AN ONLINE EM ALGORITHM FOR SOURCE EXTRACTION USING DISTRIBUTED MICROPHONE ARRAYS

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

Expectation maximization (EM)-based clustering is applied in many recent multichannel source extraction techniques. The estimated model parameters are used to compute time-frequency masks, or estimate second order statistics (SOS) of the source signals. However, in applications with moving sources where the model parameters are time-varying, the batch EM algorithm is inapplicable. We propose an online EM-based clustering of position estimates, where the model parameters are estimated adaptively. A direct-to-diffuse ratio-based speech presence probability is used to detect noisy observations and reduce diffuse and spatially incoherent noise. The desired source signal is extracted by a multichannel Wiener filter computed using SOS estimated from the time-varying model parameters. We show that the signal of a moving source can be extracted, while reducing moving interferers and background noise.

Index Terms— expectation maximization, online learning, PSD matrix estimation, distributed arrays

1. INTRODUCTION

Extracting one or more desired sources while reducing noise and interferers is required in many modern communication systems. Several recently proposed multichannel techniques make use of time-frequency (TF) masks based on spatial features such as binaural cues, signal vectors [1–3], or apply statistically optimal filters using SOS of the desired and the interfering signals [3–6]. The TF masks or the SOS of the different signals need to be estimated from the mixtures received at multiple microphones. The aforementioned methods assume that speech signals are sparse in the TF domain, implying that in each TF bin at most one source is dominant [7].

The main idea in probabilistic source segregation is to use a probabilistic model describing the distribution of certain features extracted from the microphone signals when a particular source is dominant. Common features include binaural cues [1], position [6] and signal vectors [2, 3]. The index of the source that generates the observation is a hidden variable, leading to an unsupervised estimation problem. The expectation maximization (EM) algorithm is a commonly used tool to deal with unsupervised estimation and incomplete data [8]. A position-based source segregation method which makes use of the EM algorithm was proposed by the present authors in [6], where probabilistic TF masks were used to compute the SOS and separate the different source signals by applying multichannel Wiener filters (MWFs). Nevertheless, a drawback of the EM-based approaches is that they require a training phase, and that the model parameters can not be adapted once the EM algorithm has converged. This is a limitation in scenarios where the position of the sources is time-varying. In such cases, the model parameters are time-varying as well and need to be adapted online.

Online learning of model parameters using the EM algorithm is commonly used in video processing [9, 10]. For instance, adaptive Gaussian mixtures allow for model adaptation by computing the sufficient statistics within short temporal windows and continuously updating the model parameters. The online EM variants do not require a training phase and although they represent noisy approximations of their batch counterparts, they exhibit faster convergence, as updates are done for each incoming data sample. For details about online EM variants the reader is referred to [11]. In this paper we extend the EM-based clustering of position estimates proposed in [6] to a frame-wise online version. The position estimates for each TF bin are computed by triangulating direction of arrival (DOA) estimates obtained from at least two distributed microphone arrays. In the E-step, the sufficient statistics in a short temporal window are computed, while in the M-step the model parameters are updated in the standard manner. Consequently, updates take place each frame and the posterior probabilities used for power spectral density (PSD) matrix estimation are correctly computed even for moving sources. We demonstrate that for an appropriate temporal window size, the algorithm adapts the model correctly whenever the location of a desired source or an interferer changes. Eventually, the desired source is extracted from the mixture by a MWF.

The paper is organized as follows: in Section 2 the signal model is described. Section 3 provides an overview of the MWF and the posterior probability-based estimation of the required PSD matrices. The computation of posterior probabilities and the online clustering of position estimates are described in Section 4. Section 5 provides simulation results and a discussion.

2. SIGNAL MODEL

A setup is considered where M microphones from two or more distributed arrays capture an additive mixture of a desired source, undesired spatially coherent interferers (e.g. interfering talkers) and background noise. The number of interferers I is assumed to be known, nevertheless, their positions are unknown and possibly time-variant, due to e.g. movement of an interferer. The position of the desired source is also unknown and possibly time-variant. The signal at the m-th microphone in the short-time Fourier transform (STFT) domain for time index n and frequency index k is given as

$$Y_m(n,k) = X_{0,m}(n,k) + \sum_{i=1}^{I} X_{i,m}(n,k) + V_m(n,k), \quad (1)$$

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

where $X_{0,m}$ and $X_{i,m}$, for $i = 1, 2 \ldots, I$, are the spectral coefficients of the desired source signal and the i-th interferer, respectively and V_m denote the spectral coefficients of the noise. In the following, the time and frequency index are omitted wherever possible.

Using the vector notation $\mathbf{y} = [Y_1 \ldots Y_M]^{\mathrm{T}}$, the PSD matrix of the random vector $\mathbf{y}$ is defined as $\mathbf{\Phi}_{\mathbf{y}} = \mathrm{E}\left[\mathbf{y}\,\mathbf{y}^{\mathrm{H}}\right]$, where $(\cdot)^{\mathrm{H}}$ denotes the conjugate transpose. The vectors $\mathbf{v}$ and $\mathbf{x}_i$ and the matrices $\mathbf{\Phi}_{\mathbf{v}}$ and $\mathbf{\Phi}_{\mathbf{x},i}$ for $i = 0, \ldots, I$ are defined similarly. The different source signals and the noise are assumed to be realizations of mutually uncorrelated zero-mean random processes, such that

$$\mathbf{\Phi}_{\mathbf{y}}(n) = \mathbf{\Phi}_{\mathbf{x},0}(n,k) + \sum_{i=1}^{I} \mathbf{\Phi}_{\mathbf{x},i}(n,k) + \mathbf{\Phi}_{\mathbf{v}}(n,k). \qquad (2)$$

The focus of this paper is computing an estimate of the desired source signal $\widehat{X}_{0,m}$ at the m-th microphone. This is achieved by an MWF that reduces the interferers and the noise while preserving the desired source. In contrast to previous work [3,5,6], we address non-stationary scenarios where the position of the desired source and the interferers is varying.

3. OPTIMAL LINEAR FILTERING

To compute a statistically optimum linear filter with respect to a defined criterion, the PSD matrices of the desired and the interfering signals are required [4]. Using instantaneous parametric information extracted from the microphone signals in the PSD matrix estimation results in an informed spatial filter [6, 12, 13], where the PSD matrices can be promptly updated to extract or reduce moving sources.

3.1. Multichannel Wiener filter

If the filter coefficients computed for the reference microphone m are denoted by $\mathbf{h}_m$, an estimate of the desired signal is given by

$$\widehat{X}_{0,m}(n,k) = \mathbf{h}_m^{\mathrm{H}}(n,k)\,\mathbf{y}(n,k). \qquad (3)$$

The noise-and-interference PSD matrix $\mathbf{\Phi}_{\mathbf{u}}$ is given by

$$\mathbf{\Phi}_{\mathbf{u}}(n,k) = \sum_{i=1}^{I} \mathbf{\Phi}_{\mathbf{x}_i}(n,k) + \mathbf{\Phi}_{\mathbf{v}}(n,k). \qquad (4)$$

The MWF coefficients are obtained by minimizing the minimum mean squared error (MMSE) between $X_{0,m}$ and the estimate $\widehat{X}_{0,m}$ leading to [4]

$$\mathbf{h}_m(n,k) = \frac{\mathbf{\Phi}_{\mathbf{u}}^{-1}(n,k)\mathbf{\Phi}_{\mathbf{x}_0}(n,k)}{1 + \mathrm{tr}\{\mathbf{\Phi}_{\mathbf{u}}^{-1}(n,k)\mathbf{\Phi}_{\mathbf{x}_0}(n,k)\}}\mathbf{e}_m, \qquad (5)$$

where tr$\{\cdot\}$ denotes the trace operator and

$$\mathbf{e}_m = [\underbrace{0 \ldots 0}_{m-1}\; 1\; \underbrace{0 \ldots 0}_{M-m}]^T. \qquad (6)$$

3.2. PSD matrix estimation

A common approach to PSD matrix estimation consists of performing recursive rank one updates based on the certainty that a given signal is dominant at a particular TF bin. We introduce the following hypothesis related to the activity of the sources at each TF bin

$\mathcal{H}_{\mathbf{v}}:\;\; \mathbf{y}(n,k) = \mathbf{v}(n,k),$ indicating speech absence (7a)

$\mathcal{H}_{\mathbf{x}}^i:\;\;$ indicating that the i-th source is dominant, i.e $\mathbf{y}(n,k) \approx \mathbf{x}_i(n,k) + \mathbf{v}(n,k).$ (7b)

Consequently, speech presence is indicated by

$$\mathcal{H}_{\mathbf{x}} = \mathcal{H}_{\mathbf{x}}^0 \cup \mathcal{H}_{\mathbf{x}}^1 \cup \ldots \cup \mathcal{H}_{\mathbf{x}}^I. \qquad (8)$$

Note that this estimation framework assumes that speech signals are approximately sparse in the STFT domain [7].

In state-of-the-art approaches, the update is controlled by the posterior probability of the relevant hypothesis [3, 14, 15]. In this manner, for a chosen averaging constant $0 < \tilde{\alpha}_v < 1$ the noise PSD is updated as follows

$$\widehat{\mathbf{\Phi}}_{\mathbf{v}}(n) = \alpha_v(n)\,\widehat{\mathbf{\Phi}}_{\mathbf{v}}(n-1) + [1 - \alpha_v(n)]\,\mathbf{y}(n)\mathbf{y}^{\mathrm{H}}(n), \qquad (9)$$

where α_v is computed by the multichannel speech presence probability (SPP) $p[\mathcal{H}_x \,|\, \mathbf{y}]$ according to

$$\alpha_v(n) = p[\mathcal{H}_x \,|\, \mathbf{y}(n)] + \tilde{\alpha}_v\;(1 - p[\mathcal{H}_x \,|\, \mathbf{y}(n)])\,. \qquad (10)$$

Similarly, to compute the PSD matrix of each source, we use the probabilities $p[\mathcal{H}_{\mathbf{x}}^i \,|\, \mathbf{y}]$. Let the PSD matrix $\mathbf{\Phi}_{\mathbf{x}_i+\mathbf{v}}$ be defined as

$$\mathbf{\Phi}_{\mathbf{x}_i+\mathbf{v}} = \mathbf{\Phi}_{\mathbf{x}_i} + \mathbf{\Phi}_{\mathbf{v}}. \qquad (11)$$

Using a recursive update rule as in (9), its estimate is given by

$$\widehat{\mathbf{\Phi}}_{\mathbf{x}_i+\mathbf{v}}(n) = \alpha_x(n)\,\widehat{\mathbf{\Phi}}_{\mathbf{x}_i+\mathbf{v}}(n{-}1){+}[1{-}\alpha_x(n)]\,\mathbf{y}(n)\mathbf{y}^{\mathrm{H}}(n) \qquad (12)$$

where for a chosen averaging constant $0 < \tilde{\alpha}_x < 1$

$$\alpha_x(n) = 1 - p[\mathcal{H}_x^i \,|\, \mathbf{y}(n)] + \tilde{\alpha}_x\; p[\mathcal{H}_x^i \,|\, \mathbf{y}(n)]. \qquad (13)$$

Eventually, PSD matrices of the sources are obtained by

$$\widehat{\mathbf{\Phi}}_{\mathbf{x}_i} = \widehat{\mathbf{\Phi}}_{\mathbf{x}_i+\mathbf{v}} - \widehat{\mathbf{\Phi}}_{\mathbf{v}}. \qquad (14)$$

The remaining task is to estimate the posterior probabilities used in (10) and (13) by exploiting parametric information extracted from the microphone signals.

4. ESTIMATION OF POSTERIOR PROBABILITIES AND ONLINE CLUSTERING

The posterior probability that the i-th source is dominant given the current observation $\mathbf{y}$ can be decomposed as

$$p[\mathcal{H}_{\mathbf{x}}^i \,|\, \mathbf{y}] = p[\mathcal{H}_{\mathbf{x}}^i \,|\, \mathbf{y}, \mathcal{H}_{\mathbf{x}}] \cdot p[\mathcal{H}_{\mathbf{x}} \,|\, \mathbf{y}], \qquad (15)$$

where the first factor is the probability that the i-th source is dominant, conditioned on speech presence, and the second factor represents the SPP, which can be computed using the estimator proposed in [16]. The computation of these probabilities is detailed in Sectons 4.2 and 4.1, respectively. To compute $p[\mathcal{H}_{\mathbf{x}}^i \,|\, \mathbf{y}, \mathcal{H}_{\mathbf{x}}]$, the present authors in [6] used position-based probabilities such that

$$p[\mathcal{H}_{\mathbf{x}}^i \,|\, \mathbf{y}, \mathcal{H}_{\mathbf{x}}] \approx p[\mathcal{H}_{\mathbf{x}}^i \,|\, \widehat{\mathbf{\Theta}}, \mathcal{H}_{\mathbf{x}}], \qquad (16)$$

where $\widehat{\mathbf{\Theta}}$ is the estimated position vector in a certain TF bin. The full band distribution of $\widehat{\mathbf{\Theta}}$ given that speech is present, was modeled by a Gaussian mixture (GM) as

$$p[\widehat{\mathbf{\Theta}} \,|\, \mathcal{H}_{\mathbf{x}}] = \sum_{i=0}^{I} \pi_i\, \mathcal{N}\left(\widehat{\mathbf{\Theta}}; \boldsymbol{\mu}_i, \mathbf{\Sigma}_i\right) \qquad (17)$$

where π_i denotes the i-th mixing coefficient and $\mathcal{N}\left(\widehat{\mathbf{\Theta}}; \boldsymbol{\mu}_i, \mathbf{\Sigma}_i\right)$ denotes a Gaussian distribution with mean $\boldsymbol{\mu}_i$ and covariance matrix $\mathbf{\Sigma}_i$. If the mixture parameters are known, the required posterior probabilities can be computed as

$$p[\mathcal{H}_{\mathbf{x}}^{i} \mid \widehat{\boldsymbol{\Theta}}, \mathcal{H}_{\mathbf{x}}] = \frac{\pi_i \, \mathcal{N}\left(\widehat{\boldsymbol{\Theta}}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)}{\sum_{i'=0}^{I} \pi_{i'} \, \mathcal{N}\left(\widehat{\boldsymbol{\Theta}}; \boldsymbol{\mu}_{i'}, \boldsymbol{\Sigma}_{i'}\right)}. \tag{18}$$

In Section 4.2, we propose a frame-wise online EM-based method to estimate the GM parameters.

4.1. Direct-to-diffuse ratio (DDR)-based SPP estimation

In order to detect TF bins where a speech source is present we employ the DDR-based SPP estimator proposed in [12], to compute $p[\mathcal{H}_{\mathbf{x}} \mid \mathbf{y}]$ required in (15). Therefore, the source signals that are coherent across the arrays are detected and do not leak into the noise PSD matrix estimate. Moreover, in order to improve the estimator at low frequencies where the DDR is overestimated due to small inter-microphone distances, each frame is subdivided into two frequency bands and the ratio of the two DDRs is compared. Subsequently, a binary mask is computed which is equal to zero if the ratio of direct-to-diffuse ratios (DDRs) is larger than a threshold, and one otherwise. Eventually, the a priori speech absence probability (SAP) as computed in [12] is multiplied by the binary mask. This modification improves the robustness of the subsequent online clustering.

4.2. Online EM clustering of source posteriors

In [6], the GM parameters $\mathcal{P} = \{\pi_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \ldots\}$ were computed by a batch EM algorithm using a training set of R TF bin-wise observations $\mathcal{D} = \{\widehat{\boldsymbol{\Theta}}_1, \ldots \widehat{\boldsymbol{\Theta}}_R\}$. In the E-step of the algorithm, a set of sufficient statistics is computed based on the current parameters

$$P_i = \sum_{r=1}^{R} p[\mathcal{H}_{\mathbf{x}}^{i} \mid \widehat{\boldsymbol{\Theta}}_r; \mathcal{P}] \tag{19a}$$

$$M_i = \sum_{r=1}^{R} p[\mathcal{H}_{\mathbf{x}}^{i} \mid \widehat{\boldsymbol{\Theta}}_r; \mathcal{P}] \cdot \widehat{\boldsymbol{\Theta}}_r \tag{19b}$$

$$S_i = \sum_{r=1}^{R} p[\mathcal{H}_{\mathbf{x}}^{i} \mid \widehat{\boldsymbol{\Theta}}_r; \mathcal{P}] \cdot (\widehat{\boldsymbol{\Theta}}_r - \boldsymbol{\mu}_i)(\widehat{\boldsymbol{\Theta}}_r - \boldsymbol{\mu}_i)^{\mathrm{T}} \tag{19c}$$

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

$$\boldsymbol{\mu}_i = {}^{M_i}/_{P_i}, \quad \boldsymbol{\Sigma}_i = {}^{S_i}/_{P_i}, \quad \pi_i = {}^{P_i}/\sum_{i'=0}^{I} P_{i'}. \tag{20}$$

The goal in this paper is to model the short-term distribution in a moving temporal window, hence allowing the model to be time-varying and applicable to moving sources. To achieve this, we adopt an approach proposed in the context of video processing [9, 10], where parameter updates take place for each incoming data sample and no training phase is required. Here, we propose a frame-wise online EM approach where the parameters are updated each time frame using the position estimates over all frequencies in that frame. In the E-step, the sufficient statistics are computed using position estimates from the L most recent frames. In this manner, at time frame n the sufficient statistics are given by

Algorithm 1 Online clustering of bin-wise position estimates

```
for each incoming frame $n$ do
    Compute the SPP $p[\mathcal{H}_{\mathbf{x}} \mid \widehat{\boldsymbol{\Theta}}(n,k)]$ for all frequencies $k$
    Form the set $K_n = \{k \mid p[\mathcal{H}_{\mathbf{x}} \mid \widehat{\boldsymbol{\Theta}}(n,k)] > p_{\text{min}}\}$
    if $K_n \neq \varnothing$ then
        E-step: compute (21) with the last $L$ frames for which $K_l \neq \varnothing$, where $l$ denotes the frame index
        M-step: update parameters $\mathcal{P}(n-1)$ using (20)
    end if
    Evaluate (18) with the current parameters $\mathcal{P}(n)$
end for
```

$$P_i(n) = \sum_{n'=n-L+1}^{n} \sum_{k} p[\mathcal{H}_{\mathbf{x}}^{i} \mid \widehat{\boldsymbol{\Theta}}(n',k); \mathcal{P}(n-1)] \tag{21a}$$

$$M_i(n) = \sum_{n'=n-L+1}^{n} \sum_{k} p[\mathcal{H}_{\mathbf{x}}^{i} \mid \widehat{\boldsymbol{\Theta}}(n',k); \mathcal{P}(n-1)] \cdot \widehat{\boldsymbol{\Theta}}(n',k) \tag{21b}$$

$$S_i(n) = \sum_{n'=n-L+1}^{n} \sum_{k} p[\mathcal{H}_{\mathbf{x}}^{i} \mid \widehat{\boldsymbol{\Theta}}(n',k); \mathcal{P}(n-1)] \times \left(\widehat{\boldsymbol{\Theta}}(n',k) - \boldsymbol{\mu}_i(n-1)\right)\left(\widehat{\boldsymbol{\Theta}}(n',k) - \boldsymbol{\mu}_i(n-1)\right)^{\mathrm{T}} \tag{21c}$$

The M-step is done as in (20), by using the short-term statistics given by (21) instead of the batch statistics given by (19) to obtain the new parameters $\mathcal{P}(n)$. The proposed algorithm is summarized by Algorithm 1 and the overall source extraction framework used in this paper is illustrated in Figure 1.

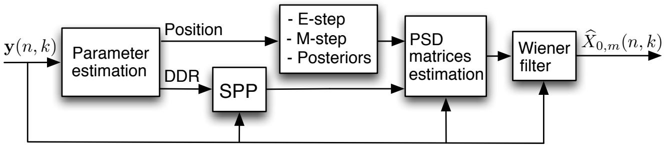


Fig. 1. Proposed online clustering framework for source extraction.

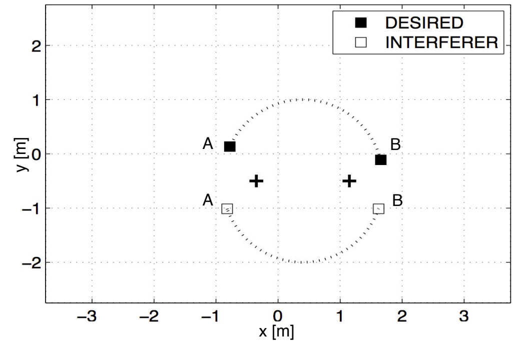
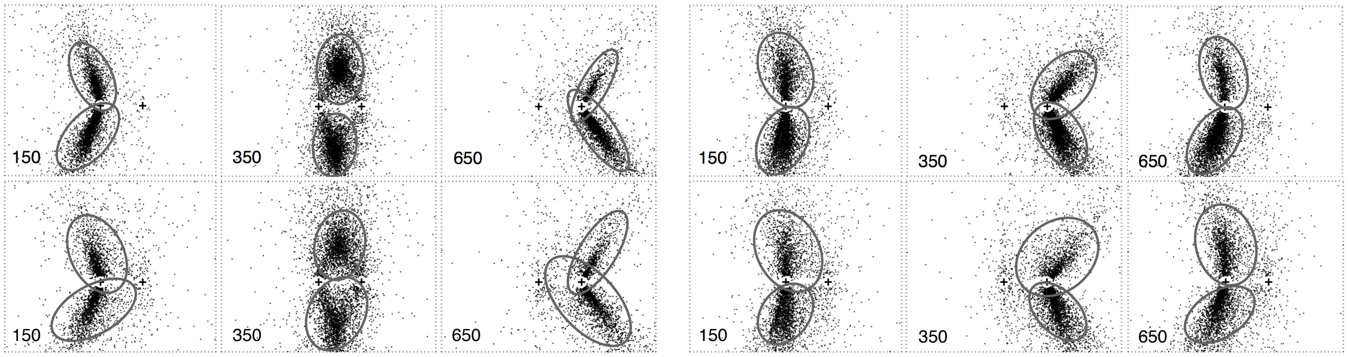


Fig. 2. Simulated scenario. The dotted lines illustrate the source trajectories. The crosses represent the array locations.

5. EXPERIMENTAL RESULTS

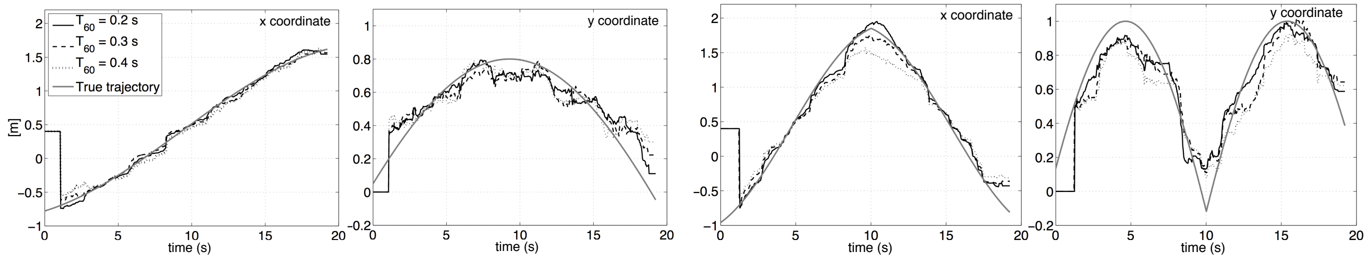
The proposed framework was evaluated using a simulated microphone signals. The signals were obtained as a sum of clean speech signals convolved with simulated room impulse responses [17], such that the source trajectory is sampled and for each point the impulse responses are recomputed and convolved with the clean speech signals. A diffuse babble noise signal with a segmental speech-to-noise ratio of 17 dB, and uncorrelated sensor noise with a segmental



(a) Scenario A-B (speed 1x)

(b) Scenario A-B-A (speed 2x)

Fig. 3. Online EM results. Top plots: T_{60} = 0.2 s. Bottom plots: T_{60} = 0.35 s. The numbers in the bottom left corner denote the time frame indices. The array locations are marked by a plus sign. The interior of each ellipse contains 85% probability mass of the respective Gaussian.



(a) Scenario A-B (speed 1x)

(b) Scenario A-B-A (speed 2x)

Fig. 4. Tracking accuracy of the EM algorithm. The x and y coordinates of the true source position and the means of the respective Gaussian distributions are plotted over time.

	T_{60} = 0.2 s		T_{60} = 0.25 s		T_{60} = 0.3 s		T_{60} = 0.35 s		T_{60} = 0.4 s	
	A-B	A-B-A	A-B	A-B-A	A-B	A-B-A	A-B	A-B-A	A-B	A-B-A
$\mathcal{S}_{i,x}$ [dB]	2.00	1.47	2.15	1.20	2.25	1.68	2.43	2.00	2.42	1.81
$\mathcal{S}_{o,v} - \mathcal{S}_{i,v}$ [dB]	4.50	3.48	4.00	3.21	3.47	2.85	3.00	2.61	2.50	2.70
$\mathcal{S}_{o,x} - \mathcal{S}_{i,x}$ [dB]	9.70	9.12	8.15	7.96	6.55	6.27	5.70	5.37	4.66	5.35
Δ-PESQ	0.70	0.68	0.57	0.49	0.45	0.40	0.35	0.34	0.30	0.29
ν_{sd}	0.07	0.16	0.11	0.23	0.16	0.27	0.24	0.30	0.29	0.32

Table 1. Performance of the MWF for Scenarios A-B and A-B-A.

speech-to-noise ratio of 40 dB were added to the microphone signals. The STFT frame length was 1024 samples, with 50% overlap and sampling frequency of 16 kHz. A scenario with two sources was simulated, employing two uniform circular arrays with three omnidirectional microphones each, a diameter 2.5 cm and an inter-array spacing of 1.5 m. The DOA was computed for each array using the estimator proposed in [18], and the position was computed by a triangulation of the DOA vectors. For the given array diameter, the DOA estimates over the used frequency range are not affected by spatial aliasing. In order to increase the algorithm robustness to noisy observations, in addition to discarding positions estimates where the SPP is below $p_{min} = 0.85$ from the sufficient statistics computation, positions within a radius of 20 cm around the array centers are discarded as well. The averaging constants in (10) and (13) were $\alpha_v = 0.9$ and $\alpha_x = 0.8$. The length L of the temporal window for online clustering was 40 frames ($\approx$ 1.3 s). The EM algorithm is initialized such that the two means are placed on the axis of symmetry between the two arrays, in positions (0.4, 0) and (0.4,-1), the mixture coefficients are set to 0.5 for both sources, and the covariance matrices are scaled identity matrices. In the following, we evaluate the tracking performance of the online EM algorithm and the source extraction performance of the MWF for reverberation times T_{60} in the range from 0.2 to 0.4 seconds and two different speeds of the moving sources. The scenario is illustrated in Fig. 2, where in a first experiment, both sources move from positions A to positions B, simultaneously (this scenario is referred to as A-B), whereas in a second experiment, each source starts from position A, goes to position B, and comes back to position A, resulting in two times faster movement (this scenario is referred to as A-B-A). The two sources are of approximately equal power. A reference microphone for the MWF is chosen from the array which is closer to the current position estimate of the desired talker (i.e., the mean of the respective Gaussian distribution).

5.1. Performance measures

We denote the input desired speech to diffuse plus sensor noise ratio by $\mathcal{S}_{i,v}$, the input desired speech to undesired speech ratio by $\mathcal{S}_{i,x}$, and the corresponding output values $\mathcal{S}_{o,v}$ and $\mathcal{S}_{o,x}$ [4]. The source extraction performance at the output of the MWF is evaluated in terms of

- Desired-speech-to-noise ratio improvement $\mathcal{S}_{o,v}$ - $\mathcal{S}_{i,v}$.

- Desired-to-undesired speech ratio improvement $\mathcal{S}_{o,x}$ - $\mathcal{S}_{i,x}$.

- Speech distortion index ν_{sd} [4].

- PESQ score improvement [19], denoted by Δ-PESQ.

The performance measures are computed segmentally, using non-overlapping frames of 20 ms, where only frames with input SIR or SNR between -40 dB and 40 dB were considered. The final values are obtained by averaging the segmental values in the logarithmic domain, except the speech distortion index which is averaged in the linear domain.

5.2. Results

The output of the online EM algorithm at different time instants is shown in Figure 3 for two different values of T_{60}, where it is visible that the distribution of position estimates associated with a particular source estimated with good accuracy. As expected, the variance of the Gaussian distribution associated with a particular source increases with increasing reverberation time and increasing the speed of the moving sources. Nevertheless, the mean of the distribution represents a good estimate of the true source location at all times. This is corroborated by the plots in Fig. 4, where the x and y coordinates of the true source position and the estimated means are plotted over time, for different values of T_{60}. The values of the estimated source position in the first second represent the initial values of the EM algorithm, which are not updated until a sufficient number of L frames (in this case 40 frames) are processed. After this initial latency, the parameters are updated for each incoming frame.

The results from the objective performance evaluation are summarized in Table 1. Although the tracking performance was similar for the different speeds and reverberation levels, the quality of the extracted source signal decreases with increasing reverberation levels . Nevertheless, increasing the speed of the sources by a factor of two leads to only slightly worse performance (except for the desired-speech-to-noise ratio and the desired-to-undesired speech ratio improvement at T_{60} = 400 ms, where slightly better results were achieved in Scenario A-B-A than in Scenario A-B), indicating that the tracking algorithm adapts the model parameters quickly to the new scenario. The most significant performance difference for the different speeds is in the speech distortion index ν_{sd}. Developing measures to keep the distortion index low and less susceptible to estimation errors in the PSD matrices is a topic of ongoing research.

6. CONCLUSIONS

An online EM-based clustering of position estimates in the TF domain was proposed, with application to power spectral density matrix estimation of moving sources in the presence of background noise. The algorithm was evaluated in a double-talk scenario, where two sources are constantly moving. The tracking and the source extraction performance was evaluated for different reverberation levels and different speeds of the moving sources. The algorithm was able to track the source position changes and accurately update the model parameters, resulting an extracted desired source signal with low distortion and good interference reduction. Nevertheless, although the tracking performance was satisfactory, a decrease in the source extraction performance was observed for increasing reverberation times. Future work includes evaluation in scenarios with more sources, and with different spatial filters which could lead to better noise and interference reduction in environments with higher reverberation levels.

7. REFERENCES

[1] 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.

[2] 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.

[3] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, “A multichannel MMSE-based framework for speech source separation and noise reduction,” *IEEE Transactions on Audio, Speech, and Language Processing*, to appear.

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

[5] D. H. Tran Vu and R. Haeb-Umbach, “An EM approach to integrated multichannel speech separation and noise suppression,” in *Proc. IWAENC*, 2012.

[6] M. Taseska and E.A.P Habets, “MMSE-based source extraction using position-based posterior probabilities,” in *Proc. IEEE ICASSP*, 2013.

[7] 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.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[9] D.-S. Lee, “Online adaptive Gaussian mixture learning for video applications,” in *Proc. Statistical Methods in Video Processing*, Prague, Czech Republic, May 2004.

[10] N. Friedman and S. Russel, “Image segmentation in video sequences: a probabilistic approach,” in *Proc. 13th Conf. Uncertainty in Artificial Intelligence*, 1997.

[11] R. M. Neal and G. E. Hinton, *Learning in Graphical Models*, chapter A view of the EM algorithm that justifies incremental, sparse, and other variants, Mit Pr, 1998.

[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. IWAENC*, Sept. 2012.

[13] O. Thiergart and E.A.P Habets, “An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates,” in *Proc. IEEE ICASSP*, 2013.

[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 WASPAA*, New Paltz, NY, 2011.

[16] 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.

[17] E. A. P. Habets, “Room impulse response generator,” Tech. Rep., Technische Universiteit Eindhoven, 2006.

[18] S. Araki, H. Sawada, R. Mukai, and S. Makino, “A novel blind source separation method with observation vector clustering,” in *Proc. IWAENC*, 2005.

[19] 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 ICASSP*, 2001, pp. 749–752.