Short-Term Spatio–Temporal Clustering Applied to Multiple Moving Speakers

Guillaume Lathoud, *Member, IEEE*, and Jean-Marc Odobez, *Member, IEEE*

***Abstract*—Distant microphones permit to process spontaneous multiparty speech with very little constraints on speakers, as opposed to close-talking microphones. Minimizing the constraints on speakers permits a large diversity of applications, including meeting summarization and browsing, surveillance, hearing aids, and more natural human–machine interaction. Such applications of distant microphones require to determine where and when the speakers are talking. This is inherently a multisource problem, because of background noise sources, as well as the natural tendency of multiple speakers to talk over each other. Moreover, spontaneous speech utterances are highly discontinuous, which makes it difficult to track the multiple speakers with classical filtering approaches, such as Kalman filtering of particle filters. As an alternative, this paper proposes a probabilistic framework to determine the trajectories of multiple moving speakers in the short-term only, i.e., only while they speak. Instantaneous location estimates that are close in space *and* time are grouped into "short-term clusters" in a principled manner. Each short-term cluster determines the precise start and end times of an utterance and a short-term spatial trajectory. Contrastive experiments clearly show the benefit of using short-term clustering, on real indoor recordings with seated speakers in meetings, as well as multiple moving speakers.**

***Index Terms*—Localization, multiple acoustic sources, short-term clustering, speech segmentation, tracking.**

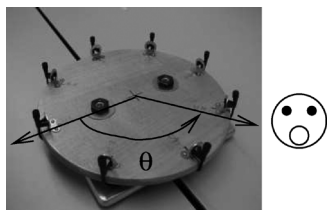

Fig. 1. Eight-microphone UCA used in the experiments (10-cm radius). θ denotes the azimuth angle of the speaker.

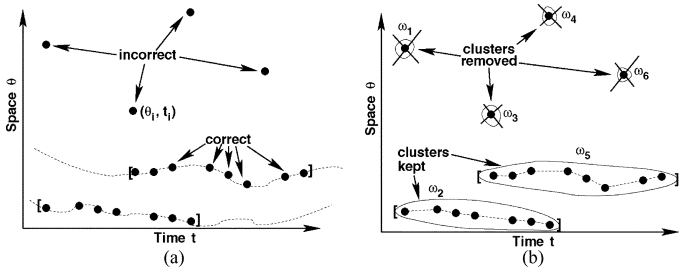

Fig. 2. Goal (a) and the proposed approach (b). Dots depict instantaneous location estimates θ_i, t_i. Dashed lines depict trajectories of the sources [true in (a), estimated in (b)]. Square brackets depict beginning and end of each speech utterance. Thin, continuous lines depict short-term clusters $\omega_1 \cdots \omega_6$.

I. Introduction

THIS paper investigates the analysis of spontaneous multiparty speech in a noninvasive manner. The goal is to estimate where and when the various speakers are talking. "Noninvasive" means distant microphones, for example a uniform cirular array or UCA (Fig. 1). Comparison between the signals received at the various microphones of the array permits to evaluate the instantaneous locations of multiple acoustic sources [1]–[3]. For example, with a UCA, the instantaneous location of a given acoustic source, at a given instant t_i, is estimated in terms of azimuth angle θ_i, i.e., the source direction in the horizontal plane (round face in Fig. 1 and dots in Fig. 2). Noninvasive methods can be opposed to very efficient but invasive methods that use close-talking microphones such as lapels [4], where one microphone is worn by each speaker, usually near the throat. Lapels permit to know precisely when each speaker is talking, because their signals are much cleaner than those received by distant microphones, due to the difference of distance [see the difference in signal-to-noise ratio (SNR) in Table I]. However, the range of applications permitted by lapels is limited, because 1) they require each user to wear a lapel, and 2) they practically provide no information about the location of each speaker.

On the contrary, distant microphones are noninvasive, thus putting much less constraints onto the users, and permit to estimate speakers' locations. These two properties allow for a wider range of applications to spontaneous speech processing, including surveillance [5], intelligent homes, offices and meeting rooms [6], hearing aids [7], hands-free speech processing in cars [8], as well as autonomous robots [9]. For example, a user browsing a meeting may be interested to jump directly at the presentation of a person, i.e., when that person stood up and moved to the screen. This would require to determine where and when each speaker is talking. The purpose of this paper is to build and evaluate an integrated system for the detection and localization of multiple speakers with distant microphones. The integrated system is designed to handle both static scenarii such as seated speakers in a meeting [10], and dynamical scenarii such as multiple moving speakers [11]. A

Manuscript received July 3, 2006; revised January 18, 2007. This work was supported by the European Union through the projects AMI and HOARSE. This work was also carried out in the framework of the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM)2. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Israel Cohen.

The authors are with IDIAP Research Institute, CH-1920 Martigny, Switzerland (e-mail: lathoud@idiap.ch; odobez@idiap.ch).

Digital Object Identifier 10.1109/TASL.2007.896667

TABLE I
AVERAGE SNR ACROSS MEETINGS AND SPEAKERS OF THE M4 CORPUS [10], IN dB DOMAIN. THE LAPELS ARE WORN BY EACH SPEAKER, BELOW THE THROAT. MEANWHILE, EACH SPEAKER IS BETWEEN 1 AND 3 m FROM THE MICROPHONE ARRAY

	Mean	Std. dev.
Lapel SNR	18.7	2.36
Mic. array SNR	10.7	2.02

generic probabilistic framework is proposed, and successfully tested on both types of scenarii.

More precisely, the task at hand is to use the instantaneous location estimates given by the microphone array, for two tasks: "short-term tracking" and "speech segmentation." "Short-term tracking" means using the instantaneous location estimates (θ_i, t_i) [dots in Fig. 2(a)] to reconstruct the spatial trajectories of the various speakers over time [dashed lines in Fig. 2(a)] while eliminating the incorrect location estimates [marked in Fig 2(a)]. "Speech segmentation" means detecting along time the beginning and end of each speech utterance [square brackets in Fig. 2(a)]. Both tasks are difficult, not only because of the lower SNR of distant microphones, but also because the number of active sources varies very often over time. In other words, spontaneous speech utterances are "sporadic" and "concurrent" events. "Sporadic" events are short, and interspersed with many silences. Instantaneous location estimates (dots in Fig. 2) are thus often unavailable, during silences and parts of speech with low energy. "Concurrent" events are simultaneous: indeed, people often talk over each other [12], and very often some background acoustic sources need to be eliminated (projector, laptops).

Tracking sporadic and concurrent events may be particulary difficult with classical filtering approaches, such as Kalman filtering [13], [14] and its extensions [15]–[17], and particle filtering [18], [19]. Although both have been successfully used to locate and track a single acoustic source [20]–[25], the fast-changing speaker turns encountered in spontaneous multiparty speech require either allowing a single-source model to switch between speakers [26], or specific multisource models [27]–[29]. Although particle filters can model multiple sources via multimodal distributions, deciding which modes are significant and which sources they belong to is an open and difficult issue [30]. Overall, while filtering approaches are interesting for modalities where events are somewhat continuously observable over relatively long durations of time (radar, video), complex birth/death rules are needed when the number of active sources varies very often along time, and difficult data association issues appear. Adding visual information, as in audiovisual speaker tracking [31]–[34], permits to circumvent these issues because each speaker can be continuously tracked using visual information, even while silent. However, the present paper considers the case where only audio information is available. With audio only, alternative approaches are thus needed to deal with sporadic and concurrent events.

This paper proposes to address "short-term spatio–temporal clustering," an intermediary task between instantaneous localization and continuous speaker tracking. The main contribution is a threshold-free, probabilistic framework for short-term clustering.[1] Instantaneous location estimates (θ_i, t_i) that are close to each other in both space and time are grouped into "short-term clusters" ω_k, as depicted by the thin continuous lines in Fig. 2(b). Each short-term cluster ω_k implicitly defines a part of the spatial trajectory of one speaker, as well as the beginning and the end of a speech utterance, as depicted by Fig. 2(b). Short-term clustering thus addresses both "short-term tracking" and "speech segmentation" tasks. This versatility could not be achieved with a purely static analysis of instantaneous location estimates for speech segmentation, as for example K-means, or the static criterion used in [36]. The aims and contributions of the present paper are thus threefold.

1) To investigate noninvasive methods for speech analysis.
2) To introduce a generic theoretical framework for short-term clustering, along with a confidence measure to detect trajectory crossings. The latter could be useful to select reliable location estimates, as a prior step to a trajectory reconstruction approach such as [37].
3) To determine its usefulness in contrastive experiments on real recordings, with multiple moving or static speakers. Indeed, the determination of an optimal partition $(\omega_1 \cdots \omega_k \cdots \omega_K)$ ground-truth would be difficult to elaborate, precluding the direct evaluation of the short-term clustering approach. On the other hand, it is conceptually simple to define a ground-truth in terms of speaker location and speech segmentation (respectively, azimuth locations and time segments).

In a previous work [38], [39], excellent speech segmentation performance was obtained, but speaker locations were assumed static, and known in advance. The present paper removes both assumptions through short-term clustering [Fig. 2(b)]. It is important to bear in mind that there will be many more short-term clusters ω_k than speakers: one short-term cluster per speech utterance. Long-term speaker clustering [40]–[42], where the target is only one cluster per speaker, is out of the scope of the present article. For speaker clustering results with distant microphones only, based on short-term clustering, the reader is referred to [43].

The rest of this paper is organized as follows. Section II introduces "maximum-likelihood" short-term clustering in a fully generic manner, considering a variable number of sources and a variable number of simultaneous location estimates (zero, one, or more). The proposed framework is probabilistic, threshold-free, does not require any random sampling. It can handle an unknown, time-varying number of observable sources, without any explicit birth/death rule. Section III describes online and offline implementations. Section IV illustrates the flexibility of short-term clustering, by using it to detect trajectory crossings in a threshold-free manner. Section V proposes an integrated multispeaker detection-localization system, based on short-term clustering. The integrated system is successfully tested for detection-localization of multiple moving speakers in highly dynamical recordings [11]. In addition, on the continuous tracking task, short-term clustering followed by Kalman filtering compares favorably to an existing particle

[1]Initial results were presented in the workshop paper [35].

filtering approach [28], [29]. Moreover, speech segmentation experiments in Section VII prove its ability to handle more static contexts such as meetings [10]. In both static and dynamical cases, it is shown beneficial to take a speech/nonspeech decision in terms of short-term clusters, as compared, e.g., to a frame-level approach. Finally, Section VIII concludes the paper. An implementation of multispeaker detection-localization with microphone arrays, including short-term clustering, is freely available at: http://mmm.idiap.ch/Lathoud/2006-multidetloc.

II. Short-Term Spatio–Temporal Clustering

This section presents the proposed short-term spatio–temporal clustering approach. The context is multiple moving sources: for each source and for each time frame, an *instantaneous* location estimate $X_i \stackrel{\text{def}}{=} (\theta_i, t_i)$ may or may not be available, where θ_i denotes the spatial location of the source, and t_i denotes time. At each instant t, there can be zero, one, or multiple location estimates $X_i = (\theta_i, t_i)$, such that $t_i = t$. The proposed approach relies on a threshold-free criterion to cluster these location estimates into short-term trajectories.

Although the approach is fully generic, throughout this paper the practical context will be one microphone array on a table (Fig. 1), recording multiparty speech in a meeting room (Fig. 10). The array is used to provide *instantaneous* audio source location information (see [1], [3], [44], and [45] for comprehensive reviews on this topic). The spatial location θ_i is for example an azimuth value in degrees. Our ultimate goal is to cluster the correct location estimates into speech utterances, and to discard the incorrect location estimates.

A. Assumption on Local Dynamics

Let $X_i = (\theta_i, t_i)$ for $i = 1 \dots N$ be all instantaneous location estimates of events emitted by the various sources. This includes the desired events (speech sounds) as well as noise. $\theta_i \in \mathbb{R}^D$ is a location in space, while $t_i \in \mathbb{N} \setminus \{0\}$ is a time frame index: $t_i \in (1, 2, 3, \dots)$. The notation $X_{1:N}$ designates the set of all location estimates: $X_{1:N} \stackrel{\text{def}}{=} \{X_1, X_2, \cdots, X_N\}$. For convenience, without loss of generality, we assume the location estimates ordered in time:

$$t_1 \leq t_2 \leq \cdots \leq t_N. \tag{1}$$

Note that there can be multiple location estimates per time frame, i.e., $t_i = t_{i+1}$.

The notation p designates a probability density function (pdf) or likelihood. The notation P designates a probability or a posterior probability. For any pair of location estimates (X_i, X_j), we define the two hypotheses.

- $H_0(i,j) \stackrel{\text{def}}{=}$ "X_i and X_j correspond to **different** sources."
- $H_1(i,j) \stackrel{\text{def}}{=}$ "X_i and X_j correspond to the **same** source."

The two hypotheses are complementary: $H_1(i,j) = \overline{H_0(i,j)}$. As a preliminary experiment, we ran instantaneous audio source localization with a Uniform Circular Array (UCA) of microphones [1] on real data [11], using the Steered Response Power with PHAse Transform (SRP-PHAT) approach [46]. For each location estimate $X_i = (\theta_i, t_i)$, θ_i is an estimate of the direction of an active acoustic source (azimuth in the horizontal plane).

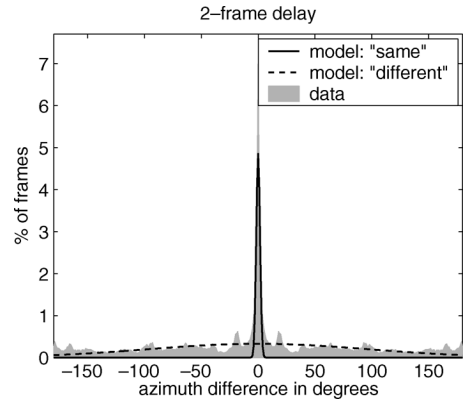

Fig. 3. Histogram of azimuth angle variations $\theta_i - \theta_j$ over a two-frame delay ($|t_i - t_j| = 2$), on real data (recording `seq01` from [11]). The super-imposed curves depict the bi-Gaussian mixture model obtained through EM training.

We observed the values of the difference $\theta_i - \theta_j$ for short delays $|t_i - t_j|$ up to T_{short}, where T_{short} is a small number of time frames (e.g., 7). Fig. 3 displays a typical histogram of location variations $\theta_i - \theta_j$ (in gray). Our interpretation is as follows: two location estimates X_i and X_j either correspond to the same source or not. In the first case (H_1), the difference $\theta_i - \theta_j$ is small: a source does not move a lot during a short time period. Hence the zero-mean central peak in the histogram. In the second case (H_0), the difference $\theta_i - \theta_j$ is random: the trajectories of two sources are independent, at least in the short-term. We therefore propose the following model for local dynamics, i.e., for $|t_i - t_j| \leq T_{\text{short}}$:

$$\begin{cases} p\left(\theta_i - \theta_j | H_0(i,j)\right) \sim \mathcal{N}\left(0, \sigma^{\text{diff}}_{|t_i - t_j|}\right) \\ p\left(\theta_i - \theta_j | H_1(i,j)\right) \sim \mathcal{N}\left(0, \sigma^{\text{same}}_{|t_i - t_j|}\right) \end{cases} \tag{2}$$

where $\forall T\ \sigma^{\text{same}}_T < \sigma^{\text{diff}}_T$, and $\mathcal{N}(\mu, \sigma)$ denotes the Gaussian pdf with mean μ and standard deviation σ. Although an intuitive choice in the case of H_0 would be a uniform distribution, we opted for a Gaussian to capture the dependency of σ^{diff}_T on the delay T. This dependency was observed on real data; examples can be found in [35].

The standard deviation σ^{same}_T accounts for short-term variations of location estimates due to both local motion and measurement imprecision. We argue that there is no need to distinguish the two, as long as the analysis is restricted to short delays $T \leq T_{\text{short}}$. For each delay T, σ^{same}_T, and σ^{diff}_T can be estimated simply, through EM training [47] of a bi-Gaussian mixture model, either on the entire data $\{\theta_i - \theta_j\}$ such that $|t_i - t_j| = T$, or in a blockwise fashion when the data is processed online, as in Section III-A. The mean of each Gaussian is fixed to zero. Although the weights are also trained during EM, they are not used in the rest of the process. Fig. 3 shows an example of bi-Gaussian mixture model.

The proposed model allows location differences $\theta_i - \theta_j$ to be close to zero, while X_i and X_j belong to two different sources: $H_0(i,j)$. Such a situation may happen in reality, whenever two sources' trajectories cross each other; see Section IV for further discussion on this topic.

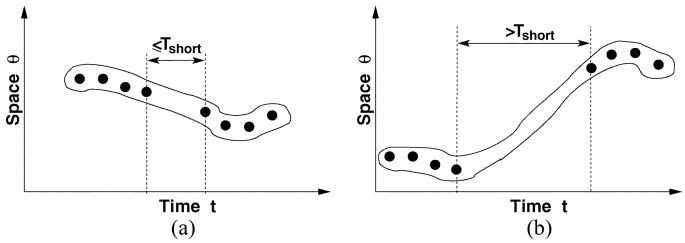

Fig. 4. Two types of clusters. This paper focuses on (a) short-term clusters, obtained with location cues only. (b) Long-term clustering requires additional cues, and is out of the scope of this paper. However, short-term clustering (STC) was shown to be useful for long-term clustering [43].

The present paper reports tests in 1-D space (azimuth). For higher dimensions, e.g., in spherical or Euclidean coordinates, one could simply replace σ_T^{same} and σ_T^{diff} with covariance matrices (diagonal should be sufficient). The rest of the approach presented below is unaffected by such a modification, because it relies on probabilities only (2).

B. Short-Term Clustering (STC)

This paper is focussed on short-term clustering (STC), using location cues alone. Given a value of T_{short}, a cluster $\omega \subset X_{1:N}$ is "short-term" if it has "time gaps" of at most T_{short} [Fig. 4(a)]. All other clusters are called "long-term clusters" [Fig. 4(b)].

Formally, a cluster ω is "short-term" iff

$$\forall t \in \left[\min_{X_i \in \omega} t_i, \max_{X_i \in \omega} t_i\right] \quad \exists X_j \in \omega \text{ s.t. } |t_j - t| \leq \frac{T_{\text{short}}}{2}. \quad (3)$$

A partition $\Omega = \{\omega_1, \cdots, \omega_k, \cdots, \omega_{K_\Omega}\}$ of the data $X_{1:N}$ is then "short-term" iff all clusters $\omega_k \in \Omega$ are "short-term." We'll denote this property with

$$\Omega \in \Gamma_{\text{ST}} \quad (4)$$

where Γ_{ST} is the set of all possible short-term partitions Ω of the data $X_{1:N}$, for a given value of T_{short}.

C. Threshold-Free Maximum Likelihood Clustering

Given the local dynamics (2), we propose to detect and track events as follows: find a short-term partition Ω of $X_{1:N}$ that maximizes the likelihood of the observed data:

$$\Omega^{\text{ML}} \stackrel{\text{def}}{=} \arg\max_{\Omega \in \Gamma_{\text{ST}}} p(X_{1:N}|\Omega). \quad (5)$$

Note that the number of clusters K_Ω has to be estimated as well. Each cluster $\omega_k \subset X_{1:N}$ contains locations for one event, e.g., a speech utterance. We are *not* trying to produce a single trajectory per source, but rather an oversplitted solution where $K_\Omega \gg 1$ is the number of individual events, for example speech utterances. The exact value of K_Ω is thus not important: we rather want to be *sure* that all location estimates within each cluster ω_k correspond to the *same* source. However, defining one cluster per location estimate obviously fulfills this constraint, although it is of little practical interest. Therefore, within each cluster, we would also like to have as many location estimates as possible, that belong to the same source. In other words, a criterion should be derived from the data-driven dynamical constraints (2), that also minimizes K_Ω as much as possible.

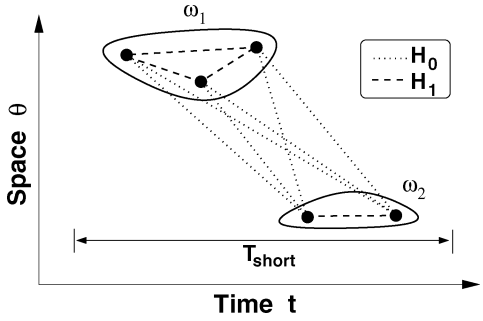

Fig. 5. This two-cluster partition $\Omega = \{\omega_1, \omega_2\}$ implicitly defines six local decisions $H_0(i,j)$ (dotted lines) and four local decisions $H_1(i,j)$ (dashed lines). In this particular case, all location estimates (dots) are within a T_{short} time window.

Over time, a source may move while being unobservable (e.g., silent, moving speaker). Using location cues alone, it is impossible to determine whether location estimates before and after the "silence" period belong to the same source. In other words, we can relate location cues in the short-term only. We therefore propose to maximize the following "short-term criterion," using a simplifying independence assumption between all pairwise differences $\theta_i - \theta_j$:

$$p_{\text{ST}}(X_{1:N}|\Omega) \propto \prod_{\substack{0 \leq i < j \leq N \\ 0 \leq |t_i - t_j| \leq T_{\text{short}}}} p\left(\theta_i - \theta_j | H^\Omega(i,j)\right) \quad (6)$$

where $H^\Omega(i,j)$ is either $H_0(i,j)$ or $H_1(i,j)$, depending on whether or not X_i and X_j belong to the same cluster ω_k in the candidate partition Ω, as depicted by Fig. 5. Each term of the product is expressed using (2). One important characteristic of this approach is that it does *not* need to explicitly model the true number of sources whose events are observed. Therefore, complex dynamical constraints and associated birth/death rules are not needed.

The proposed task, to cluster observations, fundamentally differs from particle or Kalman filtering, which estimate a hidden state variable from the observations. In addition, filtering usually relies on a conditional independence assumption between consecutive observations, given the state values [48]. On the contrary, the proposed STC precisely consists in modeling dependencies between several consecutive observations, up to the order T_{short}.

III. Optimization Algorithms

The goal is to find a short-term partition Ω of the observed location estimates $X_{1:N}$ that maximizes the criterion (6). Even short recordings contain thousands of location estimates: $N \gg 1$. It is thus untractable to try all possible short-term partitions $\Omega \in \Gamma_{\text{ST}}$. Sections III-A and III-B propose tractable, suboptimal implementations (online and offline).

A. Online: Sliding Window (SW)

We propose to find a suboptimal solution $\hat{\Omega}^{\text{ML}}$ by using a sliding analysis window, shifted at each iteration by N_{future} lo-

TABLE II
ONLINE SLIDING WINDOW (SW) MAXIMUM LIKELIHOOD ALGORITHM. THE LIKELIHOOD OF A PARTITION IS ESTIMATED WITH (6). LOCATION ESTIMATES ARE ORDERED BY INCREASING TIMES ($t_n \leq t_{n+1}$)

1) Initialization: For $T = 0 \cdots T_{\text{short}}$, initialize standard deviations σ_T^{same} and σ_T^{diff}, with unsupervised EM training on the beginning or all of $X_{1:N}$.
$n \leftarrow 1$.

2) $F \leftarrow X_{n:n+N_{\text{future}}-1}$.
Define all possible partitions of location estimates in F.
Choose the most likely partition $\hat{\Omega}_F^{\text{ML}}$.

3) $P \leftarrow \{X_i = (\theta_i, t_i) \mid t_n - T_{\text{short}} \leq t_i \leq t_n \text{ and } i < n\}$.
Define all possible merges between $\hat{\Omega}_P^{\text{ML}}$ and $\hat{\Omega}_F^{\text{ML}}$.
Choose the most likely merged partition and update $\hat{\Omega}_{P \cup F}^{\text{ML}}$.

4) Optionally, update σ_T^{same} and σ_T^{diff} using recently seen data (EM training, as in Step 1).

5) $n \leftarrow n + N_{\text{future}}$ and loop to Step 2.

cation estimates, where location estimates $X_{1:N}$ are ordered by increasing times

$$t_1 \leq t_2 \leq \cdots \leq t_N. \tag{7}$$

Table II describes the algorithm. Step 1 is the initialization: for each delay T, a bi-Gaussian model is fitted on azimuth differences $\{\theta_i - \theta_j\}$ such that $|t_i - t_j| = T$, as in Section II-A. Steps 2, 3, and the optional Step 4 constitute one iteration of the algorithm. Step 2 selects the maximum-likelihood (ML) partition of the N_{future} location estimates in the set F, independently of all other data. The set F has a fixed size (N_{future}), given by the user. Step 3 merges some clusters of the partition of F selected at Step 2, with some clusters in the past set P, again maximizing the likelihood. P contains all location estimates within T_{short} frames in the past. There can be a variable number of location estimates for each time frame, therefore the set P has a variable size. The optional Step 4 updates the bi-Gaussian models with recently seen data. The result of this algorithm is an estimate $\hat{\Omega}^{\text{ML}} \in \Gamma_{\text{ST}}$ of the ML short-term partition $\Omega^{\text{ML}} \in \Gamma_{\text{ST}}$ of all observed data $X_{1:N}$. The entire process is online, threshold-free, and can be deterministic.[2] As discussed in Section II-C, this process fundamentally differs from particle or Kalman filtering. In particular, the proposed approach models observation dependencies (2) up to the order T_{short}, even in the case $N_{\text{future}} = 1$.

One interest of this approach is bounded computational load. For both Step 2 and Step 3, evaluating a candidate partition (Step 2) or a candidate merge (Step 3) following (6) is easily implemented through a sum in the log domain over location estimates within F (Step 2) or $P \cup F$ (Step 3). The question is: how many partitions must be evaluated?

The total number of partitions evaluated at Step 2 is shown in Table III. For $N_{\text{future}} \leq 7$, there are at most 877 such partitions. As for Step 3, the worst case computational complexity was investigated in [35], in the special case where there is only one location estimate per time frame: for $T_{\text{short}} = 6$, there are at most 13 327 possible merges. However, in the general case investigated here, there can be multiple location estimates per time frame, thus many more possible merges. In practice, this situation is rarely encountered, as it corresponds to a case where most location estimates in $P \cup F$ are unrelated to each other. Two practical solutions can be used, favoring oversplitting. First, one could set a hard limit on the number of partitions that are constructed at Step 3 (e.g., 10 000), always including *at least* the case without any merge. Second, a heuristic can be used to prune out most of the "unlikely" merges, by forbidding short-term partitions Ω of the analysis window, that include "new" decisions $H^{\Omega}(i,j) = H_1(i,j)$ whenever

$$\frac{p\left(\theta_i - \theta_j | H_1(i,j)\right)}{p\left(\theta_i - \theta_j | H_0(i,j)\right)} \leq \epsilon \tag{8}$$

where ϵ is a small value, e.g., 10^{-10}. On tests with synthetic data (Section IV-B), we obtained exactly the same results with pruning or without pruning.

TABLE III
SW ALGORITHM: NUMBER OF POSSIBLE PARTITIONS, FOR EACH POSSIBLE NUMBER OF ELEMENTS (STEP 2 IN TABLE II)

Number of elements N_{future}	Number of possible partitions (Bell number [50])
1	1
2	2
3	5
4	15
5	52
6	203
7	877
8	4 140
9	21 147
10	115 975
11	678 570
>11	prohibitive

B. Offline: Simulated Annealing (SA) Optimization

Alternatively, the proposed modeling can be cast into a Markov random field framework [51], by defining a label field $E = \{E_i, i = 1 \ldots N\}$. E_i is a random variable denoting the label associated with observation X_i. The actual label values are unimportant, they can be, for example, integers. We define a graph $\langle E, \mathcal{G} \rangle$, where E represents the set of nodes, and $\mathcal{G}$ denotes the neighborhood system. E_i is a neighbor of E_j iff $|t_i - t_j| \leq T_{\text{short}}$. A graph $\langle E, \mathcal{G} \rangle$ uniquely defines a short-term

[2]A deterministic initialization of the EM training of the bi-Gaussian model can be used, similarly to [49].

TABLE IV
SA ALGORITHM: THE MRF OPTIMIZATION (IN PRACTICE $\lambda = 0.97$)

1) Initialization: $\tau \leftarrow \tau_0$, $E \leftarrow E_{init}$
2) While $\tau > \tau_{end}$

- $E \leftarrow SA(E, \tau)$
- $\tau \leftarrow \tau * \lambda$

3) Iterated Conditional Mode (ICM) optimization steps until there is no label change:
Do
$E' \leftarrow E$ and $E \leftarrow SA(E', 0)$
While $E \neq E'$

TABLE V
SA ALGORITHM: THE SIMULATED ANNEALING $SA(E, \tau)$ STEPS

1) Initialization: $I \leftarrow \{1, ..., N\}$
2) While $I \neq \emptyset$

- sample $i \in I$ uniformly.
- define candidate labels $L_i \leftarrow E_{\mathcal{G}_i} \cup \{NewLabel\}$ where $E_{\mathcal{G}_i}$ is the set of current labels from the E_j's in the neighboorhood of E_i.
- compute the posterior probabilities $p_{ik} \stackrel{\text{def}}{=} p(\, E_i = l_k \mid E \setminus E_i \,)$ over the candidate labels $l_k \in L_i$:

$$p_{ik} \propto \exp\left(-\frac{1}{\tau} \sum_{\substack{j \\ \langle i,j \rangle \in \mathcal{C}}} \beta_{ij} \delta_{l_k - E_j} \right)$$

- sample $E_i \simeq \text{Multinomial}(p_{ik})$.
- remove i from I.

partition $\Omega \in \Gamma_{\text{ST}}$. Given the observations $X_{1:N}$, the goal is to estimate the label field E that maximizes the ML criterion (6), or strictly equivalently, that maximizes the following Potts field [52]:

$$p_{\text{Potts}}(E) \stackrel{\text{def}}{=} \frac{1}{Z} \cdot e^{-U(E)} \tag{9}$$

with

$$U(E) \stackrel{\text{def}}{=} \sum_{\langle i,j \rangle \in \mathcal{C}} V_{ij}(E) \stackrel{\text{def}}{=} \sum_{\langle i,j \rangle \in \mathcal{C}} \beta_{ij} \cdot \delta_{E_i - E_j} \tag{10}$$

where $\mathcal{C}$ is the set of pairwise cliques of the neighborhood system $\mathcal{G}$, Z is the partition function (normalization factor), and δ_x is the Kronecker function, the value of which is 1 when $x = 0$, and 0 otherwise. The β_{ij} are the Potts coefficients, the values of which depend on the observations and can be derived from (2) and (6)

$$\beta_{ij} = \log\left(\frac{p(\theta_i - \theta_j | H_0(i,j))}{p(\theta_i - \theta_j | H_1(i,j))} \right). \tag{11}$$

The maximization of the probability $p_{\text{Potts}}(E)$ with respect to the label field E is equivalent to the minimization of the energy function $U(E)$ and can be conducted using standard techniques. We adopted a simulated annealing approach [52], [53], with Gibbs sampling and an exponentially decaying temperature, followed by an iterated conditional mode (ICM) [52], [53] procedure, as described in Tables IV and V.[3] We considered three alternatives, with different initializations (E_{init}).

- SA(1): The initial label field E_{init} has a single label shared by all nodes.
- SA(N): The initial label field E_{init} has one different label per node.
- SA(SW-1): The initial label field E_{init} is constructed in a sequential and causal fashion: for each new observation, we select the label that minimizes $U(E)$ given all previous observations. This initialization is strictly equivalent to SW-1: the SW algorithm with $N_{\text{future}} = 1$.

The outcome of these alternatives and on their impact on the criterion, the number of short-term clusters, and the performance, are discussed in Section VII-E.

IV. APPLICATION: THRESHOLD-FREE DETECTION OF TRAJECTORY CROSSINGS

This section defines a confidence measure for each possible individual decision $H_d(i,j)$ ($d = 0$ or 1), and explains how it allows to detect and deal with low confidence situations such as trajectory crossings. The goal is to illustrate the flexibility of the proposed probabilistic framework (6). It is relevant to contexts where the events emitted by the various sources are somewhat "continuous" (e.g., acoustic signals from vehicles [54]). The task investigated here is to extract pieces of trajectories that each belong *for sure* to a single source. Achieving this task would be useful as a first step, prior to Bayesian network tracking [37], for example.

We propose to use the posterior probability $P(H_d(i,j)|X_{1:N})$ as a confidence measure for a given local decision $H_d(i,j)$. Assuming equal priors for all possible short-term partitions $\Omega \in \Gamma_{\text{ST}}$ of the observed data $X_{1:N}$, the posterior probability of the local decision can be expressed as follows, for $d = 0$ or 1:

$$P\left(H_d(i,j)|X_{1:N}\right) \propto \sum_{\substack{\Omega \in \Gamma_{\text{ST}} \\ H^{\Omega}(i,j) = H_d(i,j)}} p(X_{1:N}|\Omega) \tag{12}$$

where $P(H_0(i,j)|X_{1:N}) + P(H_1(i,j)|X_{1:N}) = 1$. Section IV-A proposes to use this confidence measure to modify the ML optimization procedure.

A. Threshold-Free Confident Clustering

We would like to determine when trajectories cross and to split short-term clusters accordingly. Fig. 6(a) gives an example of ML partition. X_i and X_j are very close, it is thus not clear which short-term cluster X_i and X_j should ideally belong to. In such a case, there may exist a different partition with a close-to-optimal likelihood [Fig. 6(b)]. We propose here to break each short-term cluster that contains X_i or X_j into two "confident" parts, and to create two separate one-element clusters $\{X_i\}$ and $\{X_j\}$ [Fig. 6(c)].

Let us assume that the ML criterion (6) leads to the decision $H^{\hat{\Omega}^{\text{ML}}}(i,j) = H_0(i,j)$. "Confidence" in the latter, is low on Fig. 6(a), and implicitly, confidence is low as well for H_0 and H_1

[3]Note that any short-term partition configuration ($\Omega \in \Gamma_{\text{ST}}$) can be reached with a nonzero probability, which is a requirement of simulated annealing.

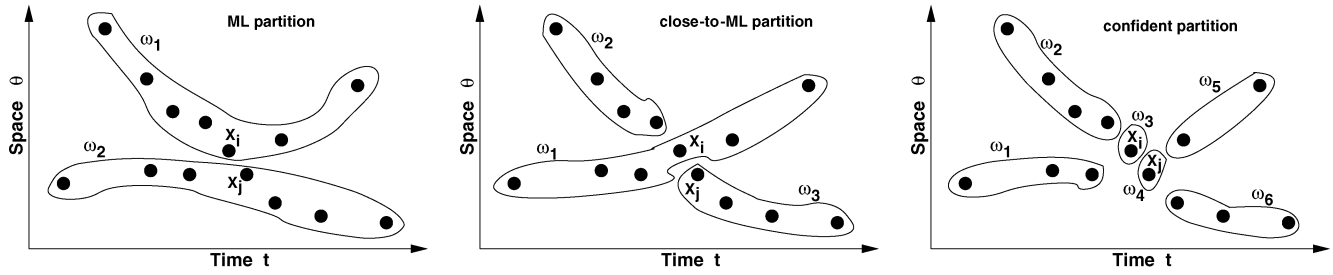

Fig. 6. Example of low confidence decision $H_0(i,j)$ at a trajectory crossing. Each dot is a location estimate. A continuous line depicts each short-term cluster ω_k.

decisions that involve X_i or X_j. We propose to detect "low confidence" in a ML decision $H_0(i,j)$, by comparing it to all ML decisions $H_1(r,s)$ in the same analysis window W. Formally, a "low confidence" $H_0(i,j)$ decision is defined as verifying

$$P\left(H_0(i,j)|X_{1:N}\right) < M_1\left(\hat{\Omega}_W^{\mathrm{ML}}\right) \tag{13}$$

where

$$M_1\left(\hat{\Omega}_W^{\mathrm{ML}}\right) \stackrel{\mathrm{def}}{=} \max_{\substack{r<s \\ (X_r,X_s)\in W\times W \\ H^{\hat{\Omega}_W^{\mathrm{ML}}}(r,s)=H_1(r,s)}} P\left(H_1(r,s)|X_{1:N}\right). \tag{14}$$

For the SW algorithm, "confident clustering" is implemented by modifying Steps 2 and 3 in Table II as follows.

- For all (X_i, X_j) in the analysis window $W = F$ (Step 2) or $W = P \cup F$ (Step 3), estimate $P(H^{\hat{\Omega}_W^{\mathrm{ML}}}(i,j)|W)$ using (12). For Γ_{ST}, we use the set of all candidate partitions in W.
- Step 2: whenever a decision $H_0(i,j)$ given by the ML algorithm has "low confidence" (13), split in two parts the short-term cluster containing X_i, at time t_i. Idem for X_j. Additional one-element clusters $\{X_i\}$ and $\{X_j\}$ are created [Fig. 6(c)].
- Step 3: whenever a decision $H_0(i,j)$ given by the ML algorithm has "low confidence" (13), forbid any merge between each of the two short-term clusters containing X_i (resp. X_j), and any other short-term cluster.

Confident clustering requires $N_{\mathrm{future}} > 1$. Otherwise, with $N_{\mathrm{future}} = 1$, cancellation of a single ML merge (Step 3) will most often be replicated in the future, thus resulting in an unnecessarily long series of one-element clusters. We verified this phenomenon on the same synthetic data as the one used in Section IV-B.

In the case of SA, since only some of the partitions are explored, a different implementation may be needed to detect trajectory crossings.

B. Multisource Tracking Examples

We generated data that simulates "sporadic" and "concurrent" events by restricting $X_{1:N}$ to have at most only one location estimate per time frame $(\forall i\, t_i < t_{i+1})$, yet with trajectories that look continuous enough so that it is still a tracking problem. In all test sequences, the number of active sources varies over time, and trajectories cross several times. The task is twofold.

- Task 1: From instantaneous location estimates $X_{1:N}$, build the various trajectories accurately.
- Task 2: Extract pieces of trajectory (clusters), where each piece *surely* belongs to a single source. This implies that no short-term cluster extends beyond any trajectory crossing.

Fig. 7 compares the result of ML clustering (SW implementation, with $T_{\mathrm{short}} = 7$ and $N_{\mathrm{future}} = 7$) with the result of the confident clustering described in Section IV-A. Although the ML clustering correctly builds the various trajectories (task 1), it produces arbitrary decisions around the points of crossing. On the contrary, confident clustering correctly splits the trajectories at all crossing points (Task 2).

Thus, confident clustering could be particularly useful to create *reliable* pieces of trajectories, which do not include any crossing point. These pieces of trajectories can then be linked using approaches such as Bayesian networks [37].

V. Application to Detection-Localization of Multiple Speakers

This section presents an integrated system for detection and instantaneous localization of multiple speakers, along with experimental results on recordings with multiple moving speakers. We show that the use of STC (Sections II and III) for speech/nonspeech (SNS) classification permits to achieve substantial improvements over frame-level approaches. The resulting integrated system is used as a platform for multispeaker tracking in Section VI, and for multispeaker segmentation in Section VII.

Since the focus of this paper is STC, the multisource detection-localization system is summarized as much as possible (Section V-A). A detailed description of this implementation is available in [43].

A. Instantaneous Multisource Detection-Localization

Zero, one, or more location estimates $X_i = (\theta_i, t_i)$ are produced at each time frame, where θ_i is the azimuth of an audio source with respect to a microphone array (Fig. 1), and t_i the frame time. "Instantaneous" means that each time frame is processed individually. A two-step approach is used, as illustrated by Fig. 8. First, a sector-based predetection step [49] limits the search space to zero, one or more sectors of space around the microphone array. Second, the SRP-PHAT [46] local maximum is found within each active sector, through scaled conjugate gradient descent [55]. Both steps rely on a Generalized Cross-Correlation with PHAse Transform (GCC-PHAT) [56] representation of the data to estimate the following.

- The bandwidth occupied by the sources in a given sector of space [8], which is then modeled in an unsupervised

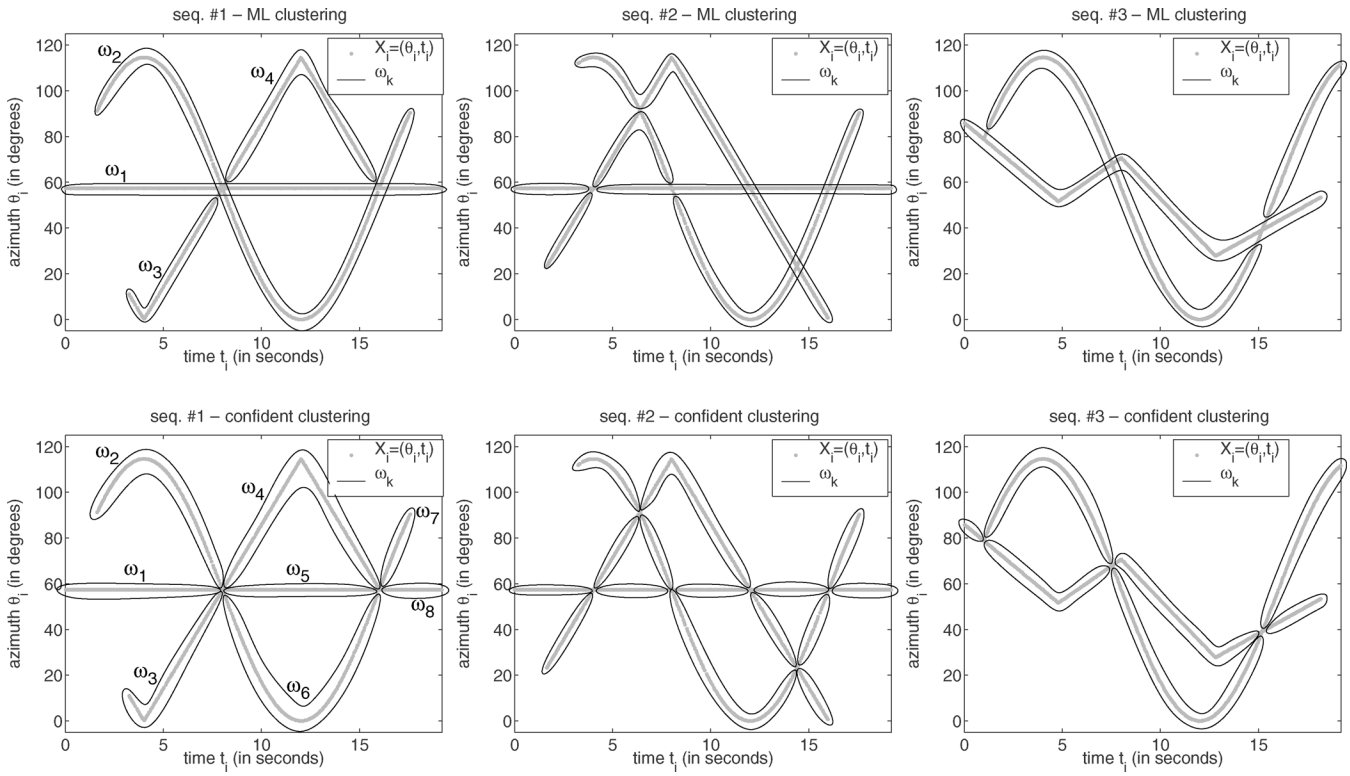

Fig. 7. Comparison ML clustering/confident clustering on multiple source cases, where the number of active sources varies over time. Gray dots: location estimates X_i (62.5 Hz). Black lines: clusters ω_k. We can see that the ML clustering algorithm takes arbitrary decisions at trajectory crossings. On the contrary, the confident clustering correctly splits the short-term clusters at each trajectory crossing. A Matlab implementation of short-term clustering (ML and confident, SW implementation) can be found on the following website, along with ten synthetic data examples: http://mmm.idiap.ch/Lathoud/2006-short-term-clustering/.

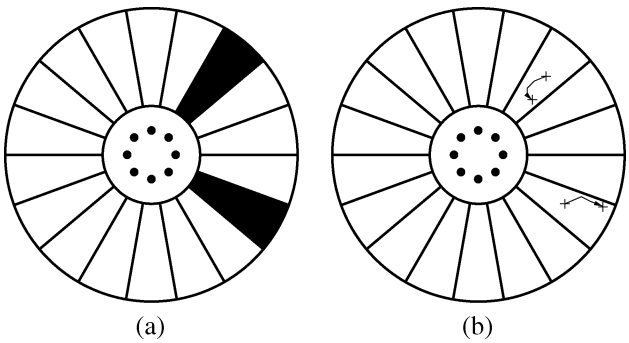

Fig. 8. Two-step implementation for multisource detection-localization [43]. The eight dots indicate the locations of the microphones. (a) sector-based detection-localization. (b) gradient descent within each active sector.

manner, using a probabilistic framework [49]. A final binary decision [Fig. 8(a)] is taken using automatic thresholding [49].

- The SRP-PHAT metric [46] at a given point of space, for gradient descent [43] [Fig. 8(b)].

The reader is referred to [43] for full details, freely available code, and tests on real data that show that this system achieves detection-localization of up to three multiple simultaneous speakers, with near real-time performance (implementation called "FAST" in [43]). We used a 32-ms frame length with 50% overlap (16-ms frame shift).

B. Speech/Nonspeech (SNS) Classification

Let us assume that we have a system for instantaneous detection and localization of multiple audio sources, as described above. "Audio sources" include not only human speakers, but also noise sources such as a projector, a laptop and the various reverberations, as shown in Fig. 11(a). However, our final task is multi-*speaker* detection-localization, so it is needed to remove the nonspeech location estimates [see the result in Fig. 11(b)]. In other words, each location estimate must be classified as speech or nonspeech. In this paper, two systems are investigated: the SNS decision is taken either at the location estimate level (X_i), i.e., not using context—or at the short-term cluster level (ω_k), i.e., using context.

"Individual SNS": SNS Decision for Each Individual Location Estimate X_i Separately: As detailed in [43], we compare to a threshold the posterior probability of having a wideband, non-noisy signal emitted by the source at location θ_i and time t_i. The threshold is determined without tuning, as in [49], to match a user-defined target of detection false alarm rate (FAR), for example FAR $= 1\%$.

"Cluster SNS": SNS Decision per Short-Term Cluster ω_k: When a short-term cluster contains more than one location estimate, it is possible to estimate the non-stationarity of the whole short-term cluster, based on a location-dependent way of extracting MFCC, detailed in [43]. A "speech cluster" ω_k is then defined as having the following.

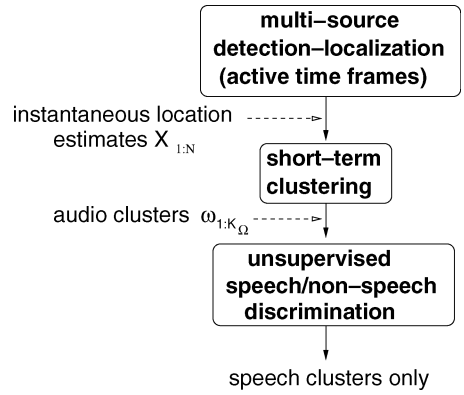

Fig. 9. Detection-localization of multiple speakers, using microphone arrays (systems SW-1, SW-7, and SA).

Fig. 10. Recording `seq45` from the AV16.3 Corpus [11], with three moving speakers. The eight-microphone array is marked with an ellipse. The ball markers were used to construct the ground-truth location of each speaker with respect to the array.

1) At least two location estimates corresponding to a wideband, non-noisy source (X_i, X_j, $i \neq j$). (Individual SNS decisions.)
2) Non-stationarity above a threshold. In practice, this threshold is fixed and does not require tuning, due to an underlying spectrum normalization [43].

Two advantages can be expected from cluster SNS over individual SNS. First, some correct location estimates that would not pass the "individual" test, called "weak but correct" estimates hereinafter, may still be part of a speech cluster, and thus be correctly detected. Thus, more speech should be retrieved, as verified in Section V-D. Second, the non-stationarity measure allows to exclude machine noise sources such as a projector or a laptop. This is useful in the meeting environment, as reported in Section VII-D.

C. Experimental Protocol

To assess whether STC is beneficial to the detection decision, we compared the two SNS classification systems (individual versus cluster) using the same underlying instantaneous multisource detection-localization system (Section V-A, top block in Fig. 9). We ran the two systems on eight real indoor recordings from the freely available AV16.3 corpus [11]. Multiple simultaneous speakers are moving around a table, with an eight-microphone, 10-cm radius, UCA on its top (Figs. 1 and 10). Three cameras were used to reconstruct the 3-D ground-truth location of each speaker, with an error inferior to 1.2 cm [11]. In the clustering case, we used the SW algorithm with $N_{\text{future}} = T_{\text{short}} = 7$.

The focus here is correct detection-localization of multiple moving speakers. For both systems, speech location estimates are compared with the closest ground-truth speaker location. We derive the following performance metrics [43] on intervals on which the ground-truth locations of all speakers are known:

- bias and standard deviation in degrees, to assess the precision of the localization.
- the percentage of detected speech that was correctly localized, i.e., within a small error margin (the margin is derived from the bias and the standard deviation, as detailed in [43]).

TABLE VI
COMPARISON BETWEEN TWO TYPES OF SNS DECISION, ON THE AV16.3 CORPUS [11], INCLUDING REAL RECORDINGS WITH MULTIPLE MOVING SPEAKERS, SIMULTANEOUSLY SPEAKING. BIAS AND STANDARD DEVIATION (STD) ARE EXPRESSED IN DEGREES

SNS decision granularity	Total detected	Correctly localized	Precision (deg.) bias	std
Individual SNS: X_i	**284.9 s**	91.85%	0.335	2.158
Cluster SNS: ω_k	**699.0 s**	92.05%	0.381	2.542

D. Results and Discussion

From Fig. 11(a) and (b), one can see that the SNS decision using short-term clusters permits to remove most of the incorrect location estimates, while keeping most of the correct location estimates. This is also visible in Fig. 11(c), which presents a three-speaker case. Note that the gaps in the ground-truth do not mean that a speaker is silent, but simply that the mouth location was occluded on a camera—and thus the ground-truth location unavailable.

Table VI presents the overall detection-localization results. The percentage of correct location estimates is very similar for both methods, but short-term clustering clearly retrieves much more speech signal.[4] Indeed, as discussed in Section V-B, each short-term *speech* cluster contains some "weak but correct" location estimates, which would not pass the "individual" test. This confirms the interest of grouping location estimates *before* rejecting noise. The price to pay is a slight decrease in localization precision, probably due to those "weak" location estimates. This loss of precision can anyway be compensated for by smoothing the trajectory described by each short-term cluster, e.g., using Kalman filtering [13], [57] or RTS smoothing [58], as shown in Section VI.

Overall, the proposed cluster SNS method allows to select a much larger amount of correct location estimates as compared to the individual SNS method, while rejecting the same proportion of incorrect ones. This could be useful as a prior step to

[4]To obtain the same "Total detected" duration as for the cluster method (699.0 s), the individual method can be made less conservative. The percentage correct then falls to 62.28%, with precision bias 0.375, std 2.869.

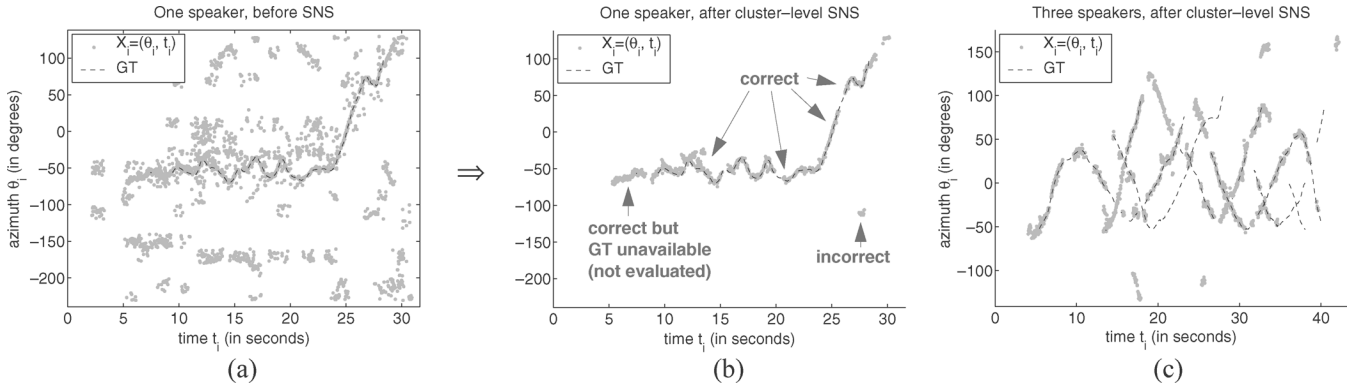

Fig. 11. Effect of the cluster-level SNS classification (Section V-D). Grey dots: location estimates X_i, dashed lines: ground-truth (GT). The GT of a speaker is only available when the mouth is visible on three cameras. Data: AV16.3 Corpus [11] (one speaker: `seq11`, three speakers: `seq45`).

trajectory analysis, as done in [37]. Section VII uses the cluster SNS method for meeting segmentation.

VI. Additional Filtering for Multispeaker Continuous Tracking

While Section V investigated the detection and the instantaneous localization of multiple speakers, the present section investigates the continuous tracking of multiple speakers. We propose to infer a filtered trajectory for each short-term speech cluster ω_k (as defined in Section V-B). Each short-term speech cluster is processed separately as a single source, thus avoiding all data association issues. The proposed deterministic approach (STC, cluster SNS, then filtering) is compared to an existing (stochastic) particle filtering approach [28], [29], which directly processes the location estimates $X_{1:N}$.

A. Cluster SNS Followed by Filtering (Cluster+Filter)

Each short-term speech cluster ω_k is filtered separately. For each short-term speech cluster ω_k, for each $X_i \in \omega_k$, we replace the spatial location estimate θ_i with a filtered estimate. We propose three fully deterministic[5] approaches:

Cluster+WMF: WMF stands for weighted mean filtering, where each filtered estimate is the seven-point weighted mean of $\theta_{i-3:i+3}$. Each weight is a probability of speech, as estimated in [43]. The window size (seven points) was not tuned.

Cluster+KF: KF stands for Kalman filtering [13], [57], in 2-D state space $[\theta_i \ \dot{\theta}_i]^{\mathrm{T}}$. The measurement noise matrix [57] was set to $\mathbf{I} \cdot \sigma_1^{\mathrm{same}}$ where $\mathbf{I}$ is the 2 by 2 identity matrix, the process noise matrix [57] was set to $\mathbf{I} \cdot \sigma_1^{\mathrm{same}} \cdot 10^{-1}$, and the error covariance matrix [57] was initialized to $\mathbf{I} \cdot \sigma_1^{\mathrm{same}} \cdot 10^{5}$. Only one parameter required tuning (10^{-1}).

Cluster+RTS: RTS stands for Rauch–Tung–Striebel smoothing [58], also known as Kalman smoothing. The parameters are exactly the same as in Cluster+KF.

B. Existing Particle Filtering Approach (PF)

PF: We implemented an existing PF approach for multiple audio sources [28], [29]. The PF explicitly addresses the data association issue, where for each time frame, the number of audio sources and their locations are estimated from the multiple instantaneous measurements (X_i). The PF also includes rules for births and deaths of audio sources. We used 1000 particles for each source. In practice, we observed the PF approach to be very sensitive to the chosen dynamical parameters (α and β in [28], [29]) as well as to the initialization of the speed when creating an audio source. For a fair comparison with the "Cluster+Filter" approaches, initialization of the speed was implemented using the next T_{short} time frames. Finally, we had to post-process the result by thresholding posterior probabilities of "existence" [28], [29], to remove spurious trajectories. We used the same parameter values as in [28], [29], except for seven parameters that required tuning. Tuning involved substantially more tests than in the "Cluster+Filter" approaches.

PF+SNS: The PF approach in [28], [29] does not distinguish between speech sources and nonspeech sources. The PF+SNS approach rejects a nonspeech source using the exact same SNS classification as for a cluster ω_k (Section V-B).

C. Results and Discussion

Results are reported in Table VII, using the same metrics as described in Section V-C. All "Cluster+Filter" approaches substantially improve the localization precision as compared to "Cluster" alone, especially in terms of standard deviation. PF and PF+SNS provide a slightly smaller bias ($-0.1°$) than "Cluster+Filter," but a much larger standard deviation ($+0.2$ to $+0.4°$). A possible reason is the stochastic nature of the PF approach, where the inferred trajectory sometimes "jumps" far away. This is not the case of the three "Cluster+Filter" approaches, which are fully deterministic.

In the case of a "jump," the PF eventually destroys the corresponding source (death) because it does not match observations anymore. The PF also creates new source(s) that match the observations (births), but it often requires a few frames before a new source is considered to be "fully existing" [28, Sec. 2.4.5]. In our understanding, this hysteresis-like behavior leads to lose much speech. This may be an explanation for the much lower "Total detected" duration for PF and PF+SNS (first column of Table VII), as compared to the "Cluster+Filter" approaches. We tried to lower the existence threshold of the PF, but it led

[5]STC is implemented deterministically (Section III-A). SNS is also deterministic [Section V-B], as well as the three filters (WMF, KF, and RTS).

TABLE VII
COMPARISON BETWEEN VARIOUS FILTERING TECHNIQUES FOR TRACKING MULTIPLE MOVING SPEAKERS, ON THE AV16.3 CORPUS [11]. WMF, KF, RTS, PF, AND PF+SNS ARE DEFINED IN SECTION VI. L = INSTANTANEOUS LOCATION ESTIMATES, P = PROBABILITY OF SPEECH ACTIVITY [43], N. PAR. = TOTAL NUMBER OF PARAMETERS, (TUNED) = NUMBER OF PARAMETERS THAT REQUIRED TUNING

Tracking system	Input	N. par. (tuned)	Total detected	Correctly localized	Precision (deg.) bias	std
Cluster (no filt.)	L	5(0)	699.0 s	92.05%	0.381	2.542
Cluster+WMF	L, P	6(0)	699.0 s	91.56%	0.378	1.854
Cluster+KF	L	8(1)	699.0 s	91.67%	0.320	2.074
Cluster+RTS	L	8(1)	699.0 s	91.01%	0.353	1.796
PF	L, P	20(7)	425.0 s	86.68%	0.257	2.222
PF+SNS	L, P	22(7)	313.4 s	95.29%	0.290	2.257

to a quick increase of the proportion of noisy location estimates. This phenomenon is visible by comparing the "Total detected" and "Correctly localized" figures of the PF and PF+SNS approaches.

Overall, the "Cluster+Filter" approach seems to be superior to the PF approach on spontaneous multiparty speech in terms of both detection and localization precision, with several possible reasons. First, "Cluster+Filter" is implemented in a fully deterministic manner, whereas the stochastic nature of PF permits erroneous "jumps." Second, in "Cluster+Filter," both STC and filtering are threshold-free, thus potentially less sensitive to tuning than PF.[6] Third, STC groups location estimates that are close to each other *before* rejecting noise and filtering, whereas PF attempts to do all at the same time.[7] Finally, STC does *not* attempt to extract long-term trajectories from the essentially sporadic spontaneous speech, whereas PF does, which leads to many births and deaths. This behavior of PF differs from the results shown in [29, Ch. 2], probably because the latter only tested continuous read speech.

VII. MEETING SEGMENTATION APPLICATION

In this section, we report experiments conducted on real meeting data recorded with a UCA, the M4 Corpus [10]. We use the system described in Section V, with cluster-level SNS classification (the filtering proposed in Section VI is not necessary for this task). A comparison with close-talking microphones is given. These experiments can be seen as a more static counterpart to the moving speaker experiments reported above. We want to determine whether *the same system* can cope with both static and dynamic contexts. In the previous section, the focus was on correct detection for precise localization of multiple moving speakers. In this section, we focus on the speech segmentation task: we have a precise time-domain ground-truth, but an approximate spatial ground-truth.

"Speech segmentation" means that we are only trying to separate the different speakers in the short-term (where? when?). The target is one short-term cluster per speech utterance. Results reported in [43] show that the proposed speech segmentation system forms a strong basis for long-term speaker clustering (who?) with distant microphones, where the target is only one cluster per speaker. However, it is out of the scope of the present article.

The differences between a previous work [59] and the approach presented here are as follows.

- We are focusing on the speech segmentation task only, not on the speaker clustering task.
- We use distant microphones only (no lapel).
- We segment each meeting independently.
- The proposed approach does not rely on a hidden Markov model (HMM).

On the contrary to the preliminary results reported in [35], all systems presented here perform automatic removal of non-speech sources (e.g., projector).

A. Test Data

The test corpus includes 21 short meetings from the publicly available M4 Corpus (http://mmm.idiap.ch). The total amounts to about 2 h of multichannel speech data. Three meetings were used as a development set to tune post-processing parameters (Section VII-D), and after that, 18 meetings were used as a test set to evaluate performance metrics.

In the data, people are seated around a table, and sometimes stand up and move to the screen for a presentation using a projector, or to the blackboard. In all meetings, an independent observer provided a very precise speech/silence segmentation. Because of this high precision, the ground-truth includes many very short segments. Indeed, more than 50% of the speech segments are shorter than 1 s, as depicted in Fig. 12.

B. Proposed Systems

We tested several variants of this system, corresponding to the different optimization algorithms introduced in Section III. The online systems use the sliding window algorithm, with $T_{\text{short}} = 7$ and either $N_{\text{future}} = 1$ (SW-1) or $N_{\text{future}} = 7$ (SW-7). $N_{\text{future}} = 7$ was not tuned, it was only chosen to keep the computational cost low (Table III). We also tested the offline systems based on SA.

In all cases we use maximum-likelihood clustering (Section II-C) for this application. The confident clustering described in Section IV-A is not necessary in the case of speech, since trajectory crossings are rarely seen due to the sporadicity of speech. Confident clustering is more relevant to cases where the signals are more continuous in time (e.g., vehicles [54]).

C. Baseline System Using Lapels

The proposed systems use distant microphones only. We compared them to a lapel-only baseline. The latter is an energy-based technique that selects the lapel with the most energy at each frame and applies energy thresholding to classify the frame as speech or silence. We tried to use zero-crossing rate (ZCR) as well, but it degrades significantly the segmentation performance. Indeed, ZCR appeared very sensitive to some noises found in meetings, such as writing on a sheet of paper. Therefore, results are reported with energy only. Note that lapels have an SNR around 18.7 dB, while distant microphones have an SNR around 10.7 dB (Table I).

[6]As for the SNS, it has one threshold, but no tuning was required (Section V-B).

[7]Grouping before denoising already explained the superiority of the "Cluster SNS" method over the "Individual SNS" method in Section V-D.

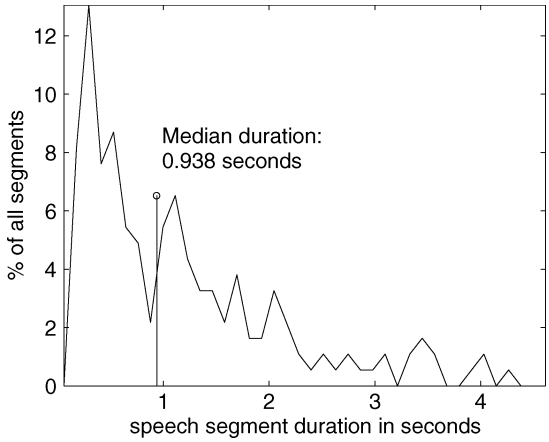

Fig. 12. Histogram of speech segment durations in the ground-truth (M4 Corpus [10]).

TABLE VIII
SEGMENTATION RESULTS ON THE M4 CORPUS. SW-1 AND SW-7 USE DISTANT MICROPHONES ONLY. VALUES ARE PERCENTAGES, RESULTS ON OVERLAPS ONLY ARE INDICATED IN BRACKETS. PRC,RCL,F: THE HIGHER THE BETTER. DER: THE LOWER, THE BETTER

	Lapel baseline	SW-1	SW-7
PRC	89.3 (67.8)	83.8 (71.8)	83.9 (71.7)
RCL	90.4 (63.8)	90.9 (82.0)	90.6 (81.6)
F	89.8 (64.6)	87.2 (75.7)	87.0 (75.5)
DER	8.2 (34.9)	11.8 (19.7)	12.0 (20.0)

D. Performance Measures

We evaluated the result of each system as follows. For the proposed systems (SW-1, SW-7 and SA),[8] for each speech location estimate, the corresponding time frame (32-ms segment) is attributed to the closest human speaker in space (the ground-truth location(s) of each speaker are known). Similarly, for the lapel baseline, for each lapel, each speech time frame is attributed to the speaker wearing the lapel. For each speaker, the resulting speech/silence segmentation is further post-processed with basic morphological operators [60]: dilation, erosion, closure, and opening, as in [39]. For each system, post-processing parameters are tuned to maximize the F-measure on the development set (three meetings). Each system is then applied on the test set (18 meetings). The performance metrics described in the following were evaluated for each meeting separately. Averages across all meetings are reported in Tables VIII–X. As opposed to previous results [35], all systems must include automatic removal of nonspeech sources such as the projector.

For each meeting, evaluation was performed as follows. For each speaker, the resulting speech/silence segmentation is compared to the ground truth (GT). Following [61], four types of durations are calculated:

- D_{TP}: total duration of all segments in a meeting where a speaker is speaking in both result and GT;
- D_{TN}: total duration of all segments in a meeting where a speaker is silent in both result and GT;
- D_{FP}: total duration of all segments in a meeting where a speaker is speaking in the result, but silent in the GT;
- D_{FN}: total duration of all segments in a meeting where a speaker is silent in the result, but active in the GT.

Following [61], six metrics are defined, with values in [0,1]:

- False Alarm Rate: $\mathrm{FAR} \stackrel{\mathrm{def}}{=} D_{\mathrm{FP}}/(D_{\mathrm{FP}} + D_{\mathrm{TN}})$;
- False Rejection Rate: $\mathrm{FRR} \stackrel{\mathrm{def}}{=} D_{\mathrm{FN}}/(D_{\mathrm{FN}} + D_{\mathrm{TP}})$;
- Half Total Error Rate: $\mathrm{HTER} \stackrel{\mathrm{def}}{=} (\mathrm{FAR} + \mathrm{FRR})/2$;
- Precision: $\mathrm{PRC} \stackrel{\mathrm{def}}{=} D_{\mathrm{TP}}/(D_{\mathrm{TP}} + D_{\mathrm{FP}})$;
- Recall: $\mathrm{RCL} \stackrel{\mathrm{def}}{=} D_{\mathrm{TP}}/(D_{\mathrm{TP}} + D_{\mathrm{FN}})$;
- F-measure: $\mathrm{F} \stackrel{\mathrm{def}}{=} 2 \times \mathrm{PRC} \times \mathrm{RCL}/(\mathrm{PRC} + \mathrm{RCL})$.

In the optimal case, FAR, FRR, and HTER are all equal to 0, and PRC, RCL, and F are all equal to 1. The F-measure is a harmonic mean of PRC and RCL; therefore, a large value of F-measure requires a large value for *both* PRC and RCL.

We also report results in terms of diarization error rate (DER), a percentage metric defined by NIST [62]. As opposed to PRC/RCL/F results, DER excludes part of the data from the evaluation: within a collar of 0.25 s around each speech segment end-point, results are not evaluated. Moreover, silences of less than 0.300 s are removed from both result and ground-truth. The DER is then defined as the percentage of speech that was wrongly attributed: $\mathrm{DER} = \mathrm{MISS} + \mathrm{FA} + \mathrm{SPKR}$, where MISS and FA are the percentages of missed speech and false alarms, respectively, and SPKR is the percentage of speech attributed to the wrong speaker. Full details can be found in [62].

In any case (PRC/RCL/F or DER), it is important to bear in mind that in this paper we are *only* evaluating the speech segmentation quality (one cluster per utterance). An evaluation of the application of STC to speaker clustering (one cluster per speaker) is reported in [43].

E. Results and Discussion

Choice of an Optimization Method: Fig. 13 presents a comparison of various instances of SA, where different initializations and different values of the initial temperature τ_0 are tried. Results are reported in terms of energy $U(E)$ (10), final number of clusters $K_{\hat{\Omega}}$, and segmentation performance F. To accommodate the various lengths of the meetings, we have normalized all three measures with respect to a reference method (SW-1):

- Normalized energy: for each meeting, $\frac{U(E) - U(E^{(SW-1)})}{N_{\mathrm{terms}}}$, where N_{terms} is the number of terms in the sum in (10);
- Normalized log number of clusters: for each meeting, $\log K_{\hat{\Omega}} - \log K_{\hat{\Omega}}^{(SW-1)}$;
- Normalized F: for each meeting, $F - F^{(SW-1)}$.

Fig. 13(a) shows that the proposed criterion is effectively related to the final segmentation performance: the lower the energy, the higher the performance. All lowest energies lead to very similar performances. It could be concluded that the dynamics (2), in conjunction with the proposed criterion (6), constrain the type of solution that can be obtained. Fig. 13(b) shows that minimizing $U(E)$ is highly correlated with minimizing $K_{\hat{\Omega}}$, which was one of the objectives announced in Section II-C. Fig. 13(c) shows that a high initial temperature τ_0 leads to a result independent from the initialization, which is similar to the well-known

[8]To have a fair comparison between online and offline implementations, in all cases we used the same σ^{diff} and σ^{same} values for each recording, obtained through EM fitting on the whole recording data $X_{1:N}$.

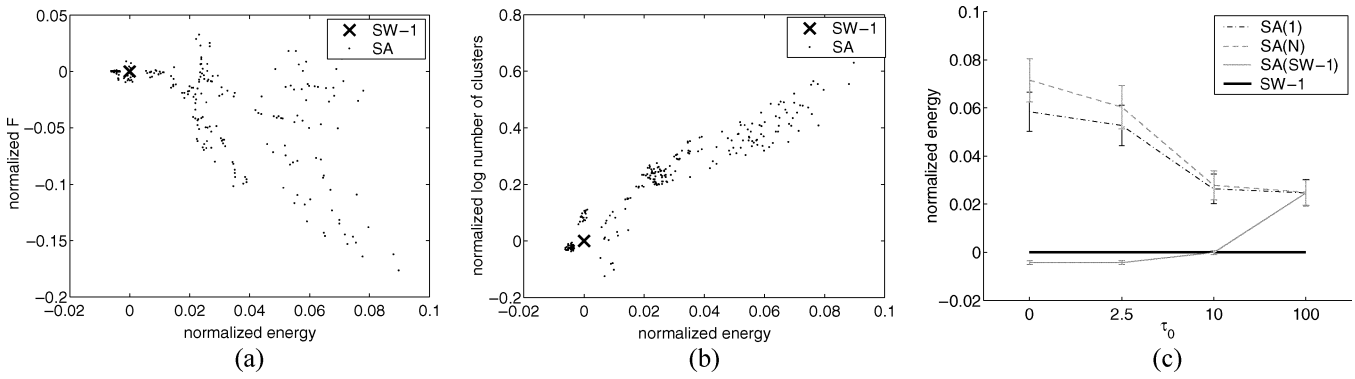

Fig. 13. Comparison between different optimization methods: SW-1, SA(1), SA(N), SA(SW-1) (Sections III and VII-E). In (a) and (b) there is one dot for each triplet (meeting, $\mathrm{SA}(^*)$ method, τ_0 value), that is $18 \times 3 \times 4$ combinations. In (c), each bar represents mean and standard deviation across the 18 meetings. All values are normalized with respect to SW-1 (Section VII-E).

property of simulated annealing when temperature decreases in a logarithmic fashion [52].

The diversity of behaviors observed for a lower initial temperature τ_0 can be explained as follows: when the initial labeling is rather bad [SA(1) and SA(N)], since the local optimization is pointwise and points are visited at random (see Table V), the procedure tends to accept too often the NewLabel tag, which ultimately result in a slightly oversegmented solution. This effect does not appear when using the SA(SW-1) solution as the labels are much more stable because the local posterior probability of the NewLabel tag is very low. Overall, results with the lowest energies are obtained using a somewhat low initial temperature τ_0, and SA(SW-1). SW-1 alone provides close-to-optimal results, in terms of energy. Thus, in the following, results are reported for SW-1 only.

Comparison With Lapels: Table VIII gives the segmentation performance on the test set for SW-1 and the lapel baseline. The proposed approach SW-1 compares well with the lapel baseline, both in terms of F-measure and DER. The proposed approach yields major improvement on overlapped speech. These results are particularly significant, given the high precision of the ground-truth and the fact that we use distant microphones only. Indeed, close-talking lapel signals are about 8-dB cleaner than distant microphone array signals, due to the difference of distance (Table I). The decrease in precision can be explained by the automatic SNS decision leading to more False Positives (D_{FP}) as compared to lapels, because the decision is taken *without* knowledge of the number of speakers. On the contrary, the number of speakers is implicitly known in the lapel baseline.

Comparison With a Previous Speaker Clustering Work: We also compared our approach to a HMM-based previous work [59], on a slightly different task: only six meetings are segmented, and the task excludes silences smaller than 2 s. The results reported in Table IX show a clear improvement. However, the previous work was attacking a wider task: speech segmentation and speaker clustering. This comparison shows that we can obtain a very good speech segmentation with location cues.

Window Size: In Table VIII, the two results SW-1 and SW-7 show that N_{future}, the size of the "future" window, has very little impact on this application. However, this may not be the case in other contexts: for example, the confident clustering approach introduced in Section IV-A *requires* $N_{\mathrm{future}} > 1$.

TABLE IX
COMPARISON WITH A PREVIOUS SPEAKER CLUSTERING WORK: SEGMENTATION RESULTS ON SIX MEETINGS, WITH A SILENCE MINIMUM DURATION OF 2 S. VALUES ARE PERCENTAGES: THE LOWER, THE BETTER

	SW-1	HMM-based
HTER	4.3	17.3

TABLE X
F-MEASURE ON THE M4 CORPUS WITH SW-1, FOR TWO TYPES OF SPEECH/NONSPEECH DECISIONS

SNS decision granularity	Result without post-processing	Result with post-processing
Individual SNS: X_i	48.1	84.6
Cluster SNS: ω_k (SW-1)	83.1	87.2

Interest of STC: As in Section V-C, the same segmentation experiments were also conducted with the speech/nonspeech decision taken for each location estimate individually—without short-term clustering. As reported in Table X, the proposed STC method clearly leads to the best results and is a lot less dependent on segmentation post-processing.. Finally, we noted that the nonstationarity test mentionned in Section V-B effectively removes all short-term clusters belonging to the projector.

Overall, the proposed STC method allows to fulfill two goals of this application: to obtain with distant microphones a segmentation performance comparable to that obtained with close-talking microphones, and to handle multiple simultaneous speakers in an appropriate manner. It can serve as a strong starting point for unsupervised speaker clustering with distant microphones only: [43] reports results superior to that of a state-of-the-art approach.

VIII. CONCLUSION

Accurate segmentation and tracking of speech in a meeting room is crucial for a number of tasks, including speech acquisition and recognition, speaker tracking, and recognition of higher-level events.

In this paper, we first described a generic, threshold-free scheme for short-term clustering of sporadic and concurrent

events. The motivation behind this approach is that with highly sporadic modalities such as speech, it may not be relevant to try to output a single trajectory for each source over the entire data, since it leads to complex data association issues. We proposed here to track in the short-term only, thus avoiding such issues. The core of our approach is a threshold-free probabilistic criterion. We described an algorithm based on a sliding-window analysis, spanning a context of several time frames at once. It is online, can be fully deterministic, and can function in real-time when using reasonable context durations (N_{future}). It is unsupervised: local dynamics are extracted from the data itself, and the short-term clustering is threshold-free. We also presented investigations on the problem of trajectory crossings, useful, e.g., in the context of acoustic vehicle tracking [54] or visual tracking [37]. In this context, experiments on synthetic data highlighted the benefit of processing several time frames at once ($N_{\text{future}} > 1$).

Second, we described speech specific applications of this algorithm. Short-term clustering was used to build a multispeaker detection-localization system with microphone arrays, which was then successfully applied to both dynamic and static recordings with multiple simultaneous speakers. In both cases, short-term clustering permits to discriminate between speech and nonspeech in a much more advantageous manner, as compared to an individual decision for each location estimate. Highly dynamic, nonlinear human motions are well handled by the short-term clustering algorithm. In particular, a comparison with offline simulated annealing optimization shows that the proposed online implementation is sufficient. In the case of multiple moving speakers, short-term clustering followed by deterministic filtering appears clearly superior to an existing multisource particle filtering approach [28], [29].

In terms of final performance, short-term clustering leads to a meeting segmentation performance with distant microphones only, close to that obtained with close-talking microphones. This result can already be considered as a success, since distant microphones are much more noisy than close-talking microphones. Moreover, since multiple speech sources are effectively "tracked in the short-term," a dramatic improvement is seen on overlapped speech, which is often found in spontaneous multiparty speech. These results validate the short-term clustering algorithm, as well as the idea of relying on location cues to obtain high precision short-term tracking and speech segmentation of multiple moving speakers. This, in turn, permits a much wider range of applications than with close-talking microphones, due to the nonintrusive aspect of distant microphones. Investigations on the unsupervised speaker clustering task with distant microphones [43] show that short-term clustering permits to obtain a speaker clustering performance superior to that of a state-of-the-art approach.

ACKNOWLEDGMENT

The authors would like to thank Dr. I. McCowan for supporting this work in its early developments, and Dr. A. Vinciarelli, Dr. F. Valente, and B. Mesot for suggestions. The authors would also like to thank the anonymous reviewers for their suggestions.

REFERENCES

[1] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.

[2] M. Brandstein and D. Ward, Eds., *Microphone Arrays*. New York: Springer, 2001.

[3] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," in *Proc. EURASIP J. Appl. Signal Process., Special Iss. Adv. Multimicrophone Process.*, 2006, pp. 1–19.

[4] S. Wrigley, G. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multi-channel audio," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 84–91, Jan. 2005.

[5] S. Cerwin, "Ears in the sky," *Technol. Today*, pp. 12–13, 2004.

[6] "A. A. for Artificial Intelligence," Smart Rooms, Smart Houses and Households Appliances, Jun. 2006, [Online]: Available: http://www.aaai.org/AITopics/html/rooms.html.

[7] A. Spriet, "Adaptive filtering techniques for noise reduction and acoustic feedback cancellation in hearing aids," Ph.D. dissertation, Faculty Eng., K.U.Leuven, Leuven, Belgium, Sep. 2004.

[8] G. Lathoud, J. Bourgeois, and J. Freudenberger, "Sector-based detection for hands-free speech enhancement in cars," in *EURASIP J. Appl. Signal Process., Special Iss. Adv. Multimicrophone Speech Process.*, 2006, pp. 1–15.

[9] Sony AIBO. Sony Corp., 2006. [Online]. Available: http://www.sony.net/Products/aibo/

[10] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, vol. 27, no. 3, pp. 305–317, Mar. 2005.

[11] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Proc. 2004 MLMI Workshop*, S. Bengio and H. Bourlard, Eds., 2005, pp. 182–195, Springer Verlag.

[12] E. Shriberg, A. Stolcke, and D. Baron, "Can prosody aid the automatic processing of multi-party meetings? Evidence from predicting punctuation and disfluencies, and overlapping speech," in *Proc. ISCA Workshop Prosody*, 2001, pp. 139–146.

[13] R. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME, J. Basic Eng.*, vol. 82, pp. 35–45, Mar. 1960.

[14] H. Sorenson, *Kalman Filtering: Theory and Application*. New York: IEEE Press, 1985.

[15] S. Julier, J. Uhlmann, and H. Durrant-Whyte, "A new approach for filtering nonlinear systems," in *Proc. 1995 Amer. Control Conf.*, 1995, pp. 1628–1632.

[16] S. Julier and J. Uhlmann, "A new extension of the Kalman filter to nonlinear systems," in *Proc. AeroSense: 11th Int. Symp. Aerospace/Defense Sensing, Simulation, Controls*, 1997, pp. 182–193, Multi Sensor Fusion, Tracking and Resource Management II, SPIE.

[17] J. LaViola, "A comparison of unscented and extended Kalman filtering for estimating quaternion motion," in *Proc. Amer. Control Conf.*, Jun. 2003, pp. 2435–2440.

[18] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *IEEE Proc. Radar Signal Process.*, vol. 140, no. 1, pp. 107–113, 1993.

[19] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for on-line non-linear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[20] T. Dvorkind and S. Gannot, "Speaker localization using the Unscented Kalman Filter," in *Proc. HSCMA Workshop*, 2005, pp. c-3–c-4.

[21] M. Wölfel, K. Nickel, and J. McDonough, "Microphone array driven speech recognition: Influence of localization on the word error rate," in *Proc. MLMI Workshop*, 2005, pp. 320–331.

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

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

[24] D. Ward, E. Lehmann, and R. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, Nov. 2003.

[25] E. Lehmann, "Particle filtering methods for acoustic source localisation and tracking," Ph.D. dissertation, Australian National Univ., Canberra, Jul. 2004.

[26] E. Lehmann, "Importance sampling particle filter for robust acoustic source localisation and tracking in reverberant environments," in *Proc. HSCMA*, Piscataway, NJ, Mar. 2005, pp. c-7–c-8.
[27] J. Larocque, J. Reilly, and W. Ng, "Particle filters for tracking an unknown number of sources," *IEEE Trans. Signal Process.*, vol. 50, no. 12, pp. 2926–2937, Dec. 2002.
[28] J. Valin, "Auditory system for a mobile robot," Ph.D. dissertation, Univ. Sherbrooke, Sherbrooke, QC, Canada, Aug. 2005.
[29] J. Valin, F. Michaud, and J. Rouat, "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," in *Proc. ICASSP*, 2006, pp. 841–844.
[30] J. Vermaak, A. Doucet, and P. Pérez, "Maintaining multi-modality through mixture tracking," in *Proc. ICCV*, 2003, vol. 2, pp. 1110–1116.
[31] H. Asoh, F. Asano, T. Yoshimura, Y. Motomura, N. Ichimura, I. Hara, and J. Ogata, "An application of a particle filter to Bayesian multiple sound source tracking with audio and video information fusion," in *Proc. 7th Int. Conf. Inf. Fusion*, 2004, vol. 2, pp. 805–812.
[32] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Proc. ICASSP*, 2004, pp. 881–884.
[33] D. Gatica-Perez, G. Lathoud, J. Odobez, and I. McCowan, "Multimodal multispeaker probabilistic tracking in meetings," in *Proc. Int. Conf. Multimodal Interfaces (ICMI)*, Oct. 2005.
[34] D. Gatica-Perez, G. Lathoud, J. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 601–616, Feb. 2007.
[35] G. Lathoud, I. McCowan, and J. Odobez, "Unsupervised location-based segmentation of multi-party speech," in *Proc. NIST Meeting Recognition Workshop*, Montreal, QC, Canada, May 2004.
[36] D. Ellis and J. Liu, "Speaker turn segmentation based on between-channel differences," in *Proc. NIST Meeting Recognition Workshop*, Montreal, QC, Canada, May 2004.
[37] P. Jorge, J. Marques, and A. Abrantes, "Estimation of the Bayesian network architecture for object tracking in video sequences," in *Proc. Int. Conf. Pattern Recognition*, 2004, vol. 2, pp. 732–735.
[38] G. Lathoud and I. McCowan, "Location based speaker segmentation," in *Proc. ICASSP*, Hong Kong, Apr. 2003, pp. 176–179.
[39] G. Lathoud, I. A. McCowan, and D. C. Moore, "Segmenting multiple concurrent speakers using microphone arrays," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 2889–2892.
[40] S. Chen and P. Gopalkrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," *IBM Tech. J.*, 1998.
[41] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proc. IEEE ASRU Workshop*, 2003, pp. 411–416.
[42] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," in *Proc. NIST RT05s*, 2005, pp. 402–414.
[43] G. Lathoud, "Further applications of sector-based detection and short-term clustering," IDIAP-RR-06 26, 2006.
[44] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. New York: Springer, 2001, ch. 8, pp. 157–180.
[45] E. D. Claudio and R. Parisi, "Multi-source localization strategies," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. New York: Springer, 2001, ch. 9, pp. 181–201.
[46] J. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments," Ph.D. dissertation, Brown Univ., Providence, RI, 2000.
[47] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc., Series B*, vol. 39, pp. 1–38, 1977.
[48] J.-M. Odobez, D. Gatica-Perez, and S. Ba, "Embedding motion in model-based stochastic tracking," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3514–3530, Nov. 2006.
[49] G. Lathoud, M. Magimai.-Doss, and H. Bourlard, "Threshold selection for unsupervised detection, with an application to microphone arrays," in *Proc. ICASSP*, 2006, pp. 285–288.
[50] E. Weisstein, "Bell Number," From Mathworld-A Wolfram Web Resource, Jun. 2006 [Online]. Available: http://mathworld.wolfram.com/BellNumber.html
[51] S. Li, *Markov Random Field Modeling in Computer Vision*. New York: Springer-Verlag, 1995.
[52] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, vol. 6, pp. 721–741, 1984.
[53] P. J. M. van Laarhoven and E. H. L. Aarts, *Simulated Annealing: Theory and Practice*. Dordrecht, The Netherlands: Kluwer, 1987.
[54] T. Pham and M. Fong, "Real-time implementation of MUSIC for wideband acoustic detection and tracking," in *Proc. SPIE AeroSence: Automatic Target Recognition VII*, 1997, pp. 250–256.
[55] M. Moller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Netw.*, vol. 6, pp. 525–533, 1993.
[56] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.
[57] G. Welch and G. Bishop, "An introduction to the Kalman filter," Dept. Comput. Sci., Univ. North Carolina, Chapel Hill, TR 95-041, 2004.
[58] H. E. Rauch, G. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *J. Amer. Inst. Aeronautics Astronautics*, vol. 3, no. 8, pp. 1445–1450, 1965.
[59] J. Ajmera, G. Lathoud, and I. McCowan, "Segmenting and clustering speakers and their locations in meetings," in *Proc. ICASSP*, 2004, pp. 605–608.
[60] S. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*, 2nd ed. San Diego, CA: California Technical Publishing, 1999.
[61] S. Bengio, J. Mariéthoz, and M. Keller, "The expected performance curve," in *Proc. ICML 2005 Workshop ROC Anal. Mach. Learning*, Bonn, Germany, 2005.
[62] National Institute of Standards and Technology, "The rich transcription spring 2003 (RT-03S) evaluation plan," 2003, Tech. Rep.

Guillaume Lathoud (M'06) received the M.S. degree in computer science and telecommunications from the Institut National des Télécommunications (INT), Evry, France, in 1999 and the Ph.D. degree in signal processing from the Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2006.

He was a member of the Digital Television team at the National Institute of Standards and Technology (NIST), Gaithersburg, MD, from 1999 to 2001, participating in terrestrial DTV standardization and implementation efforts, in collaboration with industrial partners. His research interests include microphone array processing, audio source localization, audio-visual speaker tracking, camera calibration, and multimodal processing, as well as noise-robust automatic speech recognition.

Jean-Marc Odobez (M'03) was born in France in 1968. He graduated from the Ecole Nationale Supérieure des Télécommunications de Bretagne (ENSTBr) in 1990, and received the Ph.D. degree in signal processing and telecommunications from Rennes University, Rennes, France, in 1994. He performed his dissertation research at IRISA/INRIA Rennes, France, on dynamic scene analysis (image stabilization, object detection and tracking, image sequence coding) using statistical models (robust estimation, 2-D statistical labeling with Markov Random Field).

He then spent one year as a Postdoctoral Fellow at the GRASP Laboratory, University of Pennsylvania, Philadelphia, working on visually guided navigation problems. From 1996 to 2001, he was Associate Professor at the Université du Maine, Le Mans, France. He his now a Senior Researcher at the IDIAP Research Institute, Martigny, Switzerland, working on the development of statistical methods for multimodal dynamic scene analysis, including media analysis and human tracking and activity analysis. He is the author/coauthor of more than 60 reviewed papers.