Tracking and Voice Separation of Moving Speakers based on IMM-PDA filters

Ilyas Potamitis
Wire Communications Laboratory,
Electrical and Computer Engineering Dept.,
University of Patras, 265 00 Rion, Patras,
Greece
potamitis@wcl.ee.upatras.gr

Nikos Fakotakis
Wire Communications Laboratory,
Electrical and Computer Engineering Dept.,
University of Patras, 265 00 Rion, Patras,
Greece
fakotaki@wcl.ee.upatras.gr

Abstract - *The problem addressed in this work is that of separating the voices of simultaneously active moving speakers using a single microphone array. We adapt the multi-sensor multi-target tracking theory to the context of microphone arrays, in order to form a receptive beam that locks on each moving speaker on an extended time basis and, therefore, achieves voice separation. Our approach (a) incorporates kinematical information of speakers' movement by using an interacting multiple model (IMM) estimator per speaker in order to constrain the evolution of Direction of Arrival (DOA) measurements, and (b) can directly account for measurement origin uncertainty by using the probabilistic data association (PDA) technique in conjunction with the IMM estimator. Finally, we demonstrate the function of the gate as a means to initiate and terminate track segments corresponding to phrases.*

Keywords: Speaker tracking, voice separation, microphone array applications, speech signal processing, IMM filters.

1 Introduction

The practical exploitation of voice based communication technologies is severely hampered if the input includes speech sources other than the one coming from the target speaker. The enhancement of the target signal in the case of concurrent speaker corruption is not trivial since the interference shares the same statistical characteristics with the useful signal. In general, there are two categories of techniques that are employed in order to separate the target speaker from the competing voice. The first exploits the statistical property of the independence of the signal sources, reaching multiple microphones used to record the mixtures of the sound sources in order to recover the underlying sources. The second is based on the spatial selectivity of microphone arrays that are able to form a sharp receptive beam steered on the direction of arrival (DOA) of the desired talker, while turning nulls to the undesired ones.

The first category, usually referred to as 'blind signal separation' (BSS), has Independent Component Analysis (ICA) as its main representative [1-3]. In the context of speech separation, ICA is based on information-theoretic criteria in order to find a linear transform which, when applied to the mixture of the voices, returns statistically independent sources. ICA assumes that each microphone receives a linear mixture of the sources and its application to speech separation of voices in real situations is complicated due to three main reasons: a) The movement of the speaker implies a time-varying mixing matrix, b) reflective indoor environments give rise to reverberation and, therefore, to convolutive mixtures reaching the microphones, and, c) long silence parts in utterances affect the convergence and the separation performance of any ICA separation algorithm. Currently, there are no on-line ICA based algorithms that can deal with all three aforementioned problems simultaneously.

The second category of algorithms that can provide acoustical discrimination between individuals in multispeaker environments is based on microphone arrays. Microphone arrays are electronically steerable, angle-of-arrival filters, designed to constrain their receptive field to a desired direction (i.e., beamforming process) [4-5]. Their ability to provide spatially selective speech acquisition makes them good candidates for a wide field of applications including teleconferencing, multimedia conferencing [6], speech recognition with regard to distant talkers [7] and automatic camera steering [6], [8]. The majority of reported research deals with beamforming a single speaker [3-7]. The multi-speaker case is either treated as a single beam switching to the stronger speaker or limited to specific multi-speaker scenarios. The main difficulty of the multiple speakers' case is that the direct application of DOA-estimation techniques over consecutive frames does not yield tracking of each individual speaker, as beamforming on an extended basis would require. The latter problem becomes evident in the case of moving speakers and, in particular, of moving speakers with a crossing trajectory. Moreover, multimedia, teleconference and 'smart' home applications of the future will have to deal with the complex interaction and speech overlap of participants as experienced in a normal, vivid conversation.

In this work we will present a novel, general framework that can deal with both cases, that is, tracking the direction of arrival and separating the voices of multiple, possibly moving speakers. In order to be able to cover the unpredictable movement of the speaker over time, the proposed state inference scheme, the IMM, handles the uncertainty of the speaker's motion by incorporating multiple motion models in the tracking process. The models restrict the possible evolution of measurements in time [5] and, therefore, improve the DOA estimation accuracy and reduce audio drop out due to misaim of the beamformer. A data association technique is also incorporated into the state inference scheme to efficiently reject clutter measurements and to unambiguously associate the angle observations to speakers. The latter allows each receptive beam to track and lock on an extended basis to the same speaker using a single array and, therefore, to achieve separation of voices and reduction of reverberation (due to the spatial selectivity of the receptive field). The effectiveness of the approach is illustrated by simulation study on tracking two speakers with a crossing angle in their trajectories and three static speakers having a conversation with partially overlapping speech and long pauses.

2 Tracking speakers' angles of arrival using IMM-PDA

An active speaker's trajectory can be subdivided into distinct segments, each corresponding to a different behavioral mode of movement. That is, the speaker may stand still while talking or walk around, make a turn, etc. In such cases, a single motion model can not characterize the speaker's movement throughout the time. The speaker's motion can be modeled by one of a finite number of modes, each of which represents a different speaker movement. We use angle and angle rate as the state of a speaker's motion, and a discrete regime variable which describes the distinct segments of motion. All speakers have the same set of modes (each mode corresponds to a Kalman filter or an extended Kalman filter). Each filter corresponds to a behavior mode that undergoes jumps from model *i* to model *j*, according to a Markov transition matrix. The IMM algorithm merges the state estimates produced by each model at the beginning of each cycle. The problem of IMM state estimation in the context of our work is to infer the kinematical state based on noisy DOA measurements. One cycle of tracking at the IMMs is as follows, [9-11]:

Step 0: An initial estimate of the speaker's angle position is provided by the DOA technique applied and an initial clustering of the values. Each of the speakers' state equation describing their movement and the observation equation is a linear function of the state. An IMM consists of multiple (say *r*) models, and, in our simulation setting *r*=2. We assume that the speakers' follow each model at time *k* with probability μ_j. In this work the Newtonian source motion and observational equations for each mode *j*=1,.., *r* become:

$$\mathbf{s}_j(\mathrm{k})=\mathbf{F}\mathbf{s}_j(\mathrm{k}\text{-}1)+\mathbf{v}_j(\mathrm{k}) \tag{1}$$

$$\mathbf{y}_j(\mathrm{k})=\mathbf{H}\mathbf{s}_j(\mathrm{k})+\mathbf{w}_j(\mathrm{k}) \tag{2}$$

$\mathbf{v}_j(\mathrm{k})\sim\mathrm{N}(\mathbf{0}, \mathbf{Q}_j)$ is the Gaussian zero mean process noise vector having covariance matrix $\mathbf{Q}_j$. $\mathbf{v}_\mathrm{j}(\mathrm{k})$ models angular accelerations experienced by a moving source. $\mathbf{w}_j(\mathrm{k})\sim\mathrm{N}(\mathbf{0}, \mathbf{R})$ is the measurement noise with covariance $\mathbf{R}\sim1/\sin^2\theta$.

Step 1 (Interaction for mode j): To initiate processing of a new DOA measurement, the previous state estimates for each mode are blended to produce mixed state estimates, i.e., the IMM algorithm merges the state estimates produced by each model at the beginning of each cycle. This first step accounts for mode switching between the time of the last track update and the prediction based on the current measurement. Let $\mu_\mathrm{j}(k\text{-}1)$ denote the probability that the speaker was moving according to model *j* at time instant *k-1* and $\mu_{ij}(k\text{-}1)$ represents the model mixing probability given that the speaker is in state *j* and the transition occurred from state *i*. The previous outputs $\hat{\mathbf{s}}_j$ from each mode of movement are combined (interact) through the previous model probabilities $\mu_j(k\text{-}1)$ to produce updated estimates $\hat{\mathbf{s}}_j^*$ that serve as inputs for each mode for the current cycle. That is

$$\hat{\mathbf{s}}_j^*\left(k-1\middle|k-1\right)=\sum_{i=1}^{r}\mu_{ij}\left(k\,\text{-}1\right)\hat{\mathbf{s}}_i\left(k\,\text{-}1\,|\,k\,\text{-}1\right) \tag{3}$$

$$\mathbf{P}_j^*\left(k-1\middle|k-1\right)=\sum_{i=1}^{r}\mu_{ij}\left(k-1\right)\left[\mathbf{P}_i\ \left(k-1\middle|k-1\right)+\mathbf{DP}_{ij}\ \right]$$

where $\mathbf{DP}_{ij}$ models the increased uncertainty due to disagreement between the state estimates models *i* and *j*.

$$\mathbf{DP}_{ij}\left(k-1\right)=\left[\hat{\mathbf{s}}_i\left(k\text{-}1|k\text{-}1\right)-\hat{\mathbf{s}}_j^*\left(k-1\middle|k-1\right)\right]\left[\hat{\mathbf{s}}_i\left(k\text{-}1|k\text{-}1\right)-\hat{\mathbf{s}}_j^*\left(k-1\middle|k-1\right)\right]^T$$

$$\mu_{ij}\left(k-1\right)=\Pi_{ij}\mu_i\left(k-1\right)/\,C_j\left(k-1\right)$$

where $C_j(k-1)=\sum_{i=1}^{r}\Pi_{ij}\mu_i\left(k-1\right)$ denotes the predicted model probability. Π_{ij} is the model transition matrix of the Markov chain which governs the switching between modes and is denoted by the probabilities of transition from mode *i* to mode *j*.

Step 2: (Kalman Update and Prediction for mode j). Subsequently each filter functions as an independent Kalman filter in conjunction with PDA, producing the updated estimation of the state $\hat{\mathbf{s}}_j\left(k\middle|k\right)$ given the current measurement $\mathbf{z}(k)$. The Kalman filtering process is described by:

Predicted state: $\hat{\mathbf{s}}_j\left(k\middle|k-1\right)=\mathbf{F}\hat{\mathbf{s}}_j^*\left(k\,\text{-}1\,|\,k\,\text{-}1\right)$

Covariance prediction: $\mathbf{P}_j\left(k\middle|k-1\right)=\mathbf{F}\mathbf{P}_j^*\left(k-1\middle|k-1\right)\mathbf{F}^T+\mathbf{Q}_j$

Measurement residual: $\mathbf{v}_j\left(k\right)=\mathbf{z}\left(k\right)-\mathbf{H}\hat{\mathbf{s}}_j\left(k\,|\,k\,\text{-}1\right)$

Residual covariance estimates: $\mathbf{S}_j\left(k\right)=\mathbf{H}\mathbf{P}_j\left(k\middle|k-1\right)\mathbf{H}^T+\mathbf{R}$

Filter gain: $\mathbf{W}_j\left(k\right)=\mathbf{P}_j\left(k\middle|k-1\right)\mathbf{H}^T\mathbf{S}_j\left(k\right)^{-1}$

Covariance update: $\mathbf{P}_j\left(k\middle|k\right)=\mathbf{P}_j\left(k\middle|k-1\right)-\mathbf{W}_j\left(k\right)\mathbf{S}_j\left(k\right)\mathbf{W}_j\left(k\right)^T$

However, not all measurements originate from speakers. False measurements originate mainly due to reverberation and low energy segments. A validation region for each mode *j* at time *k* is constructed around the predictions. As a new measurement is received from an array it is validated weather it lies inside a validation region e_j (also called 'gate') with a given probability denoted by P_G which is constructed around the predicted measurement $\mathbf{H}\hat{\mathbf{s}}_j\left(k\,|\,k\,\text{-}1\right)$:

$$e_j\left(k\right)=\exp\left\{-\frac{1}{2}\mathbf{v}_j^T\left(k\right)\mathbf{S}_j^{-1}\left(k\right)\mathbf{v}_j\left(k\right)\right\}\le g_j^2 \tag{4}$$

g_j^2 (known as the number of standard deviations of the gate) is determined based on a chi-square test with a

P_G=99% (or P_G =99.9%) confidence region [9]. A measurement inside the gate is considered to be possibly originated from the true speaker so as to be associated to the previous estimates forming a track. Otherwise it is rejected in order not to affect the estimation procedure. PDA computes the probability β_j of each validated measurement being generated from the speaker, (as well as the probability that no measurement is obtained from the speaker, β_{j0}). The microphone array returns two DOA measurements at time k and each track is updated after inserting each DOA in turn to the IMM. Then, the β_j and β_{j0} become:

$$\beta_j = \frac{e_j}{b+e_j},\ \beta_{j0} = \frac{b}{b+e_j},\ b{=}\lambda\sqrt{\det\left(2\pi\mathbf{S}_j\left(k\right)\right)}\frac{1\text{-}P_D P_G}{P_D}$$

wherte λ is density of the clutter [23], P_D is the detection probability and in simulation we use P_D =0.9, assuming the clutter is uniformly distributed within the gate, and the term b reflects the possibility that the measurement is not speaker originated or fell outside the gate.

The state estimate and covariance for mode j are:

$$\hat{\mathbf{s}}_j\left(k|k\right) = \hat{\mathbf{s}}_j\left(k|k-1\right) + \mathbf{W}_j\left(k\right)\tilde{\mathbf{v}}_j(k) \qquad (5)$$

where $\tilde{\mathbf{v}}_j(k)$ is the valid residual, given by $\tilde{\mathbf{v}}_j(k) = \beta\mathbf{v}_j(k)$ and $\mathbf{v}_j(k)$ is the residual for mode j. The estimated error covariance of mode j so far makes use only the IMM filter. The speaker originating probabilities of the PDA process are incorporated in the estimated error covariance [23]:

$$\mathbf{P}_j\left(k|k\right) = \beta_{j0}\mathbf{P}_j\left(k|k-1\right) + \\ \left(1-\beta_{j0}\right)\left[\mathbf{P}_j\left(k|k-1\right) - \mathbf{W}_j\left(k\right)\mathbf{S}_j\left(k\right)\mathbf{W}_j\left(k\right)^T\right] + \tilde{\mathbf{P}}_j$$

$$\tilde{\mathbf{P}}_j\left(k\right) = \mathbf{W}_j\left(k\right)\left[\beta_j\mathbf{v}_j\left(k\right)\mathbf{v}_j^T\left(k\right) - \tilde{\mathbf{v}}\left(k\right)\tilde{\mathbf{v}}^T\left(k\right)\right]\mathbf{W}_j^T\left(k\right)$$

Step 3 (Probability Update): After each mode is updated with the new measurement, the model probabilities are also updated based on the Markov chain assumption and Bayes formula:

$$\mu_j(k) = \frac{\Lambda_j(k)C_j(k-1)}{C} \text{ where } C = \sum\nolimits_j \Lambda_j(k)C_j(k-1)$$

$$\Lambda_j(k) = V_G^{-1}\left(1\text{-}P_D P_G\right) + \\ P_D\left(2\pi\right)^{-1}\left|\mathbf{S}_j(k)\right|^{-1/2}\exp\left(-\frac{1}{2}\tilde{\mathbf{v}}_j^T(k)\mathbf{S}_j(k)^{-1}\tilde{\mathbf{v}}_j(k)\right)$$

the likelihood that each measurement is from the correct Kalman-PDA filter.

Step 4 (Output mixing of modes): The state estimate and covariance at time k are computed by combining and weighting the estimates of all possible modes:

$$\hat{\mathbf{s}}\left(k|k\right) = \sum_{j=1}^{r}\mu_j\left(k\right)\hat{\mathbf{s}}_j\left(k\,|\,k\right) \qquad (6)$$

$$\mathbf{P}\left(k|k\right) = \sum_{j=1}^{r}\mu_j\left(k\right)\left\{\begin{matrix}\mathbf{P}_j\left(k|k\right) + \\ \left[\hat{\mathbf{s}}\left(k|k\right) - \hat{\mathbf{s}}_j\left(k|k\right)\right]\left[\hat{\mathbf{s}}\left(k|k\right) - \hat{\mathbf{s}}_j\left(k|k\right)\right]^T\end{matrix}\right\}$$

The algorithm iterates through step 1.

The corresponding state vector at time k for this case is $\mathbf{s}$(k)=[θ(k) θ'(k)]^T, and the measured position is $\mathbf{y}$(k)=[θ(k)] and the corresponding matrices in (1), (2) are:

$$\mathrm{F}{=}\begin{bmatrix}1 & \mathrm{T}\\ 0 & 1\end{bmatrix},\ \mathrm{Q_j}{=}\begin{bmatrix}\dfrac{\mathrm{T}^4}{4} & \dfrac{\mathrm{T}^3}{2}\\ \dfrac{\mathrm{T}^3}{2} & \mathrm{T}^2\end{bmatrix}\mathrm{q_j},\ \Pi_{\mathrm{ij}} = \begin{bmatrix}\mathrm{p}_{11} & \mathrm{p}_{12}\\ \mathrm{p}_{21} & \mathrm{p}_{22}\end{bmatrix} = \begin{bmatrix}0.6 & 0.4\\ 0.4 & 0.6\end{bmatrix}$$

and T= block_size*overlap/sampling freq. We used 512 samples FFT at 8 kHz sampling rate with 50% overlap.

The IMM we used had two modes of movement, one for the static-slowly moving case and one for tha fast changes. We have fixed the design parameters so that the non-moving and slowly moving mode possess low-level process noise (q_1=0.001) while the turning mode (manoeuvering model) possesses a much higher noise level (q_2=1).

3 The Gate as a voice activity detector

A speaker can talk and then be silent for a long period making measurement-to-tracks association difficult. To deal with this problem we rely on the fact that in the context of DOA estimation a Kalman-based tracking algorithm incorporates source motion into angle estimates and can be characterized as an angle predictor followed by an observation-dependent angle corrector. The DOAs originating from a certain speaker cannot change randomly since the speaker's movement follows the Newtonian dynamics of motion. Using the previously estimated angle and angle rate the tracker forms an area where the next DOA is possible to lay. The permissible area is a circle centred at the predicted DOA having a radius proportional to the square root of the covariance matrix. Equation (4) that describes the formation of a validation gate around the predicted DOAs serves also as a means for the initiation and termination of the tracks of each individual speaker. If a number of consecutive measurements fall outside the validation gate the track is terminated. Consistent rejection of DOA measurements for a speaker indicates that the particular speaker is silent. Therefore, the role of the gate is twofold:

a) Initiation and termination of DOA tracks (see Fig. 4).

b) Act as voice activity detector (VAD) for each speaker.

The proposed speech separation method based on incorporating multiple motion models in the tracking process and using the prediction to carry out the gating process, can supply independent voice streams, each corresponding to a different speaker. If an utterance contains long pauses or silence parts, the gating process will dissect the stream into disjoint speech segments. In the case of static speakers the disjoint segments can be associated with the same speaker (see Figs. 4). A similar association with moving speakers is not generally possible if we rely solely on the acoustic modality. The latter was anticipated as, for example, in the case of two speakers that would resume talking after switching (silently) their location. The latter switching process would be transparent to the acoustic modality regardless of the tracking or beamforming algorithm. However, in practice what is crucial for communication applications (i.e., speech or speaker recognition) is to have independent voice streams with as little competing voice interference as possible and to leave the speech or speaker recognition engine to deal with the meaningfull interpretation of the voice streams. Current speaker recognition engines could easily tag the distinct segments and associate them to the right speaker provided the speaker is active for 2-3 seconds.

The proposed method relies on the robust initiation of tracks. The initiation of tracks is based on clustering the DOAs derived from the initial speech frames (~ 500 ms).

4 Simulations

The simulation is based on the method presented in [13] to simulate the reverberation in a room for a source holding a given location at a certain time and each microphone location of the array. The impulse response for rectangular rooms is modelled as a set of delayed impulses with gradual drop in their amplitude using the method of images for sound sources in reflecting walls. The four walls have a 0.7 reflection coefficient, while the ceiling and the floor possess 0.9 and 0.4 respectively. We used an array of M=8 omni-directional microphones with d=0.1m spacing between them and their centre located at 1.6m height.

4.1 Performance evaluation for the case of two moving speakers with crossing angles

The layout of the room, a typical 6.8m×4.55m×3m room with 30 dB background noise, is shown in Fig. 1. The topology of the experiment is designed in a way that analytic derivation of the true DOAs of the speakers is possible. In order to have a diversity of the speaker motion we included a speaking-while-standing mode as well as speaking-while-walking mode, abrupt stops, turns as well as circular movements. The first speaker starts from (6.4, 0.25, 1.6)m at t=0 and performs a circular movement with angular velocity 0.1257 rad/sec. He walks for 6.25 seconds and at 45^o stops. At t=8.25 sec he moves on and at 14.5 sec stops moving at 90^o (at broadband position). At t=17.5 sec he continues the circular motion until t=30 sec. At time t=0 speaker two is located at (0.4, 3, 1.6)m and talks for 3 seconds holding this position. Then she walks for 15 seconds heading parallel to the x-axis at a speed of 0.33m/sec. At (5.4, 3, 1.6)m and t=18 sec she makes a right turn and keeps walking for 4.35 sec with the same speed. She stops at (5.4, 1.55, 1.6)m and talks for 3 sec. Then she continues walking for 4.65 sec to (5.4, 0.33, 1.6)m where he arrives at t=30 sec. The speakers are active during the whole scenario. The tracking performance of the algorithm is depicted in Fig. 2.

We used the signal-to-interference ratio as a performance measure to assess the performance of the separation. We performed 10 simulations of the same experimental settings using concatenated speech signals randomly chosen from the TIMIT database and averaged the SIR results. The signals where amplified to reach the range of the input SIR of Table 1 and each set composed of the 10 simulations was re-executed. We followed the notation and performance evaluation metrics of [14] and we denote as $y_k(t)$, $y_k^{(s)}(t)$ k=1,2 the signals of the two speakers just before and after being processed by the separation algorithm. For each speaker the other speaker is considered as interference. We define the output signal-to-interference ratio (SIR_k^{Out}) in the time domain as follows:

$$SIR_k^{Out} = 10\log\frac{\sum_t \left|y_k(t)\right|^2}{\sum_t \left|y_k(t)-y_k^{(s)}(t)\right|^2} \text{ (dB)}, \text{ where } k=1,2 \qquad (7)$$

The SIR of the wavefronts reaching the array is defined as $SIR_{1,2}^{Input}$ when we consider the second source as being the interference:

$$SIR_{1,2}^{Input} = 10\log\frac{\sum_t\sum_{i=1}^{N} \left|y_{i1}(t)\right|^2}{\sum_t\sum_{i=1}^{N} \left|y_{i2}(t)\right|^2} \text{ (dB)}, N\text{=no. of microphones} \qquad (8)$$

and $SIR_{2,1}^{Input}$ if we consider the first source as interference, where the order of $y_{i1}^{(s)}$ and $y_{i2}^{(s)}$ is switched. We use SIR=SIR_k^{Out} - $SIR_{1,2}^{Input}$ as a performance measure of the achieved separation. The results for N=8 are depicted in Table 1.

The proposed voice separation technique does not rely on a specific DOA estimation or beam shaping technique, although, as expected, the tracking performance as well as the achieved separation are greatly assisted by the use of a precise DOA estimation and beamforming algorithm. For the truth of concept the depicted experimental results of Tables 1-2 are derived by using wideband MUSIC and minimum variance beamforming (MVB) (see e.g., [15-16]). In [15] we demonstrated some preliminary results solely on moving speakers and in [16] we provide the implementation code of several DOA estimation and all beamforming methods applied (i.e., delay and sum (DS), minimum variance (MV) and generalized sidelobe canceling (GSC)) [17].

4.2 Performance evaluation for three speakers with partially overlapping speech and long silence parts

We evaluated the proposed speech separation technique on a multi-speaker speaker scenario taking place in 5x3x3 enclosure, where the utterances of the speaker contain large silence parts in between. We performed 10 simulations of the same experimental settings using concatenated digit signals randomly chosen from the NOISEX database and averaged the SIR results. For each voice the rest were considered as interference. Since there are many amplification combinations by mean of which the signals reach the range of the input SIR of Table 2, we used the same amplification coefficient for each interference signal to reach the desired SIR and re-executed each simulation. The room layout is depicted in Fig. 3. The initial DOAs of the speakers are calculated from clustering the angles derived from the first 0.5 secs of speech. If a number of consecutive DOA measurements (three in our case) do not fall inside the predicted DOA gate constructed by the predicted mean and variance of the

IMM filter, the track is terminated. Track termination entails blocking of the microphone array from the direction of the speaker whose track is terminated. Subsequently, the initiation of both track and IMM filter is based on nulling the initial velocity and validating the new angle against the gate.

The SIR results are based on (7)-(8) and are depicted in Table 2. Again we observe large improvement in the separation results, which is due to two factors: a) the directivity of the reception lobe and, b) the inactivation of the microphones from the direction of the silent speaker. The latter is achieved by nulling the voice stream of the inactive speaker as long as its corresponding DOAs do not fall in the predicted gate, therefore, preventing the interfering voices of the other speakers to leak in the the silence segment of the stream of any inactive speaker. The functioning of the gating process as a VAD is illustrated in Fig. 4b and requires a buffer of three frames each of 256 samples at 8 kHz sampling frequency. Sample recordings of the separation results are included in [16].

5 The applicability of multi-target multi-sensor techniques for voice separation

Much work has to be done in order to apply multi-target tracking techniques (mostly developed for radars that use simple narrowband signals) to take into account the idiosyncrasies of the speech signal which is a broadband, non-stationary signal. As there are distinct differences between speakers and moving targets, human motion and conversational attitude need to be taken into account; in polite conversations most speakers do not speak simultaneously, while target signals coexist for most of the tracking time. A speaker can talk and then be silent for a long period making hard the association of distinct track segments. However, the proposed method can derive independent streams corresponding to the different voices. Currently we are experimenting on the incorporation of a speaker recognition module that calculates the probability that a segment belongs to a specific speaker and to incorporate it in (4) to perform track initiation, termination and measurement-to-track association for complicated human interaction scenarios involving multiple speakers.

Although the versatility of human motion and the variability of the environment make quantitative analysis of speaker tracking problematic, extensive experimentation using the IMM filter strongly indicated that it is adequate to track any kind of speaker's motion. The use of more advanced data association techniques more suitable for the multi-speaker scenario as joint probabilistic data association or multiple hypothesis tracking [11] is expected to be beneficial.

Although DOA estimation and voice activity detection were achieved by employing only the acoustic modality, association of segments is expected to be assisted by data fusion in the IMM framework based on integrating observations from a variety of sensors like infrared, vision [18] and laser trackers [19], [20] to handle the measurement origin uncertainty. If these sensors are used to observe the same (unknown) state, the incorporation of their measurements in our proposed IMM framework is straightforward.

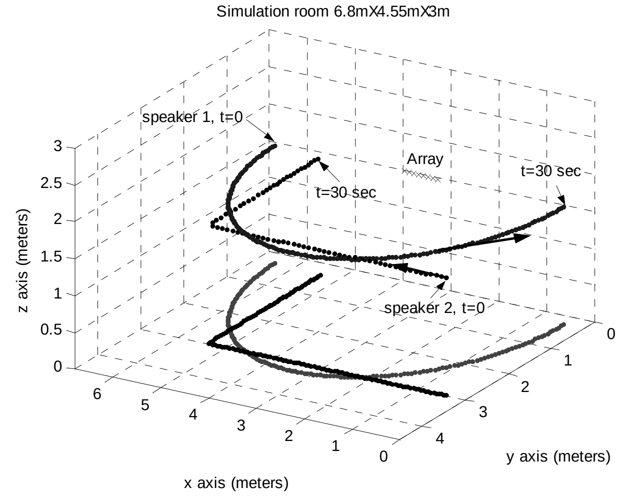

Fig. 1. Speakers' movement in the enclosure.

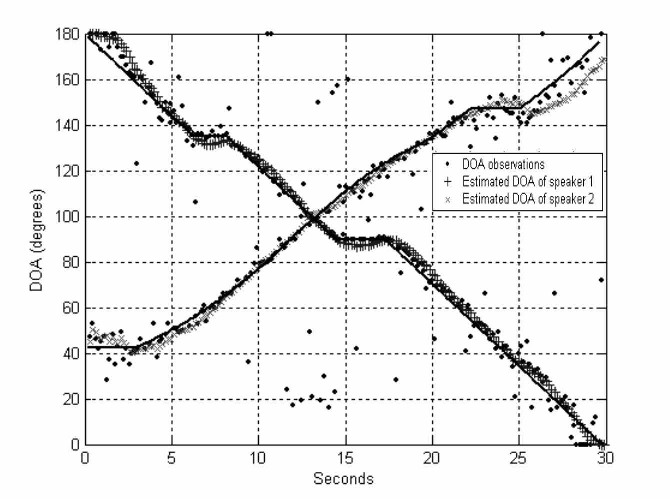

Fig. 2. (.): DOA observations (DOAs extracted using wideband MUSIC at 30 dB background noise)
Solid line: true DOA trajectory
(+): Estimated DOAs associated to speaker 1 (circular motion). Speaker 1 stops moving at t=6.25 sec and moves on at t=8.25 sec. Subsequently he stops at t=14.5 sec and continues at t=17.5 sec.
(x): Estimated DOAs associated to speaker 2 (linear motion and abrupt turn). Speaker 2 is not moving for 3 secs and then she continues her motion. At t=22.35 secs she stops moving. At t=25.35 she continues her motion.

Table 1: Two moving speakers case, Signal to Interference Ratio (SIR) improvement. SIR_In2 is the SIR input in dB considering speaker 1 as target speaker and speaker 2 as interference. SIR_Imp{i} is the improvement in dB on speaker {i} considering speaker the other as interference.

SIR_In2 (dB)	SIR_Imp1 (dB)	SIR_ Imp2 (dB)
-10	17,21	21,14
-5	15,11	19,59
0	12,27	17,42
5	9,17	14,82
10	5,82	11,15

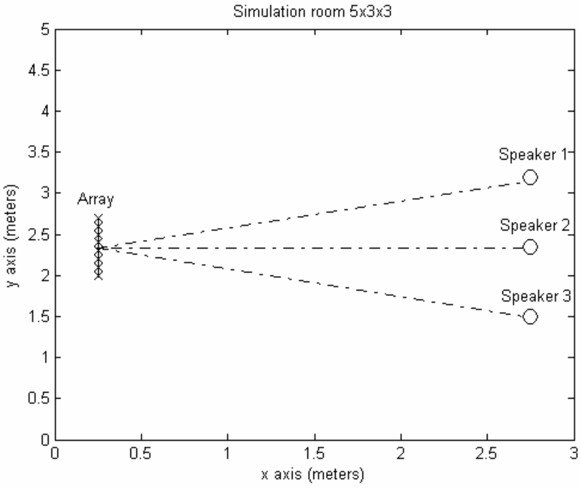

Fig. 3. Speakers' positions in the enclosure.

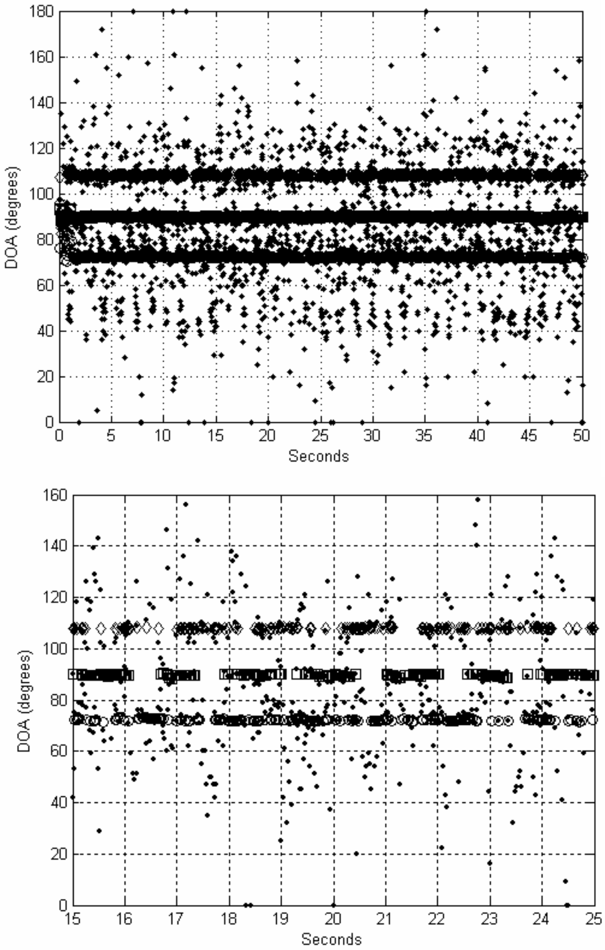

Fig. 4. Three partially overlapping speakers uttering random numbers (static case).

Top: Tracking the DOAs of the full conversation. The three speakers' case is clearly detected and the erroneous DOAs are efficiently rejected.

Bottom: Detail of Fig. 4 - top depicting the distinct DOA clusters corresponding to words.

(.): DOA observations (extracted using wideband MUSIC at 30 dB background noise)

(◊): Estimated DOAs associated to speaker 1.

(□): Estimated DOAs associated to speaker 2.

(○): Estimated DOAs associated to speaker 3.

Table 2: Three static speakers case, Signal to Interference Ratio (SIR) improvement. SIR_In1 is the SIR considering speaker 1 as the target speaker and the rest of the speakers as interference. SIR_Imp (*i*) is the SIR improvement on speaker *i* (the rest are interference).

SIR_In1 (dB)	SIR_ Imp_1	SIR_ Imp_2	SIR_ Imp_3
-10	14.57	28.34	20.65
-5	12.78	27.69	18.45
0	11.31	27.55	17.21
5	8.73	25.82	16.99
10	6.57	23.62	15.66

6 Conclusions

IMM and PDA techniques can be efficiently integrated with DOA estimation techniques to track speakers in reverberant environments, where speakers may change their motion behavior while talking without imposing unrealistic constraints on their motion or velocity. The hybrid state estimation with PDA to account for measurement origin uncertainty has the potential to unify the spatial sound selectivity and allow the receptive beam to follow each moving speaker without misclassifying speech segments on an extended time basis using a single microphone array. We demonstrated that techniques commonly found in the multi-target multi-sensor tracking area (see also [21-23]), provide a consistent and coherent way to reduce uncertainty and ambiguity of angle measurements, and, therefore, reduce audio drop out due to misaim. Moreover, in the static-speakers case we have demonstrated how the gate of the IMM estimator can serve as a means for initiation and termination of tracks, voice activity detection and speech segmentation.

Acknowledgements

We kindly acknowledge the financial support provided by the EC project INSPIRE: INfotainment management with SPeech Interaction via REmote-microphones and telephone interfaces IST-2001-32746. We are also grateful for the inspiring discussions with Dr. Huimin Chen and his help.

References

[1] Lee Te-Won. Independent Component Analysis. *Kluwer Academic Publishers*, 1998.

[2] Roberts S., Everson R., Independent Component Analysis: Principles and Practice. *Cambridge University Press*, 2001.

[3] Asano F., Ikeda S., Ogawa M., Asoh H., Kitawaki N., Combined Approach of Array Processing and Independent Component Analysis for Blind Separation of Acoustic Signals. *IEEE Trans. Speech Audio Processing*, Vol.11, No. 3, pp. 204-215, 2003.

[4] Krim H., Viberg M., Two Decades of Array Signal Proc. Research. *IEEE Signal Processing Magazine*, pp. 67-93, 1996.

[5] Johnson D., Dudgeon D., Array Signal Processing: Concepts and Techniques. *Prentice Hall*, 1993.

[6] Huang Y., Benesty J., Elko G., Mersereau R., Real-time passive source localization: an unbiased linear-correction least-squares approach. *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943-956, 2001.

[7] Yamada T., Nakamura S., Shikano K., Distant-talking speech recognition based on a 3-D Viterbi search using a microphone array. *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 2, pp. 48-56, 2002.

[8] Brandstein M., Silverman H., A practical methodology for speech source localization with microphone Arrays. *Computer Speech and Language*, Vol. 2, pp. 91-126, 1997.

[9] Blom H., Bar-Shalom Y., The Interacting Multiple Model Algorithm for Systems with Markovian Switching Coefficients. *IEEE Trans. on Automatic Control*, Vol. 33, No. 8, pp. 780-783, 1988.

[10] Mazor E., Averbuch A., Bar-Shalom Y., Dayan J., IMM methods in target tracking. *IEEE Trans. on Aerospace and Electronics Systems*, Vol. 34, No.1, pp. 103-123, 1998.

[11] Blackman S., Popoli R., Design and analysis of modern tracking systems. *Artech House*, 1999.

[12] Bar-Shalom Y., Li X., Kirubarajan T., Estimation with application to tracking and navigation. *Wiley*, 2001.

[13] Allen J., Berkley D., Image method for efficiently simulating small-room acoustics. *Journal of the Acoust. Society of America*, Vol. 65, No. 4, pp. 943-950, 1979.

[14] Mukai R., Sawada H., Araki S., Makino S., Robust real-time blind source separation for moving speakers in a room. *IEEE Proc. of ICASSP*, Vol. 5, pp. 469-473, 2003.

[15] Potamitis I., Tremoulis G., Fakotakis N., Multi-Speaker DOA Tracking Using Interactive Multiple Models and Data Association Techniques. *Proc. of EUROSPEECH 2003*, 8th European Conference on Speech Communication and Technology, Vol. I, pp. 517-520, 2003.

[16] http://slt.wcl.ee.upatras.gr/potamitis/IMMPDA_SMC.zip

[17] Griffiths L., Jim C., An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. on Antennas and Propagation*, vol. 30, 1, pp. 24-27, 1982.

[18] Strobel N., Spors S., Rabenstein R., Joint audio-video object localization and tracking. *IEEE Signal Processing Magazine*, vol. 18 (1), pp. 22-31, 2001.

[19] Schulz D., Fox D., Hightower J., People Tracking with Anonymous and ID-Sensors Using Rao-Blackwellised Particle Filters. *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2003.

[20] Fod A., Howard A., Mataric M., Laser-Based People Tracking. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA-02)*, pp. 3024-3029, 2002.

[21] Sturim D., Brandstein M., Silverman H., Tracking Multiple Talkers using Microphone Array Measurements. *IEEE Proc. of ICASSP*, 1997.

[22] Potamitis I., Tremoulis G., Fakotakis N., Multi-Array fusion for beamforming and localization of moving speakers. *Proc. of EUROSPEECH 2003*, 8th European Conference on Speech Communication and Technology, Vol. II, pp. 1721-1724, 2003.

[23] Bar-Shalom Y., Blair W., “Multitarget-multisensor tracking, Applications and advances – Volume III,” *Artech House*, 2000.