Online Localization of Multiple Moving Speakers in Reverberant Environments

Xiaofei Li, Bastien Mourgue, Laurent Girin, Sharon Gannot, Radu Horaud

Online Localization of Multiple Moving Speakers in Reverberant Environments

Xiaofei Li[1], Bastien Mourgue[1], Laurent Girin[1,2], Sharon Gannot[3] and Radu Horaud[1]
[1]*INRIA Grenoble Rhône-Alpes,* [2]*Univ. Grenoble Alpes, Grenoble-INP, GIPSA-lab, France*
[3]*Faculty of Engineering, Bar-Ilan University, Israel*

***Abstract*—This paper addresses the problem of online multiple moving speakers localization in reverberant environments. The direct-path relative transfer function (DP-RTF), as defined by the ratio between the first taps of the convolutive transfer function (CTF) of two microphones, encodes the inter-channel direct-path information and is thus used as a localization feature being robust against reverberation. The CTF estimation is based on the cross-relation method. In this work, the recursive least-square method is proposed to solve the cross-relation problem, due to its relatively low computational cost and its good convergence rate. The DP-RTF feature estimated at each time-frequency bin is assumed to correspond to a single speaker. A complex Gaussian mixture model is used to assign each observed feature to one among several speakers. The recursive expectation-maximization algorithm is adopted to update online the model parameters. The method is evaluated with a new dataset containing multiple moving speakers, where the ground-truth speaker trajectories are recorded with a motion capture system.**

***Index Terms*—sound-source localization, multiple moving speakers, reverberant environments**

I. Introduction

In the real world, online multiple-speaker localization is a challenging task due to the influence of interfering speakers, reverberation and ambient noise. Moreover, *short sentences* and *pauses* are quite common in a natural conversation, which leads to frequent speech turns among speakers.

Time difference of arrival (TDOA) is widely used for single source localization [1]. Most TDOA estimators, such as the generalized cross-correlation method [2], are based on the direct-path propagation model, and thence do not perform well in reverberant environments. To overcome this, several TDOA estimators based on system identification were proposed in [3]–[6]. For multiple-speaker localization, the W-disjoint orthogonality (WDO) of the speech sources [7] is widely employed. The audio signal is assumed to be dominated by only one speaker in each small region of the time-frequency (TF) domain, because of the natural sparsity of speech signals in this domain. Applying the short-time Fourier transform (STFT), or any TF decomposition, inter-channel localization features (e.g. interaural phase difference [7]) can be extracted for each TF bin. In [7], multiple-speaker localization is based on histograms of the inter-channel features. Given a grid of candidate locations, a complex Gaussian mixture model (CGMM) is used in [8], with each CGMM component representing one candidate location. After maximizing the likelihood of the features, the weight of each component represents the probability that there is an active speaker at the corresponding candidate location. Therefore, the counting and localization of active speakers is jointly carried out by selecting the components with large weights.

The inter-channel features mentioned above are based on the direct-path propagation model, thence poorly suited for reverberant environments. In [9], we proposed to use the direct-path relative transfer function (DP-RTF) as a TF-domain localization feature robust against reverberation, and we exploited those features in a CGMM model similar to [8]. The DP-RTF feature estimation is based on the identification of the STFT-domain representation of the room impulse response (RIR), i.e. the CTF [10], [11]. Overall, this method integrates the merits of system identification based TDOA estimators [3]–[6] and the TF-domain WDO assumption, and thus allows multiple-speaker localization in reverberant environments.

To localize moving speakers, a tracking scheme based on Bayesian techniques estimates the posterior distribution of source locations given a sequence of instantaneous estimates of localization features (or of speaker locations) and a dynamic model of source movement, e.g. [12]–[14]. To tackle speech turns, speaker birth and death processes [15] and/or a model of speech activity [16] can be included. In a different line, a CGMM model similar to [8] (but with one CGMM per predefined speaker) was used in [17] and plugged into a recursive EM algorithm to update online the CGMM component weights.

The CTF identification used for DP-RTF extraction in [9] was formulated in batch mode, and speakers were considered static. In the present work, we exploit an adaptive CTF identification method based on the recursive least-square (RLS) algorithm. RLS has a better convergence rate than the least mean square algorithm used in [3], [4], which is especially important for moving speakers scenario. In addition, we extend the two-channel method presented in [9], [18] to a more general multi-channel framework. To count and localize speakers, we combine the CGMM model [8] with the recursive EM [17]. At each time step, the CGMM weights encode the number and locations of active speakers.

This research has received funding from the ERC Advanced Grant VHIA (#340113).

II. Recursive Multichannel DP-RTF estimation

A. RLS for DP-RTF estimation

To simplify the presentation, let us first consider the noise-free single-speaker case. In the time domain, the i-th microphone signal, $i \in \{1,\dots,I\}$, is: $x^i(n) = s(n) \star a^i(n)$, where n is the time index, $s(n)$ is the source signal, $a^i(n)$ is the RIR from the source to the i-th microphone, and $\star$ denotes the convolution. Applying the STFT, and using the CTF approximation, we have for each frequency index $k \in \{0,\dots,K-1\}$: $x^i_{p,k} = s_{p,k} \star a^i_{p,k}$, where $x^i_{p,k}$ and $s_{p,k}$ are the STFT coefficients of the corresponding signals, and the CTF $a^i_{p,k}$ is a subband representation of $a^i(n)$. Here, the convolution is executed with respect to the frame index p ($p \in \{1,\dots,P\}$ for $x^i_{p,k}$ and $s_{p,k}$, and $p \in \{0,\dots,Q-1\}$ for $a^i_{p,k}$, with $Q \ll P$). The number of CTF coefficients Q is related to the reverberation time of the RIR. The first CTF coefficient $a^i_{0,k}$ mainly consists of the direct-path information, thence the DP-RTF is defined as the ratio between the first CTF coefficients of two channels: $a^i_{0,k}/a^r_{0,k}$, where channel r is the reference channel.

Based on the cross-relation method [19], using the CTF model of one microphone pair (i,j), we have: $x^i_{p,k} \star a^j_{p,k} = x^j_{p,k} \star a^i_{p,k}$. This can be written in vector form as $\mathbf{x}^{i\top}_{p,k}\mathbf{a}^j_k = \mathbf{x}^{j\top}_{p,k}\mathbf{a}^i_k$, with $\mathbf{a}^i_k = (a^i_{0,k},\dots,a^i_{Q-1,k})^\top$ and $\mathbf{x}^i_{p,k} = (x^i_{p,k},\dots,x^i_{p-Q+1,k})^\top$, where $^\top$ denotes matrix/vector transpose. There is a total of $I(I-1)/2$ distinct microphone pairs, indexed by (i,j) with $i \in \{1,\dots,I-1\}$ and $j \in \{i+1,\dots,I\}$. For each pair, we construct a cross-relation equation in terms of the CTF of all channels, i.e. $\mathbf{a}_k = (\mathbf{a}^{1\top}_k,\dots,\mathbf{a}^{I\top}_k)^\top$. For this aim, we define:

$$\mathbf{x}^{ij}_{p,k} = (\underbrace{0,\dots,0}_{(i-1)Q}, \mathbf{x}^{j\top}_{p,k}, \underbrace{0,\dots,0}_{(j-i-1)Q}, -\mathbf{x}^{i\top}_{p,k}, \underbrace{0,\dots,0}_{(I-j)Q})^\top,\ j > i. \tag{1}$$

Then we have:

$$\mathbf{x}^{ij\top}_{p,k}\mathbf{a}_k = 0. \tag{2}$$

To avoid a trivial solution, i.e. $\mathbf{a}_k = \mathbf{0}$, we constrain the first CTF coefficient of the reference channel, e.g. $r = 1$, by dividing both sides of (2) by $a^1_{0,k}$ and moving the first entry of $\mathbf{x}^{ij}_{p,k}$, denoted by $-y^{ij}_{p,k}$, to the right side, rewriting (2) as:

$$\tilde{\mathbf{x}}^{ij\top}_{p,k}\tilde{\mathbf{a}}_k = y^{ij}_{p,k}, \tag{3}$$

where $\tilde{\mathbf{x}}^{ij}_{p,k}$ is $\mathbf{x}^{ij}_{p,k}$ with the first entry removed, and $\tilde{\mathbf{a}}_k$ is the relative CTF vector:

$$\tilde{\mathbf{a}}_k = \left(\frac{\tilde{\mathbf{a}}^{1\top}_k}{a^1_{0,k}}, \frac{\mathbf{a}^{2\top}_k}{a^1_{0,k}}, \dots, \frac{\mathbf{a}^{I\top}_k}{a^1_{0,k}}\right)^\top. \tag{4}$$

In the above equation, $\tilde{\mathbf{a}}^1_k$ is $\mathbf{a}^1_k$ with the first entry removed, i.e. $\tilde{\mathbf{a}}^1_k = [a^1_{1,k},\dots,a^1_{Q-1,k}]^\top$. The DP-RTFs appear in (4) as the first entries of $\frac{\mathbf{a}^{i\top}_k}{a^1_{0,k}}$, for $i \in \{2,\dots,I\}$. Therefore, the DP-RTF estimation amounts to solving (3).

Equation (3) is defined for one microphone pair at one frame. In batch mode, the equation terms $\tilde{\mathbf{x}}^{ij\top}_{p,k}$ and $y^{ij}_{p,k}$ can be concatenated along microphone pairs and frames to construct a least square problem. For the online case, we would like to update the estimate of $\tilde{\mathbf{a}}_k$ using the current frame, say p. For notational convenience, let $m = 1,\dots,M$ denote the index of microphone pair, where $M = I(I-1)/2$. Then let the superscript ^{ij} be replaced with ^{m}. The fitting error of (3) is

$$e^m_{p,k} = y^m_{p,k} - \tilde{\mathbf{x}}^{m\top}_{p,k}\tilde{\mathbf{a}}_k. \tag{5}$$

At the current frame p, for the microphone pair m, RLS aims to minimize the error

$$J^m_{p,k} = \sum\nolimits_{p'=1}^{p}\sum\nolimits_{m'=1}^{m} \lambda^{p-p'} e^{m'}_{p',k} e^{m'*}_{p',k}, \tag{6}$$

where * denotes complex conjugate. The forgetting factor $\lambda \in (0,1]$ gives exponentially less weight to older frames, whereas at one frame, all the microphone pairs have the same weight. To minimize $J^m_{p,k}$, we set its complex derivative with respect to $\tilde{\mathbf{a}}^*_k$ to zero, and obtain an estimate at frame p for microphone pair m as:

$$\begin{aligned}\tilde{\mathbf{a}}^m_{p,k} = \mathbf{R}^{m\,-1}_{p,k} r^m_{p,k} = &\left(\sum\nolimits_{p'=1}^{p}\sum\nolimits_{m'=1}^{m} \lambda^{p-p'} \tilde{\mathbf{x}}^{m'}_{p',k}{}^{*} \tilde{\mathbf{x}}^{m'\top}_{p',k}\right)^{-1} \\ &\times \left(\sum\nolimits_{p'=1}^{p}\sum\nolimits_{m'=1}^{m} \lambda^{p-p'} \tilde{\mathbf{x}}^{m'}_{p',k}{}^{*} y^{m'}_{p',k}\right),\end{aligned} \tag{7}$$

which can be recursively computed based on the rank-one modification of the covariance matrix $\mathbf{R}^m_{p,k}$. The recursion procedure is summarized in Algorithm 1, where $\mathbf{P}^m_{p,k} = \mathbf{R}^{m\,-1}_{p,k}$, and $\mathbf{g}$ is the *gain vector*. The current frame p is initialized by the previous frame $p-1$. At the first frame, we initialize $\tilde{\mathbf{a}}^0_{1,k}$ as zero, and $\mathbf{P}^0_{1,k}$ as identity. At one frame, all the microphone pairs are related to the same CTF vector that corresponds to the current speakers' location, which thence should be simultaneously used to estimate the CTF vector of the current frame. This can be easily realized by concatenating the microphone pairs in batch mode. However, in RLS, to satisfy the rank-one modification of the covariance matrix, we need to process the microphone pair one by one as shown in (6) and Algorithm 1. After the recursions for all microphone pairs, $\tilde{\mathbf{a}}^M_{p,k}$ is the CTF estimation of the current frame, and is used for speaker localization. The DP-RTF estimates, denoted as $\tilde{c}^i_{p,k}$, are obtained from $\tilde{\mathbf{a}}^M_{p,k}$, $i \in \{2,\dots,I\}$. Note that implicitly we have $\tilde{c}^1_{p,k} = 1$.

Algorithm 1 RLS at frame p

Input: $\tilde{\mathbf{x}}^m_{p,k}$, $y^m_{p,k}$, $m = 1,\dots,M$
Initialization: $\tilde{\mathbf{a}}^0_{p,k} \leftarrow \tilde{\mathbf{a}}^M_{p-1,k}$, $\mathbf{P}^0_{p,k} \leftarrow \lambda^{-1}\mathbf{P}^M_{p-1,k}$
for $m = 1$ to M **do**
  $e^m_{p,k} = y^m_{p,k} - \tilde{\mathbf{x}}^{m\top}_{p,k}\tilde{\mathbf{a}}^{m-1}_{p,k}$
  $\mathbf{g} = \mathbf{P}^{m-1}_{p,k}\tilde{\mathbf{x}}^{m}_{p,k}{}^{*}/(1 + \tilde{\mathbf{x}}^{m\top}_{p,k}\mathbf{P}^{m-1}_{p,k}\mathbf{x}^{m}_{p,k}{}^{*})$
  $\mathbf{P}^m_{p,k} = \mathbf{P}^{m-1}_{p,k} - \mathbf{g}\tilde{\mathbf{x}}^{m\top}_{p,k}\mathbf{P}^{m-1}_{p,k}$
  $\tilde{\mathbf{a}}^m_{p,k} = \tilde{\mathbf{a}}^{m-1}_{p,k} + e^m_{p,k}\mathbf{g}$
end for
Output: $\tilde{\mathbf{a}}^M_{p,k}$, $\mathbf{P}^M_{p,k}$

B. Multiple Moving Speakers

So far, the proposed online DP-RTF estimation method has been presented for the noise-free single-speaker case. We now extend it to the noisy multiple-speaker case. This extension has already been formulated in [9], [20], but only for the batch mode and for the two-channel case. In the following, we will summarize the principles, and specify the online and multichannel version of this extension.

1) Multiple moving speakers: We assume that the speakers are static over a short time, and that in a small region of the TF plane only one source is active due to the WDO assumption. Therefore, the CTF can be assumed to be locally time-invariant, and be estimated using a small number of recent frames. However, to efficiently estimate $\tilde{\mathbf{a}}_{p,k}^M \in \mathbb{C}^{(IQ-1)\times 1}$, $\rho(IQ-1)$ equations are required, with a large factor ρ, that is we need $\bar{P} = \frac{\rho(IQ-1)}{I(I-1)/2} \approx \rho\frac{2Q}{I-1}$ frames. The parameter ρ should be empirically set to achieve a good tradeoff between the validity of the above assumptions and a robust estimate of $\tilde{\mathbf{a}}_{p,k}^M$. The number of frames used to estimate $\tilde{\mathbf{a}}_{p,k}^M$ can be reduced by increasing the number of microphones. To approximately have a memory of $\bar{P}$ frames, we can set $\lambda = \frac{\bar{P}-1}{\bar{P}+1}$.

2) Noisy signals: Even in a low-noise case, many TF bins are dominated by noise due to the sparsity of speech spectra. Therefore, we need to classify the frames into noise frames and speech frames, and to suppress the noise from the speech frames. An inter-frame spectral subtraction algorithm was proposed in [20], [21]. The cross- and auto-PSD of the microphone signals $\mathbf{x}_{p,k}^i$ and $x_{p,k}^1$, denoted as $\boldsymbol{\phi}_{p,k}^i$, are first computed by averaging the cross- and audio-periodograms over frames. In the present work, we use the recursive averaging: $\boldsymbol{\phi}_{p,k}^i = \beta\boldsymbol{\phi}_{p-1,k}^i + (1-\beta)\mathbf{x}_{p,k}^i x_{p,k}^{1\,*}$, where β is a smoothing factor. The noise frames and speech frames are classified based on the minimum statistics of the PSD of $x_{p,k}^1$. Then the cross/auto-PSD of noise frames are subtracted from the cross/auto-PSD of speech frames. After spectral subtraction, let $\hat{\boldsymbol{\phi}}_{p,k}^i$ denote the cross/auto-PSD vector of speech frame, which is assumed to be noise-free. Instead of using $\mathbf{x}_{p,k}^i$, we use $\hat{\boldsymbol{\phi}}_{p,k}^i$ to construct (1). Correspondingly, we have a new (2), which is still valid, since it is equivalent to that, with noise removed, taking the cross/auto-PSD between both sides of the original (2) and $x_{p,k}^1$. In the RLS process, noise frames are skipped, and a speech frame with a preceding noise frame is initialized by the latest speech frame.

3) Consistency test: In practice, a DP-RTF estimate can sometimes be unreliable. The possible reasons are that in a small frame region, i) the CTF is time-varying due to a fast movement of the speakers, ii) multiple speakers are present, iii) only noise is present due to a wrong noise-speech classification, and iv) only reverberation is present at the end of a speech occurrence. In [9], a consistency test was proposed to tackle this problem: If a small frame region corresponds to an actual active speaker, the DP-RTFs estimated using different reference channels are consistent, otherwise the DP-RTFs are biased, with inconsistent bias values. In the present work, we use the first and second channels as reference, and obtain the DP-RTF estimates $\tilde{c}_{p,k}^i$ (with $\tilde{c}_{p,k}^1 = 1$) and $\bar{c}_{p,k}^i$ (with $\bar{c}_{p,k}^2 = 1$), respectively. Then $\tilde{c}_{p,k}^i$ and $\bar{c}_{p,k}^i/\bar{c}_{p,k}^1$ are two estimates of the same DP-RTF $a_{0,k}^i/a_{0,k}^1$. For each channel $i \in \{2,\dots,I\}$, if the similarity of the two estimates is large, they are said to be consistent. They are then averaged and normalized as done in [9], resulting in a final complex-valued feature $\hat{c}_{p,k}^i$ with module in $[0,1]$. The estimates that do not pass the consistency test are simply skipped.

At frame p, we obtain a set of features $\mathcal{C}_p = \{\{\hat{c}_{p,k}^i\}_{i\in\mathcal{I}_k}\}_{k=0}^{K-1}$, where $\mathcal{I}_k \subseteq \{2,I\}$ denotes the set of microphone indices that pass the consistency test. Note that $\mathcal{I}_k$ is empty if, p is a noise frame at frequency k, or if no channel passes the consistency test. Each of the features is assumed to be associated with a single speaker.

III. Recursive EM for Online Localization of Multiple Moving Speakers

In order to recursively assign the DP-RTF features in $\mathcal{C}_p$ to speakers, we adopt the CGMM formulation proposed in [8] and the recursive EM algorithm proposed in [17]. We define a set $\mathcal{S}$ of S candidate source locations. Let $s \in \{1,S\}$ denote the location index. An observed feature $\hat{c}_{p,k}^i$, emitted by a sound source located at candidate location s, is assumed to be drawn from a complex-Gaussian distribution with mean $c_k^{i,s}$ and variance σ^2, i.e. $\hat{c}_{p,k}^i|s \sim \mathcal{N}_c(c_k^{i,s},\sigma^2)$. The mean $c_k^{i,s}$ is the predicted feature at frequency k for channel i, and is precomputed based on the direct-path propagation model from the s-th candidate location to the microphones. The variance σ^2 is empirically set as a constant value. The marginal density of an observed feature $\hat{c}_{p,k}^i$ (taking into account all candidate locations) is a CGMM with each component corresponding to a candidate location:

$$P(c_{p,k}^i|\mathcal{S}) = \sum\nolimits_{s=1}^{S} \alpha^s \mathcal{N}_c(c_{p,k}^i; c_k^{i,s}, \sigma^2), \tag{8}$$

where $\alpha^s \geq 0$ is the prior probability (component weight) of the s-th component, with $\sum_{s=1}^S \alpha^s = 1$. The component weights are the only free model parameters.

Recursive EM [17], at frame p, first calculates the posterior probability of each candidate location s for each new observation $\hat{c}_{p,k}^i$ in $\mathcal{C}_p$, using the parameter estimates at the previous frame, i.e. α_{p-1}^s:

$$\eta_{p,k}^{i,s} = \frac{\alpha_{p-1}^s \mathcal{N}_c(\hat{c}_{p,k}^i; c_k^{i,s}, \sigma^2)}{\sum_{s'=1}^S \alpha_{p-1}^{s'} \mathcal{N}_c(\hat{c}_{p,k}^i; c_k^{i,s'}, \sigma^2)}, \quad \forall s \in \mathcal{S}, \forall \hat{c}_{p,k}^i \in \mathcal{C}_p. \tag{9}$$

Then the parameters are updated using all the new observations in $\mathcal{C}_p$, as:

$$\alpha_p^s = (1-\gamma)\alpha_{p-1}^s + \gamma \frac{1}{|\mathcal{C}_p|} \sum\nolimits_{\hat{c}_{p,k}^i \in \mathcal{C}_p} \eta_{p,k}^{i,s}, \quad \forall s \in \mathcal{S}, \tag{10}$$

where the smoothing factor γ controls the update rate of parameters, $|\mathcal{C}_p|$ denotes the cardinality of $\mathcal{C}_p$. When $\mathcal{C}_p$ is empty, such as during a silent period, α_p^s is updated by no-information value, i.e. $1/S$. For each component, the weight smoothly varies along time due to its recursive update, which indicates a tracking scheme. At each time frame, a plot of the weights as a function of the candidate location index exhibits a quite smooth curve with a few peaks that should correspond to active speakers. Therefore, the counting and localization of active speakers can be jointly carried out by selecting the peaks with a weight larger than a predefined threshold.

IV. Experiments

Experiments with real data are conducted using a version 5 NAO robot whose head has four microphones in a horizontal plane [22]. Thence we only perform 360° azimuth localization. The data are recorded in INRIA's Kinovis room [23], of size 10.19 m × 9.87 m × 5.6 m with $T_{60} = 0.53$ s. The speakers were moving around the robot with a speaker-to-robot range between 1.5 m and 3.5m. A motion capture system records their trajectories using a head-mounted infrared marker. Fourteen sequences were recorded with up to three participants (from 0 to 3 actively speaking along the sequence), for a total length of about 500 s. The sampling rate is 16 kHz and the STFT frame length is 16 ms with a 8 ms frame shift. The CTF length is $Q = 8$ frames. The RLS forgetting factor λ is computed using $\rho = 1$. The PSD smoothing factor is $\beta = 0.875$. The set of candidate locations $\mathcal{S}$ comprises $S = 72$ azimuths every 5° in $[-175°, 180°]$. For each azimuth candidate, the CGMM mean $c_k^{i,s}$ is computed using the HRTF of NAO. The EM smoothing factor is $\gamma = 0.92$.

Fig. 1 shows the result for one sequence. The first half is a natural conversation between three speakers, in which the participants take speech turns with a small overlap. It can be seen that the recursive update of the CGMM weights is able to monitor the moving, appearance, and disappearance of active speakers with only a small time lag. Consequently, the counting and localization of active speakers can be efficiently carried out by selecting the peaks. The vertical gray bars in the top figure represent the no-information case, namely when all the weights are close to $1/S$. The second half of the sequence comprises two simultaneously speaking participants. Here many gray bars are observed as well. The possible reasons are: i) there are many short pauses even during the speaking time, and ii) there is no DP-RTF estimate passing the consistency test due to the speech spectral overlap.

The proposed method is compared with SRP-PHAT [24] which uses the same STFT configuration, candidate locations and peak selection scheme just described. The frame-wise steered-response power is recursively smoothed with a smoothing factor of 0.92. A speaker is considered to be successfully localized if the localization error is smaller than 15°. For quantitative evaluation, we count the miss detection (MD) (speaker active but not detected) and false alarm (FA)

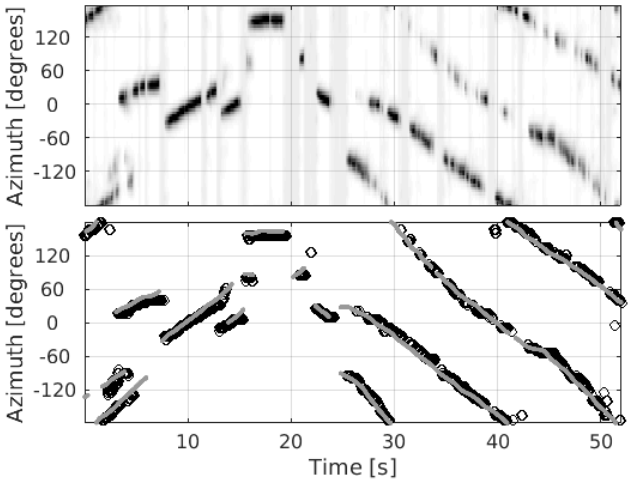

Fig. 1: An example of multiple-speaker tracking. **top** The CGMM weights along time. **bottom** The black circles represent the detected speakers by selecting the peaks with a weight larger than 0.02. The gray curves represent the ground-truth trajectories of active speakers.

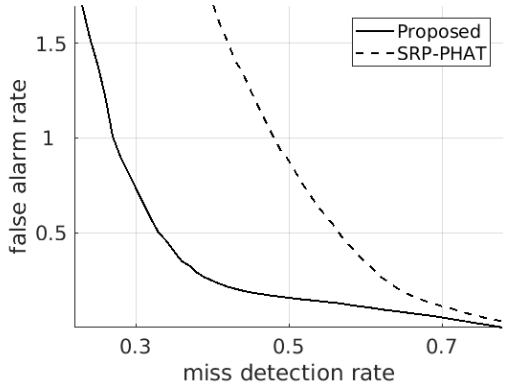

Fig. 2: ROC curve

(speaker detected but not active) at each STFT frame (i.e. each 8 ms). Then the MD and FA rates are computed as the percentage of the total MDs and FAs out of the total number of actual speakers, respectively. The receiver operating characteristic (ROC) curve (MD rate versus FA rate) is plotted in Fig. 2 at various peak selection threshold settings. It can be seen that the proposed method achieves a much better ROC curve than SRP-PHAT, since the proposed DP-RTF feature is robust against reverberation. Note that, the FA rate could even be larger than 1, since false detected speakers could be more than the actual speakers when the threshold is too small.

V. Conclusion

In this work, an online multiple-speaker localization method has been proposed, as an extension of the multiple-speaker localization method [9] which was based on batch processing. An RLS-based adaptive CTF identification method is developed for online DP-RTF feature estimation. The CGMM model [8] and the recursive EM [17] are combined with the proposed method for jointly counting and localizing the moving speech sources.

REFERENCES

[1] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on applied signal processing*, vol. 2006, pp. 170–170, 2006.

[2] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[3] Y. Huang and J. Benesty, "Adaptive multichannel time delay estimation based on blind system identification for acoustic source localization," in *Adaptive Signal Processing*, pp. 227–247, Springer, 2003.

[4] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 1110–1124, 2003.

[5] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, 2005.

[6] K. Kowalczyk, E. A. Habets, W. Kellermann, and P. A. Naylor, "Blind system identification using sparse learning for TDOA estimation of room reflections," *IEEE Signal Processing Letters*, vol. 20, no. 7, pp. 653–656, 2013.

[7] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing,*, vol. 52, no. 7, pp. 1830–1847, 2004.

[8] Y. Dorfan and S. Gannot, "Tree-based recursive expectation-maximization algorithm for localization of acoustic sources," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1692–1703, 2015.

[9] X. Li, L. Girin, R. Horaud, and S. Gannot, "Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1997–2012, 2017.

[10] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[11] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

[12] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 728–739, 2008.

[13] C. Evers, A. H. Moore, P. A. Naylor, J. Sheaffer, and B. Rafaely, "Bearing-only acoustic tracking of moving speakers for robot audition," in *IEEE International Conference on Digital Signal Processing (DSP)*, pp. 1206–1210, 2015.

[14] Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Exploiting the complementarity of audio and visual data in multi-speaker tracking," in *ICCV Workshop on Computer Vision for Audio-Visual Media*, vol. 3, 2017.

[15] S. Ba, X. Alameda-Pineda, A. Xompero, and R. Horaud, "An on-line variational Bayesian model for multi-person tracking from cluttered scenes," *Computer Vision and Image Understanding*, vol. 153, pp. 64–76, 2016.

[16] I. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal Bayesian fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[17] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 392–402, 2014.

[18] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[19] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Transactions on signal processing*, vol. 43, no. 12, pp. 2982–2993, 1995.

[20] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 2171–2186, 2016.

[21] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 320–324, 2015.

[22] X. Li, L. Girin, F. Badeig, and R. Horaud, "Reverberant sound localization with a robot head based on direct-path relative transfer function," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2819–2826, IEEE, 2016.

[23] https://kinovis.inria.fr/inria-platform.

[24] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays* (M. S. Brandstein and D. Ward, eds.), pp. 157–180, Springer, 2001.