MULTI-SPEAKER DOA TRACKING USING INTERACTIVE MULTIPLE MODELS AND PROBABILISTIC DATA ASSOCIATION

Ilyas Potamitis, George Tremoulis, Nikos Fakotakis

Wire Communications Laboratory, Electrical and Computer Engineering Dept.,
University of Patras, 261 10 Rion, Patras, Greece, Tel:+30 2610 991722, Fax:+30 2610 991855
e-mail: potamitis@wcl.ee.upatras.gr

Abstract

The general problem addressed in this work is that of tracking the Direction of Arrival (DOA) of active moving speakers in the presence of background noise and moderate reverberation level in the acoustic field. In order to efficiently beamform each moving speaker on an extended basis we adapt the theory developed in the context of Multi-target Tracking for military and civilian applications to the context of microphone array. Our approach employs Wideband MUSIC and Interacting Multiple Model (IMM) estimators to estimate the DOAs of the speakers under different kinds of motion and sudden change in their course. Probabilistic Data Association (PDA) is used to disambiguate and resolve DOA measurements. The efficiency of the approach is illustrated on simulated and real room experiments dealing with the crossing trajectories of two speakers.

1. Introduction

Microphone arrays are electronically steerable, angle-of-arrival filters, designed to constrain their receptive field to a desired direction. Their ability to provide spatially selective speech acquisition makes them good candidates for hands-free speech applications and a potential replacement of headset or directive microphones. The field of applications is wide and includes videoconferencing [1], multimedia conferencing [2], speech recognition with regard to distant talkers, etc. These kinds of applications are greatly simplified if speech communication is based on hands-free acoustic interfaces that can focus on desired talkers while simultaneously suppressing noise and interfering talkers.

The beamformer output is a filtered estimate of the time-aligned microphone outputs where coherence of signals is achieved by using the DOA of the speakers. The DOAs can be estimated by high resolution spectral estimation techniques like MUSIC and Minimum Variance [3], or by finding the maximum power of a steered response over a range of angles [4] or by first estimating the time delay of arrival between sets of microphones [1, 2]. This work deals with the multi-speaker situation in reverberant enclosures. Since speakers are most of the time in close proximity we employed a high resolution technique for the DOA estimation task.

Beamforming on a moving speaker introduces more complications than those faced in the case of a speaker in a fixed position. High resolution angle estimation techniques are based on the concept of the spatial correlation matrix derived from an ensemble average of the microphone outputs. In order to avoid unrealistic stationarity assumptions one has to allow the derivation of the spatial correlation matrix from few frames. However, reverberation and the lack of long data segments lead to the estimation of unreliable angle fixes. In the case of multiple moving speakers things become even more complicated due to the fact that repeated application of MUSIC over consecutive frames does not yield tracking of targets [2, 4]. This is because DOAs are selected by peak-peeking the spatial spectrum. Thus the succession of peaks from frame to frame is highly dependent on the spectral content of each source, preventing successive DOAs from being associated with particular speakers. In addition to spurious DOAs due to reverberation and DOA ambiguity, angular resolution can degrade depending on the speakers' location, the number of simultaneous sources present in the receptive field and their angular spacing, thus producing wildly inconsistent measurements (referred to as clutter).

This paper aims at incorporating kinematic models in the application of DOA estimation of wideband signals to reduce audio drop out due to misaim. The proposed system is based on an estimation technique known as IMM that consists of multiple Kalman filters functioning in parallel, where each filter is being associated with tracking a certain type of movement. Kalman filters can distinguish between speakers' DOAs and spurious DOAs because the motion model restricts the evolvement of measurements in time [4]. The PDA algorithm uses the prediction of the covariance of the Kalman filters to construct a validation region (gate) around the predicted measurement of each model, to reject clutter measurements and associate DOAs to specific speakers.

2. System Overview

The outline of the proposed DOA tracker and beamforming system is shown in Fig. 1. The system is composed of four distinct stages. The first stage is responsible for applying wideband MUSIC to estimate the multiple DOAs of active speakers. DOA estimates at this stage include many spurious measurements due to reasons mentioned in the introduction (see Fig. 3). Additionally, since MUSIC requires that the number of sources is known it will always return a fixed number of angle estimates per speech frame, regardless of whether all speakers are active or not in a particular frame. In principle, one could employ algorithms that estimate the number of active sources on a per frame basis. However, these algorithms return unreliable estimates in the presence of medium reverberation and their application would greatly complicate the performance of track maintenance. In this work the number of speakers is assumed to be known and the erroneous DOAs associated with a non-active speaker are treated as clutter. The second stage involving the application of Kalman filtering is actually in close connection and inseparable to data association (third stage). The fourth stage examines standard beamforming techniques applied to estimated speakers' tracks to reconstruct the separated voices.

Our research was supported by INSPIRE (IST32746) project.

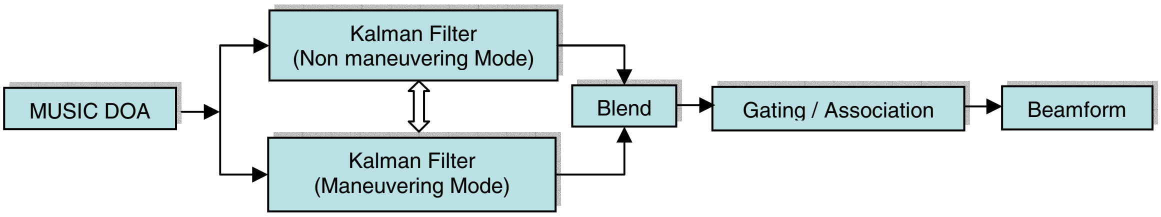

MUSIC DOA
Kalman Filter (Non maneuvering Mode)
Kalman Filter (Maneuvering Mode)
Blend
Gating / Association
Beamform

Fig. 1: Schematic diagram of DOA tracking and beamforming process.

2.1. Wideband MUSIC

In order to adapt narrowband MUSIC to be applicable to speech signals one has to deal with the non-stationary and wideband characteristics of the speech signals. To overcome the nonstationary nature of the speech the microphone outputs are segmented into fixed blocks and signal stationarity is assumed for each block and each FFT bin is considered a narrowband signal. The output from a linear microphone array of M sensors receiving N far-field wideband signals sampled at time instances t=1,..,T and bin f is $\mathbf{x}(f,t)=[X_1(f,t),..,X_M(f,t)]^T \in C^{Mx1}$ where, $X_m(f,t)$ m=1,..,M is the short-time FFT coefficient of the signal received at the m^{th} microphone. The array observation vector is assumed to satisfy:

$$\mathbf{x}(f,t)=\mathbf{A}(f,\theta)\mathbf{s}(f,t)+\mathbf{n}(f,t) \tag{1}$$

where,

$$\mathbf{A}(f,\theta)=[\mathbf{a}_1(f,\theta_1)..,\mathbf{a}_N(f,\theta_N)] \quad \in C^{MxN} \tag{2}$$

$$\mathbf{s}(f,t)=[S_1(f,t),..,S_N(f,t)]^T \quad \in C^{Nx1} \tag{3}$$

$$\mathbf{n}(f,t)=[N_1(f,t),..,N_M(f,t)]^T \quad \in C^{Mx1} \tag{4}$$

the matrix $\mathbf{A}(f,\theta)$ is referred to as the *steering matrix* having as columns the *steering vector of each source n* in frequency f from direction θ_n, and $\mathbf{a}_n(f,\theta_n)=[1, e^{-jkd\cos(\theta n)},..,e^{-jkd(M-1)\cos(\theta n)}]^T$. Let L the window size and also the size of the FFT then $f=f_s l/L$ l=0,..,L-1 where f_s is the sampling frequency, $k=2\pi f/c$, d microphone spacing, c sound velocity ($c \approx 343$ m/sec).

$S_n(f,t)$ is the short-time FFT of the n^{th} source and $N_m(f,t)$ is the short-time FFT of additive noise to the m microphone.

MUSIC is based on forming signal and noise subspaces through the eigendecomposition of the spatial covariance matrix [3]. In practice, the *Spatial Correlation Matrix* $\mathbf{R}(f)$ is estimated from the samples of an ensemble of N_F frames:

$$\mathbf{R}(f)=\frac{1}{N_F}\sum_{t=1}^{N_F}\mathbf{x}(f,t)\mathbf{x}^H(f,t)=\mathbf{A}(f)\mathbf{P}(f)\mathbf{A}^H(f)+\sigma^2\mathbf{I}=\mathbf{U_s}(f)\mathbf{\Lambda_s}\mathbf{U}_s^H(f)+\mathbf{U_n}(f)\mathbf{\Lambda_n}\mathbf{U}_n^H(f)$$

where, $P(f)=E\{\mathbf{s}(f,t)\mathbf{s}^H(f,t)\}$. The eigenvalue decomposition partitions $\mathbf{R}(f)$ into the signal subspace $\mathbf{U}_s(f)$ and noise subspace $\mathbf{U}_n(f)$ with $\mathbf{\Lambda_s}=(\lambda_1,.., \lambda_N)$, $\mathbf{\Lambda_n}=(\lambda_{N+1},.., \lambda_M)$ and λ_i i=1,.., N being the largest N eigenvalues of $\mathbf{R}(f)$. The N columns of $\mathbf{U}_s(f)$ form an orthonormal basis for the estimated signal subspace in each f and the M-N columns of $\mathbf{U}_n(f)$ form the orthogonal counterpart. Since Eq. 2 assumes certain angle fixes and the speakers can be moving we cannot make use of large number of frames in order to derive $\mathbf{R}(f)$ and for our case N_F is limited to five half overlapping frames. Noise is assumed to be additive and uncorrelated between microphones.

Let B the set of frequencies corresponding to the k_b narrowband components with the largest Power values. The k_b narrowband MUSIC spectra are averaged across each $f \in B$ to form the incoherent MUSIC 'spectrum'.

$$P_{MUSIC}(\theta)=\sum_{f\in B}\frac{\mathbf{a}^H(f,\theta)\mathbf{a}(f,\theta)}{\mathbf{a}^H(f,\theta)\mathbf{U}_n(f)\mathbf{U}_n^H(f)\mathbf{a}(f,\theta)} \tag{5}$$

The N angles where the 'spatial spectrum' of Eq. 5 is maximized are the estimated DOAs of the speech sources.

2.2. Interacting Multiple Models estimation

A Kalman-based tracking algorithm in the context of DOA estimation incorporates source motion into angle estimates and can be characterized as an angle predictor followed by an observation-dependent corrector. The source dynamic equation and the observation equation are linear functions of the current state where $\mathbf{s}(k)=[\theta(k)\ \theta'(k)]^T$ is the state vector at time k composed by the angle θ and the angular velocity θ'. IMM approach employs multiple Kalman filters operating in parallel; each corresponding to a behavior mode that undergoes jumps from model i to model j according to a set of transition probabilities. The mode transition probability is governed by a first-order homogeneous Markov chain $p_{ij}=\{m_j(k+1)|m_i(k)\}$, i,j=1,2 the kinematics' behavior models.

In this work the source motion and observational equations become:

$$\mathbf{s}(k)=\mathbf{Fs}(k-1)+\mathbf{w}_i(k)$$

$$\mathbf{y}(k)=\mathbf{H}s(k)+\mathbf{u}_i(k)$$

$$F=\begin{bmatrix}1 & T\\ 0 & 1\end{bmatrix},\ Q=\begin{bmatrix}\frac{T^4}{4} & \frac{T^3}{2}\\ \frac{T^3}{2} & T^2\end{bmatrix}q,\ p_{ij}=\begin{bmatrix}p_{11} & p_{12}\\ p_{21} & p_{22}\end{bmatrix}=\begin{bmatrix}0.6 & 0.4\\ 0.4 & 0.6\end{bmatrix}$$

$\mathbf{w}_i(k)\sim N(\mathbf{0}, \mathbf{Q}_i)$ is the Gaussian zero mean process noise vector having covariance matrix $\mathbf{Q}_i$. $\mathbf{w}_i(k)$ models angular accelerations experienced by the moving source. We have fixed the design parameters so that the non-maneuvering model (i=1) possesses low-level process noise (q_1=0.001) while the maneuvering model possesses a much higher noise level (q_2=100). T denotes the time step between two consecutive DOAs. The observation is $\mathbf{y}(k)=[\theta(k)]$, $\mathbf{H}=[1\ 0]$. $\mathbf{u}_i(k)\sim N(\mathbf{0}, \mathbf{R}_i)$ is the measurement noise having covariance $\mathbf{R}_i$ calculated from $\mathbf{R}_i \sim 1/\sin^2\theta$.

MUSIC provides an initial estimate of DOAs and the initial velocity vector is assumed to be the zero vector [6].

A IMM cycle consists of three steps for each possible model i, i.e. mixing, filtering and a weighted combination of all state estimates [5]. $\hat{s}_i(k|k)$ and $\mathbf{P}_i(k|k)$ are the state estimate and its associated covariance for filter i at time k and $\hat{s}_{0i}(k|k)$, $\mathbf{P}_{0i}(k|k)$ the combined estimates. $\mu_i(k)$, $\mu_{i|i}(k)$, $\Lambda_i(k)$ is the mode probability, mixing probability and likelihood of filter i at time k respectively.

Mixing: Given the previous estimates $\hat{s}_i(k-1|k-1)$ of the state vectors, the error covariance matrices $\mathbf{P}_i(k-1|k-1)$ for each Kalman filter and the probability of the models $\mu_i(k-1)$ at time step k-1 for each model i, at time k one can calculate the combined estimations $\hat{s}_{0j}$, $\mathbf{P}_{0j}$ for each model and p_{ij}.

$$\hat{s}_{0j}(k-1|k-1)=\sum_i\mu_{i|j}(k-1|k-1)\hat{s}_i(k-1|k-1) \tag{6}$$

$$\mu_{i|j}(k-1|k-1)=p_{ij}\mu_i(k-1)/c_j, \qquad c_j=\sum_i p_{ij}\mu_i(k-1) \tag{7}$$

$$\mathbf{P}_{0j}(k-1|k-1)=\sum_i\mu_{i|j}(k-1|k-1)[\mathbf{P}_i(k-1|k-1)+[\hat{\mathbf{s}}_i(k-1|k-1)-\hat{\mathbf{s}}_{0j}(k-1|k-1)][\hat{\mathbf{s}}_i(k-1|k-1)-\hat{\mathbf{s}}_{0j}(k-1|k-1)]^T] \tag{8}$$

Filtering: Calculation of the state estimates and covariances conditioned on a mode being in effect performed in parallel for each mode.

$$\hat{\mathbf{s}}_j(k|k\text{-}1)=\mathbf{F}\hat{\mathbf{s}}_{0j}(k\text{-}1|k\text{-}1) \tag{9}$$

$$\mathbf{P}_j(k|k\text{-}1)=\mathbf{F}\mathbf{P}_{0j}(k\text{-}1|k\text{-}1)\mathbf{F}^T+\mathbf{Q}_j(k) \tag{10}$$

$$\hat{\mathbf{y}}_j(k|k\text{-}1)=\mathbf{H}\hat{\mathbf{s}}_j(k|k\text{-}1) \text{ (predicted easurements)} \tag{11}$$

$$\mathbf{r}_j(k)=\mathbf{y}(k)\text{-}\hat{\mathbf{y}}_j(k|k\text{-}1) \text{ (innovation)} \tag{12}$$

$$\mathbf{S}_j(k)=\mathbf{H}\mathbf{P}_j(k|k\text{-}1)\mathbf{H}^T+\mathbf{R}_j(k) \tag{13}$$

$$\mathbf{W}_j(k)=\mathbf{P}_j(k|k\text{-}1)\mathbf{H}_j(k)^T\mathbf{S}(k)^{-1} \tag{14}$$

$$\mathbf{\Lambda}_j(k)=\mathbf{N}(\mathbf{r}_j(k);0,\mathbf{S}_j(k)) \tag{15}$$

$$\mu_j(k)=1/c\mathbf{\Lambda}_j(k)\textstyle\sum_i p_{ij}\mu_i(k\text{-}1)=\mathbf{\Lambda}_j(k)c_j/c \tag{16}$$

$$c=\textstyle\sum_j \mathbf{\Lambda}_j(k)c_j \tag{17}$$

$$\hat{\mathbf{s}}_j(k|k)=\hat{\mathbf{s}}_i(k|k)+\mathbf{W}_j(k)\mathbf{r}_j(k) \tag{18}$$

$$\mathbf{P}_j(k|k)=\mathbf{P}_j(k|k\text{-}1)\text{-}\mathbf{W}_j(k)\mathbf{S}(k)\mathbf{W}_j(k)^T \tag{19}$$

Combination: A weighted sum of the updated state estimates of all filters yields the output state and covariance estimates.

$$\hat{\mathbf{s}}(k|k)=\textstyle\sum_j \mu_j(k)\hat{\mathbf{s}}_j(k|k) \tag{20}$$

$$\mathbf{P}(k|k)=\textstyle\sum_j \mu_j(k)[\mathbf{P}_j(k|k)+[\hat{\mathbf{s}}_j(k|k)\text{-}\hat{\mathbf{s}}(k|k)][\hat{\mathbf{s}}_j(k|k)\text{-}\hat{\mathbf{s}}(k|k)]^T] \tag{21}$$

2.3. Probabilistic Data Association Estimation

The PDA algorithm uses model consistency tests to evaluate measurements for assignment to previous tracks. A validation region for each mode *i* at time k is constructed around the prediction of Eq. 12. A measurement y_k is accepted only when it is inside the acceptance region with probability P_G fulfilling:

$$(\mathbf{y}(k)\text{-}\hat{\mathbf{y}}_j(k|k\text{-}1))(\mathbf{S}_j(k))^{-1}(\mathbf{y}(k)\text{-}\hat{\mathbf{y}}_j(k|k\text{-}1))\leq g_j^2 \tag{22}$$

where $\mathbf{S}_j(k)$ is given from (13) and g_j^2 (known as the number of standard deviations of the gate) is determined by P_G (P_G>0.99) and the dimension of the state from a chi-squared table (see [5] for a thorough review of IMM-PDA). Resolving of DOAs is achieved by assigning the MUSIC DOA measurement (if it is accepted by Eq. 22) to the validation gate that contains it.

3. Evaluation

3.1. Simulation Experiments

The simulation is based on the method of images [7] and takes place in a typical 6.8m×4.55m×3m room with 30 dB background noise. We made use of an array of *M*=8 omni-directional microphones with *d*=0.1m spacing between microphones. The topology of the experiment is designed in a way that analytic derivation of the DOAs is possible and includes various types of movement. The movement of the speakers is presented in Fig. 2. During their movement on opposite directions both speaker are always active uttering random TIMIT utterances. The array is located at 1.6m height. At time t=0 speaker one is located at (0.4, 3, 1.6)m (its mouth that is) and talks for 3 seconds holding this position. Then he walks for 15 seconds heading parallel to the y-axis at a speed of 0.33m/sec. At (5.4, 3, 1.6)m and t=18 sec makes a right turn and keeps walking for 4.35 sec with the same speed. He stops at (5.4, 1.55, 1.6)m and talks for 3 sec. Then he continues walking for 4.65 sec to (5.4, 0.33, 1.6)m where he arrives at t=30 sec. The second 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 moves on and at 14.5 sec stops moving at 90^o with regards to the endfire of the array. At t=17.5 sec continues the circular motion until t=30 sec. For each speaker location the true DOAs were calculated and the absolute error of their values is demonstrated in Fig. 4. MUSIC estimates are derived from a block of five half-overlapping 64 ms hamming-windowed frames (T=160 msec) using a 512 points FFT. Fig. 4 shows that the larger angle errors from 100 Monte Carlo runs occur for DOAs near the endfire line. However, the Kalman filters effectively deal with these measurements by assigning a large variance to predictions (unreliable DOA).

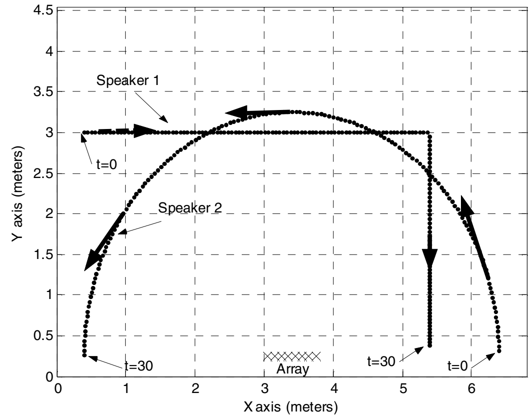


Fig. 2: DOA tracking of 2 speakers.

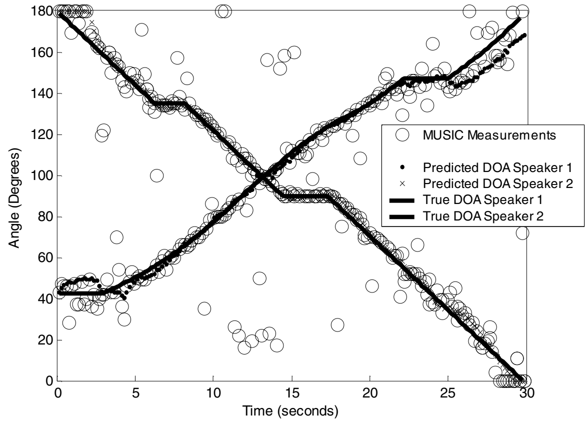


Fig. 3: MUSIC, True and Predicted DOA measurements.

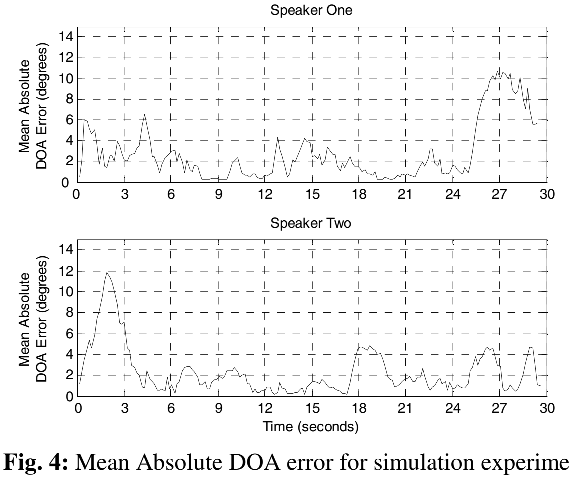
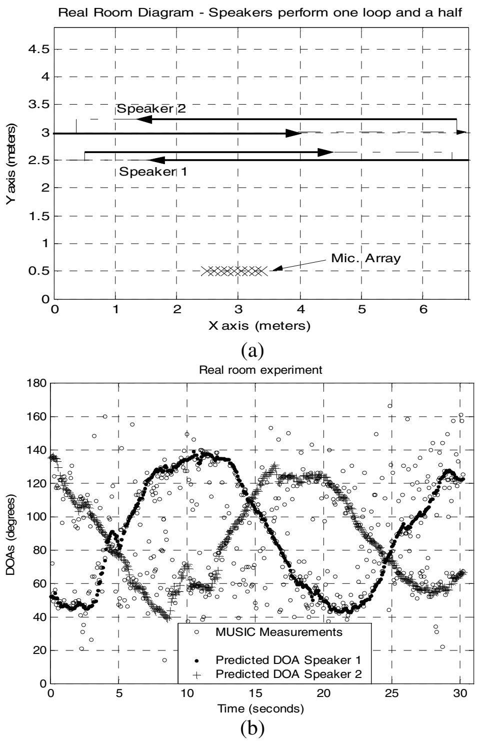


Fig. 4: Mean Absolute DOA error for simulation experiment.

3.2. Real Room Experiments

The experiment takes place in a 6.75mx4.9mx3m room with reverberation time of 0.3 sec. Two persons are walking according to the scenario demonstrated in Fig. 5. Our microphone array consists of an 8 omni electret condenser



(a)

(b)

Fig. 5: a) Diagram of the real room, b) DOA measurements.

Panasonic capsules WM54BT Ø 9,7 mm and an Aardvark Q10 Pro soundcard with internal pre-amps, 8-channel A/D converter for the signal acquisition and 0.05m spacing between microphones. The signals are sampled at 44.1 kHz with 16 bits per sample and downsampled to 8 kHz. The back of the array is covered with low-reflection, high-absorption acoustic foam. A band bass filter [300, 3400] is applied to reduce low frequency noise from computer fans.

4. Implementation Details and Discussion

We have employed three different beamforming techniques in order to evaluate with our system, namely Delay-Sum (DSB), the Minimum Variance (MVB) and the Generalized Sidelobe Canceller (GSC) beamformer. For a concise introduction to these techniques the reader is referred to [3, 4, 8]. Recordings of simulations and real room experiments as well as implementation code can be freely downloaded from [9].

The GSC beamformer uses adaptive noise cancellation (RLS filters in frequency domain). Consideration is given in designing the experiments to have crossing targets in order to study the problem of source identification after re-separation.

To summarize the results of the simulations and real experiments, the combination of MUSIC and IMM filters returns reliable DOA estimation even for crossing speakers and speakers at close proximity. The best results in terms of quality of perception are attained for the MVB beamformer and especially when speakers are well separated. However a gradual drop in its separation performance is demonstrated as speakers approach each other. This is due to the inability of the beam size to provide attenuation when both speakers are inside the beam. A possible solution to this problem is the use of multiple arrays to track speakers from different perspectives so that at least for one array the reception field can be more focused to each independent speaker. The GSC technique does not surpass the performance of MVB. However GSC proves effective especially while a speaker is not moving. This is due to the adequate convergence of the RLS filters when they are given enough time to converge. The worst performance is observed for DS whose broad receptive beam restrains the system from achieving good separation.

5. Conclusions

IMM and PDA techniques can be efficiently integrated with super-resolution spectral estimation techniques to track speakers that may change their motion behaviour while talking. We successfully applied a combination of this approach and beamforming techniques to the problem of speech separation of moving speakers in reverberant environments. The movement is *not* restricted to slow motion. Much work has to be done in order to adapt and transfer Multi-Target tracking theory (mostly developed for radar sensors) to take into account the idiosyncracies of the speech signal, of human motion and conversational attitude. To name a few distinct differences between speakers and moving targets; 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 measurement-to-tracks association difficult. At least as regards in-home situations people either sit while talking or have small speed and accelerations that do not match well the concepts of maneuver and coordinated turn of an aircraft. Further investigations are looking at incorporating amplitude information as well as information from speaker recognition systems into the motion models, in order to initiate and terminate DOA tracks. Multi-array tracking is also a possibility where DOA information from a network of arrays could be effectively integrated [10]. Last but not least one could also consider the idea of using active acoustical arrays that would emit a chirp signal at ultrasound frequencies in the manner of bats to estimate the number of potential speakers in an enclosure and track their range, azimuth and heading.

6. References

[1] 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.

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

[3] Krim H., Vibeg M., "Two Decades of Array Signal Proc. Research", *IEEE Signal Proces. Mag.,* pp. 67-93, 1996.

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

[5] 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.

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

[7] 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.

[8] 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.

[9] http://slt.wcl.ee.upatras.gr/potamitis/IMMPDA_DOA.zip

[10] Chen H., et al., "Multiple Target Tracking with Multiple Finite Resolution Sensors", *5th International Conference on Information Fusion*, 2002.