MULTIPLE CONCURRENT SPEAKER SHORT-TERM TRACKING USING A KALMAN FILTER BANK

Youssef Oualil and Dietrich Klakow

Spoken Language Systems, Saarland University, Saarbrücken, Germany
youssef.oualil@lsv.uni-saarland.de

ABSTRACT

This paper presents a novel filtering approach for tracking multiple concurrent speakers with a microphone array. In this framework, a Kalman filter bank that evolves in time according to a temporal Hidden Markov Model (HMM) is proposed. This approach was designed to overcome two major problems that occur in spontaneous speech; namely, 1) the speaker overlap. This problem is solved using a bank of parallel Kalman filters that track multiple simultaneous speakers, and 2) the high discontinuity of spontaneous speech caused by short breaks and silences. This is solved using an HMM that allows speakers to change their state (speaking, silent, etc.) over time. The actual active speakers number and locations are extracted from the active filters using a second Kalman filter. Experiments on the AV16.3 showed an average tracking rate improvement of 8% compared to a short-term clustering approach, while being 7 times faster.

Index Terms— Microphone array, multiple speaker tracking, Kalman filter, hidden Markov model

1. INTRODUCTION

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

Despite their relative success, these approaches were mainly designed to overcome few classical problems of multiple object tracking, such as the non-linearity of the state space model dynamics [4, 8, 10], the robustness to noise [2, 12], and the correct estimation of the number of speakers [13]. These approaches however, did not address two main problems related to the speech nature, namely, 1) the high discontinuity of spontaneous speech, where an active speaker becomes frequently inactive for a short time (100-300ms), and 2) the suppression problem, were the dominant speaker masks the remaining speakers. These two problems reduce the speaker detection rate, and thereby makes the tracking of acoustic sources possible **only in short-term** i.e., while a speaker is talking without being suppressed. To overcome this problem, Lathoud et al. [14] proposed a short-term clustering (STC) approach, which extracts the speakers trajectories as short-term location clusters.

Following a line of thought similar to [14], we propose a novel multiple speaker short-term tracking framework, which consists of a bank of parallel KFs tracking multiple instantaneous speakers. More particularly, the state of each filter is updated according to a temporal Hidden Markov Model (HMM) that models 1) the frequent and short transitions in a speaker state (silent, speaking, etc.), as it models 2) the time-varying number of speakers, by allowing new speakers to appear (birth state) and existing speakers to disappear (final state). In doing so, the proposed approach presents a more realistic and flexible model to the multiple speaker tracking problem. this approach overcomes the above mentioned problems using short-term processing, similarly to [14], but proposes a more realistic model through use of the KF bank and the integrated HMM.

In the remaining part of this paper, we proceed by reviewing the location measurements detector that we have previously developed [15, 16, 17] (Section 2). Section 3 presents the single object tracking framework. Then, we introduce the proposed multiple speaker tracking framework in Section 4. Section 5 demonstrates the effectiveness of the proposed filter by means of an experimental study conducted on the AV16.3 corpus [18], including a comparison to the STC approach [14]. Finally, we conclude in Section 6.

2. MULTIPLE LOCATION MEASUREMENT DETECTOR

The location measurements detector aims at providing multiple instantaneous location estimates at each time frame. These measurements are then processed by the proposed tracking framework, which filters them over time to estimate the short-term speakers trajectories. In this work, we use our previously developed multiple speaker localization framework as a measurement detector [15, 16, 17]. This framework consists of 1) a multiple instantaneous location estimator [15, 16] that extracts a *fixed* number of potential location estimates per frame, followed by 2) an unsupervised Bayesian classifier [17], that controls the noise rate by classifying the resulting estimates into noise/speaker.

2.1. Multiple Instantaneous Location Estimator

In a recent work [15, 16], we have proposed a novel approach to the multiple source localization problem. This framework interprets each normalized Generalized Cross Correlation function (GCC) as a Probability Density Function (pdf) of the TDOA. This pdf is then approximated by a Gaussian mixture (GM) distribution using either the Weighted Expectation Maximization (WEM) algorithm from [16] or its practical approximation in [15]. The resulting TDOA Gaussian

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

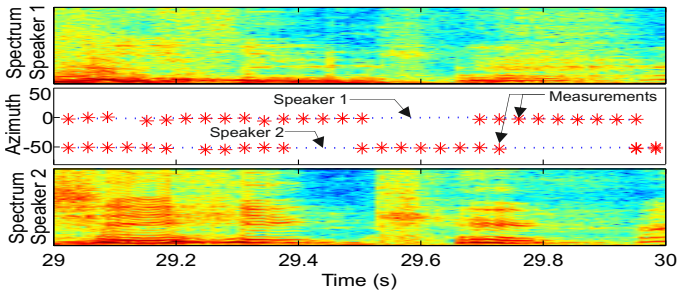

Fig. 1: One second of spontaneous speech showing an example, where the instantaneous location detector fails in producing location measurements (stars) during short silence/low energy frames.

mixtures are mapped to the location space using the location-TDOA mapping given by (1). The approach proposed in [15] combines the GMs using a probabilistic interpretation of the Steered Response Power (SRP_{prob}), whereas the approach proposed in [16] maximizes the TDOA joint pdf in the location space. The rest of this section presents a brief introduction to the approach proposed in [15], which is used in this work as a measurement detector.

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

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

where c denotes the speed of sound in the air.
The GM approximating the normalized GCC function (interpreted as a pdf of the TDOA) of the q-th microphone pair, is given by

$$p(\tau^q) = \sum_{k=1}^{K^q} w_k^q \cdot \mathcal{N}_k^q(\tau^q, \mu_k^q, (\sigma_k^q)^2) \quad (2)$$

where μ_k^q, σ_k^q and w_k^q denote the mean, standard deviation and mixture weight of the k-th component, $k = 1, \ldots, K^q$, respectively. The probabilistic SRP of a given location $\mathbf{s}$ is given by [15]

$$SRP_{prob}(\mathbf{s}) \propto \sum_{q=1}^{Q} \sum_{k=1}^{K^q} w_k^q \cdot \mathcal{N}_k^q(\tau^q(\mathbf{s}), \mu_k^q, (\sigma_k^q)^2) \quad (3)$$

The source location estimate $\mathbf{s}_e$ is obtained by 1) extracting from each GM distribution the Gaussian component $(w_{\mathbf{s}_e}^q, \mu_{\mathbf{s}_e}^q, \sigma_{\mathbf{s}_e}^q)$ where the source is dominant. Then, 2) calculating the restriction of (3) on the space region $\mathcal{S}_e$ where $\mathbf{s}_e$ is dominant. Finally, 3) the optimal location estimate is obtained via numerical optimization (see [15, 16] for more details).

2.2. 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 (BC) [17] that uses two location features to classify the location measurements to noise/speaker. More precisely, we calculate, for each location estimate $\mathbf{s}_e$, the Cumulative SRP (CSRP) feature given by

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

and the Maximum Likelihood Error (MLE) feature defined as

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

The EM algorithm is then used to estimate the probability distribution of each feature separately as a 2-component mixture distribution (noise+speaker). The resulting distributions are then combined using a naive Bayesian classifier that classifies each of the location estimates to noise/speaker (see [17] for more details).

3. SINGLE OBJECT TRACKING 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 estimation 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 $y_t = h(s_t, w_t)$.
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 $Y_t = y_t$.

v_t and w_t are, respectively, the process and measurement noise. The dynamics f, h and the initial posterior distribution form what is known as the *Dynamic State Space Model* (DSSM). The recursion of the above mentioned transformations 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 (this is the case in our problem). In this case, all the involved random variables remain Gaussian at all times and the posterior distribution $p(s_t|y_{1:t})$ can be obtained as a conditional Gaussian distribution. This solution is generally known as *Kalman filter*.

In this work, we propose to track the speaker location s_t using this recursive Bayesian framework on the following DSSM

$$\text{Process Model} \quad : \; s_t = f(s_{t-1}, v_t) = s_{t-1} + v_t \quad (6)$$

$$\text{Measurement Model} : \; y_t = h(s_t, w_t) = s_t + w_t \quad (7)$$

The proposed DSSM assumes that the speaker is stationary at each time transition. This assumption is reasonable given the short time frame that is considered in this work (32ms).

Section 4 introduces a generalization of this framework to a special multiple measurement/object case, where objects switch state from active to inactive (and vice versa) for a short period of time.

4. PROPOSED KALMAN FILTER BANK

Multi-party spontaneous speech utterances can be looked at as a sequence of *sporadic* and *concurrent* events [14, 19]. More precisely, 1) speech utterances are generally short and interspersed with many short silences, which results in a sequence of short and isolated segments of speech [14]. Furthermore, the sporadic nature of spontaneous speech increases in the multiple concurrent speaker scenario, where the dominant speaker suppresses the remaining speakers. This property automatically decreases the performance of classical tracking approaches. More precisely, these approaches often require that the object of interest is continuously observable over, relatively, a long period of time. This assumption is violated in the spontaneous speech case, where the instantaneous location estimates (from Section 2) are often unavailable during silences and during the speech

segments with low energy (Fig. 1). Moreover, the fast-changing speaker turns and the varying number of active speakers encountered in multi-party speech require very complex models, that allow the fast and concurrent transitions in the speaker turns.
The remaining part of this section presents a novel short-term filtering approach that incorporates these two characteristics. This is done using a KF bank that 1) models the multiple concurrent speaker scenario, and 2) allows speakers to change their state (speaking, silent,...etc) according to a HMM.

4.1. Short-Term Tracking Filter

The Short-Term Tracking (STT) filter proposes to track multiple speaker using a dynamic bank of KFs running independently and in parallel. Each filter in this bank estimates a single speaker short-term trajectory using the DSSM and the recursive Bayesian estimation framework from Section 3. Furthermore, the state of each filter is updated according to a temporal HMM (Fig. 2 is a simplified illustration of the proposed HMM). More precisely, a filter can be

1. In the hidden "Birth" state (B). In this state, the filter is initialized to track potential emerging targets.
2. Active (A), this hidden state corresponds to filters that are tracking the current active targets in the scene. These include 1) speakers from the previous frame that remained active, 2) speakers that went inactive for a short period of time (100-300ms) and became active again and 3) the new targets that just appeared in the scene.
3. Inactive (I), this hidden state models the short silence/break time frames as well as frames with low speech energy (see example in Fig. 1). This phenomenon causes a lack of measurements. Therefore, the filter becomes inactive.
4. Dead (D). This *final* state models filters that went inactive for a long period of time. This mainly occurs when speakers change turns or when a speaker stops talking. Filters that reach this state are automatically **removed** from the filter bank.

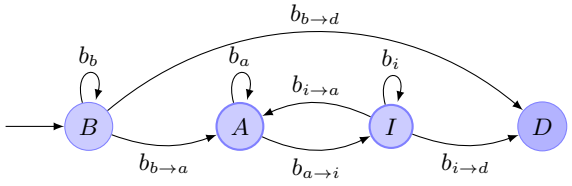

Fig. 2: A simplified HMM illustrating the filter state update at time t, given the observed filter activity.

4.2. Multiple Speaker Tracking Framework

This section introduces the mathematical formulation of the multiple speaker short-term tracking framework. Let $\mathcal{B}_t = \{\mathcal{F}_{t,k}\}_{k=1}^{N_t}$ be a bank of N_t KF running in parallel at time t. $\mathcal{B}_t$ can be divided to three disjoint banks according to each filter state

$$\mathcal{B}_t = \{\mathcal{F}^a_{t,k}\}_{k=1}^{N^a_t} \bigcup \{\mathcal{F}^i_{t,k}\}_{k=1}^{N^i_t} \bigcup \{\mathcal{F}^b_{t,k}\}_{k=1}^{N^b_t} \quad (8)$$

where $\mathcal{B}^a_t = \{\mathcal{F}^a_{t,k}\}_{k=1}^{N^a_t}$, $\mathcal{B}^i_t = \{\mathcal{F}^i_{t,k}\}_{k=1}^{N^i_t}$ and $\mathcal{B}^b_t = \{\mathcal{F}^b_{t,k}\}_{k=1}^{N^b_t}$ are the bank of active, inactive and potential (new speakers) filters, respectively. N^a_t, N^i_t and N^b_t are their respective cardinality. Let $\mathcal{B}_{t-1}$ be the filter bank at time $t-1$ and let s_t and y_t be the (location) state and observation random variables at time t, respectively. The goal here is to estimate the updated posterior distribution $p^k(s_t|y_{1:t})$ of each filter $\mathcal{F}_{t,k}, k = 1, \ldots, N_t$ in the filter bank $\mathcal{B}_t$ at time t. This time propagation of the posterior distribution is done in four steps :

Step 1. State prediction step: This step uses the process model given by (6) to calculate the prior distribution $p^k(s_t|y_{1:t-1}), k = 1, \ldots, N_t$ of each filter $\mathcal{F}_{t,k} \in \mathcal{B}_t$.

Step 2. Joint predictive distribution: In this step, we propagate the predicted prior distribution, calculated in the previous step, from the state space to the augmented joint state-observation space according to the measurement model given by (7). We obtain then N_t joint predictive distributions $p^k(s_t, y_t|y_{1:t-1}), k = 1, \ldots, N_t$.
In fact, these two steps run the classical Bayesian tracking steps 1 and 2 from Section 3 on N_t parallel Kalman filters.

Step 3. Confidence region estimation: For each filter $\mathcal{F}_{t,k}, k = 1, \ldots, N_t$, the joint predictive distribution $p^k(s_t, y_t|y_{1:t-1})$ is marginalized on the state space to obtain the predicted observation distribution $p^k(y_t|y_{1:t-1})$, which characterizes the most likely region to contain the next measurement. This distribution is then used to define the measurement confidence region $\mathcal{C}^k_t$ of the filter $\mathcal{F}_{t,k}$

$$\mathcal{C}^k_t = \text{Gate} = \left\{ Y_t \in \text{location space} | p^k(Y_t|y_{1:t-1}) \geq p_{confid} \right\} \quad (9)$$

p_{confid} is the confidence threshold (a probability).

Step 4. Target-measurement association and filter bank update: Let $\mathcal{Y}_t = \{Y^1_t, \ldots, Y^{M_t}_t\}$ be the M_t measurements received at time t, and let $\mathcal{A}_{t,k}$ be the target-measurement binary random variable associated to $\mathcal{F}_{t,k}$. The measurement Y^m_t is associated to the target $\mathcal{F}_{t,k_m}$ ($\mathcal{A}_{t,k_m} = 1$) if and only if $Y^m_t \in \mathcal{C}^{k_m}_t$. Then, the corresponding posterior distribution $p^{k_m}(s_t|y_{1:t})$ is updated according to step 3 of the single object Bayesian tracking framework (Section 3).

After the target-measurement association step, the observations (if there is any) $\bar{Y}^l_t, l = 1 \ldots, \bar{N}_t$ that were not associated to any target are used to initialize potential new speakers. More precisely, $\bar{N}_t$ Gaussian distributions $\mathcal{N}(s_t, Y_t, \Sigma_{init})$, where the means are the observations, are added to the filter bank $\mathcal{B}^b_t$. These filters are considered to be at the **birth** state (Fig. 2).

4.3. Update of the Filters State

Once we propagate the posterior distribution of all filters in $\mathcal{B}_t$, we proceed to the update of each filter state according to the proposed HMM (see illustration in Fig. 2). The new state of each filter is estimated based on its observed *activity* $t_{a,k}$, which is calculated on a context/history window of duration T_c. Formally, let L_f be the frame length in seconds, we calculate the *active* duration of $\mathcal{F}_{t,k}$ at time t according to $\Delta\ t_{a,k} = L_f \cdot (\sum_{j=t-T_c}^{t} \mathcal{A}_{j,k})$, whereas its *inactive* duration is given by $\Delta\ t_{i,k} = T_c - \Delta\ t_{a,k}$. The ***filter activity*** is defined as $t_{a,k} = max(\Delta t_{a,k} - \Delta t_{i,k}, 0)$.

Let $T^t_{a,k}$ be the *observed filter activity* at time t. The new state of the filter $\mathcal{F}_{t,k}$ is the one that maximizes the following probabilities

$$b_{b\to a} = \begin{cases} 1 & \text{if } \int_0^{T^t_{a,k}} f_b(\theta_b, \mathbf{x}) \cdot \mathrm{d}\mathbf{x} \geq p_{birth} \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

$$b_a = b_{i\to a} = \mathcal{A}_{t,k} \quad (11)$$

$$b_{a\to i} = 1 - \mathcal{A}_{t,k} \quad (12)$$

$$b_b = b_i = p_{survival} = \int_0^{T^t_{a,k}} f_s(\theta_s, \mathbf{x}) \cdot \mathrm{d}\mathbf{x} \quad (13)$$

$$b_{i\to d} = b_{b\to d} = p_{death} = 1 - p_{survival} \quad (14)$$

$f_x(\theta_x, .)$ ($x \in \{b, s\}$) are two pdfs (with parameters θ_x) modeling the birth and survival processes, respectively. Following the classical use of the exponential pdf as distribution modeling the life duration of objects, these two pdfs are considered to be two exponential distributions with respective means μ_b and μ_s.

Table 1 : Precision rate p_s, trajectory estimation rate t_r and real-time factor t

	seq11-1p-0100			seq18-2p-0101			seq24-2p-0111			seq40-3p-0111			seq37-3p-0001		
	p_s	t_r	t	p_s	t_r	t	p_s	t_r	t	p_s	t_r	t	p_s	t_r	t
STT	**92.2**	**78.4**	**4.8**	**94.4**	**90.7**	**4.7**	**83.7**	59.4	**4.7**	92.2	**86.3**	**4.8**	**93.6**	**90.1**	**4.7**
STC	87.9	69.8	33.4	85.0	81.5	42.0	81.6	**63.7**	32.0	**94.1**	75.7	37.8	90.6	82.2	36.4

Table 2 : Speaker detection rate (d_r) and average root-mean-square error (degree)

	seq11-1p-0100		seq15-1p-0100		seq18-2p-0101		seq24-2p-0111		seq40-3p-0111		seq37-3p-0001	
	STT	STC	STT	STC	STT	STC	STT	STC	STT	STC	STT	STC
d_r of speaker 1	**78.4**	69.8	**41.4**	40.6	**61.5**	51.6	48.5	**55.0**	**44.9**	39.2	**37.3**	28.8
d_r of speaker 2	—	—	—	—	**56.9**	53.1	**42.5**	34.3	**44.7**	38.5	**72.0**	66.2
d_r of speaker 3	—	—	—	—	—	—	—	—	**64.4**	56.8	**48.7**	46.7
Average d_r	**78.4**	69.8	**41.4**	40.6	**59.2**	52.3	**45.5**	44.6	**51.3**	44.8	**52.7**	47.9
Average RMSE	3.14	**2.90**	**1.13**	1.48	2.10	**1.96**	**2.54**	3.07	**4.95**	6.56	2.48	**2.47**

The update of the filters state according to the proposed HMM leads to a new bank of **active** filters $\mathcal{B}_t^a = \{\mathcal{F}_{t,k}^a\}_{k=1}^{N_t^a}$. Although $\mathcal{B}_t^a$ can be considered to be the final set of active speakers, the independent update of the filters, at each time frame, leads to a high perturbation in the number of active filters over time. This is often undesirable. Therefore, we use the estimated number of **active** filters $\mathcal{B}_t^a$ as a measurement in a second KF that smooths the number of active speakers over time.

5. EXPERIMENTAL SETUP AND RESULTS

We evaluate the proposed approach using the AV16.3 corpus [18], where 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 3-D error $\approx$ 1.2cm [18]. The AV16.3 corpus proposes a variety of scenarios, such as stationary and quickly moving speakers, varying number of simultaneous speakers, etc. In the experiments reported below, the signal was divided into frames of 512 samples (32ms). The instantaneous location estimates [15] and the speaker/noise classification task [17] were accomplished using the same setting proposed in [17]. We also use the same evaluation method proposed in [16], which estimates a 2-components GM $\mathcal{G}_n + \mathcal{G}_s$ that separates the "noise+speaker(s)" tracking estimates.

The evaluation statistics are derived from the component representing the speaker estimates. More precisely, the results are reported in terms of 1) the precision rate p_s, 2) the tracking rate t_r, this is calculated as the correct tracking duration w.r.t. the duration of frames with a (at least one) ground truth location, 3) the individual speaker detection rate d_r, 4) the average Root-Mean-Square Error (RMSE), and finally 5) the real-time factor t of the complete framework, on a standard Pentium(R) Quad-Core i5-3550 CPU clocked at 3.30GHz. Similarly to the work proposed in [14, 19], the tracking is limited to the azimuth angle. This is due to the far-field assumption as well as to the small size of the microphone array. The proposed approach however is general and can be applied to 3-D tracking problems with other types of microphone arrays, such as the distributed arrays. The tracking parameter setting is as follows, the birth mean is set to $\mu_b = 0.3$s whereas $\mu_s = 0.1$s. The latter aims at excluding filters with a decreasing activity near to 0. The birth probability $p_{birth} = 0.8$, the confidence probability is $p_{confid} = 10^{-3}$, whereas the duration of the context/history window is $T_c = 1$s.

Table 1 and Table 2 present the performance of the proposed short-term tracking (STT) approach on different sequences from the AV16.3 corpus, and compares it to the complete short-term clustering (STC) framework proposed in [14, 19]. 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 a short-term clustering approach. Finally, 4) a speech/non-speech classification is performed to discard estimates from non-speech frames (more details can be found in the PhD. thesis [19]). The STC results were generated using the public/free original code [19], using the same parameter setting explained above.

Table 1 shows a clear improvement of the STT over the STC approach. More precisely, the STT achieves longer correct tracking trajectories (the increased correct tracking duration rate t_r) while achieving comparable or improved precision rate p_s. Moreover, the time-factor t shows that the STT is 7-8 times faster than the STC. We can also conclude from this table that the proposed approach achieves a very satisfying tracking rate (average $t_r \approx 81\%$) and that it mostly tracks the correct acoustic sources (average $p_s \approx 91\%$).

Table 2 analyzes the distribution of the precision p_s and the tracking rate t_r results from Table 1 on the individual instantaneous speakers. We can see clearly that the proposed approach highly increases the speaker detection rate d_r without compromising the RMSE, which is comparable for both approaches. We can also see that for sequences which contain very long and frequent intentional segments of silence. Namely, *seq15-1p-0100* and *seq24-2p-0111*. For these sequences, the performance of the STT decreases and becomes comparable to the performance of the STC. This is mainly due to the absence of a speech/non-speech classifier that uses speech cues to reject the noise estimates during long silence/noise frames. As a result, the STT tracks noise sources during these long segments of silence/noise. The STC however, integrates such a classifier. Table 2 shows also that the detection rates d_r of the multiple speaker sequences are low compared to the corresponding tracking rate t_r. This is mainly due to the absence of the simultaneous speaker measurements caused by the speaker suppression problem, as well as the high active/inactive transition rate.

6. CONCLUSION

We have proposed a novel multiple speaker short-term tracking framework that incorporates the spontaneous/conversational speech properties. This approach consists of a Kalman filter bank that evolves in time according to a hidden Markov model. Experiments on the AV16.3 showed a clear improvement compared to a short-term clustering framework. The proposed approach however does not learn the HMM parameters, nor does it investigate the HMM structure, which can highly affect the tracking performance. This will be part of the future work.

7. REFERENCES

[1] 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.

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

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

[4] 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.

[5] D. B. Ward and R. C. Williamson, "Particle filter beamforming for acoustic source localization in a reverberant environment," in *Proc. ICASSP*, May 2002, vol. 2, pp. 1777–1780.

[6] 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.

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

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

[9] 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.

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

[11] A. Masnadi-Shirazi and B.D. 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.

[12] X. Zhong and J.R. Hopgood, "Nonconcurrent multiple speakers tracking based on extended kalman particle filter," in *Proc. ICASSP*, 2008, pp. 293–296.

[13] A. Quintan and F. Asano, "Tracking a varying number of speakers using particle filtering," in *Proc. ICASSP*, 2008, pp. 297–300.

[14] 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, pp. 15, July 2007.

[15] 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.

[16] 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.

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

[18] 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.

[19] G. Lathoud, *Spatio-Temporal Analysis of Spontaneous Speech with Microphone Arrays*, Ph.D. thesis, École Polytechnique Fédérale de Lausanne, Switzerland, Dec. 2006.