse of each frequency's corresponding matrix $\boldsymbol{A}(f)$, and applying the Inverse DFT to the collection of the acquired pseudo-inverses $\boldsymbol{A}^\dagger(f)$ for $f = 0,\ldots,T-1$, where $(.)^\dagger$ denotes the Moore-Penrose pseudo-inverse, we could determine estimates of the inverse-channel matrices $\widehat{\boldsymbol{W}}(\tau)$, where $\tau = 0,\ldots,T-1$. Therefore the BSS problem boils down to the problem of estimating the matrices $\boldsymbol{A}(f), f = 0,\ldots,T-1$.

It is well known that the matrices $\boldsymbol{A}(f), f = 0,\ldots,T-1$ can only be specified up to an individual permutation and scaling of their columns (spatial signatures),that vary arbitrarily from one frequency bin to another. This constitutes a serious problem since only consistent permutations and scaling across frequencies will correctly reconstruct the sources. Hence a second major problem arises, that of resolving the frequency-dependent permutation and scaling issue.

3. PARALLEL FACTOR ANALYSIS

By dividing the whole data block of N snapshots into P sub-blocks, with each sub-block corresponding to a time interval over which the source signals are assumed stationary, the measured snapshots within any pth sub-block correspond to the following autocorrelation matrix

$$R_x(f,t_p) = \boldsymbol{A}(f)D_s(f,t_p)\boldsymbol{A}^H(f) + D_n(f,t_p) \quad (5)$$

for $f = 0,\ldots,T-1$, in accordance with (4). Using all P sub-blocks, we will have P different autocorrelation matrices $\{R_x(f,t_1),\ldots,R_x(f,t_P)\}$ for each frequency. By neglecting the noise component for the moment, we observe that for each frequency, these matrices differ from each other only because the source signal autocorrelation matrices $D_s(f,t_p)$ differ from one sub-block to another.

Let us stack the P matrices $R_x(f,t_p) - D_n(f,t_p), p = 1,\ldots,P$ together to form a three-way array $\underline{R}_x(f)$. The (j,l,p)th element of such an array can be written as

$$r_{j,l,p} \triangleq [\underline{R}_x(f)] = \sum_{i=1}^{I} \alpha_{j,i}(f)v_i(f,p)\alpha^*_{l,i} \quad (6)$$

for $f = 0,\ldots,T-1$, where $v_i(f,p) \triangleq [D_s(f,t_p)]_{i,i}$ is the power-spectral-density of the ith source in the pth sub-block and $(.)^*$ denotes the complex conjugate. Defining the matrices $\boldsymbol{P}(f) \in \mathbb{C}^{P\times I}, f = 0,\ldots,T-1$ as

$$\boldsymbol{P}(f) \triangleq \begin{bmatrix} v_1(f,1) & \ldots & v_I(f,1) \\ \vdots & \ddots & \vdots \\ v_1(f,P) & \ldots & v_I(f,P) \end{bmatrix} \quad (7)$$

we can write the following relationship between $D_s(f,t_p)$ and $\boldsymbol{P}(f)$

$$D_s(f,t_p) = \mathcal{D}_p\{\boldsymbol{P}(f)\} \quad (8)$$

for all $p = 1,\ldots,P$ and all $f = 1,\ldots,T$. In (8), $\mathcal{D}_p\{.\}$ is the operator which makes a diagonal matrix by selecting the pth row and putting it on the main diagonal while putting zeros elsewhere.

Equation (6) implies that $r_{j,l,p}$ is a sum of rank-1triple products; this equation is known as (conjugate-symmetric) *parallel factor* (PARAFAC) analysis of $r_{j,l,p}$ [5]. If I is sufficiently small (6) represents a low-rank decomposition of $\underline{R}_x(f)$. Therefore, the problem of estimating the matrix $\boldsymbol{A}(f)$ for a specific frequency f can be reformulated as the problem of low-rank decomposition of the three-way autocorrelation array $\underline{R}_x(f)$. By solving a similar problem separately for every frequency we obtain the entire collection of the frequency-domain mixing matrices $\boldsymbol{A}(f), f = 0,\ldots,T-1$.

The uniqueness (up to inherently unresolvable source permutation and scale ambiguities) of all source spatial signatures, given the exact frequency-domain autocorrelation data is an issue known as the identifiability of the model. Identifiability conditions i.e., conditions under which the trilinear decomposition of $\underline{R}_x(f)$ is unique, are discussed in [5].

In practice, the exact autocorrelation matrices $R_x(f,t_p)$ are unavailable but can be estimated from the array snapshots $\boldsymbol{x}(t), t = 1,\ldots,N$. If we define $K = \lfloor \frac{N_s}{T} \rfloor$, the sample autocorrelation matrix estimates are given by

$$\widehat{R}_x(f,t_p) = \frac{1}{K}\sum_{k=0}^{K-1} \boldsymbol{x}(f,t_p+kT)\boldsymbol{x}^H(f,t_p+kT) \quad (9)$$

for $p = 1,\ldots,P$.

In our problem, PARAFAC fitting at each frequency was based on the implementation of a fast and monotonically convergent least squares separation algorithm, known as trilinear alternating least squares (TALS) technique [5], which is used to estimate the matrices $\boldsymbol{A}(f)$, up to a frequency-dependent permutation and scaling ambiguity. Exactly how these ambiguities are resolved (to yield our final separation solution), will be discussed in section 4. For more information on TALS we refer the interested reader to the above references.

4. RESOLVING THE SCALING AND PERMUTATION AMBIGUITIES

In this section, we propose a novel approach for resolving the scaling and permutation ambiguities that arise in the speech separation context. In the first subsection, our method of dealing with the scaling problem is presented, while the second subsection describes how the frequency-dependent permutation ambiguity is resolved.

1. Resolving the frequency-dependent scaling problem

Assume, for the moment, that the frequency-dependent permutation ambiguity is resolved, that is, the only remaining problem is the frequency-dependent scaling ambiguity of columns of the estimated mixing matrices. The resolution of the latter is based upon the following result, whose proof is omitted for brevity

Result 1: Consider a full rank matrix $\boldsymbol{A} \in \mathbb{C}^{J\times I}$, with $J > I$. Let $\boldsymbol{A}_n \in \mathbb{C}^{J\times I}$ denote the matrix whose ith column $\boldsymbol{A}_{n(:,i)}$ is the ith column of $\boldsymbol{A}$, $\boldsymbol{A}_{(:,i)}$, divided by the element $\alpha_{j_i,i}$ of $\boldsymbol{A}_{(:,i)}$, with $j_i \in \{1,\ldots,J\}$, i.e., $\boldsymbol{A}_{n(:,i)} = \frac{1}{\alpha_{j_i,i}}\boldsymbol{A}_{(:,i)}, i = 1,\ldots,I$. Then $\boldsymbol{A}_n^\dagger\boldsymbol{A} = diag\{[\alpha_{j_1,1},\ldots,\alpha_{j_I,I}]^T\}$, where $diag\{\boldsymbol{x}\}$ forms a diagonal matrix from vector $\boldsymbol{x}$.
Let

$$\begin{aligned} \boldsymbol{A}_s(f) &= [c_1(f)\boldsymbol{\alpha}_1(f),\ldots,c_I(f)\boldsymbol{\alpha}_I(f)] = \\ &\boldsymbol{A}(f)diag\left\{[c_1(f),\ldots,c_I(f)]^T\right\} = \boldsymbol{A}(f)\boldsymbol{C}(f) \end{aligned} \quad (10)$$

where $c_i(f) \in \mathbb{C}$ for all $i = 1, \ldots, I$, represent a scaled version of the true mixing matrix $\boldsymbol{A}(f)$ at frequency f. Divide the columns of matrix $\boldsymbol{A}_s(f)$ by their respective elements $\{c_1(f)\alpha_{j_1,1}(f), \ldots, c_I(f)\alpha_{j_I,I}(f)\}$, with $j_i \in \{1, \ldots, J\}$, to yield matrix $\boldsymbol{A}_{s,n}(f)$

$$\boldsymbol{A}_{s,n}(f) = \boldsymbol{A}_s(f) diag\left\{ \frac{1}{c_1(f)\alpha_{j_1,1}(f)}, \ldots, \frac{1}{c_I(f)\alpha_{j_I,I}(f)} \right\}^T =$$
$$\boldsymbol{A}(f)\boldsymbol{C}(f)\boldsymbol{C}^{-1}(f) diag\left\{ \frac{1}{\alpha_{j_1,1}(f)}, \ldots, \frac{1}{\alpha_{j_I,I}(f)} \right\}^T =$$
$$\boldsymbol{A}(f) diag\left\{ \frac{1}{\alpha_{j_1,1}(f)}, \ldots, \frac{1}{\alpha_{j_I,I}(f)} \right\}^T = \boldsymbol{A}_n(f) \quad (11)$$

where $\boldsymbol{A}_n(f)$ denotes the matrix whose ith column $\boldsymbol{A}_n(f)_{(:,i)}$ is the ith column of the true mixing matrix $\boldsymbol{A}(f)$, $\boldsymbol{A}(f)_{(:,i)}$, divided by element $\alpha_{j_i,i}(f)$ of $\boldsymbol{A}(f)_{(:,i)}$.

Let us multiply (3) (we neglect the noise term) from the left with the pseudo-inverse $\boldsymbol{A}^{\dagger}_{s,n}(f) = \boldsymbol{A}^{\dagger}_n(f)$. Invoking Result 1, we have

$$\boldsymbol{A}^{\dagger}_{s,n}(f)\boldsymbol{x}(f,t) \approx \boldsymbol{A}^{\dagger}_n(f)\boldsymbol{A}(f)\boldsymbol{s}(f,t) =$$
$$diag\{[\alpha_{j_1,1}(f), \ldots, \alpha_{j_I,I}(f)]^T\}\boldsymbol{s}(f,t) =$$
$$[\alpha_{j_1,1}(f)s_1(f,t), \ldots, \alpha_{j_I,I}(f)s_I(f,t)]^T \quad (12)$$

We see from (12), that each source signal $s_i(f,t)$ is multiplied by the element $\alpha_{j_i,i}(f)$, for all $i = 1, \ldots, I$. But $\alpha_{j_i,i}(f)$ is just the total frequency response that corresponds to the link between the ith speaker and the j_ith microphone, at frequency f, which is approximately constant across frequency in (quasi-) non reverberant environments. Thus, the frequency-dependent scaling of the individual sources that would result if we had used the pseudo-inverses of the scaled matrices (10) in (12), now is reduced to a single, frequency-independent scaling, i.e., the frequency-dependent scaling ambiguity problem is resolved.

Intuitively, the filtering process of the ith speaker's speech signal $s_i(t)$ through $\alpha_{j_i,i}(f)$, that yields the separated signal $\widehat{s}_i(t) = \alpha_{j_i,i}(t) * s_i(t)$, means that after separation, we essentially listen to the signal $s_i(t)$ the way it is captured by the j_ith microphone, as if there were no other interfering speakers.

2. Resolving the frequency-dependent permutation problem

In the previous subsection we assumed that the frequency-dependent permutation ambiguity was already resolved. Thus, prior to applying the method presented above for dealing with the scaling problem, the frequency-dependent permutation ambiguity of the columns of the matrices $\boldsymbol{A}(f), f = 0, \ldots, T-1$, must be lifted. We resolve the permutation problem by exploiting two inherent attributes of the system under study. Proper interpretation and utilization of each of these attributes allows us to transform the frequency-dependent permutation problem into an optimization problem with each attribute providing us with a different optimization criterion. We also formulate a joint optimization criterion that incorporates these two constituent criteria.

•Criterion 1: Let the collection of vectors $\{\boldsymbol{\alpha}_{n_i}(f), f = 0, \ldots, T-1, i = 1, \ldots, I\}$ denote the entire collection of the (normalized with respect to an arbitrary reference microphone) spatial signatures that are acquired after the application of the TALS algorithm. In non-reverberant (or, mildly reverberant) environments, where the recordings have been conducted using microphones with (almost) flat frequency responses, the magnitudes of the elements of the normalized spatial signatures, are approximately independent of frequency. Thus the aforementioned collection of normalized spatial signatures can be divided into I separate clusters, each of the latter being associated with a different source.

We use vector quantization [7] to determine the I centroids $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_I\}$, with $\boldsymbol{c}_i \in \mathbb{C}^J$, but other clustering techniques could also be used. With the I centroids at our disposal, the frequency-dependent permutation problem boils down to the following Integer Least Squares (ILS) minimization problem

$$\min_{\{s_f\}_{f=1}^T \in \{1,\ldots,I!\}^T} \sum_{f=1}^{T} \|\boldsymbol{C} - \boldsymbol{A}_n(f)\boldsymbol{\Pi}_1 1_{\{s_f=1\}}$$
$$- \boldsymbol{A}_n(f)\boldsymbol{\Pi}_2 1_{\{s_f=2\}} - \ldots - \boldsymbol{A}_n(f)\boldsymbol{\Pi}_{I!} 1_{\{s_f=I!\}}\|_F^2 \quad (13)$$

where $\boldsymbol{C} \in \mathbb{C}^{J \times I}$ is the matrix whose columns are the centroids $\{\boldsymbol{c}_i, i = 1, \ldots, I\}$, $\boldsymbol{A}_n(f) \in \mathbb{C}^{J \times I}$ is the matrix whose columns are the arbitrarily permuted vectors $\{\boldsymbol{\alpha}_{n_i}(f), i = 1, \ldots, I\}$, $\boldsymbol{\Pi}_l \in \mathbb{C}^{I \times I}$ for $l = 1, \ldots, I!$ is one of the total number of $I!$ permutation matrices of dimensionality $I \times I$. Note that $1_{\{x=\alpha\}} = 0$ if $x \neq \alpha$ and $1_{\{x=\alpha\}} = 1$ if $x = \alpha$.

•Criterion 2: For adjacent frequency bins, $f_1, f_2 = f_1 + 1$, $\boldsymbol{\alpha}_i(f_1) \approx \boldsymbol{\alpha}_i(f_2)$ and consequently $\boldsymbol{\alpha}_{n_i}(f_1) \approx \boldsymbol{\alpha}_{n_i}(f_2)$. This is valid for dense FFT grids even in reverberant environments, by continuity of the Fourier transform. Based on this fact, we formulate our second ILS minimization criterion

$$\min_{\{s_f\}_{f=1}^T \in \{1,\ldots,I!\}^T} \sum_{f=1}^{T} \|\boldsymbol{A}_n(f)\boldsymbol{\Pi}_1 1_{\{s_f=1\}} + \ldots + \boldsymbol{A}_n(f)\boldsymbol{\Pi}_{I!} 1_{\{s_f=I!\}}$$
$$- \boldsymbol{A}_n(f-1)\boldsymbol{\Pi}_1 1_{\{s_{f-1}=1\}} - \ldots - \boldsymbol{A}_n(f-1)\boldsymbol{\Pi}_{I!} 1_{\{s_{f-1}=I!\}}\|_F^2 \quad (14)$$

where $\boldsymbol{A}_n(0) = \boldsymbol{0}^{J \times I}$, by convention.

The criteria of (13) and (14) can be combined into one overall ILS minimization criterion

$$\min_{\{s_f\}_{f=1}^T \in \{1,\ldots,I!\}^T} \sum_{f=1}^{T} \{\|\boldsymbol{C} - \boldsymbol{A}_n(f)\boldsymbol{\Pi}_1 1_{\{s_f=1\}} - \ldots$$
$$- \boldsymbol{A}_n(f)\boldsymbol{\Pi}_{I!} 1_{\{s_f=I!\}}\|_F^2 + \lambda\|\boldsymbol{A}_n(f)\boldsymbol{\Pi}_1 1_{\{s_f=1\}} + \ldots$$
$$+ \boldsymbol{A}_n(f)\boldsymbol{\Pi}_{I!} 1_{\{s_f=I!\}} - \boldsymbol{A}_n(f-1)\boldsymbol{\Pi}_1 1_{\{s_{f-1}=1\}} - \ldots$$
$$- \boldsymbol{A}_n(f-1)\boldsymbol{\Pi}_{I!} 1_{\{s_{f-1}=I!\}}\|_F^2 \quad (15)$$

where λ is a weighting factor. When $\lambda \to 0$ (practically, for $\lambda << 1$), the combined criterion (15) coincides with criterion 1, while for $\lambda \to \infty$ (practically, for $\lambda >> 1$), (15) coincides with criterion 2.

The ILS minimization criterion of (15) proves to be satisfactory for resolving the permutation problem that arises in the majority of real world BSS problems, as will be shown in section 5. The ILS problem in (15) is NP-hard, however, a plethora of polynomial complexity approximate solutions have been developed for it in the multiuser detection literature. We chose to employ the Successive Interference Cancellation - Iterative Least Squares (SIC-ILS) method [6], since it exhibits the best performance-complexity trade-off in our context, where the decision vectors are of very high dimensionality (=FFT length).

5. EXPERIMENTAL RESULTS

In our experiments, we used a subset of the PEACH multi-microphone database [8] using $I = 2$ loudspeakers as sources and $J = 4$ omni-directional microphone as sensors. The measurements were separated into three sessions of three experiments each, with each session corresponding to a different rectangular geometric configuration. The loudspeakers and the microphones were uniformly-spaced on two parallel lines at distance d_1. The

Table I. BSS performance of proposed PARAFAC-ILS method compared to Parra's method [1] in terms of output SIR and SIR improvement in dB for each session, speaker pair and utterance (source S1 or S2).

Experiment:	*Session I: d_1=150cm d_2=10cm*						*Session II: d_1=150cm d_2=30cm*						*Session III: d_1=100cm d_2=30cm*					
	1st pair		*2nd pair*		*3rd pair*		*1st pair*		*2nd pair*		*3rd pair*		*1st pair*		*2nd pair*		*3rd pair*	
	S1	*S2*	*S1*	*S2*	*S1*	*S2*	*S1*	*S2*	*S1*	*S2*	*S1*	*S2*	*S1*	*S2*	*S1*	*S2*	*S1*	*S2*
Raw Data Mean SIR	-3.8	3.8	-4.1	4.1	0.5	-0.5	-4.6	4.6	-3.0	3.0	-0.2	0.2	-6.7	6.7	-1.0	1.0	-1.2	1.2
Parra's Method SIR	-3.0	5.5	3.3	6.0	5.3	4.8	-4.2	6.4	1.9	-0.1	3.8	3.8	-3.3	4.6	1.0	1.8	1.8	0.4
SIR Improvement	**0.8**	1.7	**7.4**	**1.9**	**4.8**	**5.3**	0.4	1.8	4.9	-3.1	**4.0**	**3.6**	3.4	-2.1	2.0	0.8	3.0	-0.8
PARAFAC-ILS SIR	-3.9	6.9	0.7	5.9	1.5	1.8	-1.2	9.7	2.5	5.3	2.1	3.2	-3.2	12.3	3.5	3.4	3.5	9.5
SIR Improvement	-0.1	**3.1**	4.8	1.8	1.0	2.3	**3.4**	**5.1**	**5.5**	**2.3**	2.3	3.0	**3.5**	**5.6**	**4.5**	**2.4**	**4.7**	**8.3**

Table II. Overall BSS performance of PARAFAC-ILS and Parra's methods in terms of average SIR improvement in dB for each experimental session.

Session	*PARAFAC-ILS SIR impr. mean*	*Parra's Method SIR impr. mean*
I: 150cm-10cm	2.15	3.65
II: 150cm-30cm	3.60	1.93
III: 100cm-30cm	4.83	1.05
Overall	3.53	2.21

distance between microphones was d_2. For session I: $d_1 = 150cm, d_2 = 10cm$, for session II: $d_1 = 150cm, d_2 = 30cm$ and for session III: $d_1 = 100cm, d_2 = 30cm$. For all experiments, the transcription of the utterance played back from each loudspeaker was the same. The same three sets of speakers were used for each of the three session giving a total of 9 experiments. For more information on the measurement setup see [8].

In order to quantify the separation performance, we measured the average Signal to Interference Ratio (SIR) across microphones (mean SIR of raw data in Table I) and the SIR for each of the unmixed signals that are the output of the BSS algorithms. SIR is measured following the method proposed in [2]. A separate SIR value was measured for each source ("S1" and "S2" in Table I).

Throughout all BSS experiments using our proposed PARAFAC-ILS algorithm, the number of time segments the speech signals were divided into, was set to $P = 5$. For the solution of the permutation problem, the two ILS minimization criteria in Eq. (15) were weighted equally, i.e., $\lambda = 1$. The frequency-domain sample autocorrelation matrix estimates were calculated according to Eq. (9) using $T = 512$ FFT points in each segment. The performance of our method was compared with the method presented in [1] referred to as "Parra's method" in Table I. The parameters for Parra's method are identical to the PARAFAC-ILS parameters; in addition, the length of the inverse channel impulse response in Parra's method was set to $Q = 128$ as suggested in [1]. In Table I, the raw data SIR is presented for each experiment and signal source along with the SIR of the two BSS algorithms. Improvement over raw data SIR is also shown for the two methods (marked as "SIR Improvement"). The overall performance of the BSS algorithms in terms of average SIR improvement in dB for each session is reported in Table II[1].

We see from Table II, that our method outperforms the method proposed in [1] in sessions II and III. Furthermore, the overall score of the PARAFAC-ILS method is higher by more than 1.3 dB compared to Parra's method. In the first experimental set, however, Parra's method exhibits better separation performance. Bearing in mind that the experimental sets correspond to different geometric configurations, we conclude that the separation performance is a function of the geometry of the measurement setup.

[1]A subset of the recorded speech mixtures along with the corresponding separated signals, can be found in the following website: *http://www.speech.tuc.gr/res/research_bss.html*

6. CONCLUSIONS

We have presented a new two-step approach for the blind speech separation problem. The new approach employs PARAFAC to separate signals in the frequency domain and a context-specific ILS method to resolve the frequency-dependent permutation ambiguity, which is the key problem in this setting. Our approach offers guaranteed convergence, unlike [1], and often considerably better separation performance, measured both quantitatively and in subjective tests. The link to the conjugate symmetric PARAFAC model established here, also shows that a much higher number of signals can be separated than what was suggested in [1]. In the future, we will investigate the separation performance of our algorithm as a function of the source-sensor geometry and the permutation minimization criterion weight λ.

7. REFERENCES

[1] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 320-327, May 2000.

[2] K. Rahbar and J.P. Reilly, "A Frequency Domain Method for Blind Source Separation of Convolutive Audio Mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 832-844, Sep. 2005.

[3] D.T. Pham and J. Cardoso, "Blind source separation of instantaneous mixtures of nonstationary sources," *IEEE Transactions on Signal Processing*, vol. 49, no. 9, pp. 1837-1848, Sep. 2001.

[4] H. Sahlin and H. Broman, "MIMO signal separation for FIR channels: a criterion and performance analysis," *IEEE Transactions on Signal Processing*, vol. 48, no. 3, pp. 642-649, Mar. 2000.

[5] Y. Rong, S.A. Vorobyov, A.B. Gershman, and N.D. Sidiropoulos "Blind Spatial Signature Estimation via Time-Varying User Power Loading and Parallel Factor Analysis," *IEEE Transactions on Signal Processing*, vol. 53, pp. 1697-1710, May 2005.

[6] T. Li and N.D. Sidiropoulos, "Blind digital signal separation using successive interference cancellation iterative least squares," *IEEE Transactions on Signal Processing*, vol. 48, no. 3, pp. 810-823, Apr. 2000.

[7] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1992.

[8] PEACH multi-microphone database: *http://shine.itc.it*