An EM Approach to Integrated Multichannel Speech Separation and Noise Suppression

Dang Hai Tran Vu and Reinhold Haeb-Umbach
Department of Communications Engineering, University of Paderborn, 33098 Paderborn, Germany
{tran,haeb}@nt.uni-paderborn.de

***Abstract*—In this contribution we provide a unified treatment of blind source separation (BSS) and noise suppression, two tasks which have traditionally been considered different and for which quite different techniques have been developed. Exploiting the sparseness of the sources in the short time frequency domain and using a probabilistic model which accounts for the presence of additive noise and which captures the spatial information of the multi-channel recording, a speech enhancement system is developed which suppresses noise and simultaneously separates speakers in case multiple speakers are active. Source activity estimation and model parameter estimation form the E-step and the M-step of the Expectation Maximization algorithm, respectively. Experimental results obtained on the dataset of the Signal Separation Evaluation Campaign 2010 demonstrate the effectiveness of the proposed system.**

I. Introduction

Exploiting the sparseness of source signals in the short time frequency domain is a popular approach for blind source separation (BSS). In this approach each time-frequency (T-F) slot can be assigned to a single dominant source using spatial cues, e.g. [1] [2]. Recently, methods based on the Expectation Maximization (EM) framework for the detection of the dominant source and the mixing system identification emerged showing good separation results even under reverberant conditions, [3] [4].

In sparseness based BSS it is common to assume the absence of additive noise. Hence, the system performance can severely deteriorate if the noise level is significant. Actually stationary noise cannot be considered as an additional source, since noise is typically present in all T-F slots thus violating the sparseness assumption. In case of diffuse noise it also has very different spatial characteristics compared to the sources.

On the other hand the estimation of noise is the fundamental issue treated in the speech enhancement literature where complex spatial and spectral filtering approaches have been developed, e.g. [5]. A crucial issue is the estimation of the speech presence probability which is required to estimate the noise spectrum. This can be achieved by using solely spectral cues [6] or spatial and spectral cues in the case of multi-channel recordings, e.g. [7].

In this paper we propose to use the same modeling framework for dominant source detection and speech presence probability estimation. We model the normalized short time Fourier transform (STFT) coefficients of the microphone signal as a mixture of $(P+1)$ complex Watson distribution, where P is the number of active speech sources. We have proposed a Watson mixture model earlier [4] as it allows to capture the spatial information important for source separation, while disregarding the amplitudes, which are mainly determined by the scalar source signal. The additional $(P+1)$-st distribution is omnidirectional and meant to model the additive noise after spatial whitening. We employ the EM algorithm where in the E-step the speech presence probability or dominant source identity is estimated, while the parameters, such as source-to-microphone transfer functions ratios (TFR), are estimated in the M-step. These estimates are then employed for spatial beamforming and further spectral speech enhancement.

II. Framework

We consider a mixture of P independent source signals $S_i(m,k), i = 1,\dots,P$, captured by D microphones as $X_j(m,k), j = 1,\dots,D$ in the STFT domain, where in a reverberant environment the signal path from source i to microphone j is characterized by a multiplicative transfer function (MTF) $H_{ij}(k)$. Here, m is the time frame index and k denotes the frequency bin. Additionally, the microphone signals $X_j(m,k)$ are corrupted by additive noise $N_j(m,k)$. Using vector notation we have

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

where $\mathbf{X} = [X_1,\dots,X_D]^T$ is the observation vector, $\mathbf{H}_i = [H_{i1},\dots,H_{iD}]^T$ is the vector of MTFs, and $\mathbf{N} = [N_1,\dots,N_D]^T$ is the noise vector.

For sparse signals such as speech it is reasonable to assume that at any T-F slot (m,k) only a single source is active. Consequently, the observation model (1) can be reformulated as a doubly stochastic process with a hidden random variable $Z(m,k) \in \{0,\dots,P\}$:

$$\begin{aligned} &Z(m,k) = 0: \\ &\quad \mathbf{X}(m,k) = \mathbf{N}(m,k) \qquad (2) \\ &Z(m,k) = i, \qquad i = 1,\dots,P: \\ &\quad \mathbf{X}(m,k) = \mathbf{H}_i(k) S_i(m,k) + \mathbf{N}(m,k), \qquad (3) \end{aligned}$$

where $Z(m,k) = i$ with $i = 1,\dots,P$ indicates that source i is active and $Z(m,k) = 0$ indicates that only noise is present in a given T-F slot (m,k).

The proposed system processes each bin k independently. Therefore, we will drop the argument k in the next subsections for the sake of a simpler notation.

III. Separation using EM Algorithm

An important intermediate goal in BSS is to uncover the hidden variable $Z(m,k)$ and to identify the unknown mixing system solely from the observations $\mathbf{X}$. This can elegantly be achieved by the EM algorithm.

Let $\mathbf{\Phi_{NN}} = \mathrm{E}\left[\mathbf{N}\mathbf{N}^H\right]$ be the power spectral density (PSD) matrix of the stationary noise vector which can be estimated in speech absence periods. Here, $(\cdot)^H$ is the conjugate transpose operator. Our statistical modeling involves spatial whitening and unit-norm normalization

$$\tilde{\mathbf{X}}(m) = \mathbf{\Phi}_{\mathbf{NN}}^{-\frac{1}{2}}\mathbf{X}(m) \tag{4}$$

$$\mathbf{Y}(m) = \tilde{\mathbf{X}}(m) \Big/ \left\|\tilde{\mathbf{X}}(m)\right\|, \tag{5}$$

where $\|\cdot\|$ denotes the Euclidean norm. The whitening ensures that the conditional PSD matrix of the whitened observation vectors $\tilde{\mathbf{X}}|Z=i$ has the form

$$\mathrm{E}\left[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^H \middle| Z=i\right] = \sigma_i^2 \cdot \mathbf{W}_i\mathbf{W}_i^H + \mathbf{I} \tag{6}$$

where $\sigma_i^2 = \mathrm{E}\left[|S_i|^2\right] \cdot ||\mathbf{\Phi}_{\mathbf{NN}}^{-\frac{1}{2}}\mathbf{H}_i||^2$ and $\mathbf{I}$ is the identity matrix. The normalized complex vector $\mathbf{W}_i$ contains the spatial informations and is given by

$$\mathbf{W}_i = \mathbf{\Phi}_{\mathbf{NN}}^{-\frac{1}{2}}\mathbf{H}_i \Big/ \left\|\mathbf{\Phi}_{\mathbf{NN}}^{-\frac{1}{2}}\mathbf{H}_i\right\|. \tag{7}$$

The normalization to unit length corresponds to a mapping onto the unit hypersphere in D-dimensional complex vector space. By this, the important spatial information is kept, while variations due to the scalar signal source are disregarded. The spatial diversity of the sources assures the formation of distinct clusters on the hypersphere, where clusters in antipodal locations correspond to the same source, since the sign is due to the scalar source signal S_i.

Recently, we have proposed to model the statistics of the feature vectors $\mathbf{Y}$ for speech present T-F slots with a mixture of complex Watson probability density functions (PDF) [4]. While in [4] and also in the common BSS literature, e.g. [1] [2] [3], noise only T-F slots are ignored we suggest to model this case by an uniform distribution on the complex hypersphere in this paper. This is reasonable due to the spatial whitening in (4) which constrains the conditional PSD matrix of noise only slots to $\mathrm{E}\left[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^H \middle| Z=0\right] = \mathbf{I}$. Consequently, we have

$$p\left(\mathbf{Y};\mathbf{\Theta}\right) = \sum_{i=0}^{P} c_i \, p\left(\mathbf{Y}|Z=i;\mathbf{W}_i,\kappa_i\right), \tag{8}$$

where $\mathbf{\Theta} = \{c_0,\ldots,c_P,\mathbf{W}_0,\ldots,\mathbf{W}_P,\kappa_0,\ldots,\kappa_P\}$ is the unknown parameter set and $c_i = P(Z=i)$ are non-negative mixture weights which sum to one. The mixture components are given by complex Watson PDFs [8]

$$p\left(\mathbf{Y}|Z=i;\mathbf{W}_i,\kappa_i\right) = \frac{(D-1)!}{2\pi^D \, M\left(1,D,\kappa_i\right)} \, e^{\kappa_i \left|\mathbf{W}_i^H\mathbf{Y}\right|^2}, \tag{9}$$

where the mean orientation $\mathbf{W}_i$ is defined in (7), κ_i is the concentration parameter and $M(a,b,z)$ is the confluent hypergeometric function of the first kind. The greater the value of κ_i, the more the observations $\mathbf{Y}$ are concentrated around the mean orientation $\pm\mathbf{W}_i$. For $Z=0$ we set the concentration parameter to $\kappa_0 = 0$ with an arbitrary $\mathbf{W}_0$ to get a uniform distribution.

The E-step is equivalent to the estimation of the complete-data sufficient statistics which is given by the source dependent PSD matrices

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

and the expected class occurrence probability

$$\tilde{c}_i = \frac{1}{T}\sum_{m=1}^{T} \gamma_i^{(\nu)}(m). \tag{11}$$

Here

$$\begin{aligned}\gamma_i^{(\nu)}(m) &= P\left(Z(m)=i \middle| \mathbf{Y}(m);\hat{\mathbf{\Theta}}^{(\nu)}\right) \\ &= \frac{p\left(\mathbf{Y}(m) \middle| Z(m)=i;\hat{\mathbf{W}}_i^{(\nu)},\hat{\kappa}_i^{(\nu)}\right)\ \hat{c}_i^{(\nu)}}{\sum_{l=0}^{P} p\left(\mathbf{Y}(m) \middle| Z(m)=l;\hat{\mathbf{W}}_l^{(\nu)},\hat{\kappa}_l^{(\nu)}\right)\ \hat{c}_l^{(\nu)}}\end{aligned} \tag{12}$$

is the a posteriori probability based on the parameter set $\hat{\mathbf{\Theta}}^{(\nu)}$, and ν is the iteration counter.

Parameter reestimation in the M-step can now by carried out using (10) and (11). The update of the mixing weights is trivially obtained by

$$\hat{c}_i^{(\nu+1)} = \tilde{c}_i. \tag{13}$$

The update equation for the mean orientation is given by eigenvalue equations

$$\mathbf{\Phi}_{\mathbf{YY},i}^{(\nu)} \, \hat{\mathbf{W}}_i^{(\nu+1)} = \upsilon_i^{(\nu+1)} \, \hat{\mathbf{W}}_i^{(\nu+1)}, \tag{14}$$

where the eigenvectors corresponding to the largest eigenvalues υ_i of the source dependent PSD matrices $\mathbf{\Phi}_{\mathbf{YY},i}^{(\nu)}$ are efficiently computed with the power iteration method [9]. Estimates for the concentration parameters are recalculated by

$$\kappa_i^{(\nu+1)} = \eta_D^{-1}\left(\frac{\upsilon_i^{(\nu+1)}}{\hat{c}_i^{(\nu+1)}}\right), \tag{15}$$

where $\kappa = \eta_D^{-1}(\mu)$ is the inverse of the function

$$\mu = \eta_D(\kappa) = \frac{M\left(2,D+1,\kappa\right)}{D \cdot M\left(1,D,\kappa\right)}. \tag{16}$$

Since $\eta_D(\kappa)$ is a ratio of confluent hypergeometric functions there is no analytical inverse function. Hence, we fall back on spline based function approximations for low concentrations and the approximation $\eta_D^{-1}(\mu) \approx (D-1)/(1-\mu)$ for high concentrations [8]. In Fig. 1 the curves of $\eta_D^{-1}(\mu)$ for various numbers of sensors D are plotted.

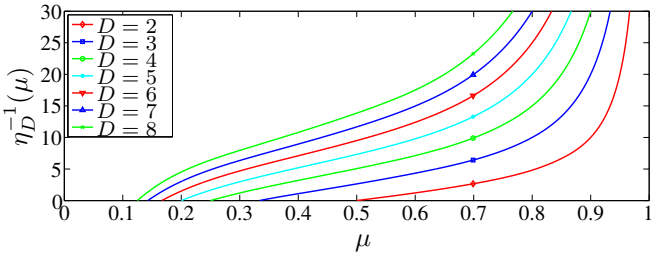

Fig. 1. Shape of $\eta_D^{-1}(\mu)$ for various number of sensors D

Starting with an initial guess, the E-step (10) - (12) and

the M-step (13) - (15) are iterated until no significant changes of the estimates occur. We will denote the estimates of the parameter set in steady state by $\hat{\boldsymbol{\Theta}}^{(\infty)}$.

IV. Permutation alignment and Beamforming

Since the EM algorithm is carried out in each frequency bin separately we are suffering from arbitrary ordering of the clusters in each bin. In order to properly reconstruct separated signals in the time-domain, clusters originating from the same source in frequency-domain should be aligned together. This well known problem in frequency domain BSS can be solved by finding mappings which minimize the inter frequency correlation among the posteriors $\gamma_i(m,k)$ of different clusters. We refer to [10] and assume in the following that the permutation problem is solved.

Besides the revelation of the hidden variable Z another benefit of the EM algorithm is that it provides a maximum likelihood blind mixing system identification as we can obtain estimates of the mixing system by

$$\hat{\mathbf{H}}_i = \frac{\boldsymbol{\Phi}_{\mathbf{NN}}^{\frac{1}{2}} \hat{\mathbf{W}}_i^{(\infty)}}{\left\| \boldsymbol{\Phi}_{\mathbf{NN}}^{\frac{1}{2}} \hat{\mathbf{W}}_i^{(\infty)} \right\|}, \qquad i = 1, \ldots, P. \tag{17}$$

Note that these are just estimates of the source-to-sensor transfer function ratios (TFR), not the transfer functions themself, as an eigenvector may have an arbitrary scaling [11].

For reconstruction of the source signals we first apply a spatial filtering with

$$\tilde{S}_i(m) = \mathbf{F}_i^H \mathbf{X}(m). \tag{18}$$

The TFR estimates are used to compute the minimum variance beamforming coefficients by

$$\mathbf{F}_i = \frac{\boldsymbol{\Phi}_{\mathbf{XX}}^{-1} \hat{\mathbf{H}}_i}{\hat{\mathbf{H}}_i^H \boldsymbol{\Phi}_{\mathbf{XX}}^{-1} \hat{\mathbf{H}}_i}, \tag{19}$$

where $\boldsymbol{\Phi}_{\mathbf{XX}} = \mathrm{E}\left[\mathbf{X}\mathbf{X}^H\right]$. In a practical setup it is necessary to apply diagonal loading with a small constant on $\boldsymbol{\Phi}_{\mathbf{XX}}$ for stability reasons and to control the effect of noise and interference suppression.

V. Postfiltering and System overview

To further enhance the signals produced by spatial filtering we apply a spectral subtraction based postfiltering which requires an estimator for residual noise and crosstalk power λ_i present in $\tilde{S}_i(m)$. A common practice for this is to apply recursive averaging

$$\lambda_i(m) = \big(1 - \alpha_i(m)\big)\lambda_i(m-1) + \alpha_i(m) \left| \tilde{S}_i(m) \right|^2, \tag{20}$$

where the time variant smoothing factor $\alpha_i(m)$ is dependent on the target speech presence probability. A reasonable value for $\alpha_i(m)$ can be obtained as follows

$$\alpha_i(m) = \alpha_{\max} \left(1 - \gamma_i^{(\infty)}(m)\right), \tag{21}$$

where $\alpha_{\max}$ is some maximum learning rate. So, if the i-th source is dominant ($\gamma_i^{(\infty)}$ is large), the current beamforming output $\tilde{S}_i(m)$ is disregarded for the update of the estimate of the distortion in the i-th beamformer output. Observe that while transient noise components are problematic in single channel speech enhancement our system can benefit from spatial informations to cope with these distortions.

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

$$\hat{S}_i(m) = G_i(m)\ \tilde{S}_i(m), \tag{22}$$

where $G_i(m)$ is the gain function. In this paper we employ the Wiener filter gain

$$G_i(m) = \max\left\{ \frac{\xi_i(m)}{1 + \xi_i(m)}, G_{\min} \right\}, \tag{23}$$

where ξ_i is the a priori SNR. The lower bound $G_{\min}$ of the gain has to be chosen as a compromise between reduction of musical tones and suppression of noise and interferes. The a priori SNR is estimated in the well known decision-directed way [12]

$$\xi_i(m) = \beta \frac{\left| \hat{S}_i(m-1) \right|^2}{\lambda_i(m-1)} + (1-\beta)\max\left\{\zeta_i(m) - 1, 0\right\}, \tag{24}$$

where the weighting factor β controls the trade-off between residual perturbation suppression and distortion of speech transients. $\zeta_i(m)$ is the a posteriori SNR

$$\zeta_i(m) = \frac{\left| \tilde{S}_i(m) \right|^2}{\lambda_i(m)}. \tag{25}$$

In Fig. 2 a flow chart of the proposed integrated BSS and noise suppression system is depicted. If we run the system in single source mode, $P = 1$, the permutation alignment block can be omitted. A voice activity detection (VAD) is required for estimation of the noise-only PSD matrix $\boldsymbol{\Phi}_{\mathbf{NN}}$ in the spatial whitening step (4). Note that the VAD is much simpler than the frequency bin wise determination of $\gamma_0^{(\infty)}(m)$ performed by the EM algorithm and still serves our purpose well, since only a coarse detection on a frame basis is required.

VI. Evaluation

The algorithm is evaluated on the "Source separation in the presence of real-world background noise" task of the second signal separation evaluation campaign (SiSEC2010) [13]. The SiSEC2010 task consists of three different live recorded scenarios: Square (Sq), Cafeteria (Ca) and Subway (Su). In all scenarios the noise is spatially and spectrally nonstationary. We focus here on the development dataset with $D = 4$ linearly arranged microphones with $8.5\,\mathrm{cm}$ spacing, since we need access to the contribution of each source and that of noise in order to evaluate performance. The number of sources are $P = 1$ and $P = 3$ to demonstrate single source signal enhancement mode and BSS mode. All signals are $10\,\mathrm{s}$ long and sampled at $16\,\mathrm{KHz}$. In Table I system parameters are summarized.

The system performance was evaluated in terms of signal-to-interference-ratio (SIR), signal-to-noise-ratio (SNR), signal-to-distortion-ratio (SDR) and signal-to-artifacts-ratio (SAR) as proposed in [14].

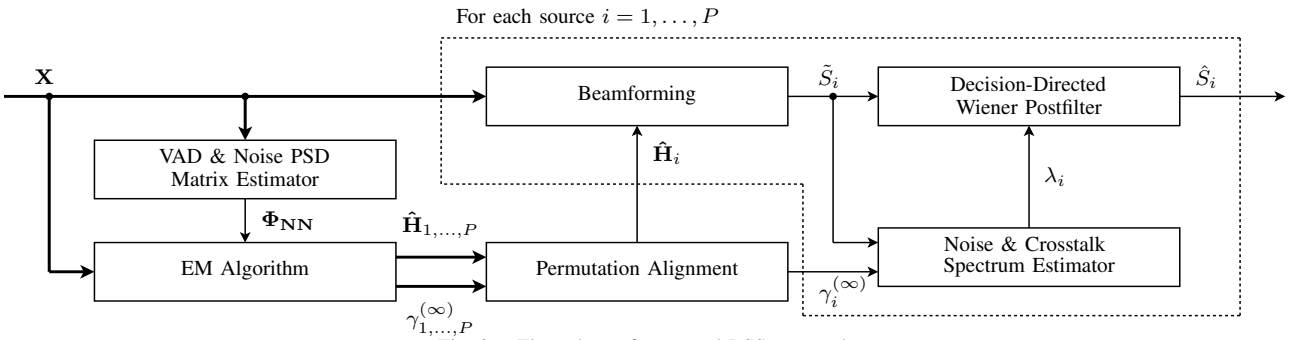

Fig. 2. Flow chart of proposed BSS approach

Audio examples and detailed system performance of the separation and speech enhancement are available at [15]. In Table II system performance is summarized. For proper assessment we also provide baseline input SIR and input SNR measured at the second microphone. We observed excellent source separation performance at the output in low reverberation conditions in the (Sq) scenario. As expected separation performance degrades for higher reverberation in (Ca) and (Su). The consistent good noise suppression in all conditions, despite their different reverberation time, could be an indication that noise is mostly diffuse in the dataset. Informal listening tests confirm good audio quality indicated by SDR and SAR, where in some cases musical tones are present.

Frame length:	1024	$\alpha_{\text{max}} =$	0.05
Frame shift:	256	$\beta =$	0.92
Window type:	Hamming	$G_{\text{min}} =$	0.1

TABLE I
SYSTEM PARAMETERS

Scenario	Input		Output			
	SIR	SNR	SIR	SNR	SDR	SAR
Square (3 sources)	-3.49	4.14	24.04	20.54	11.37	12.73
Cafeteria (3 sources)	-3.30	8.19	12.59	23.56	7.62	10.63
Subway (3 sources)	-3.81	1.75	9.17	21.62	5.70	10.63
Square (1 source)	-	-0.37	-	14.87	9.94	12.15
Cafeteria (1 source)	-	5.26	-	19.25	12.12	13.24
Subway (1 source)	-	0.67	-	21.06	13.29	14.22

TABLE II
AVERAGED PERFORMANCE RESULTS ON THE DEVELOPMENT DATASET

VII. CONCLUSION

In this paper, we proposed a novel unified approach for blind source separation and speech enhancement. Based on a sparseness model for the observations we derived an EM algorithm for blind system identification and dominant source activity detection. This information is exploited in a beamforming and postfiltering system for spatial and spectral filtering. The algorithm demonstrates its ability to cope with the challenging conditions of SiSEC2010 where a significant amount of noise is present.

In future research we will focus on the examination of under-determined source separation, usage of more sophisticated gain functions and additional tracking for highly spatially non-stationary noise in speech present frames. Additionally, a block-online implementation of the EM algorithm and tuning towards low latency system behavior is preferable for many applications.

ACKNOWLEDGMENT

This work was in part supported by the German Research Foundation (DFG) under contract number HA 3455/4-1.

REFERENCES

[1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, 2004.
[2] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, 2007.
[3] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007.
[4] D. H. Tran-Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. ICASSP*, 2010.
[5] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.
[6] T. Gerkmann, M. Krawczyk, and R. Martin, "Speech presence probability estimation based on temporal cepstrum smoothing," in *Proc. ICASSP*, 2010.
[7] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function gsc and postfiltering," *IEEE Trans. Speech and Audio Processing*, 2004.
[8] K. V. Mardia and I. L. Dryden, "The complex watson distribution and shape analysis," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 61, no. 4, 1999.
[9] J. Karhunen, "Adaptive algorithms for estimating eigenvectors of correlation type matrices," in *Proc. ICASSP*, 1984.
[10] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain bss," in *Proc. ISCAS*, 2007.
[11] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio Speech Language Processing*, vol. 15, no. 5, 2007.
[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, no. 6, 1984.
[13] F. Theis, G. Nolte, and S. Araki, "Signal separation evaluation campaign," *http://sisec.wiki.irisa.fr/*.
[14] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 4, 2006.
[15] *http://nt.upb.de/index.php?id=bss_iwaenc10*.