MMSE-BASED SOURCE EXTRACTION USING POSITION-BASED POSTERIOR PROBABILITIES

Maja Taseska and Emanuël A.P. Habets

International Audio Laboratories Erlangen*
Am Wolfsmantel 33, 91058 Erlangen - Germany
{maja.taseska,emanuel.habets}@audiolabs-erlangen.de

ABSTRACT

A scenario with multiple talkers and additive background noise is considered, where some talkers are active simultaneously and the activity of the talkers changes with time. We propose an MMSE-based method to blindly extract any talker using bin-wise position estimates obtained from distributed microphone arrays. In order to distinguish between different talkers, the position estimates are clustered using the expectation maximization algorithm. The resulting posterior probabilities allow to estimate the PSD matrices of the talkers and compute an MMSE-optimal linear filter for extracting each talker. We evaluate the performance of the proposed method in terms of noise and interference reduction and distortion of the desired speech signal at the output of a multichannel Wiener filter.

Index Terms— speech separation, PSD matrix estimation, distributed arrays, expectation maximization

1. INTRODUCTION

Most of the recently proposed approaches for source extraction in multi-talker scenarios rely on the sparsity of speech signals in the time-frequency (TF) domain [1]. Under the sparsity assumption, each TF bin can be associated with a single dominant source. This information can e.g. be used to compute a TF mask [2–4], or to estimate second order statistics of the sources, required for minimum mean-squared error (MMSE)-optimal linear filters [5–7].

Recently, several approaches for dominant source detection based on the expectation maximization (EM) algorithm have been proposed [2–7]. For instance, clustering of observation vectors in the signal domain has been proposed in [4, 6, 7], whereas the authors in [3] use inter-aural parameters extracted from the microphone signals. In addition, spectral features [5] or temporal correlation information [8] can be included to improve the clustering performance. Commonly, the EM-based approaches do not consider background noise, which can degrade clustering performance, and hence the subsequent source extraction. EM-based approaches aiming at joint source separation and noise reduction were recently developed in [5, 6]. The noise was also considered in [2], by including the noise signal in the likelihood function [9]. Nevertheless, the latter approach does not aim at noise reduction.

In this paper, we propose a source extraction method using EM-based clustering of position estimates obtained from distributed microphone arrays. The information about the dominant source for each TF-bin is used to estimate the power spectral density (PSD) matrices required for MMSE-based filters, such as the multichannel Wiener filter (MWF) [10]. The position estimates for each TF bin are obtained by triangulation of direction-of-arrival (DOA) estimates from at least two microphone arrays. In contrast to the approaches in [4–7] where clustering is done per frequency bin, we apply full-band clustering, as in [2, 3, 16]. Similar framework that uses as a feature the signal vector at one array was proposed in [5, 7]. In this work, we employ distributed arrays and TF bin-wise position estimates as features. The position has lower dimension than the signal vector, resulting in a low computational complexity and fast convergence of the EM algorithm. Moreover, in contrast to [5, 7], the noisy observations are removed from the EM by using a direct-to-diffuse ratio (DDR)-based multichannel speech presence probability (SPP) proposed by the present authors in [12]. The noise PSD matrix required for the MMSE-based filtering is estimated using this SPP.

The paper is organized as follows: in Section 2 the problem is formulated. Section 3 provides a brief overview of SPP-based noise PSD matrix estimation. The main contribution of the paper is described in Section 4, where the PSD matrices of the talkers are computed using position-based EM clustering. In Section 5, the performance of the algorithm is evaluated. Section 6 concludes the paper.

2. PROBLEM FORMULATION

Consider a scenario where M microphones from two or more distributed arrays capture an additive mixture of J talkers and background noise. In this work, we assume that the number of talkers J is known. The signal at the m-th microphone is given in the short-time Fourier transform (STFT) domain as follows

$$Y_m(n,k) = \sum_{j=1}^{J} X_m^{(j)}(n,k) + V_m(n,k), \tag{1}$$

where $X_m^{(j)}$, for $j = 1, 2 \ldots, J$, and V_m denote the complex spectral coefficients of the different talkers and the background noise, respectively, and n and k are the time and frequency indices, respectively. For brevity, we omit the time and frequency indices in the following, wherever possible. The microphone signals are written in vector notation as $\mathbf{y}(n) = [Y_1(n) \ldots Y_M(n)]^{\mathrm{T}}$ and the PSD matrix of $\mathbf{y}(n)$ is defined as $\mathbf{\Phi}_{\mathbf{y}}(n) = \mathrm{E}\left[\mathbf{y}(n)\,\mathbf{y}^{\mathrm{H}}(n)\right]$, where $(\cdot)^{\mathrm{H}}$ denotes the conjugate transpose of a vector or a matrix. The vectors $\mathbf{x}^{(j)}$ and $\mathbf{v}$ and the matrices $\mathbf{\Phi}_{\mathbf{x}}^{(j)}$ and $\mathbf{\Phi}_{\mathbf{v}}$ are defined similarly. The different speech signals and the noise are modelled as mutually uncorrelated, zero-mean random processes, such that

$$\mathbf{\Phi}_{\mathbf{y}}(n) = \sum_{j=1}^{J} \mathbf{\Phi}_{\mathbf{x}}^{(j)}(n) + \mathbf{\Phi}_{\mathbf{v}}(n). \tag{2}$$

In order to describe the activity of the different talkers in every

*A joint institution of the University Erlangen-Nuremberg and Fraunhofer IIS, Germany.

TF bin, the following hypotheses are introduced

$$\mathcal{H}_{\mathbf{v}}: \ \mathbf{y}(n) = \mathbf{v}(n), \text{ indicating speech absence} \tag{3a}$$

$$\mathcal{H}_{\mathbf{x}}: \ \text{indicating speech presence,.} \tag{3b}$$

$$\mathcal{H}_{\mathbf{x}}^{j}: \ \text{indicating that the } j\text{-th source is dominant, i.e} \tag{3c}$$

$$\mathbf{y}(n) \approx \mathbf{x}^{(j)}(n) + \mathbf{v}(n).$$

Assuming sparsity of speech in the STFT domain [1] and mildly reverberant environment, it holds that whenever $\mathcal{H}_{\mathbf{x}}$ is true, exactly one of the J hypotheses in (3c) is also true. In other words, most of the energy contribution during speech corresponds to one talker.

If a desired talker is denoted by an index $d \in \{1, 2, \ldots, J\}$, the goal is to estimate the signal $X_m^{(d)}$. In order to achieve this by a linear MMSE-optimum spatial filter, the PSD matrices of the noise and the different talkers are required.

3. SPP-BASED NOISE PSD MATRIX ESTIMATION

State-of-the-art multichannel noise PSD estimation methods employ a recursive update based on a SPP [11]. Let $q(n) = p[\mathcal{H}_{\mathbf{v}}(n)]$ denote the *a priori* speech absence probability (SAP) and

$$\xi(n) = \mathrm{tr}\{\mathbf{\Phi}_{\mathbf{v}}^{-1}(n)\, \mathbf{\Phi}_{\mathbf{x}}(n)\}, \tag{4}$$

$$\beta(n) = \mathbf{y}^{H}(n)\mathbf{\Phi}_{\mathbf{v}}^{-1}(n)\mathbf{\Phi}_{\mathbf{x}}(n)\mathbf{\Phi}_{\mathbf{v}}^{-1}(n)\mathbf{y}(n), \tag{5}$$

where $\mathrm{tr}\{\cdot\}$ denotes the trace operator. If the spectral coefficients of the speech and the noise signals are modelled as complex Gaussian vectors [13], the multichannel SPP is given by

$$p[\mathcal{H}_{\mathbf{x}} \,|\, \mathbf{y}(n)] = \left\{1 + \frac{q(n)}{1-q(n)}[1+\xi(n)]\, e^{-\frac{\beta(n)}{1+\xi(n)}}\right\}^{-1}, \tag{6}$$

where $\mathbf{\Phi}_{\mathbf{x}} = \mathbf{\Phi}_{\mathbf{y}} - \mathbf{\Phi}_{\mathbf{v}}$. In this paper, we use the direct-to-diffuse ratio-based a priori SAP q proposed by the present authors in [12]. In this manner, the speech signals that are coherent across the arrays are detected and do not leak into the noise PSD matrix estimate.

The recursive update is obtained as a weighted sum of the noisy spectral power values from the current frame and an estimate of the noise PSD from the previous frame [14, 15] as follows

$$\widehat{\mathbf{\Phi}}_{\mathbf{v}}(n) = (1 - p[\mathcal{H}_{\mathbf{x}} \,|\, \mathbf{y}])\, \left(\alpha_v\, \widehat{\mathbf{\Phi}}_{\mathbf{v}}(n-1) + (1-\alpha_v)\, \mathbf{y}\mathbf{y}^{\mathrm{H}}\right) + p[\mathcal{H}_{\mathbf{x}} \,|\, \mathbf{y}]\, \widehat{\mathbf{\Phi}}_{\mathbf{v}}(n-1), \tag{7}$$

where $\widehat{\mathbf{\Phi}}_{\mathbf{v}}$ is the estimated noise PSD matrix and $0 \leq \alpha_v < 1$ is a chosen smoothing constant. The SPP in (6) is computed using the noise PSD matrix from the previous frame, followed by an update of the current PSD matrix, as given by (7).

4. SOURCE EXTRACTION

In order to estimate the PSD matrix of each talker, the dominant talker in each TF bin where speech is present needs to be identified. We propose a method that employs EM-based clustering of bin-wise position estimates. Similarly as in [2, 3, 16], the clustering is performed in a fullband manner, such that the training set for the EM algorithm consists of position estimates collected over certain time interval for all frequency bins.

4.1. Source PSD matrix estimation

Let $p[\mathcal{H}_{\mathbf{x}}^{j} \,|\, \mathbf{y}(n)]$ denote the posterior probability that the j-th source is dominant. Moreover, let the PSD matrix $\mathbf{\Phi}_{\mathbf{x}+\mathbf{v}}^{(j)}(n)$ be defined as

$$\mathbf{\Phi}_{\mathbf{x}+\mathbf{v}}^{(j)}(n) = \mathbf{\Phi}_{\mathbf{x}}^{(j)}(n) + \mathbf{\Phi}_{\mathbf{v}}(n). \tag{8}$$

Following the recursive update given by (7), the PSD matrix $\mathbf{\Phi}_{\mathbf{x}}^{(j)}$ of the j-th talker can be estimated in two steps as follows

1. Recursively estimate $\mathbf{\Phi}_{\mathbf{x}+\mathbf{v}}^{(j)}(n)$ according to

$$\widehat{\mathbf{\Phi}}_{\mathbf{x}+\mathbf{v}}^{(j)}(n) = \ p[\mathcal{H}_{x}^{j} \,|\, \mathbf{y}(n)]\ \left[\alpha_x\ \widehat{\mathbf{\Phi}}_{\mathbf{x}+\mathbf{v}}^{(j)}(n-1) + (1-\alpha_x)\, \mathbf{y}\mathbf{y}^{\mathrm{H}}\right] + \left(1 - p[\mathcal{H}_{x}^{j} \,|\, \mathbf{y}(n)]\right)\ \widehat{\mathbf{\Phi}}_{\mathbf{x}+\mathbf{v}}^{(j)}(n-1) \tag{9}$$

where $0 < \alpha_x < 1$ is a chosen smoothing constant;

2. Subtract the noise PSD matrix estimate $\widehat{\mathbf{\Phi}}_{\mathbf{v}}(n)$ [see Section 3]:

$$\widehat{\mathbf{\Phi}}_{\mathbf{x}}^{(j)}(n) = \widehat{\mathbf{\Phi}}_{\mathbf{x}+\mathbf{v}}^{(j)}(n) - \widehat{\mathbf{\Phi}}_{\mathbf{v}}(n). \tag{10}$$

In order to realize the recursive update (9) for each talker j, the bin-wise posterior probabilities $p[\mathcal{H}_{\mathbf{x}}^{j} \,|\, \mathbf{y}]$ are required.

4.2. Estimation of posterior probabilities

The posterior probability $p[\mathcal{H}_{\mathbf{x}}^{j} \,|\, \mathbf{y}]$ can be expressed as follows

$$p[\mathcal{H}_{\mathbf{x}}^{j} \,|\, \mathbf{y}] = p[\mathcal{H}_{\mathbf{x}}^{j} \,|\, \mathbf{y}, \mathcal{H}_{\mathbf{x}}] \cdot p[\mathcal{H}_{\mathbf{x}} \,|\, \mathbf{y}]. \tag{11}$$

The second factor represents the SPP described in Section 3, while the first factor allows to distinguish the different talkers. We propose the following position-based approximation for every TF-bin

$$p[\mathcal{H}_{\mathbf{x}}^{j} \,|\, \mathbf{y}, \mathcal{H}_{\mathbf{x}}] \approx p[\mathcal{H}_{\mathbf{x}}^{j} \,|\, \widehat{\mathbf{\Theta}}, \mathcal{H}_{\mathbf{x}}], \tag{12}$$

where the position $\widehat{\mathbf{\Theta}}$ is computed by triangulating DOA estimates from e.g., two distributed arrays. The distribution of $\widehat{\mathbf{\Theta}}$ when speech is present, is modelled by a Gaussian mixture (GM) [9], i.e.,

$$p[\widehat{\mathbf{\Theta}} \,|\, \mathcal{H}_{\mathbf{x}}] = \sum_{j=1}^{J}\ \pi_j\ \mathcal{N}\left(\widehat{\mathbf{\Theta}}; \boldsymbol{\mu}_j, \mathbf{\Sigma}_j\right) \tag{13}$$

where $\mathcal{N}\left(\widehat{\mathbf{\Theta}}; \boldsymbol{\mu}_j, \mathbf{\Sigma}_j\right)$ denotes a Gaussian distribution with mean $\boldsymbol{\mu}_j$ and covariance matrix $\mathbf{\Sigma}_j$ and π_j are the mixing coefficients. Note that the goal is to model the position distribution only during speech activity. Therefore, the EM training set comprises only TF bins where the SPP is above a threshold $p_{\min}$. Existing methods that detect noisy observations include VAD as done in [17], or using a "garbage source" component as proposed by Mandel *et al.* [3].

Given a training set of N observations $\mathcal{D} = \{\widehat{\mathbf{\Theta}}_1, \ldots \widehat{\mathbf{\Theta}}_N\}$, the GM parameters $\mathcal{P} = \{\pi_j, \boldsymbol{\mu}_j, \mathbf{\Sigma}_j, \ldots\}$ are found by maximizing the weighted log likelihood

$$\ln p(\mathcal{D} \ |\mathcal{P}) = \sum_{n=1}^{N}\ w(\widehat{\mathbf{\Theta}}_n)\, \ln \sum_{j=1}^{J}\ \pi_j\ \mathcal{N}\left(\widehat{\mathbf{\Theta}}_n; \boldsymbol{\mu}_j, \mathbf{\Sigma}_j\right), \tag{14}$$

where the power-based weighting function is computed as

$$w(\widehat{\mathbf{\Theta}}_n) = \frac{\mathrm{tr}[\mathbf{y}(n)\mathbf{y}^{\mathrm{H}}(n)]}{\sum_{i=1}^{N} \mathrm{tr}[\mathbf{y}(i)\mathbf{y}^{\mathrm{H}}(i)]}. \tag{15}$$

In the E-step of the algorithm, the posterior probabilities are computed using the current model parameters according to

$$p[\mathcal{H}_{\mathbf{x}}^{j} \mid \widehat{\mathbf{\Theta}}, \mathcal{H}_{\mathbf{x}}] = \frac{\pi_j\ \mathcal{N}\left(\widehat{\mathbf{\Theta}}; \boldsymbol{\mu}_j, \mathbf{\Sigma}_j\right)}{\sum_{i=1}^{J}\ \pi_i\ \mathcal{N}\left(\widehat{\mathbf{\Theta}}; \boldsymbol{\mu}_i, \mathbf{\Sigma}_i\right)}, \tag{16}$$

whereas in the M-step the mixture parameters are updated as follows

$$\boldsymbol{\mu}_j \longleftarrow \ \frac{\sum_{n=1}^{N}\ p[\mathcal{H}_{\mathbf{x}}^{j} \,|\, \widehat{\mathbf{\Theta}}_n] \cdot w(\widehat{\mathbf{\Theta}}_n)\widehat{\mathbf{\Theta}}_n}{\sum_{n=1}^{N}\ p[\mathcal{H}_{\mathbf{x}}^{j} \,|\, \widehat{\mathbf{\Theta}}_n] \cdot w(\widehat{\mathbf{\Theta}}_n)} \tag{17}$$

$$\mathbf{\Sigma}_j \longleftarrow \ \frac{\sum_{n=1}^{N}\ p[\mathcal{H}_{\mathbf{x}}^{j} \,|\, \widehat{\mathbf{\Theta}}_n] \cdot w(\widehat{\mathbf{\Theta}}_n)(\widehat{\mathbf{\Theta}}_n - \boldsymbol{\mu}_j)(\widehat{\mathbf{\Theta}}_n - \boldsymbol{\mu}_j)^{\mathrm{T}}}{\sum_{n=1}^{N}\ p[\mathcal{H}_{\mathbf{x}}^{j} \,|\, \widehat{\mathbf{\Theta}}_n] \cdot w(\widehat{\mathbf{\Theta}}_n)} \tag{18}$$

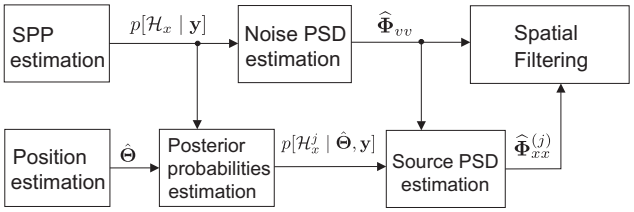


Fig. 1. Block diagram of the proposed framework

$$\pi_j \longleftarrow \frac{1}{N}\sum_{n=1}^{N} w(\widehat{\mathbf{\Theta}}_n)\; p[\mathcal{H}_{\mathbf{x}}^j \,|\, \widehat{\mathbf{\Theta}}_n]. \tag{19}$$

After computing the GMM parameters in the training step, the posterior probabilities can be computed for unseen data using (16), assuming that the geometry of the scenario is fixed. The proposed position-based source extraction framework is illustrated in Fig. 1.

4.3. Source extraction

Once the PSD matrices are computed, the d-th source, $d \in \{1, \ldots, J\}$ at the m-th microphone can be extracted by a MWF according to

$$\widehat{X}_m^{(d)}(n) = \mathbf{h}_{m,d}^{\mathrm{H}}(n)\; \mathbf{y}(n). \tag{20}$$

The MWF filter coefficients $\mathbf{h}_{m,d}$ are given by

$$\mathbf{h}_{m,d}(n) = \frac{\mathbf{\Phi}_{\mathbf{u}}^{-1}(n)\mathbf{\Phi}_{\mathbf{x}}^{(d)}(n)}{1 + \mathrm{tr}\{\mathbf{\Phi}_{\mathbf{u}}^{-1}(n)\mathbf{\Phi}_{\mathbf{x}}^{(d)}(n)\}}\mathbf{e}_{m_d}, \tag{21}$$

where $\mathbf{\Phi}_{\mathbf{u}}$ represents the noise-and-interference PSD matrix and

$$\mathbf{e}_{m_d} = [\underbrace{0 \ldots 0}_{m_d - 1}\; 1\; \underbrace{0 \ldots 0}_{M - m_d}]^T \tag{22}$$

where m_d denotes the reference microphone for extracting the desired source d. The reference microphone was computed using

$$m_d = \min_j \|\boldsymbol{\mu}_d - \mathbf{r}_j\|, \tag{23}$$

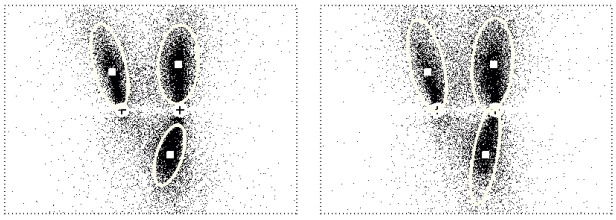
where $\boldsymbol{\mu}_d$ is the mean of the Gaussian distribution corresponding to source d, and $\mathbf{r}_j$ is the position of the j-th microphone. An estimate of the noise-and-interference PSD matrix $\mathbf{\Phi}_{\mathbf{u}}$ is given by

$$\widehat{\mathbf{\Phi}}_{\mathbf{u}}(n) = \widehat{\mathbf{\Phi}}_{\mathbf{v}}(n) + \sum_{j \neq d} \widehat{\mathbf{\Phi}}_{\mathbf{x}}^{(j)}(n), \tag{24}$$

5. PERFORMANCE EVALUATION

To evaluate the algorithm, microphone signals were simulated as a sum of speech signals with approximately equal power, convolved with simulated room impulse responses [18], a diffuse babble noise signal [19] with a segmental speech-to-noise ratio of 22 dB, and uncorrelated sensor noise with a segmental speech-to-noise ratio of 50 dB. The reverberation time was $T_{60} = 250$ ms.

The sampling frequency was 16 kHz and the STFT frame length $L = 1024$ samples, with 50% overlap. For the position estimation, two uniform circular arrays were used with three omnidirectional microphones, a diameter 2.5 cm and an inter-array spacing of 1.5 m. The DOA was computed for each array using instantaneous observation vectors, as proposed in [20], and the position was computed by a triangulation of the DOA vectors. Note that for the given diameter, the DOA estimates over the full frequency range are not affected by spatial aliasing. In addition to removing positions obtained from the TF bins where the SPP is below $p_{\min}$, positions within a radius of 20 cm around the microphone array centres are likely to correspond to noise-only frames [21] and were therefore discarded from the EM training set.



(a) Training during single-talk (b) Training during triple-talk

Fig. 2. Output of the EM algorithm (3 iterations). The actual source positions are denoted by white squares. The array location is marked by a plus symbol. The interior of each ellipse contains 85% probability mass of the respective Gaussian.

The averaging constants used in (7) and (9) were chosen as $\alpha_v = 0.9$ and $\alpha_x = 0.9$. The training sets comprised position estimates with SPP of at least $p_{\min} = 0.7$.

5.1. Performance measures

We denote the input fullband segmental signal-to-noise ratio (SNR), signal-to-interference ratio (SIR) and signal-to-noise-and-interference ratio (SINR) by $\mathcal{S}_{i,v}$, $\mathcal{S}_{i,b}$ and $\mathcal{S}_{i,u}$ respectively, and the corresponding output values by $\mathcal{S}_{o,v}$, $\mathcal{S}_{o,b}$ and $\mathcal{S}_{o,u}$ [10]. The time-domain speech interference signal at microphone m is denoted by $b_m(t)$. The output of the MWFs was assessed in terms of

1. Segmental SNR improvement $\mathcal{S}_{o,v} - \mathcal{S}_{i,v}$.
2. Segmental SIR improvement $\mathcal{S}_{o,x} - \mathcal{S}_{i,x}$.
3. Segmental speech distortion index ν_{sd}, as defined in [10].
4. Segment-wise interference reduction (SegIR).
5. PESQ score improvement [22], denoted by Δ-PESQ.

The segmental measures were computed by averaging over non-overlapping frames of 20 ms where only frames with input SIR or SNR between -25 dB and 40 dB were considered. The SegIR for the i-th segment at microphone m was computed as

$$\mathrm{SegIR}(i) = \frac{\sum_t b_m^2(t) \cdot w_i(t)}{\sum_t \widehat{b}_m^2(t) \cdot w_i(t)}, \tag{25}$$

where $\widehat{b}_m$ is the residual interference and w_i is a rectangular window equal to one during segment i and zero elsewhere. Δ-PESQ represents the difference of the PESQ scores of the inverse STFT of $\widehat{X}_m^{(d)}(n,k)$ and the inverse STFT of the mixture $Y_m(n,k)$.

5.2. Results

To demonstrate the outcome of the EM-based position clustering, the output after 3 iterations is shown in Fig. 2, for single-talk and triple-talk training. The training in both cases was performed using the position estimates over 4.5 s. Notably, even when training is done during constant triple-talk, the estimated model accurately reflects the distribution of the sources.

The extracted signals at the output of the MWFs were evaluated in two scenarios: a constant triple-talk scenario where 3 talkers are simultaneously active, and a more realistic meeting scenario where triple-talk is present only during short periods. The GM parameters estimated over 4.5 s long training segments are used to compute the posterior probabilities during the whole evaluated segments, as described by (11), (12) and (16). The mixtures, the original source signals and the extracted signals for are illustrated in Fig. 3, where in both cases the desired sources are successfully extracted.

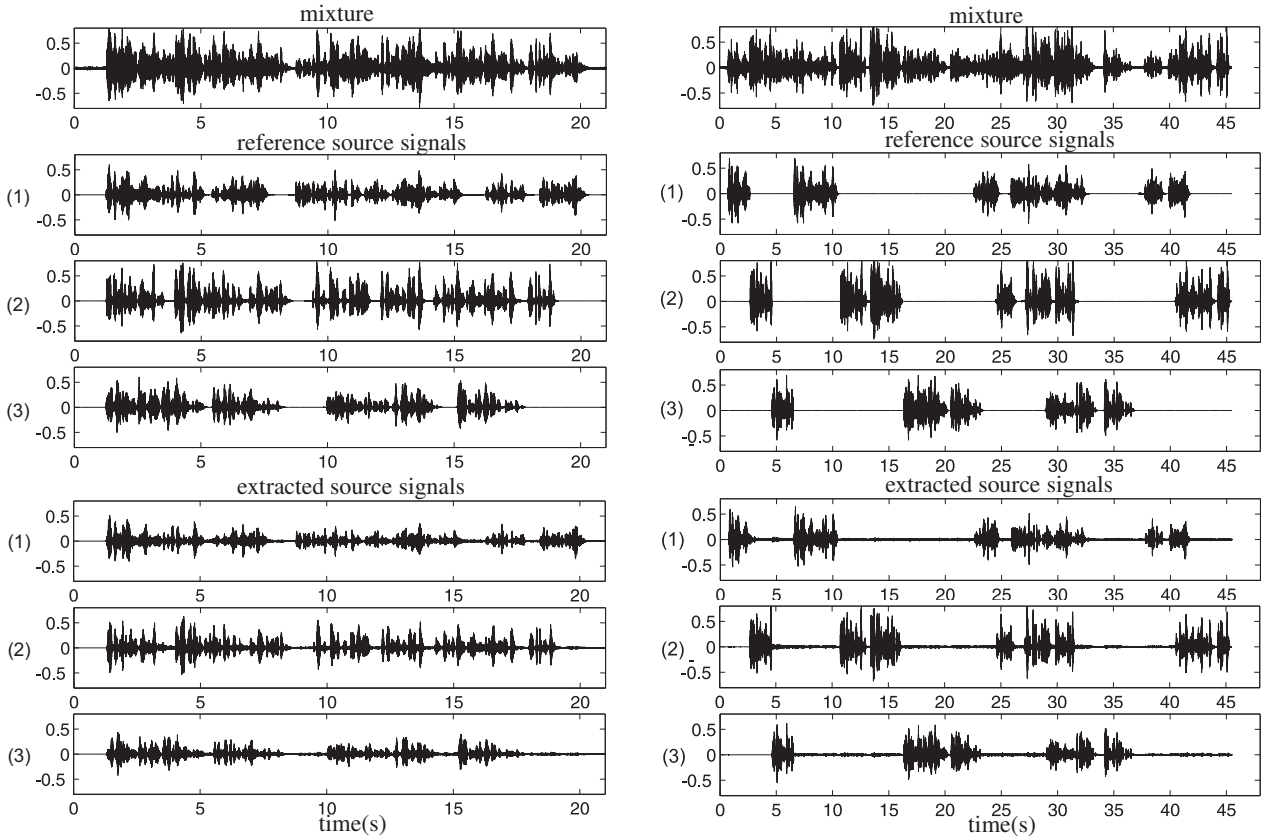


Fig. 3. Mixture, reference signals and extracted signals. Left: constant triple-talk scenario. Right: mainly single-talk scenario. The corresponding audio files are available at http://home.tiscali.nl/ehabets/publications/Taseska2013.html.

The performance of the proposed framework is compared to the performance when the dominant source is known *a priori*, such that ideal binary masks (IBM) [1] are used in (9) instead of the position-based posterior probabilities (PPP) $p[\mathcal{H}^j \mid \widehat{\mathbf{\Theta}}]$. The results are given in Tables 1 and 2. In the mostly single-talk scenario (Table 2), we computed the segmental interference reduction (Fig. 4), instead of the segmental SIR improvement. Note that in terms of all measures, the performance when using the PPP approaches the performance when using IBM. This indicates that the position-based estimation PSD matrix estimation is quite accurate, as corroborated by the good interference reduction shown in Fig. 4, even during a segment with a triple-talk. It should be mentioned that the interference reduction can be further improved by e.g. incorporating the posterior probabilities in a parametric multichannel Wiener filter.

6. CONCLUSIONS

A MMSE-based framework for source extraction using distributed arrays was proposed. The SPP and the posterior probabilities obtained by an EM-based position clustering were used to estimate the PSD matrices of the different talkers and the background noise. Eventually, each talker was extracted by a MWF. It was shown that even during triple-talk, the EM algorithm converges with only a few iterations to the desired solution and that good interference reduction is achieved at the cost of low speech distortion. In future work, the performance will be evaluated in more reverberant environments, and with different spatial filters other than the MWF. Moreover, an online implementation of the EM algorithm is to be considered.

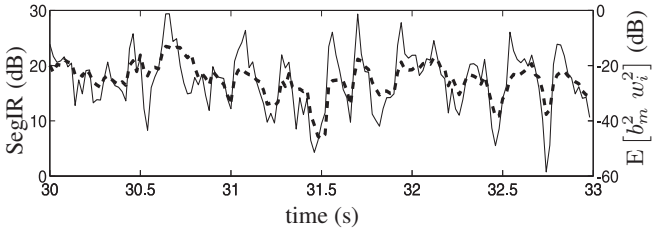


Fig. 4. Segmental interference reduction (dashed line), source 2 is desired. The interference power is shown on the right vertical axis (solid line).

	Source 1		Source 2		Source 3	
	IBM	***PPP***	***IBM***	***PPP***	***IBM***	***PPP***
$\mathcal{S}_{i,u}$	-0.4	-0.4	1	1	-4	-4
$\mathcal{S}_{o,x} - \mathcal{S}_{i,x}$	11.8	10.5	10.2	9.6	13.1	12
$\mathcal{S}_{o,v} - \mathcal{S}_{i,v}$	3.5	3	3.1	2.6	2.9	2.7
Δ-PESQ	0.97	0.8	0.88	0.75	1.06	0.95
ν_{sd}	0.05	0.09	0.07	0.13	0.12	0.16

Table 1. Performance evaluation, triple-talk scenario.

	Source 1		Source 2		Source 3	
	IBM	***PPP***	***IBM***	***PPP***	***IBM***	***PPP***
$\mathcal{S}_{o,v} - \mathcal{S}_{i,v}$	3.5	3.2	2.6	2.6	3	3.2
ν_{sd}	0.04	0.05	0.04	0.06	0.06	0.06

Table 2. Performance evaluation, mainly single-talk scenario.

7. REFERENCES

[1] Ò. Yilmaz and S. Rickard, "Blind separation of speech mixture via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1830–1847, 2004.

[2] O. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007.

[3] M. Mandel, R. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 382–394, 2010.

[4] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 516–527, 2011.

[5] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for joint blind source separation and noise reduction," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[6] D. H. Tran Vu and R. Haeb-Umbach, "An EM approach to integrated multichannel speech separation and noise suppression," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, 2012.

[7] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "Simultaneous speech source sseparation and noise reduction via clustefiltering MMSE-based filtering," in *Proc. IEEE Intl Conf. Signal Processing, Communications and Computing (ICSPCC)*, 2011.

[8] D. H. Tran Vu and R. Haeb-Umbach, "Exploiting temporal correlations in joint multichannel speech separation and noise suppression," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, 2012.

[9] C. M. Bishop, *Pattern recognition and machine learning*, Springer New York, 2006.

[10] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer-Verlag, Berlin, Germany, 2008.

[11] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2159–2169, 2011.

[12] M. Taseska and E. A. P. Habets, "MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sept. 2012.

[13] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 1072–1077, July 2010.

[14] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[15] T. Gerkmann and R. C. Hendriks, "Noise power estimation base on the probability of speech presence," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2011.

[16] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, pp. 1833–1847, 2007.

[17] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[18] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, 2006.

[19] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating non-stationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.

[20] S. Araki, H. Sawada, R. Mukai, and S. Makino, "A novel blind source separation method with observation vector clustering," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, 2005.

[21] Giovanni Del Galdo, Oliver Thiergart, Tobias Weller, and E. A. P. Habets, "Generating virtual microphone signals using geometrical information gathered by distributed arrays," in *Proc. Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Edinburgh, United Kingdom, May 2011.

[22] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2001, pp. 749–752.