EUSIPCO 2013 1569742637

BLIND SPEECH SEPARATION EXPLOITING TEMPORAL AND SPECTRAL CORRELATIONS USING 2D-HMMS

Dang Hai Tran Vu and Reinhold Haeb-Umbach

University of Paderborn, Dept. of Communications Engineering

{tran,haeb}@nt.uni-paderborn.de

ABSTRACT

We present a novel method to exploit correlations of adjacent time-frequency (TF)-slots for a sparseness-based blind speech separation (BSS) system. Usually, these correlations are exploited by some heuristic smoothing techniques in the post-processing of the estimated soft TF masks. We propose a different approach: Based on our previous work with one-dimensional (1D)-hidden Markov models (HMMs) along the time axis we extend the modeling to two-dimensional (2D)-HMMs to exploit both temporal and spectral correlations in the speech signal. Based on the principles of turbo decoding we solved the complex inference of 2D-HMMs by a modified forward-backward algorithm which operates alternatingly along the time and the frequency axis. Extrinsic information is exchanged between these steps such that increasingly better soft time-frequency masks are obtained, leading to improved speech separation performance in highly reverberant recording conditions.

1. INTRODUCTION

The goal of BSS is to extract individual target speech signals from a noisy mixture captured by a sensor array. The BSS technique for speech dealt with in this paper can be used in many applications of multichannel speech enhancement including hands-free telecommunication and automatic meeting note taking.

Let us assume a mixture of I independent source signals $S_i(m,k)$, $i = 1{:}I$, captured by D microphones as $X_j(m,k), j = 1{:}D$ in the N-point short-time Fourier transform (STFT)-domain, where $m = 1{:}M$ is the time frame index and $k = 1{:}K, K = N/2{+}1$ denotes the frequency index. We collect the sensor signals in a $D \times 1$ observation vector $\mathbf{X} := [X_j]_{j=1:D}$ where we used the notation $[\cdot]_{j;i}$ to define the element on the j-th row and i-th column of a matrix. Note, that if the matrix reduces to a column vector we omit the column index. In a noisy and reverberant environment the observation vector can approximated by

$$\mathbf{X}(m,k) \approx \sum_{i=1}^{I} \mathbf{H}_i(k) S_i(m,k) + \mathbf{N}(m,k), \qquad (1)$$

This work was supported by DFG under contract number HA 3455/8-1.

where $\mathbf{H}_i := [H_{ij}]_{j=1:D}$ is the $D{\times}1$ vector consisting of multiplicative transfer functions modeling the signal path from source i to microphone j and $\mathbf{N} := [N_j]_{j=1:D}$ is the $D \times 1$ noise vector.

Since speech signals are sparse in the STFT domain a common assumption for BSS is that at any TF-slot (m,k) only a single source is active. Based on this assumption the observation model (1) can be reformulated by

$$\mathbf{X}(m,k) = \begin{cases} \mathbf{N}(m,k) & \text{if } Z(m,k){=}0 \\ \mathbf{H}_i(k) S_i(m,k) + \mathbf{N}(m,k) & \text{if } Z(m,k){=}i \end{cases} \qquad (2)$$

where the discrete hidden random variable $Z(m,k){=}i$ for $i \in \{1,\ldots,I\}$ indicates that source i is active and $Z(m,k){=}0$ indicates that only noise is present in a given TF-slot.

The pivotal issue of sparseness-based BSS approaches is to jointly estimate parameters of the mixing system and compute the posterior probability (PP), e.g. to uncover the identity of $Z(m,k)$. Conventional methods compute the PP solely based on information in the current TF-slot $P\left(Z(m,k)|\mathbf{X}(m,k)\right)$, e.g. [1, 2, 3]. Thus, they ignore the strong correlations among adjacent TF-slots both in time and frequency of speech signals. In [4] we employ a 1D-HMM for each frequency bin independently to exploit the temporal correlations by computing the PP $P\left(Z(m,k)|\mathbf{X}(1{:}M,k)\right)$ with the forward-backward algorithm (FBA). An extension to also exploit the spectral correlations asks for the use of 2D-HMMs and consequently the computation of the PP based on information in all observations $P\left(Z(m,k)|\mathbf{X}(1{:}M,1{:}K)\right)$. Unfortunately, exact computation of the PP in large 2D-HMMs is computationally infeasible. Recently, we proposed an iterative algorithm for computation of the PP which operates by alternating between the time axis and the frequency axis where information is exchanged between these steps [5]. In this paper we review this algorithm and show how to use this iterative decoding algorithm for noisy BSS.

2. OBSERVATION MODEL AND 2D-HMM

Due to the highly non-stationary behavior of the speech sources it is difficult to obtain an accurate and computationally tractable observation model $p\left(\mathbf{X}(m,k)|Z(m,k)\right)$ which is necessary for statistical interference. We propose to split

the information contained in the observation vector into two independent feature parts:

$$p\left(\mathbf{X}|Z\right) = p\left(\varphi|Z\right) p\left(\mathbf{Y}|Z\right). \tag{3}$$

The first part is the averaged *a-posteriori* signal-to-noise ratio (SNR) of the sensor array:

$$\varphi(m,k) := \frac{1}{D}\mathbf{X}^{\mathrm{H}}(m,k)\mathbf{\Phi}_{\mathbf{NN}}^{-1}(k)\mathbf{X}(m,k), \tag{4}$$

where $\mathbf{\Phi}_{\mathbf{NN}}(k) = \mathrm{E}[\mathbf{N}(k)\mathbf{N}^{\mathrm{H}}(k)]$ is the power spectral density (PSD) matrix of the stationary noise vector which can be estimated in speech absence periods. This random variable can be modeled by scaled chi-squared distributions [6]:

$$p\left(\varphi|Z{=}i;\xi_i\right) = c_{\mathcal{X}}(r,\xi_i)\ \varphi^{\frac{r}{2}-1} \exp\left\{\frac{-\varphi r}{2(1+\xi_i)}\right\}, \tag{5}$$

where $c_{\mathcal{X}}(r,\xi_i)$ is the normalization constant, $r = 2D$ is the degree of freedom and $\xi_0 = 0$ for noise-only and $\xi_1 = \cdots = \xi_I = \xi$ are the fixed average *a-priori* SNR over all sources and all time indices.

The second part is the frequency and length normalized observation vector (NOV) suggested in [1]. Arbitrarily selecting the first sensor as the reference the NOV is given by:

$$\tilde{Y}_j(m,k) := \left|X_j(m,k)\right| \exp\left\{\mathrm{i}\frac{\arg[X_j(m,k)X_1^*(m,k)]}{2(k{-}1)f_s d_{\max}(Kc_v)^{-1}}\right\} \tag{6}$$

$$\mathbf{Y}(m,k) := \tilde{\mathbf{Y}}(m,k) \Big/ \left\|\tilde{\mathbf{Y}}(m,k)\right\| \tag{7}$$

where $\tilde{\mathbf{Y}} := [\tilde{Y}_j]_{j=1,\ldots,D}$, f_s is the sampling frequency, c_v is the wave propagation velocity and $d_{\max}$ is the maximum distance between the reference sensor and all other sensors. The benefit of frequency normalization is that it allows us to tie all frequencies components together but requires approximately linear phase response and thus the absence of spatial aliasing.

The statistics of the NOV can be modeled by a complex Watson distribution:

$$p\left(\mathbf{Y}|Z{=}i;\mathbf{W}_i,\kappa_i\right) = c_{\mathcal{W}}(D,\kappa_i)\ \exp\left\{\kappa_i \left|\mathbf{W}_i^{\mathrm{H}}\,\mathbf{Y}\right|^2\right\}, \tag{8}$$

where $c_{\mathcal{W}}(D,\kappa_i)$ is the normalization constant, $\kappa_0 = 0$ with an arbitrary $\mathbf{W}_0$ for the noise-only case and $\kappa_1 = \cdots = \kappa_I = \kappa$ for the case that speech is present cases. Due to the frequency normalization the $D{\times}1$ dimensional normalized mean orientation vector $\mathbf{W}_i$ is constant for all frequency bins which is different from [4]. This simplification avoids the inner permutation problem and facilitates exploitation of spectral correlations.

To exploit temporal and spectral correlations we consider $Z(m,k)$ as a 2D random Markov process as depicted in Fig. 1. A homogeneous and ergodic Markov process in equilibrium is assumed. Thus, the $(I{+}1)\times 1$ *a priori* probability (APP) vector for each hidden state is $\boldsymbol{\pi} := [P\left(Z(m,k){=}i\right)]_{i=0:I}\ \forall m,k$. The 2D-HMM requires the specification of a 3D transition matrix with ${}_{3\mathrm{D}}t(j_1,j_2,i) := P(Z(m,k){=}i|Z(m{-}1,k){=}j_1, Z(m,k{-}1){=}j_2)$. We reduce the complexity of the model by assuming that this transition matrix is *separable*, i.e. it can be decomposed into a product of horizontal transitions ${}_{\mathcal{H}}t(j,i) := P\left(Z(m,k){=}i|Z(m{-}1,k){=}j\right)$ and vertical transitions ${}_{\mathcal{V}}t(j,i) := P\left(Z(m,k){=}i|Z(m,k{-}1){=}j\right)$. Hence, we have

$${}_{3\mathrm{D}}t(j_1,j_2,i) = \frac{{}_{\mathcal{H}}t(j_1,i)\,{}_{\mathcal{V}}t(j_2,i)}{\sum_{\tilde{i}=0}^{I} {}_{\mathcal{H}}t(j_1,\tilde{i})\,{}_{\mathcal{V}}t(j_2,\tilde{i})}. \tag{9}$$

The transition probabilities are collected in a $(I{+}1)\times(I{+}1)$ horizontal transitions (HT)-matrix ${}_{\mathcal{H}}\mathbf{T} := [{}_{\mathcal{H}}t(j,i)]_{j=0:I;i=0:I}$ and a vertical transition (VT)-matrix ${}_{\mathcal{V}}\mathbf{T} := [{}_{\mathcal{V}}t(j,i)]_{j=0:I;i=0:I}$ of the same size. Note, that while temporal correlations are stored in the HT-matrix, spectral correlations are stored in VT-matrix.

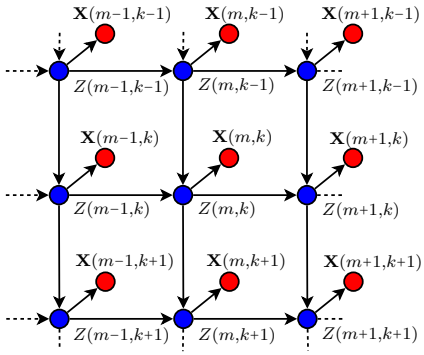

Fig. 1. *Bayesian model of 2D-HMM*

3. EXPECTATION MAXIMIZATION

In order to use the model above for BSS, we need to jointly reveal the identity of $Z(m,k)$ and to estimate the unknown parameters of the observation model solely from the observations. We apply the Expectation Maximization (EM) formalism to achieve this.

Let $\mathbf{\Theta}^{(\nu)} := \{\mathbf{W}_1^{(\nu)},\ldots,\mathbf{W}_I^{(\nu)}\}$ denote the set of unknown parameters to be estimated, where ν is the iteration index. In the E-step we have to collect the source dependent matrices

$$\mathbf{\Phi}_{\mathbf{YY},i}^{(\nu)} := \frac{\sum_{m=1}^{M}\sum_{k=1}^{K}\gamma_i^{(\nu)}(m,k)\mathbf{Y}(m,k)\mathbf{Y}^{\mathrm{H}}(m,k)}{\sum_{m=1}^{M}\sum_{k=1}^{K}\gamma_i^{(\nu)}(m,k)} \tag{10}$$

using the posterior probability

$$\gamma_i^{(\nu)}(m,k) := P\left(Z(m,k){=}i|\mathbf{X}(1{:}M,1{:}K);\mathbf{\Theta}^{(\nu)}\right). \tag{11}$$

Unfortunately, no efficient exact algorithms are known to compute the PP in a large 2D-HMM and we have to fall back to an approximate algorithm which is discussed in the next section.

For the M-step we obtain the eigenvalue equation

$$\mathbf{\Phi}_{\mathbf{YY},i}^{(\nu)}\ \mathbf{W}_i^{(\nu+1)} = v_i^{(\nu+1)}\ \mathbf{W}_i^{(\nu+1)}. \tag{12}$$

Since we want to maximize the likelihood the mean orientation $\mathbf{W}_i^{(\nu+1)}$ is updated by the eigenvector corresponding to the largest eigenvalue of the matrix $\mathbf{\Phi}_{\mathbf{YY},i}^{(\nu)}$. This computation can be done very efficiently with the power iteration.

Initializing $\mathbf{W}_i^{(0)}$ using a modified version the algorithm proposed in [7] which accounts for the frequency normaliza-

tion, the E-step and the M-step are iterated until convergence. We denote the final PP by $\gamma_i^{(\infty)}(m,k)$.

4. TURBO DECODING

To facilitate a compact notation we introduce the following vector operators: The binary operators $\circ$ and $\oslash$ are the element-wise product and the element-wise division of two vectors, respectively. The binary rescaling operator of two column vectors, denoted by the symbol $\propto$, is defined as $\mathbf{a} \propto \mathbf{b} := \mathbf{b}/(\mathbf{a}^{\mathrm{T}}\mathbf{b})$. If the first operand is a scalar then it will be expanded to a vector of the same size as the second operand by repetition. Thus, $1 \propto \mathbf{b}$ rescales the vector $\mathbf{b}$ so that all elements sum up to 1. The rescaling operator has the lowest precedence.

To derive an approximate algorithm for computation of the PP, i.e. to decode the 2D-HMM, we propose to split the decoding into horizontal and vertical processing steps and let the steps exchange information between each other [5].

Let us derive the vertical processing (VP)-step where we decode the 2D-HMM lattice column-by-column, but also account for information in the rows. For the m-th column we ignore the vertical dependencies in all other columns. Therefore, the VP-steps are independent from each other. This simplification allows us to factorize the joint PDF

$$\begin{aligned} p\left(Z(m,k), \mathbf{X}(1{:}M,1{:}K)\right) = P\left(Z(m,k)\right) \cdot \\ p\left(\mathbf{X}(1{:}M,1{:}k{-}1)|Z(m,k)\right) \cdot p\left(\mathbf{X}(1{:}M,k{+}1{:}K)|Z(m,k)\right) \cdot \\ p\left(\mathbf{X}(1{:}m{-}1,k)|Z(m,k)\right) \cdot p\left(\mathbf{X}(m{+}1{:}M,k)|Z(m,k)\right) \cdot \\ p\left(\mathbf{X}(m,k)|Z(m,k)\right). \end{aligned} \tag{13}$$

Fig. 2 depicts the statistical dependencies of VP-step in m-th column.

Let us introduce the following $(I{+}1)\times 1$ vectors

$$\mathbf{o}(m,k) := \boldsymbol{\pi} \propto \left[p\left(\mathbf{X}(m,k)|Z(m,k){=}i\right)\right]_{i=0:I}, \tag{14}$$

$${}_{\mathcal{V}}\boldsymbol{\gamma}(m,k) := \left[P\left(Z(m,k){=}i|\mathbf{X}(1{:}M,1{:}K)\right)\right]_{i=0:I}, \tag{15}$$

$${}_{\mathcal{V}}\boldsymbol{\alpha}(m,k) := \left[\frac{p\left(\mathbf{X}(1{:}M,1{:}k{-}1), Z(m,k){=}i\right)}{p\left(\mathbf{X}(1{:}M,1{:}k{-}1)\right)}\right]_{i=0:I}, \tag{16}$$

$${}_{\mathcal{V}}\boldsymbol{\beta}(m,k) := \left[\frac{p\left(\mathbf{X}(1{:}M,k{+}1{:}K)|Z(m,k){=}i\right)}{p\left(\mathbf{X}(1{:}M,k{+}1{:}K)\right)}\right]_{i=0:I}, \tag{17}$$

$${}_{\mathcal{V}}\mathbf{u}(m,k) := \left[\frac{p\left(\mathbf{X}(1{:}m{-}1,k)|Z(m,k){=}i\right)}{p\left(\mathbf{X}(1{:}m{-}1,k)\right)} \cdot \frac{p\left(\mathbf{X}(m{+}1{:}M,k)|Z(m,k){=}i\right)}{p\left(\mathbf{X}(m{+}1{:}M,k)\right)}\right]_{i=0:I}, \tag{18}$$

where $\mathbf{o}(m,k)$ is the observation evidence vector and ${}_{\mathcal{V}}\boldsymbol{\gamma}(m,k)$ is the vertical PP vector. The auxiliary variables ${}_{\mathcal{V}}\boldsymbol{\alpha}(m,k)$ is the vertical forward prediction vector (FPV), ${}_{\mathcal{V}}\boldsymbol{\beta}(m,k)$ is the vertical backward vector (BV) and ${}_{\mathcal{V}}\mathbf{u}(m,k)$ is the vertical junction vector (JV). As suggested by the factorization in (13) the PP can be easily computed by

$${}_{\mathcal{V}}\boldsymbol{\gamma}(m,k) = 1 \propto \mathbf{o}(m,k) \circ {}_{\mathcal{V}}\mathbf{u}(m,k) \circ {}_{\mathcal{V}}\boldsymbol{\alpha}(m,k) \circ {}_{\mathcal{V}}\boldsymbol{\beta}(m,k) \tag{19}$$

if the auxiliary variables are given.

The vertical FPV and vertical BV can be recursively computed by a slightly modified version of the FBA:

$${}_{\mathcal{V}}\boldsymbol{\alpha}(m,k) = 1 \propto {}_{\mathcal{V}}\mathbf{T}^{\mathrm{T}}\left({}_{\mathcal{V}}\boldsymbol{\alpha}(m,k{-}1) \circ \mathbf{o}(m,k{-}1) \circ {}_{\mathcal{V}}\mathbf{u}(m,k{-}1)\right), \tag{20}$$

$${}_{\mathcal{V}}\boldsymbol{\beta}(m,k) = \boldsymbol{\pi} \propto {}_{\mathcal{V}}\mathbf{T}\left({}_{\mathcal{V}}\boldsymbol{\beta}(m,k{+}1) \circ \mathbf{o}(m,k{+}1) \circ {}_{\mathcal{V}}\mathbf{u}(m,k{+}1)\right), \tag{21}$$

where ${}_{\mathcal{V}}\boldsymbol{\alpha}(m,1) = \boldsymbol{\pi}$ and ${}_{\mathcal{V}}\boldsymbol{\beta}(m,K) = [1,\ldots,1]^{\mathrm{T}}\ \forall m$. If there is no information in the temporal chains, which corresponds to ${}_{\mathcal{V}}\mathbf{u}(m,k) = [1,\ldots,1]^{\mathrm{T}}$, then the modified FBA is equivalent to the ordinary FBA along the spectral dependencies.

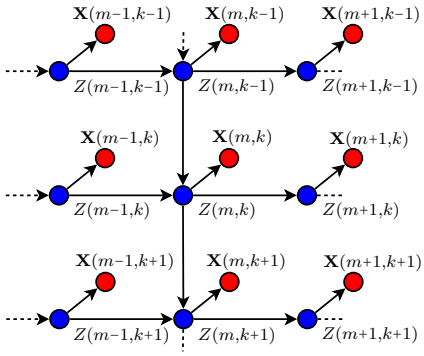

Fig. 2. *Statistical dependencies of vertical processing*

Since the 2D-HMM is symmetric a similar set of equations can be derived for a horizontal processing (HP) ignoring all other horizontal dependencies except for the considered row by substituting the indices in formulas of the VP, see Fig. 3. Analogous to (15) - (18) we define the horizontal FPV ${}_{\mathcal{H}}\boldsymbol{\alpha}(m,k)$, the horizontal BV ${}_{\mathcal{H}}\boldsymbol{\beta}(m,k)$, the horizontal JV ${}_{\mathcal{H}}\mathbf{u}(m,k)$ and the horizontal PP ${}_{\mathcal{H}}\boldsymbol{\gamma}(m,k)$. The modified FBA for the horizontal processing is given by:

$${}_{\mathcal{H}}\boldsymbol{\alpha}(m,k) = 1 \propto {}_{\mathcal{H}}\mathbf{T}^{\mathrm{T}}\left({}_{\mathcal{H}}\boldsymbol{\alpha}(m{-}1,k) \circ \mathbf{o}(m{-}1,k) \circ {}_{\mathcal{H}}\mathbf{u}(m{-}1,k)\right), \tag{22}$$

$${}_{\mathcal{H}}\boldsymbol{\beta}(m,k) = \boldsymbol{\pi} \propto {}_{\mathcal{H}}\mathbf{T}\left({}_{\mathcal{H}}\boldsymbol{\beta}(m{+}1,k) \circ \mathbf{o}(m{+}1,k) \circ {}_{\mathcal{H}}\mathbf{u}(m{+}1,k)\right), \tag{23}$$

$${}_{\mathcal{H}}\boldsymbol{\gamma}(m,k) = 1 \propto \mathbf{o}(m,k) \circ {}_{\mathcal{H}}\mathbf{u}(m,k) \circ {}_{\mathcal{H}}\boldsymbol{\alpha}(m,k) \circ {}_{\mathcal{H}}\boldsymbol{\beta}(m,k), \tag{24}$$

where ${}_{\mathcal{H}}\boldsymbol{\alpha}(1,k) = \boldsymbol{\pi}$ and ${}_{\mathcal{H}}\boldsymbol{\beta}(M,k) = [1,\ldots,1]^{\mathrm{T}}\ \forall k$.

The key to improve the modified FBA are the JVs ${}_{\mathcal{V}}\mathbf{u}(m,k)$ and ${}_{\mathcal{H}}\mathbf{u}(m,k)$. It is easy to verify that the JV for the VP ${}_{\mathcal{V}}\mathbf{u}(m,k)$ can be computed by the FBA of the HP by

$${}_{\mathcal{V}}\mathbf{u}(m,k) = \left({}_{\mathcal{H}}\boldsymbol{\alpha}(m,k) \oslash \boldsymbol{\pi}\right) \circ {}_{\mathcal{H}}\boldsymbol{\beta}(m,k) \tag{25}$$

if we set ${}_{\mathcal{H}}\mathbf{u}(m,k) = [1,\ldots,1]^{\mathrm{T}}\ \forall m,k$. The JV for the HP ${}_{\mathcal{H}}\mathbf{u}(m,k)$ can be computed by the FBA of the VP by

$${}_{\mathcal{H}}\mathbf{u}(m,k) = \left({}_{\mathcal{V}}\boldsymbol{\alpha}(m,k) \oslash \boldsymbol{\pi}\right) \circ {}_{\mathcal{V}}\boldsymbol{\beta}(m,k) \tag{26}$$

if we set ${}_{\mathcal{V}}\mathbf{u}(m,k) = [1,\ldots,1]^{\mathrm{T}}\ \forall m,k$. Obviously, the JVs are suboptimal since they are computed by ignoring the other horizontal or vertical dependencies in the simplified model in Fig. 2 and Fig. 3.

Now it seems reasonable to iterate the modified FBA in

VP and HP to obtain increasingly better estimates. Arbitrarily starting with ${}_{\mathcal{H}}\mathbf{u}(m,k) = [1,\ldots,1]^{\mathrm{T}}\ \forall m,k$ we apply the VP-step with the modified FBA in the equations (20) - (21). Then, we compute the horizontal JV with eq. (26) to prepare a HP-step using the equations (22) - (23). Now, we recompute the augmented vertical JV with eq. (25) and start over with a VP-step again. Thus, we arrived at an iterative decoding scheme where we use the JVs to exchange *extrinsic* information between VP and HP cycles. In practice the PPs are stable only after several VP and HP cycles.

The principle to exchange *extrinsic* information between alternating FBA processing steps is also known from turbo decoding. The term *extrinsic* stresses the requirement that the JVs should be an independent source of information. This, however, can be only guaranteed in the first iteration, since the interleaver, which ensures independence between the partial coders/decoders in turbo codes, is not available in our application.

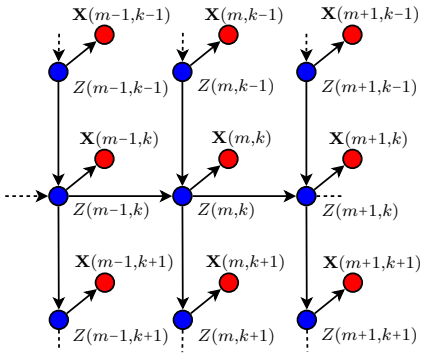

Fig. 3. *Statistical dependencies of horizontal processing*

5. SOURCE SIGNAL EXTRACTION

We propose a spatio-temporal filtering by beamforming followed by a single channel post-filter to recover each source signal where we use the final PP $\gamma_i^{(\infty)}(m,k)$ as an adaptation control. For each source $i \in \{1,\ldots,I\}$ we compute the target PSD matrix

$$\mathbf{\Phi}_{\text{target},i}(k) := \frac{\sum_{m=1}^{M} \gamma_i^{(\infty)}(m,k)\mathbf{X}(m,k)\mathbf{X}^{\mathrm{H}}(m,k)}{\sum_{m=1}^{M} \gamma_i^{(\infty)}(m,k)} \tag{27}$$

and the distortion PSD matrix

$$\mathbf{\Phi}_{\text{dist},i}(k) := \frac{\sum_{m=1}^{M}(1-\gamma_i^{(\infty)}(m,k))\mathbf{X}(m,k)\mathbf{X}^{\mathrm{H}}(m,k)}{\sum_{m=1}^{M}(1-\gamma_i^{(\infty)}(m,k))}. \tag{28}$$

According to the MaxSNR-beamforming approach [8] the coefficients which are used for computing the intermediate spatial filtering output

$$\tilde{S}_i(m,k) := A_i(k)\ \mathbf{F}_i^{\mathrm{H}}(k)\mathbf{X}(m,k). \tag{29}$$

are given by the principal eigenvector $\mathbf{F}_i$ of the generalized eigenvalue equation

$$\mathbf{\Phi}_{\text{target},i}(k)\ \mathbf{F}_i(k) = \upsilon_i(k)\ \mathbf{\Phi}_{\text{dist},i}(k)\ \mathbf{F}_i(k). \tag{30}$$

The required gain normalization $A_i(k)$ can be found by minimizing the difference between the averaged signal power at all sensors and the beamforming output at the TF-slots where the target signal is considered active:

$$\Bigg(\sum_{m=1}^{M}\gamma_i^{(\infty)}(m,k)\ \left|\tilde{S}_i(m,k)\right|^2 - \frac{1}{D}\sum_{m=1}^{M}\gamma_i^{(\infty)}(m,k)\ \mathrm{tr}\left(\mathbf{X}(m,k)\mathbf{X}^{\mathrm{H}}(m,k)\right)\Bigg)^2 \rightarrow \min \tag{31}$$

This results in the following gain normalization factor:

$$A_i(k) := \sqrt{\frac{\mathrm{tr}\left(\mathbf{\Phi}_{\text{target},i}(k)\right)}{D\ \mathbf{F}_i{}^{\mathrm{H}}(k)\mathbf{\Phi}_{\text{target},i}(k)\mathbf{F}_i(k)}}. \tag{32}$$

The subsequent spectral subtraction based post-filtering requires an estimate of the residual noise and crosstalk power λ_i present in $\tilde{S}_i(m,k)$. A well known solution for this is to apply recursive averaging

$$\lambda_i(m,k) := \big(1-\mu_i(m,k)\big)\lambda_i(m-1,k) + \mu_i(m,k)\left|\tilde{S}_i(m,k)\right|^2, \tag{33}$$

where the time variant learning factor $\mu_i(m,k)$ is dependent on the target speech presence probability. The value of $\mu_i(m,k)$ itself is driven by the posterior state probability

$$\mu_i(m,k) := \mu_{\max}\left(1-\gamma_i^{(\infty)}(m,k)\right), \tag{34}$$

where $\mu_{\max}$ is some maximum learning rate. Thus, the learning factor $\mu_i(m,k)$ is high if the probability that the source i is active in the considered TF-slot is low.

The final estimate of the clean target signal STFT is given by

$$\hat{S}_i(m,k) := G_i(m,k)\ \tilde{S}_i(m,k), \tag{35}$$

where $G_i(m,k)$ is the gain function. Here, we employ the Wiener filter gain

$$G_i(m,k) := \max\left\{\frac{\xi_i(m,k)}{1+\xi_i(m,k)}, G_{\min}\right\}, \tag{36}$$

where $\xi_i(m)$ is the instantaneous *a-priori* SNR. The lower bound $G_{\min}$ of the gain has to be chosen as a trade-off between reduction of musical tones and suppression of noise and interferers. The instantaneous *a-priori* SNR is estimated in the well known decision-directed way

$$\xi_i(m,k) = \mu_{\text{AP}}\frac{\left|\hat{S}_i(m-1,k)\right|^2}{\lambda_i(m-1,k)} + (1-\mu_{\text{AP}})\max\left\{\zeta_i(m,k)-1,0\right\}, \tag{37}$$

where the weighting factor μ_{AP} controls suppression of speech transients and $\zeta_i(m,k) := |\tilde{S}_i(m,k)|^2/\lambda_i(m,k)$ is the single channel *a-posteriori* SNR.

6. SIMULATION RESULTS AND CONCLUSION

We experimentally evaluate the proposed BSS method in a simulated reverberant enclosure in a setup similar to [1] with

$I = 3$ speech source signals taken from the TIMIT database (5 male and 5 female). Several utterances of one speaker were concatenated to obtain signal lengths of $10\,\mathrm{s}$ each. A sensor array with four sensors arranged at the vertices of regular tetrahedron with lateral length of $2\,\mathrm{cm}$ is used. The speech sources were randomly positioned around the microphone array in 3 different locations. To simulate coherent noise recordings of the fan noise of a video projector is placed in the reverberant enclosure. The enclosed angles between the positions of all sources are at least 30° to ensure spatial diversity. The power ratio of the sources and the coherent noise was about $10\,\mathrm{dB}$. To every microphone white noise at the level of $-20\,\mathrm{dB}$ below signal power was added. The time domain signals sampled at $16\,\mathrm{kHz}$ was converted into STFT domain with 1024-point Blackman window with a 75% overlap.

The system performance was evaluated in terms of the gain in signal-to-interference-ratio (SIR), signal-to-noise-ratio (SNR) and signal-to-distortion-ratio (SDR) [9] between the signals components at a reference sensor and the signal components at the system output. To demonstrate the effectiveness of the application of 2D-HMMs we compare the system performance with the case of using a 1D-HMM along the time axis, i.e. exploiting temporal correlations only [4], and with the case where the hidden states are assumed to be independent and identically distributed (i.i.d.), i.e. neglecting all correlations. Note, that [4] is modified to also use the frequency normalization (6). In Fig. 4 the performance is depicted as a function of the reverberation time T_{60}.

A focus of this paper is to demonstrate the benefits of exploiting correlations of adjacent TF-slots for noisy BSS. These benefits can be clearly seen in the performance curves in Fig. 4. Exploiting correlations results in improved performance in all cases and w.r.t. all measures. For low reverberation times the performance of all cases is very high and only very small advantages are gained from using 1D-HMMs or 2D-HMMs. Although the model complexity of the 1D-HMMs which were used in this evaluation are rigorously reduced compared to the 1D-HMMs we observe small but consistent performance improvements for this case. If reverberation time is increasing the performance advantage from using 2D-HMMs becomes apparent. Thus, we can conclude that exploiting temporal and spectral correlations of adjacent TF-slots is worthwhile for noisy BSS in highly reverberant recording conditions.

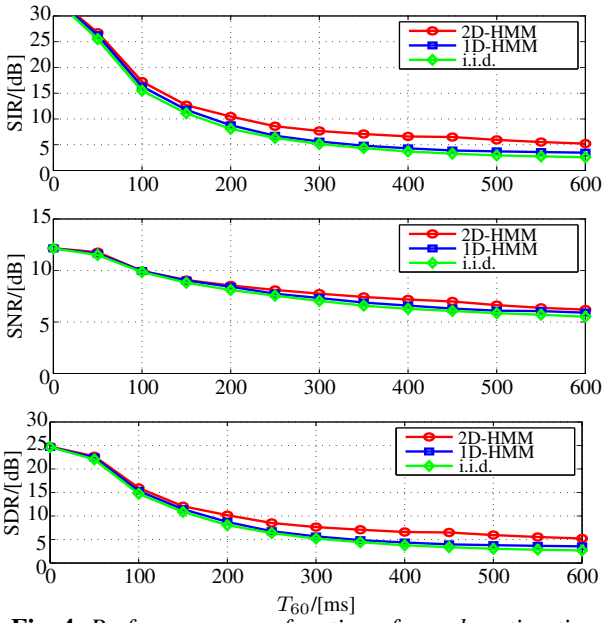

Fig. 4. *Performance as a function of reverberation time*

7. REFERENCES

[1] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, 2007.

[2] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Speech and Audio Processing*, vol. 19, no. 3, 2011.

[3] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Speech and Audio Processing*, vol. 18, no. 7, 2010.

[4] D. H. Tran Vu and R. H. Haeb-Umbach, "Exploiting temporal correlations in joint multichannel speech separation and noise suppression using hidden Markov models," in *Proc. IWAENC*, 2012.

[5] D. H. Tran Vu and R. Haeb-Umbach, "Using the turbo principle for exploiting temporal and spectral correlations in speech presence probability estimation," in *Proc. ICASSP*, 2013.

[6] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Speech and Audio Processing*, vol. 16, no. 5, 2008.

[7] D. H. Tran Vu and R. Haeb-Umbach, "On initial seed selection for frequency domain blind speech separation," in *Proc. Interspeech*, 2011.

[8] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, 2007.

[9] E. Vincent, C. Fevotte, and Gribonval R., "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 4, 2006.