Improving Overlapping Speaker Detection Using Multiple Speaker Tracking Information

Youssef Oualil, Rahil Mahdian Toroghi, Dietrich Klakow
Spoken Language Systems, Saarland University
Saarbrücken, Germany
youssef.oualil@lsv.uni-saarland.de

***Abstract*—Traditionally, multiple speaker tracking consists of two stages, namely, 1) detection of location measurements, followed by 2) a multiple object tracking approach. In general, these two steps are performed separately, and the tracking performance is highly dependent on the measurement detection rate. The performance of the widely used Steered Response Power (SRP)-based measurement detectors, however, drastically decreases in the overlapping speech scenario, where the dominant speaker frequently masks the low-energy speakers. To overcome this problem, we propose an approach that enhances the probabilistic SRP-based measurement detector, using the multiple speaker information obtained in the tracking step. In doing so, this approach tightly couples the two stages, and increases the detection rate of low-energy speakers during overlapping speech segments. Experiments conducted on the AV16.3 corpus showed a significant improvement of the detection and tracking performance, when the proposed approach is integrated into a Kalman-based multiple speaker tracking framework.**

Index Terms: Speaker overlap, multiple speaker tracking, steered response power, Kalman filter, conversational speech.

I. Introduction

Multiple speaker tracking using microphone arrays has become an essential tool to develop robust solutions to a large number of signal and speech processing problems, such as speech separation/enhancement, multi-party distant speech recognition, etc. More particularly, speaker overlap detection can be crucial for some systems, such as diarization [1], [2]. Classical acoustic source tracking approaches consist of two stages: 1) Detecting the measurements, which can be either Time Differences Of Arrival (TDOA) at the sensor pairs [3], [4], or noisy location estimates obtained with a Steered Response Power (SRP)-based technique [5], [6]. 2) These measurements are then processed by a filtering approach, such as Particle Filters (PF) [7], [8] or Kalman Filter (KF)-based approaches [9], [10]. In the multiple speaker scenario, the filtering framework is extended to include a multimodal estimation approach, which allows the tracking of multiple instantaneous speakers. Such extensions include the joint probabilistic data association filter [11], and the multiple model particle filter [12].

These two steps, namely, the measurement detection followed by filtering, are generally performed iteratively and independently. Furthermore, the multiple object tracking performance is highly dependent on the measurement detection rate, which drastically decreases in multi-party conversational/spontaneous speech. More precisely, in overlapping speech segments, the dominant speaker tends to mask, and therefore suppress, the secondary speakers causing the measurement detector to fail in detecting multiple instantaneous locations. This problem becomes more complex in noisy and/or highly reverberant environments, where the ambient noise sources become competitive to the desired source(s), leading to an increase in the clutter detection rate. We have recently proposed an approach [13] that counteracts the noise/reverberation problem by enhancing the TDOA measurement detection using tracking information. This approach, however, deals only with the **single** speaker problem, and was designed only for TDOA-based tracking approaches.

Motivated by the idea presented in [13], we propose in this paper a novel approach that improves the measurement detection of multiple overlapping speakers using tracking information. More precisely, at each time frame, the proposed approach 1) estimates the Probabilistic SRP (PSRP) [4], which is used as measurement detector in this work. This is followed by 2) the estimation of the predicted tracking distributions of all confirmed speakers. These pdfs characterize the most likely regions to contain the next measurements. 3) The resulting Gaussians are then used to update the mixture weights of the PSRP, by measuring the "similarity" between each Gaussian component in the PSRP and the predicted pdfs. Finally, (4) the enhanced PSRP is used to estimate the location measurements, which are processed by the multiple speaker tracking framework. Experiments conducted on the AV16.3 corpus show that enhancing the PSRP using a Kalman filter-based multiple speaker tracking framework improves significantly the overlap detection and the tracking rates, without any noticeable degradation of the angular error or the precision rate.

We proceed in this paper by reviewing the PSRP location measurement detector in Section II. Then, Section III shows how the tracking information can be used to improve the measurement detection stage. The performance of the proposed approach is shown in Section IV. Finally, we conclude in Section V.

II. Location measurement detector

This section reviews the PSRP-based multiple speaker localization, followed by a brief overview of the unsupervised Bayesian classifier, which is used to control the noise rate. These two approaches constitute the PSRP-based measurement detector (step (i) in Fig. 1a).

A. Multiple speaker localization approach

The TDOA that an acoustic source introduces at a microphone pair is estimated as the time difference alignment which maximizes the Generalized Cross Correlation (GCC) function of the signals [3]. Hence, the higher the GCC value is, the more likely it is that the alignment is the "true" TDOA [4], [6]. From this point of view, the **normalized** cross-correlation of two signals can be interpreted as a pdf of the TDOA, as it can be regarded as a set of observations sampled from a hidden distribution. This distribution is generally approximated by a Gaussian Mixture (GM) model using either the Weighted Expectation Maximization (WEM) algorithm from [14] or its practical approximation in [6]. The GM choice is justified by the multi-modality of the GCC function in noisy/reverberant environments as well as in the multiple speaker scenario, whereas the Gaussianity assumption of the TDOA error has been proven to be a valid assumption in speaker tracking approaches [10], [15].

The Probabilistic SRP (PSRP) approach [6] combines the resulting microphone pair GMs using a probabilistic interpretation of the SRP. The latter is typically expressed as a sum of the different microphone pair GCC functions [5]. The rest of Section II presents a brief introduction to the mathematical formulation of the PSRP approach.

Formally, let M and Q denote the number of microphones and corresponding pairs, respectively, and let $\mathbf{m}_h$ denote the position of the microphones, $h = 1, \ldots, M$. The location-TDOA mapping between the location $\mathbf{s}$ and the TDOA $\tau^q(\mathbf{s})$, introduced by $\mathbf{s}$ at the sensor-pair $q = \{\mathbf{m}_g, \mathbf{m}_h\}$, is given by

$$\tau^q(\mathbf{s}) = (\|\mathbf{s} - \mathbf{m}_h\| - \|\mathbf{s} - \mathbf{m}_g\|) \cdot c^{-1} \quad (1)$$

c denotes the speed of sound in air.
In the PSRP approach, the normalized GCC function (interpreted as a pdf of the TDOA) of the q-th microphone pair, $q = 1, \ldots, Q$, is approximated by a GM (see example Fig. 1c) given by

$$p^q_{gcc}(\tau^q) = \sum_{k=1}^{K^q} w^q_k \cdot \mathcal{N}(\tau^q, \mu^q_k, (\sigma^q_k)^2) = \sum_{k=1}^{K^q} w^q_k \cdot \mathcal{N}^q_k(\tau^q) \quad (2)$$

where μ^q_k, σ^q_k and w^q_k denote the mean, standard deviation and mixture weight of the k-th Gaussian $\mathcal{N}^q_k$, $k \in \{1, \ldots, K^q\}$, respectively. The PSRP of a given location $\mathbf{s}$ is obtained by 1) calculating the TDOA introduced by that location at all microphone pairs using (1). Then, 2) replacing each GCC contribution in the SRP function [5] with its GM approximation given by (2) (Fig. 1b shows a PSRP example with two speakers). This yields

$$PSRP(\mathbf{s}) \propto \sum_{q=1}^{Q} p^q_{gcc}(\tau^q(\mathbf{s})) = \sum_{q=1}^{Q} \sum_{k=1}^{K^q} w^q_k \cdot \mathcal{N}^q_k(\tau^q(\mathbf{s})) \quad (3)$$

The location measurement $\mathbf{s}_e$ is obtained by 1) extracting from each GM approximation the Gaussian component $\mathcal{N}^q_e$ where the potential source is dominant. Then, 2) calculating the restriction of (3) on the region of dominance associated to $\{\mathcal{N}^q_e\}_{q=1}^{Q}$. Finally, 3) the optimal location estimate is obtained via numerical optimization. This process is repeated until the number of desired instantaneous estimates is reached. For more details, the reader is referred to [4], [6].

B. Noise Rate Control

The multiple speaker localization approach provides a fixed number of instantaneous estimates (6 estimates per frame in this work). Given that the number of active speakers changes over time, a classification step is required to exclude the unlikely measurements. This is done using an unsupervised Bayesian classifier [16], which uses two location features to classify the location measurements to noise or speaker. More precisely, we calculate, for each location estimate $\mathbf{s}_e$, the Cumulative SRP (CSRP) feature, which is calculated on the region of dominance $\mathcal{S}_e$ associated to $\mathbf{s}_e$ according to

$$CSRP(\mathbf{s}_e) = \int_{\mathcal{S}_e} PSRP(\mathbf{s}) \cdot \mathrm{d}\mathbf{s} \approx \sum_{q=1}^{Q} w^q_{\mathbf{s}_e} \quad (4)$$

and the Maximum Likelihood Error (MLE) feature defined as

$$\epsilon(\mathbf{s}_e) = \sum_{q=1}^{Q} \left(\frac{\tau^q(\mathbf{s}_e) - \mu^q_{\mathbf{s}_e}}{\sigma^q_{\mathbf{s}_e}} \right)^2 \quad (5)$$

The EM algorithm is used to estimate the likelihood distribution of each feature separately as a 3-component mixture distribution modeling noise+speaker. The resulting likelihood distributions are then combined using a naive Bayesian classifier, which classifies each location estimate to noise/speaker (see [16] for more details).

III. PSRP Enhancement using Tracking Information

This section shows how the tracking information can be used to improve the measurement detection stage. Section III-A and Section III-B will introduce the mathematical formulation of the Bayesian tracking framework, whereas Section III-C will present the proposed PSRP enhancement approach using tracking information.

A. Bayesian filtering framework

The problem of tracking a time-varying system state S_t based on a sequence $y_{1:t} = \{y_1, \ldots, y_t\}$ of corresponding measurements is usually formulated as a Bayesian problem in which

1) A process model $S_t = f(S_{t-1}, V_t)$ is used to construct a prior $p(S_t|y_{1:t-1})$ for the state estimation problem at time t.
2) Then, the joint predictive distribution $p(S_t, Y_t|y_{1:t-1})$ of state and observation is constructed according to a measurement model $Y_t = h(S_t, W_t)$ (prediction step).
3) Finally, the posterior distribution $p(S_t|y_{1:t})$ is obtained by conditioning the joint predictive density $p(S_t, Y_t|y_{1:t-1})$ on the measured observation y_t (update step).

V_t and W_t are, respectively, the process and measurement noise random variables. The dynamics f, h and the initial posterior distribution form what is known as the *Dynamic State Space Model* (DSSM). The recursion of these steps form the Bayesian tracking framework. This framework has a closed form solution in the case where f, h are linear and V_t, W_t are Gaussian. In this case, the posterior distribution $p(S_t|y_{1:t})$ can be obtained as a conditional Gaussian distribution. This solution is known as Kalman Filter.

B. Generalization to the multiple speaker scenario

The generalization of the Bayesian tracking framework to the multiple object case can be done by jointly tracking **all** targets, as it is done in [11], [17], [18], or by tracking each object **separately** and **in parallel**, such as [17], [19]. The generalization approach proposed in [19] was particularly designed to overcome a few problems that occur in multi-party spontaneous speech. This approach tracks multiple concurrent speakers using a bank of parallel KF that evolve in time according to an HMM. The proposed HMM models the short and frequent active/inactive transitions of each speaker state. Therefore, this approach is used in this work to estimate the Kalman-based predicted distributions, which are used to enhance the PSRP. This novel approach will be referred to as Kalman SRP (KSRP).

The location posterior distribution $p^n(S_t|y_{1:t})$ of the n-th active speaker is updated according to the Bayesian framework above, and following the DSSM proposed in [19]

$$\text{Process Model} \quad : \; S_t = f(S_{t-1}, V_t) = S_{t-1} + V_t \quad (6)$$

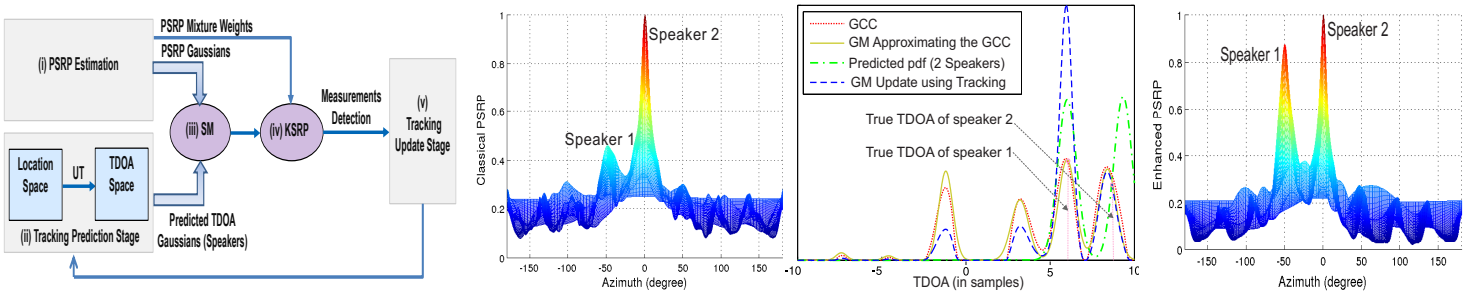
$$\text{Measurement Model} : \; Y_t = h(S_t, W_t) = S_t + W_t \quad (7)$$

The enhancement of the PSRP using the Kalman-based speaker information is explained in the next section.

C. PSRP update using tracking prediction stage

In classical tracking approaches, the observation detection step and tracking are performed separately. The presence of prior information of the targets, however, can efficiently improve the measurement detection. This idea was first investigated in [20], where the predicted pdf is used to reduce the measurement search space to a few likely regions. As an alternative to the space reduction, and following a line of thoughts similar to [13], we propose to enhance the PSRP of the most likely regions using the tracking information. This is achieved by 1) calculating the predicted location pdf of all confirmed targets at time t. Then, 2) mapping these location distributions to the TDOA space using the Unscented Transform (UT) and the location-TDOA function (1). This is followed by 3) calculating the similarity scores between each GCC-based TDOA GM approximation in the PSRP and the obtained TDOA pdfs of all confirmed speakers.

Formally, let N_t be the number of **confirmed** speakers at time t, and let S be the location random variable. We first calculate, for each speaker n, the predicted Gaussian (location) distribution $\mathcal{G}^n_S$



(a) Block Diagram of the Proposed Approach (b) Classical PSRP (c) TDOA Enhancement using Tracking (d) Enhanced KSRP

Fig. 1: *Fig. 1a is a block diagram of the proposed approach. Fig. 1b shows a classical PSRP with two speakers at azimuth -50° and 0°, respectively (step (i) in Fig. 1a). Fig. 1c shows how the tracking information is used to update the GCC-based Gaussian mixture and thereby enhances the low-energy speaker for a single microphone pair (steps (ii) and (iii) in Fig. 1a). Finally, Fig. 1d shows the enhanced KSRP after combining the updated GM of all microphone pairs. The low-energy speaker at azimuth -50° is clearly enhanced (step (iv) in Fig. 1a).*

which is expected by the tracking algorithm. This is done according to steps 1 and 2 of the tracking framework (Section III-A), followed by the marginalization over the state space. The predicted distributions $\mathcal{G}_S^n, n = 1, \ldots, N_t$ are then mapped to the TDOA space using UT.

1) Unscented transform for location-TDOA mapping: The enhancement of the measurement detection using tracking information is based on re-weighting the PSRP Gaussian mixture weights in (3). This is done by evaluating "how close" (and thereby how relevant) each component in the PSRP mixture to the predicted pdfs. The main issue here, however, is that the predicted pdfs are estimated in the location space, whereas the PSRP Gaussian components are in the TDOA space. To circumvent this problem, we propose to transform the predicted pdfs using the location-TDOA function given in (1). Unfortunately, this function is not linear, therefore, we propose to use the UT to propagate the mean and covariance of each predicted distribution through the nonlinear transformation (1) to the TDOA space (step (ii) in Fig. 1a). For ease of notation, the speaker index n is dropped in the rest of this section.

Let d be the location space dimension. In the UT approach, each speaker predicted (location) distribution $\mathcal{G}_S(\mathbf{s}) = \mathcal{N}_S(\mathbf{s}, \mu_S, \Sigma_S)$ is represented as a weighted empirical distribution of $2d+1$ weighted sigma points $\{\mathcal{S}_i, \mathcal{W}_i\}_{i=0}^{2d}$, calculated according to

$$\begin{array}{l|l} \mathcal{S}_0 = \mu_{\mathbf{s}} & \mathcal{W}_0 = \kappa/\lambda \\ \mathcal{S}_{2i+1} = \mu_{\mathbf{s}} + \sqrt{\lambda} R_i & \mathcal{W}_{2i+1} = 1/(2\lambda) \\ \mathcal{S}_{2i+2} = \mu_{\mathbf{s}} - \sqrt{\lambda} R_i & \mathcal{W}_{2i+2} = 1/(2\lambda) \end{array} \tag{8}$$

$i = 1, \ldots, (d-1)$, where $\lambda = d + \kappa$ for an arbitrary $\kappa \in \mathbb{R}$. In fact, κ specifies how much weight is placed on the mean, μ_S, and is set to $1/2$, which leads to a weight of $1/d$ for all sigma points. Furthermore, R_i are the rows of the matrix R, result of the Cholesky decomposition $R^T R$ of the covariance Σ_S.

The resulting sigma points are then mapped to the TDOA space according to the location-TDOA function (1)

$$\mathcal{T}_i^q = \tau^q(\mathcal{S}_i), \quad i = 0, \ldots, 2d, \quad q = 1, \ldots, Q \tag{9}$$

The mean and covariance of the Gaussian distribution $\mathcal{G}_{Tq}(\tau) = \mathcal{N}(\tau, \mu_{Tq}, \Sigma_{Tq})$, approximating the transformed (predicted) TDOA pdf at the q-th microphone pair, are calculated according to

$$\mu_{Tq} = \sum_{i=0}^{2d} \mathcal{W}_i \mathcal{T}_i^q, \quad \Sigma_{Tq} = \sum_{i=0}^{2d} \mathcal{W}_i (\mathcal{T}_i^q - \mu_{Tq})^2 \tag{10}$$

2) Similarity measure and PSRP update: The UT step above leads to N_t predicted TDOA distributions, and that is for each microphone pair $q \in \{1, \ldots, Q\}$. Each of these pdfs characterizes the most likely TDOA estimate to be generated by the speaker at the microphone pair q in the next time frame. Thus, these distributions can be used as a prior to enhance the most likely Gaussian components in the GM approximating the GCC function in the PSRP. This is done in this work through the calculation of similarity scores (step (iii) in Fig. 1a).

Formally, let $\mathcal{N}_k^q$ be the k^{th} Gaussian component in the GM approximating the q-th GCC function given in (2). Similarly to [13], the similarity score between $\mathcal{N}_k^q$ and $\mathcal{G}_{Tq}^n$; the predicted TDOA distribution of the n-th speaker at microphone pair q, can be calculated according to two different Similarity Measures (SM)

$$SM_1(\mathcal{G}_{Tq}^n, \mathcal{N}_k^q) = \frac{1}{1 + KLD(\mathcal{G}_{Tq}^n || \mathcal{N}_k^q)} \tag{11}$$

$$SM_2(\mathcal{G}_{Tq}^n, \mathcal{N}_k^q) = \int \sqrt{\mathcal{G}_{Tq}^n(\tau) \mathcal{N}_k^q(\tau)}\, \mathrm{d}\tau \tag{12}$$

where, $KLD(\mathcal{G}_{Tq}^n || \mathcal{N}_k^q)$ is the *Kullbak-Leibler Divergence* between the two Gaussians. The second SM is the *Bhattacharyya Coefficient* [21]. These two SM have closed form solutions for Gaussian distributions (see [13] for more details).

The resulting similarity scores are then used to enhance the mixture weights of all Gaussians in the PSRP distribution (3) before proceeding to the measurement detection step. More precisely, the new mixture weight $\bar{w}_k^q$ of the k-th Gaussian component in the GM approximating the q-th GCC function (2) is calculated according to

$$\bar{w}_k^q = \frac{w_k^q}{Z} \cdot \sum_{n=1}^{N_t} SM(\mathcal{G}_{Tq}^n, \mathcal{N}_k^q) \tag{13}$$

Z is the normalization term. Fig. 1c shows an example of enhancing a TDOA GM using the tracking information of two active speakers.

The new mixture weights incorporate the tracking prior of **all confirmed** speakers at time t into the detection step. In fact (as shown in Fig. 1c), the update step smoothes out the unlikely components in the GM and enhances the ones which are close to the predicted TDOA distributions. The enhanced $PSRP^{track}$ is given by

$$PSRP^{track}(\mathbf{s}) \propto \sum_{q=1}^{Q} \sum_{k=1}^{K^q} \bar{w}_k^q \cdot \mathcal{N}_k^q(\tau^q(\mathbf{s}), \mu_k^q, (\sigma_k^q)^2) \tag{14}$$

The updated $PSRP^{track}$ integrates only the tracking information of the **confirmed** speakers, whereas it smoothes out the space regions which do not contain an active target. These regions, however, may contain new emerging speakers (at the birth state of the tracking framework) that will be suppressed in the update step. To counteract this problem, we propose to preserve the new potential targets information provided by the PSRP according to (see example Fig. 1d)

$$KSRP = \alpha \cdot PSRP^{track} + (1 - \alpha) \cdot PSRP \tag{15}$$

α is a confidence factor characterizing how much trust is placed on the tracking information, and how unlikely it is that new speakers appear in the scene, α can also be learned as a time-dependent factor.

Table 1: Precision rate p_s, trajectory estimation rate t_r, speaker detection rate d_r and RMSE in degrees (see definitions below)

	seq18 (2 speakers)			seq24 (2 speakers)			seq40 (3 speakers)			seq37 (3 speakers)		
	STC	STT	KSRP	STC	STT	KSRP	STC	STT	KSRP	STC	STT	KSRP
p_s	85.0	**99.0**	97.2	**81.6**	81.1	80.9	94.1	**94.3**	92.9	90.6	**94.3**	92.9
t_r	81.5	90.4	**92.9**	63.7	66.8	**72.4**	75.7	86.6	**88.9**	82.2	84.2	**87.8**
d_r of Speaker 1	53.1	61.1	**67.5**	54.9	59.0	**62.7**	39.2	49.7	**53.3**	28.8	29.9	**32.2**
d_r of Speaker 2	51.6	54.8	**62.5**	34.3	37.9	**46.5**	38.4	40.0	**45.0**	66.2	71.0	**75.2**
d_r of Speaker 3	—	—	—	—	—	—	56.8	62.6	**66.3**	**46.7**	40.2	44.5
Average d_r	52.3	58.0	**65.0**	44.6	48.4	**54.6**	44.8	50.8	**54.9**	47.9	47.0	**50.6**
Average RMSE (°)	**1.96**	2.34	2.27	3.07	**3.04**	3.11	6.56	4.70	**4.59**	2.47	2.30	**2.25**

Table 2: Overlap detection rate (%) of N simultaneous speakers, N=1 is the percentage of frames with a single speaker estimate (no overlap)

N (Number of Overlapping Speakers)	seq18 (2 speakers)			seq24 (2 speakers)			seq40 (3 speakers)			seq37 (3 speakers)		
	STC	STT	KSRP	STC	STT	KSRP	STC	STT	KSRP	STC	STT	KSRP
0 (No Detection)	30.5	23.4	**21.1**	68.3	66.7	**64.0**	36.1	26.7	**24.8**	25.2	23.4	**20.2**
1 (No Overlap)	50.2	**55.6**	48.0	**25.6**	25.3	**25.6**	28.5	**33.6**	30.0	50.2	**52.0**	51.1
2 Speakers	19.3	21.0	**30.9**	6.12	8.05	**10.4**	30.6	37.6	**40.7**	23.6	23.7	**27.3**
3 Speakers	—	—	—	—	—	—	**4.83**	2.18	4.55	0.99	0.92	**1.40**

IV. Experiments and results

We evaluate the proposed approach using the publicly available AV16.3 corpus [22]. In this corpus, human speakers have been recorded in a smart meeting room (approximately 30m^2 in size) with a 20cm 8-channel circular microphone array. The sampling rate is 16 kHz and the real mouth position is known with a 3D error ≈ 1.2cm [22]. The AV16.3 corpus proposes a variety of scenarios, such as stationary and quickly moving speakers, varying number of overlapping speakers, etc. In these experiments, the signal was divided into frames of 512 samples (32ms). The PSRP-based instantaneous location estimator [6] and the speaker/noise classification task [16] were accomplished using the same setting proposed in [16]. We also use the same evaluation method proposed in [14], which estimates a 2-component GM which separates the "noise+speaker(s)" tracking estimates. The evaluation statistics are derived from the component representing the speaker estimates. More precisely, we report 1) the precision rate p_s, which represents the percentage of correct estimates, 2) the tracking rate t_r, which is calculated as the correct tracking duration with respect to the duration of frames with (at least one) ground truth location, 3) the individual and average speaker detection rate d_r, and finally 4) the average azimuth Root-Mean-Square Error (RMSE) in degrees. The speaker overlap detection rate of N simultaneous speakers is reported as the ratio of the number of detected frames with N correct simultaneous speaker estimates to the total number of frames. Similarly to the work proposed in [23], [24], the tracking is limited to the azimuth angle. This is due to the far-field assumption and the small size of the microphone array. The proposed approach, however, is general and can be applied to 3D tracking problems with other types of microphone arrays, such as distributed microphone arrays.

In the experiments reported below, the multiple speaker Short-Term Tracking (STT) approach proposed in [19] is used to estimate the tracking information, which is necessary to enhance the PSRP (Section III). The proposed approach, however, can be integrated into any multiple speaker tracking framework. The tracking confidence factor α is set to 0.9, and KLD is used to calculate the similarity scores. The STT parameter setting is the same as the one proposed in [19], except for the target-measurement confidence probability p_{confid}, which is set to 10^{-2} (see [19] for more details).

Table 1 and Table 2 present the performance of the original STT approach, which uses the PSRP approach as measurement detector to track multiple overlapping speakers, and compares it to the proposed approach (KSRP), which uses the tracking information provided by the STT to enhance the PSRP as described in this paper. Moreover, the results are compared to the complete multiple speaker Short-Term Clustering (STC) framework proposed in [23], [24]. This framework consists of 1) an instantaneous detection-localization approach, followed by 2) an automatic threshold that controls the false alarm rate. The obtained estimates are then 3) clustered into speech utterances using an STC approach. Finally, 4) a speech/non-speech classification is performed to discard estimates from non-speech frames (more details can be found in [24]).

Table 1 shows a clear improvement of the KSRP over the STT and the STC approaches. We can see that the KSRP achieves longer correct tracking trajectories (the increased correct tracking duration rate t_r), as well as higher individual and average speaker detection rates, and that is for most multiple speaker sequences from the AV16.3 corpus. More precisely, the KSRP approach achieves an average detection rate improvement of about 18.7% and 10.5% compared to STC and STT approaches, respectively, whereas the average trajectory estimation rate improvement is 13.0% and 4.5%. These results show that the main improvement of the KSRP approach is due to 1) the increased speaker detection in frames with low-energy, which is reflected by the improved trajectory rate, and to 2) the increased detection of (overlapping) speakers in low-energy frames or frames where the dominant speaker masks the secondary speakers. This is reflected in the Table 1 by the improved detection rates. This conclusion is also confirmed by Table 2, which shows that the KSRP significantly reduces the fraction of frames with no detection, which are typically low-energy/silence frames. Moreover, KSRP achieves a higher percentage of frames with (two or three) overlapping speakers, whereas the percentage of frames with a single speaker (no overlap detection) is decreased. The low overlap detection of three simultaneous speakers shows that speaker overlap mostly occurs between two speakers in spontaneous speech.

We can also conclude that the three approaches achieve comparable RMSE. The precision rate, however, shows a negligible degradation. The slight degradation introduced by the KSRP is mainly due to the absence of a speech/non-speech classifier, which uses speech cues to reject noise estimates during long silence/noise frames. As a result, the KSRP also enhances the noise trajectories during these frames leading to this slight degradation.

V. Conclusion and future work

We have proposed a novel multiple overlapping speaker detection approach, which couples the detection and tracking stages to enhance the detection of low-energy speakers. This approach uses the tracking information, obtained in the tracking prediction step, to enhance the PSRP-based overlapping speaker detector. This approach, however, does not include any speech features to increase its robustness to noise trajectories. This will be part of future work.

REFERENCES

[1] X. A. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

[2] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Automatic Speech Recognition and Understanding (ASRU). IEEE Workshop on*, Nov. 2003, pp. 411–416.

[3] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[4] Y. Oualil, F. Faubel, and D. Klakow, "A multiple hypothesis Gaussian mixture filter for acoustic source localization and tracking," in *Proc. IWAENC*, Sep. 2012.

[5] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, 2000.

[6] Y. Oualil, M. Magimai.-Doss, F. Faubel, and D. Klakow, "Joint detection and localization of multiple speakers using a probabilistic interpretation of the steered response power," in *Statistical and Perceptual Audition Workshop*, Sep. 2012.

[7] M. S. Arulampalam, S. Maskell, and N. Gordon, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, pp. 174–188, 2002.

[8] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. ICASSP*, vol. 5, May 2001, pp. 3021–3024.

[9] S. Gannot and T. G. Dvorkind, "Microphone array speaker localizers using spatial-temporal information," *EURASIP Journal on Applied Signal Processing*, pp. 174–174, 2006.

[10] U. Klee, T. Gehrig, and J. McDonough, "Kalman filters for time delay of arrival-based source localization," *EURASIP Journal on Applied Signal Processing*, pp. 167–167, 2006.

[11] T. Gehrig and J. McDonough, "Tracking multiple speakers with probabilistic data association filters," in *Proc. CLEAR*, 2007, pp. 137–150.

[12] A. Masnadi-Shirazi and B. Rao, "Separation and tracking of multiple speakers in a reverberant environment using a multiple model particle filter glimpsing method," in *Proc. ICASSP*, 2011, pp. 2516–2519.

[13] Y. Oualil, F. Faubel, M. Magimai.-Doss, and D. Klakow, "A TDOA Gaussian mixture model for improving acoustic source tracking," in *Proc. EUSIPCO*, Aug. 2012, pp. 1339 –1343.

[14] Y. Oualil, M. Magimai.-Doss, F. Faubel, and D. Klakow, "A probabilistic framework for multiple speaker localization," in *Proc. ICASSP*, May 2013, pp. 3962–3966.

[15] A. Levy, S. Gannot, and A. P. Habets, "Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments," *IEEE Trans. Acoust., Speech, Signal Process.*, 2010.

[16] Y. Oualil, F. Faubel, and D. Klakow, "An unsupervised Bayesian classifier for multiple speaker detection and localization," in *Proc. INTERSPEECH*, Aug. 2013.

[17] Y. Bar-Shalom and X.-R. Li, *Multitarget-Multisensor Tracking : Principles and Techniques*, 1995.

[18] R. P. S. Mahler, "A theoretical foundation for the stein-winter probability hypothesis density (phd) multi-target tracking approach," in *Proceedings of the National Symposium on Sensor and Data Fusion*, 2002.

[19] Y. Oualil and D. Klakow, "Multiple concurrent speaker Short-Term tracking using a Kalman filter bank," in *Proc. ICASSP - Sensor Array and Multichannel (ICASSP2014 - SAM)*, Florence, Italy, May 2014, pp. 1455–1459.

[20] Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*. Academic Press, 1988.

[21] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Transactions on Communication Theory*, vol. 15, pp. 52–60, 1967.

[22] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Proc. MLMI 04 Workshop*, May 2006, pp. 182–195.

[23] G. Lathoud and J. M. Odobez, "Short-term spatio-temporal clustering applied to multiple moving speakers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, p. 15, July 2007.

[24] G. Lathoud, "Spatio-temporal analysis of spontaneous speech with microphone arrays," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Switzerland, Dec. 2006.