ON THE EFFECTIVENESS OF PARAFAC-BASED ESTIMATION FOR BLIND SPEECH SEPARATION

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

Dept. of Electronic and Computer Engineering, Technical Univ. of Crete, 73100 Chania, Greece
{nikos,kleanthis,potam}@telecom.tuc.gr

ABSTRACT

This work establishes the effectiveness of parallel factor (PARAFAC) analysis in blind speech separation (BSS) problems. The BSS problem is formulated as a conjugate-symmetric PARAFAC model that is fitted optimally, using an efficient alternating least-squares algorithm that converges monotonically. The identifiability properties of the model are also presented, revealing the much broader identifiability potential of joint-diagonalization-based BSS methods. In order to focus on estimation performance, perfect resolution of the permutation ambiguity is assumed. Simulations under varying reverberation conditions and comparison with previous estimation methods that are widely used in BSS problems demonstrate significant performance gains. Signal-to-interference (SIR) ratio improvement of over 27 dB is achieved using PARAFAC. Average SIR gains of 2.5 and 6.3 dB are achieved compared to state-of-the-art FastICA[2] and FDSOS (Parra's)[5] estimation algorithms, respectively.

Index Terms— Blind speech separation, parallel factor analysis, estimation method, non-stationary signals.

1. INTRODUCTION

The objective of the BSS problem is to separate multiple speech signals originating from different speakers/sources in a room, using only the signals captured by microphones that are deployed in the room, i.e., without assuming knowledge of the mixing acoustic channels. Blind speech separation has a wide range of applications such as speech enhancement for teleconferencing and speech recognition systems, as well as mobile telephony applications. In the above applications, the speech signals are mixed in a convolutive fashion in (mildly or strongly) reverberant environments, as opposed to the simpler and unrealistic case of instantaneous mixtures.

In recent years, many algorithms have been proposed to tackle the convolutive BSS problem. These algorithms can be divided into two major categories depending on the domain in which the separation of the signals is realized. Time-domain (TD) BSS methods are applied directly to the convolutive mixture model whereas frequency-domain (FD) methods begin by mapping the convolutive mixture problem to the frequency domain in order to decompose it into multiple independent instantaneous BSS problems, one at each frequency bin.

The relatively high implementation complexity of TD-BSS methods renders them unappealing for on-line applications in contrast with FD-BSS methods that afford far simpler implementation. The drawback of FD-BSS methods is that they have to cope with an inherent frequency-dependent permutation and scaling ambiguity problem, which does not arise in TD methods. Thus, FD-BSS methods achieve the separation of the speech signals in two stages. In the first stage, an algorithm that estimates the unmixing system up to frequency-dependent permutation and scaling ambiguities is employed, while in the second stage the problem of matching the arbitrary permutations is addressed via a pertinent method[1]. Both stages of FD-BSS algorithms greatly affect the quality of the separated speech. In this paper, only the first stage of FD-BSS methods is investigated namely the estimation of the unmixing system up to the aforementioned indeterminacies, henceforth referred to simply as *estimation method*. Regarding the permutation problem several good algorithms have been proposed in recent years [7], [4].

As mentioned earlier, in FD-BSS methods the estimation of the unmixing system is conducted independently for each frequency bin's corresponding instantaneous mixture. The two most popular categories of estimation methods for the BSS problem are: (i) Independent Component Analysis (ICA) methods [2] which perform the estimation of the unmixing system by capturing higher-order statistics of the speech signals and try to exploit the independence and/or non-Gaussianity of the speech signals through them. (ii) Joint-diagonalization-based methods which, in addition to uncorrelatedness, exploit the information provided by the non-stationary nature of the speech signals, utilizing second-order statistics [5].

The estimation method based on PARAFAC analysis belongs to the second category, and has been successfully employed in various communications problems (e.g. [6]) in the past. The BSS problem is formulated as a conjugate-symmetric PARAFAC model that is fitted optimally, using an efficient alternating LS algorithm that converges monotonically. The conjugate-symmetric PARAFAC model exhibits strong identifiability properties that allows the separation of a much higher number of signals than what was suggested in former references. PARAFAC was introduced for use in speech separation problems in our earlier publication [4]. This paper establishes the appropriateness of the method for speech separation via pertinent simulations and also discusses the identifiability potential of the method in terms of the number of speech signals that can be separated given a fixed number of microphones.

2. PROBLEM STATEMENT

Assume I mutually uncorrelated speech signals $\boldsymbol{s}(t) = [s_1(t), \ldots, s_I(t)]^T$. These signals are convolved and mixed in a linear medium leading to J microphone 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 speaker 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

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

[1] The scaling problem is trivial; e.g. one can easily remedy it with appropriate vector normalizations [4]

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 original speech signals $\widehat{\boldsymbol{s}}(t)$

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

Under mild conditions (see e.g. [5]), the convolutive mixture in (1) can be approximated as multiple instantaneous mixtures, one at each frequency

$$\boldsymbol{x}(f,t) \approx \boldsymbol{A}(f)\boldsymbol{s}(f,t) + \boldsymbol{n}(f,t) \tag{3}$$

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 speaker in the frequency domain, at frequency f.

Before proceeding further we make the following main assumptions: (i) The speech signals $\boldsymbol{s}(t)$ 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, (ii) The number of speakers is known, and (iii) The contribution of the noise term $\boldsymbol{n}(t)$ is negligible compared to the contribution of the speaker signals. In our context, this is an accurate approximation in many real-world cases. When noise is not negligible, its power can be estimated from silence periods and subtracted from correlation matrix estimates, which is the only place where noise comes into play in our algorithm.

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

$$\begin{aligned}\boldsymbol{R}_x(f,t) &= E[\boldsymbol{x}(f,t)\boldsymbol{x}^H(f,t)] \\ &\approx \boldsymbol{A}(f)E[\boldsymbol{s}(f,t)\boldsymbol{s}^H(f,t)]\boldsymbol{A}^H(f) \\ &= \boldsymbol{A}(f)\boldsymbol{D}_s(f,t)\boldsymbol{A}^H(f)\end{aligned} \tag{4}$$

where $E[.]$ denotes the expectation operator and $(.)^H$ denotes the Hermitian transpose. Since we assume mutually uncorrelated speech signals we postulate diagonal autocorrelation matrix $D_s(f,t)$, while by assumption (iii) the autocorrelation matrix of the noise vector $\boldsymbol{n}(f,t)$ has been neglected.

Now assume that the matrices $\boldsymbol{A}(f), f = 0,\ldots,T-1$ are available and $rank[\boldsymbol{A}(f)] \geq I$. 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 $\widehat{\boldsymbol{W}}(\tau)$. When $rank[\boldsymbol{A}(f)] < I$ (in particular, when $J < I$) perfect separation is not possible; substantial reduction of crosstalk is still possible, however, using matched filtering to the columns of $\boldsymbol{A}(f)$, or more sophisticated array processing methods, like Capon beamforming. In either case, the BSS problem boils down to the problem of estimating the matrices $\boldsymbol{A}(f), f = 0,\ldots,T-1$. Next we present the solution to the estimation problem using PARAFAC analysis (assuming that the permutation problem is solved).

3. ESTIMATION OF THE MIXING MATRICES USING PARAFAC 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 speech signals are assumed stationary, the measured snapshots within any pth sub-block correspond to the following autocorrelation matrix

$$\boldsymbol{R}_x(f,t_p) = \boldsymbol{A}(f)\boldsymbol{D}_s(f,t_p)\boldsymbol{A}^H(f) \tag{5}$$

for $f = 0,\ldots,T-1$, in accordance with (4). Using all P sub-blocks, we have P different autocorrelation matrices $\{\boldsymbol{R}_x(f,t_1),\ldots,\boldsymbol{R}_x(f,t_P)\}$ for each frequency. Observe that for each frequency, these matrices differ from each other only because the source signal autocorrelation matrices $\boldsymbol{D}_s(f,t_p)$ differ from one sub-block to another.

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

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

for $f = 0,\ldots,T-1$, where $v_i(f,p) \triangleq [\boldsymbol{D}_s(f,t_p)]_{i,i}$ is the power-spectral-density of the ith speech signal 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

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

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

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

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 $r_{j,l,p}$ [6]. If I is sufficiently small (6) represents a low-rank decomposition of $\underline{\boldsymbol{R}}_x(f)$. 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 $\underline{\boldsymbol{R}}_x(f)$. By solving a similar problem separately for every frequency we obtain the entire collection of the frequency-domain mixing matrices $\boldsymbol{A}(f), f = 0,\ldots,T-1$.

In practice, the exact autocorrelation matrices $\boldsymbol{R}_x(f,t_p)$ 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

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

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 [6], which is used to estimate the matrices $\boldsymbol{A}(f)$, up to inherently unresolvable frequency-dependent permutation and scaling ambiguity of its columns. For more information on TALS we refer the interested reader to the above references.

1. Identifiability

By identifiability we mean the uniqueness (up to permutation and scaling ambiguities) of all speaker spatial signatures at a given frequency, given the exact frequency-domain autocorrelation data at that frequency.

Using (8) we can rewrite (5) as

$$\boldsymbol{R}_x(f,t_p) = \boldsymbol{A}(f)\mathcal{D}_p\{\boldsymbol{P}(f)\}\boldsymbol{A}^H(f), \text{ for } p = 1,\ldots,P \tag{10}$$

To establish identifiability, we have to obtain under which conditions matrices $\boldsymbol{P}(f)$ and $\boldsymbol{A}(f)$ are the unique (up to the scaling and permutation ambiguities) matrices that give rise to the data $\{\boldsymbol{R}_x(f, t_p), p = 1, \ldots, P\}$ of (10).

Alternative uniqueness conditions can be derived if random component matrices are considered, giving rise to the concept of almost sure-identifiability. The best known result on almost sure-identifiability conditions for the conjugate-symmetric PARAFAC model has been established in [9] and is presented in the next theorem

Theorem 1: Suppose that the elements of $\boldsymbol{A}(f)$ and $\boldsymbol{P}(f)$ are drawn from a jointly continuous distribution. If

$$\frac{I(I-1)}{2} \leq \frac{J(J-1)}{4}\left[\frac{J(J-1)}{2}+1\right] - \left(\begin{array}{c} J \\ 4 \end{array}\right) 1_{\{J \geq 4\}} \quad (11)$$

where $1_{\{J\geq 4\}} = \left\{\begin{array}{ll} 0 & \text{if } J < 4 \\ 1 & \text{if } J \geq 4 \end{array}\right.$, then for $I \leq 30$, $\boldsymbol{A}(f)$ and $\boldsymbol{P}(f)$ are almost surely unique up to inherently unresolvable permutation and scaling of columns.

In our context J corresponds to the number of microphones we use while I corresponds to the number of speakers whose spatial signatures can be identified. In Table I below, we give for values $J = 2, \ldots, 8$ the upper bound for I, according to the formula given in (11). Observe that with 6 microphones it is possible to estimate the spatial signatures of up to 15 speakers, whereas with 8 microphones it is possible to estimate the spatial signatures of up to 26 speakers - more than 3 times the number of microphones, in theory.

Note that Table I provides theoretical bounds on the number of speakers whose spatial signatures can be identified, relying on the link to the conjugate-symmetric PARAFAC model. Given the spatial signatures, the speech signals can be separated, or crosstalk can be suppressed, depending on I, J.

This result reveals the much broader identifiability potential of joint-diagonalization-based BSS methods, which went unrecognized in the past. Prior references assumed the number of speakers to be at most equal to the number of microphones used, $J \geq I$ [5].

Table I. Values of J and upper bounds for I, which follow from (11)

J	2	3	4	5	6	7	8
Upper bound for I	2	4	6	10	15	20	26

4. EXPERIMENTAL RESULTS

In order to assess the separation performance of the proposed method, multichannel recordings were simulated by convolving impulse responses that were generated using the Roomsim toolbox [10], with speech signals created by concatenating various speech segments from the TIMIT speech database. We chose to use synthetic speech mixtures for the following reasons: first, a more systematic study of the proposed method can be realized regarding its behavior under varying reverberation conditions, and second, the exact knowledge of the mixing filters allows us to solve the permutation problem perfectly and therefore determine the optimal performance of our estimation method.

The Roomsim toolbox implements the image technique [1] in order to perform a simulation of the acoustics of a simple empty "shoebox" room and subsequently provide the room impulse responses and reverberant speech signals. In our experiments the room's dimensions are $4.5m \times 3.6m \times 2.5m$ and the geometric configuration of the four microphones and the two loudspeakers is shown in Fig. 1. The first speech signal corresponds to a female English speaker and the second to a male English speaker; both speech signals' duration is approximately 30 seconds sampled at 16 KHz, and were played back through the virtual loudspeakers at approximately the same sound volume.

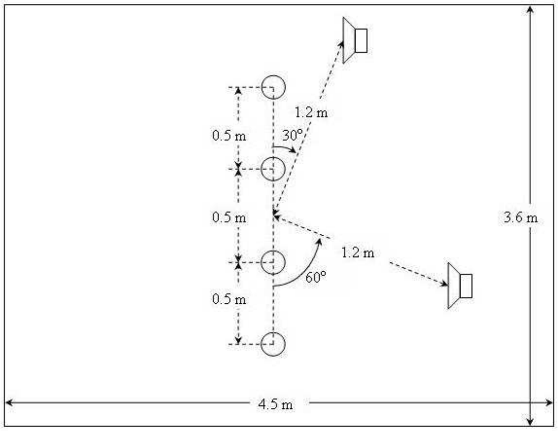


Fig. 1. Geometric configuration of the simulated room.

We have conducted experiments with three different room reverberation characteristics, which were attained by properly adjusting the absorption coefficients of the room's walls. The reverberation times for the three experiments were $50, 130$ and 215 ms correspondingly, and were measured using the method described in [8].

In order to quantify the separation performance, we measured the average Signal to Interference Ratio (SIR) across microphones and the SIR for each of the unmixed signals that are the output of the BSS algorithm. As was mentioned above, the influence of the permutation problem was avoided by utilizing the information on the mixing filters [3], thus resolving perfectly the permutation ambiguity. Therefore the results obtained are not affected by the permutation problem, allowing us to assess the full potential of PARAFAC analysis and other BSS estimation methods. Note that SIR is measured following the method proposed in [3].

The average improvement in SIR (averaged over the two speakers) achieved by PARAFAC is reported in Fig. 2(a)-(c) for various sizes of the FFT frame. The duration of the stationary segments the speech mixtures were divided into was set to 1 sec, a value that provides the highest performance in most cases. For comparison purposes, we also report the corresponding results (i.e. with the permutation problem solved perfectly) obtained by a widely used ICA algorithm, named FastICA [2], as well as the results of a popular joint-diagonalization-based algorithm using second-order statistics [5], which we will henceforth refer to as FDSOS (Frequency Domain Second Order Statistics) method. The parameters for FastICA[2] and FDSOS [3] were chosen so that each method exhibits the highest possible performance.

As shown in Fig. 2, the proposed method clearly outperforms FDSOS for all three reverberation conditions. Regarding FastICA, the single case where its performance reaches PARAFAC's levels is observed in strongly reverberant environments with high frequency resolution[4]. In all other cases, FastICA's performance is inferior. Averaging over all three experiments and all FFT sizes, PARAFAC delivers 20.3 dB's of SIR improvement as opposed to 17.8 dB's of FastICA and 14.0 dB's of FDSOS, thus establishing PARAFAC analysis as one of the most effective estimation methods for blind

[2] FastICA did not improve when overlapping FFT windows were used, in order to increase the number of temporal samples.

[3] For FDSOS, best estimation performance is achieved without any frequency truncation [3].

[4] Note that in Fig 2(a), the performance for 2048 FFT bins decreases due to the relatively short length of the impulse response of the room (approx. 600 samples at 16kHz).

speech separation in terms of estimation performance[5].

5. CONCLUSIONS

We have argued for the effectiveness of parallel factor (PARAFAC) analysis in blind speech separation (BSS) problems. Simulations under varying reverberation conditions and comparison with previous state-of-the-art methods described in [2], [5] show that, assuming ideal resolution of the permutation ambiguity, PARAFAC-based estimation significantly outperforms competing state-of-the-art methods. Average SIR gains of 2.5 and 6.3 dB in separation performance are achieved compared to the methods described in [2] and [5] respectively. Our approach offers guaranteed convergence, unlike [5]. 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 previous references. In future work, we will investigate the solution to the FD-BSS permutation problem in conjunction with the PARAFAC estimation method presented here.

6. REFERENCES

[1] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *JASA*, vol. 65, no. 4, pp. 943-950, 1979.

[2] A. Hyvarinen, "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis," *IEEE T-NN*, vol. 10, no. 3, pp.626-634, 1999.

[3] M.Z. Ikram and D.R. Morgan, "Permutation inconsistency in blind speech separation: Investigation and solutions," *IEEE T-SAP*, vol. 13, Jan. 2005.

[4] K.N. Mokios, N.D. Sidiropoulos, and A. Potamianos, "Blind speech separation using PARAFAC Analysis and Integer Least Squares," in *Proc. ICASSP*, May, 2006.

[5] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE T-SAP*, vol. 8, May 2000.

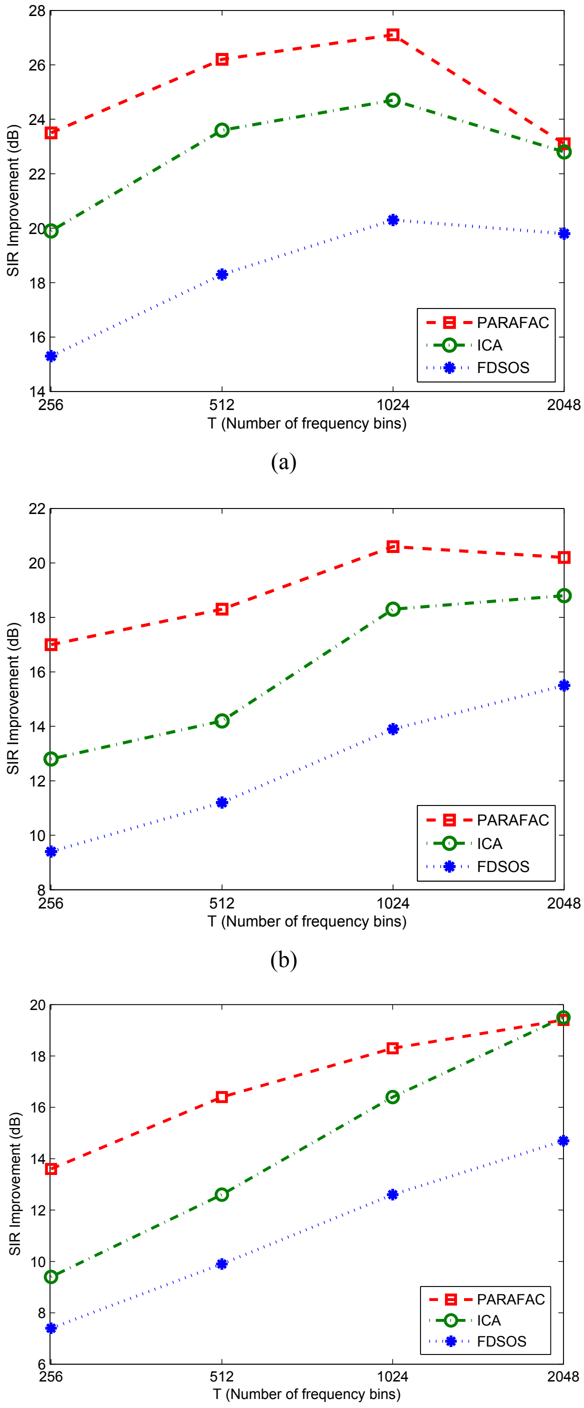
[6] Y. Rong, S.A. Vorobyov, A.B. Gershman, and N.D. Sidiropoulos, "Blind Spatial Signature Estimation via Time-Varying User Power Loading and Parallel Factor Analysis," *IEEE T-SAP*, vol. 53, May 2005.

[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 T-SAP*, vol. 12, no. 5, Sep. 2004.

[8] M. Schroeder, "New method for measuring reverberation time,"*JASA*, vol. 37, pp. 409-412, 1965.

[9] A. Stegeman, J.M.F. ten Berge, and L. De Lathauwer, "Sufficient conditions for uniqueness in CANDECOMP/PARAFAC and INDSCAL with random component matrices," *Psychometrika*, vol. 71, no. 2, pp. 219-229, Jun. 2006.

[10] Roomsim: *http://media.paisley.ac.uk/ ˜campbell/Roomsim/*



(a)

(b)

(c)

Fig. 2. Average SIR improvement vs number of frequency bins for the PARAFAC, FastICA and FDSOS estimation methods. The reverberation time of the room is: (a) 50ms, (b) 130ms, and (c) 215ms.

[5] In term of complexity, FastICA and PARAFAC have similar computation time, while FDSOS is significantly slower.