PERSON TRACKING IN ENHANCED COGNITIVE CARE: A PARTICLE FILTERING APPROACH

Aristodemos Pnevmatikakis, Fotios Talantzis

Autonomic and Grid Computing, Athens Information Technology
PO Box 68, 19.5 km, Markopoulo Avenue, Peania 19002, Athens, Greece
email: {apne, fota}@ait.edu.gr

ABSTRACT

We propose an audiovisual system for detecting the active speaker in cluttered and reverberant environments where more than one person speaks and moves. The feasibility of the systems is examined in the context of smart-houses supporting elderly living alone. Rather than using only audio, the system utilizes audiovisual information from a minimal setup comprising of three microphones and one camera feeding separate audio and visual tracking modules. The audio module operates using a Particle Filter (PF) in order to provide accurate acoustic source location under reverberant conditions. The visual module combines with a second PF video cues generated by colour and face measurements. The final decision is performed through the employment of a fusion mechanism that combines the estimate of each modality according to the current observations. Results indicate that the performance of the proposed multi-modal tracking can potentially enable services for the elderly in their domestic environment if those require knowledge of speaker.

1. INTRODUCTION

Due to the rising longevity phenomenon, we are witnessing a growing interest for pervasive context-aware services which target elderly users. Ambient assisted living solutions for the elderly target a variety of assistive functionalities such as social integration and decentralized communication support [5], as well as e-health and e-care (e.g., facilitating caretakers and minimizing the need for hospitalization) [9]. In all of these scenarios signal processing is of great importance since it provides means to implement various perceptual components that are used to provide services.

Central requirement to the above scenarios is the problem of detecting the location of the active speaker in an environment with many people. This can facilitate creation of location dependent services like targeted audio, emergency detection, and pre-filtering for speech recognition (e.g. beamforming). Because of the potentially large number of people moving and speaking in such cluttered environments the problem remains challenging. Additionally, employing such systems in actual domestic environments typically involves installation of expensive and sizable infrastructure.

Typical solutions to the problem employ multiple microphones in the enclosure and the use of an Acoustic Source Localization and Tracking (ASLT) system. Time Delay Estimation (TDE) methods, like the Generalized Cross Correlation (GCC) [7] remain the most popular variants for feeding the systems that locate the active speaker. The ASLT system then combines such estimates to return the actual location. Three dimensional (3D) visual person tracking [8] from multiple cameras is considered to be more accurate than audio based localization but evidently fails to detect by itself the active speaker. Nevertheless, there have been efforts to fuse the two modalities in the general scenario of person tracking where each modality deals with the weaknesses of the other one. Most of these approaches require a large number of sensors that are difficult to install and generate large amounts of data with corresponding processing power requirements.

In this paper we present a real-time active speaker detection system that utilizes both audio and video cues. In addition we assume use of only one camera and three microphones in order to make the system easier to employ in a real environment like the house of an elderly. In this context, we first propose an ASLT system that uses a state-space approach based on particle filters (PF) to recursively estimate the probability density of the active speaker location. The PF assumes that the source moves according to a specific model that has a specific consistency from a time frame to the next one. The functionality of this new ASLT system in detecting the active speaker is extended by the use of a visual tracking system that employs face and color measurements in a partitioned sampling PF [8]. The fusion of the audio and visual tracks determines whether speech is present and the location of the speaker.

The paper is organized as follows. In Section 2 we present the audio module, followed by the visual one in Section 3. Section 4 discusses how the combination of the video and audio cues detects the presence of speech and the active speaker. In Section 5 the performance of the system is derived, showing ample improvement of the audiovisual system over the audio one. Finally, conclusions are drawn in Section 6.

2. AUDIO TRACKING MODULE

An ASLT system considers M microphones in a multi-path environment. The sound source that the system attempts to locate and track is assumed to be in the far field of the microphones. Assuming a single source, the discrete signal recorded at the m^{th} microphone ($m = 1, 2, \ldots, M$) at time k is:

$$r_m(k) = h_m(k) * s(k) + n_m(k), \tag{1}$$

where $s(k)$ is the source signal, $h_m(k)$ is the room impulse response between the source and m^{th} microphone, $n_m(k)$ is additive noise, and $*$ denotes convolution. The length of $h_m(k)$, and thus the number of reflections, is a function of the reverberation time T_{60} (defined as the time in seconds for the reverberation level to decay to 60 dB below the initial level) of the room and expresses one of the main problems when attempting to track an acoustic source. This is because when the system is used in reverberant environments the source location estimate could occur in a spurious location created by the ensuing reflections.

Given that ASLT systems typically operate in real-time, we assume that data at each sensor m are collected over t frames $\mathbf{r}_m^{(t)} = [r_m(tL), r_m(tL+1), \ldots, r_m(tL+L-1)]$ of L samples. At frame t the representation of the microphone data is as follows:

$$\mathbf{y}_{1:t} = \begin{bmatrix} \mathbf{r}_1^{(1)}\ \mathbf{r}_1^{(2)} \ldots \mathbf{r}_1^{(t)} \\ \mathbf{r}_2^{(1)}\ \mathbf{r}_2^{(2)} \ldots \mathbf{r}_2^{(t)} \\ \vdots \\ \mathbf{r}_m^{(1)}\ \mathbf{r}_m^{(2)} \ldots \mathbf{r}_m^{(t)} \end{bmatrix} \tag{2}$$

Most localization systems ignore the concatenation of frames as seen in Eq. (2) and attempt to estimate the source location using data from the current frame only i.e. using a single column from

This work has been partly sponsored by the FP7 project, HERMES.

Eq. (2). Most ASLT systems are based on TDE. In this case the microphones are arranged in P pairs. Since the microphones of each pair p reside in different spatial locations, their corresponding recordings will be delayed with respect to each other by a relative time delay τ_p. TDE methods estimate the time difference between the two microphones of each pair p. Using all estimated τ_p the localizer can then provide an estimate of the current source location. Traditional systems typically do this by converting τ_p to a line along which the estimated source position is. The problem of localization then reduces to finding the location which minimizes the distance to each intersection points of the bearing lines [4].

In the context of the present work we use an alternative approach as described in [12]. These approaches use PFs to allow us to integrate the properties of human motion as well as the tracking history provided by Eq. (2). The following paragraphs describe the sub-systems of the audio tracker. These include the general PF framework for localization, an extension to tackle competing and interchanging speakers as well as the TDE function.

2.1 State-Space Estimation Using Particle Filters

Assuming a first order model for the acoustic-source dynamics the source state at any frame t is given as:

$$\mathbf{x}_t = [x_t, y_t, z_t, \dot{x}_t, \dot{y}_t, \dot{z}_t]^T \tag{3}$$

where $\mathbf{s}_t = [x_t, y_t, z_t]$ is the current source location estimate and $[\dot{x}_t, \dot{y}_t, \dot{z}_t]$ the corresponding source velocity. If we calculate the conditional probability density $p(\mathbf{x}_t|\mathbf{y}_{1:t})$, we could then find the source location by choosing the state that is more likely given the sensor data until frame t. We can perform this by using the following relationship [8]:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \tag{4}$$

where for clarity we have assumed $\mathbf{y}_t \equiv \mathbf{y}_{t:t}$. $p(\mathbf{y}_t|\mathbf{x}_t)$ is called the likelihood function and expresses the means with which we value the states. $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$ is known as the prediction density and it is given as [8]:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})d\mathbf{x}_{t-1} \tag{5}$$

where $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is the state transition density, and $p(\mathbf{x}_{t-1}|\mathbf{r}_{1:t-1})$ is the prior filtering density. The solution to (4) and (5) can be found using a Monte-Carlo simulation of a set of particles with associated discrete probability masses that estimate the source state [8]. For this we require a model of how the source propagates from $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$. To keep consistent with the literature we will use the Langevin model [10]. For the x_t-coordinate this is defined as:

$$\dot{x}_t = \alpha_x \dot{x}_{t-1} + \beta_x G_x \tag{6}$$

$$x_t = x_{t-1} + \Delta T \dot{x}_t \tag{7}$$

$$\eta_x = e^{-\beta_x \Delta T} \tag{8}$$

$$\beta_x = \upsilon_x \sqrt{1-\eta_x^2} \tag{9}$$

where G_x is a normally distributed random variable, $\Delta T = L/f_s$ is the time separating two location estimates and f_s is the sampling frequency. Also, υ_x refers to the steady-state velocity. Corresponding equations apply for y_t and z_t.

We also need to decide on the likelihood functions that will operate on the microphone data.

There are occasions where reverberation or noise sources can lead the particles to get trapped in a spurious location. Given this, we use the concept of an *external* particle filter $\boldsymbol{e}_t$ that has the same architecture as the main one $\mathbf{x}_t$ but it is initialized repeatedly at every frame t. The particles of $\boldsymbol{e}_t$ are distributed randomly across the entire room. Thus, if these new particles estimate a source location that is d_e m away from the main particle filter for a significant amount of time T_e s then we can reset the locations of the particles of the main PF to those of the external. This also proves useful for scenarios where competing speakers are placed far-apart. All three possible combinations of the 3 microphones are used for TDE.

The general structure of the proposed PF framework can be itemized as follows:

1. Start with a set of particles $\mathbf{x}_0^{(\iota)}, \iota = 1 \dots N$ with uniform weights $w_0^{(\iota)}, \iota = 1 \dots N$. For every new frame of data perform steps 2-8.
2. Resample the particles from state $\mathbf{x}_{t-1}^{(\iota)}$ using some resampling method (we used the *residual resampling* algorithm) and form the resampled set of particles $\widetilde{\mathbf{x}}_{t-1}^{(\iota)}, \iota = 1 \dots N$.
3. Using the Langevin model, propagate $\widetilde{\mathbf{x}}_{t-1}^{(\iota)}$ to predict the current set of particles $\mathbf{x}_t^{(\iota)}$.
4. Take a set of frames of L samples from each microphone i.e. $\mathbf{r}_m^{(t)}, m = 1, 2, \dots M$ and convert them into the frequency domain using an L-point Short Time Fourier Transfor (STFT) to get $\mathbf{X}_m^{(t)} = [X_m(\omega_0), X_m(\omega_1), \dots, X_m(\omega_{L-1})], m = 1, 2, \dots M$. ω_l denotes the l^{th} discrete frequency bin with $l = 0, 1, \dots L-1$.
5. Using a localization function convert the set of $\mathbf{X}_m^{(t)}$ into a localization measurement i.e. a TDE measurement.
6. Weight the particles using the likelihood function i.e. $w_t^{(\iota)} = p(\mathbf{y}_t|\mathbf{x}_t^{(\iota)}), \iota = 1 \dots N$ and normalize the weights so that they add up to unity.
7. The source location for the current frame $\mathbf{s}_t$ is then given as the weighted average of the particles: $\mathbf{s}_t = \sum_{\iota=1}^{N} w_t^{(\iota)} \mathbf{l}_t^{(\iota)}$. In the last expression, $\mathbf{l}_t^{(\iota)}$ denotes the location vector of the ι^{th} particle.
8. If the external PF $\boldsymbol{e}_t^{(\iota)}$ returns a source estimate that remains at a distance greater than d_e m from the estimate of $\mathbf{x}_t^{(\iota)}$ for more than T_e sec then set $\mathbf{x}_t^{(\iota)} = \boldsymbol{e}_t^{(\iota)}, \iota = 1 \dots N$.

2.2 Time Delay Estimation

In this case microphones are organised in P pairs. Consider two microphones i, q belonging to the same pair p. Since the microphones reside in different spatial locations, their corresponding recordings will be delayed with respect to each other by a relative time delay τ_p. A variety of methods like the GCC [7] method (or one of its variants) exist for TDE. For any pair p the GCC-Phase Transform (GCC-PHAT) variant $R_t(\tau)$ is defined as the cross correlation of $\mathbf{r}_i^{(t)}$ and $\mathbf{r}_q^{(t)}$, filtered by a weighting function for a range of delays τ. In the frequency domain this is given as:

$$R_t(\tau) = \frac{1}{2\pi} \sum_{\omega_l} G(\omega_l) X_i(\omega_l) X_q^*(\omega_l) e^{j\omega_l \tau} \tag{10}$$

with $G(\omega_l) = (|X_i(\omega_l)X_q^*(\omega_l)|)^{-1}$. Ideally, $R_t(\tau)$ exhibits a global maximum at the lag value which corresponds to the correct τ.

The GCC-PHAT algorithm is able to return accurate estimates of the relative delay when the environment is anechoic. However, it has a major drawback when used in an environment described by (1). In that case, reflections result in decreased system robustness since the peak provided by $R_t(\tau)$ may not always be the global maximum. This is often the case when T_{60} is not relatively low.

At every frame t and after the microphones are organised in pairs, $R_t(\tau)$ is evaluated only at a set of candidate delays defined by the location of every particle ι. For two microphones i, j belonging in the same pair p the delay is given as:

$$\tau_p(\mathbf{x}_t^{(\iota)}) = \frac{\left\|\mathbf{l}_t^{(\iota)} - \mathbf{m}_i\right\| - \left\|\mathbf{l}_t^{(\iota)} - \mathbf{m}_q\right\|}{c} \tag{11}$$

where $\mathbf{m}_m$ denotes the location of the m^{th} microphone at the p^{th} pair, c the speed of sound (typically defined as $343m/s$). The $\|.\|$

operator denotes the Euclidean distance. The likelihood function for particle ι when TDE is used is given as:

$$p(\mathbf{y}_t|\mathbf{x}_t^{(\iota)}) = \prod_{p=1}^{P} R_t(\tau_p(\mathbf{x}_t^{(\iota)})) \tag{12}$$

3. VISUAL TRACKER

The face bounding boxes are tracked on the image plane by means of a Particle Filter (PF) tracker that employs two measurement cues:

- Face detection measurement: Face detection [13,14] offers very precise localization of the face bounding box, but it is not always present. Adverse poses, illumination and expressions can cause a face to be missed by the detector, or may lead to ill-framing it on the camera plane.
- Color model matching: Matching the colors of the target model against the colors of the pixels comprising the target will always yield a match, especially if the small variations in the target caused by illumination and its pose are learnt into the model. On the other hand color matching is not guaranteed to offer precise localization in cases where the background has similar colors to the target, especially since color modeling abstracts away all structure of the colors in the target.

Hence the visual tracker employed in this system benefits both from the precise localization of the face detection measurement and the ubiquitous presence of color model matching by combining both cues using the partitioned sampling approach [8]. To do so, a measurement model $p(\mathbf{y}|\mathbf{x})$ describing the likelihood of a measurement $\mathbf{y}$ given the state $\mathbf{x}$ is derived for both cues.

3.1 Face likelihood

The face detector employed comprises a boosted cascade of simple classifiers [11], each classifier comprising stages of Haar-like features whose number increases for classifiers down the cascade. The features are selected using Adaboost during the training stage. The implementation of the detector found in OpenCV [3] is used, and the cascade is trained using 9,000 positive and 18,000 negative samples, minimum feature size 0, 99.9% hit rate and 50% false alarm per cascade stage, horizontal and 45-degrees tilted Haar-like features, non-symmetric faces, four splits and gentle AdaBoost learning. The positive samples are selected from various face databases, all cropped slightly above the eyebrows to offer insensitivity to hairstyles. Illumination insensitivity is increased both by using face samples with illumination changes and by linearly equalizing illumination in every candidate region, before applying the detector.

When the Viola-Jones frontal face detector is applied on an image, a multitude of candidate face bounding boxes are returned. Most of them are in groups, with minor variations of their location and scale, bounding actual faces. Some other can be found in much smaller density around non-frontal faces and even around false alarms. All these say N bounding boxes form the measurement vector $\mathbf{y}^{(\text{face})} = \left[[\mathbf{y}_p^{(1)}, y_w^{(1)}], \ldots, [\mathbf{y}_p^{(N)}, y_w^{(N)}]\right]^T$ where $\mathbf{y}_p^{(i)}$ is the two-dimensional position of the i-th bounding box and $y_w^{(i)}$ is its width. Note that the height needs not be specified, since the detector has the same aspect ratio for all the faces it reports.

Given the state vector $\mathbf{x} = [\mathbf{x}_p, x_w]^T$, the likelihood for the face measurement $\mathbf{y}^{(\text{face})}$ receives contributions from candidate face bounding boxes. These contributions should be larger as the bounding box locations $\mathbf{y}_p^{(i)}$ approach the state location $\mathbf{x}_p$ and as their width $y_w^{(i)}$ approaches the state width x_w. Both goals are achieved by defining the likelihood as

$$p\left(\mathbf{y}^{(\text{face})}|\mathbf{x}\right) = \sum_{i=1}^{N} \frac{w_f^{(i)}}{\left(\| \mathbf{x}_p - \mathbf{y}_p^{(i)} \|_2 / x_w^2\right)^{K/2} + 1}. \tag{13}$$

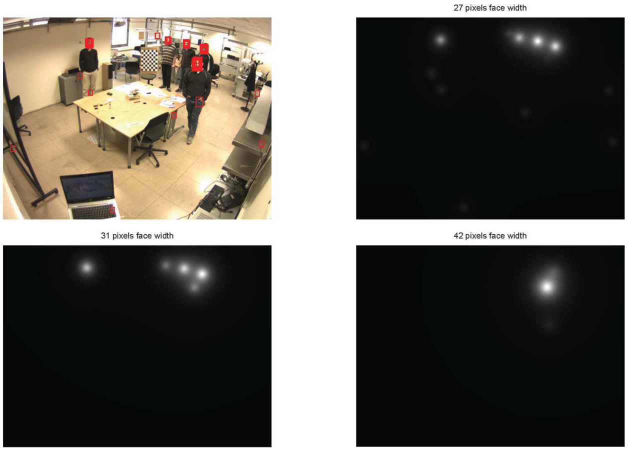

Figure 1: Image and associated face likelihoods according to eq. (13), evaluated across all the image plane for three face widths.

Note that each term $\left(\left(\| \mathbf{x}_p - \mathbf{y}_p^{(i)} \|_2 / x_w^2\right)^{K/2} + 1\right)^{-1}$ makes a large contribution to the likelihood, close to unity if the distances of the state and the bounding box positions are close to zero. The exponent K governs how fast the contributions are attenuated as these distances increase. A good choice is $K = 2$. The contributions are weighted in the likelihood summation by the weights $w_f^{(i)}$:

$$w_f^{(i)} = \exp\left(- \mid x_w - y_w^{(i)} \mid / 2\sigma_w^2\right) \tag{14}$$

to penalize the contribution of candidate face bounding boxes that are quite different in width than x_w. The reason the difference in width is not included in the norm at the denominator of eq. (13) is that differences in bounding box locations are not comparable in scale to differences in bounding box widths. Putting width differences in the same norm as location differences would scale down the importance of the former relative to the latter. Also note that the weights in eq. (14) are not scaled to sum to unity. This is chosen so that more than one similar detections would increase the likelihood compared to a single detection. A necessary penalty to pay is that the likelihood values in eq. (13) are not bounded by unity.

The face likelihood obtained by eq. (13) is a three-dimensional function, one for each dimension of the state. An example is given in Figure 1. Evaluating the likelihood for different state locations $\mathbf{x}_p$ results to the values of the different pixels on the likelihood plane evaluating for the different state widths $y_w^{(i)}$ results to the different likelihood planes. Note how the likelihood peaks in the vicinity of the bounding boxes and does more so as more bounding boxes are located nearby. Also note the effect of the weights in (14) in selecting the faces of the wanted width.

3.2 Color likelihood

Color matching is evaluated using the similarity of a model histogram to a histogram extracted from a candidate region $R_\mathbf{x}$ corresponding to the state $\mathbf{x}$. To alleviate the limitations of color modeling regarding precise localization, color is modeled in subregions to add structure to the model and the effect of similarly-colored background is attenuated with an immediate background histogram.

Let the target contain n_r subregions with known spatial arrangement within $R_\mathbf{x}$. Then n_r reference histograms $\mathbf{h}_{\text{ref}}^{(i)}$, $i = 1, \ldots, n_r$, are trained using N_h bins per color component and are compared to the target histograms $\mathbf{h}_\mathbf{x}^{(i)}$ using the Bhattacharyya distance:

$$D^{(i)} = 1 - \sum_{n=0}^{N_h^3-1} \sqrt{\mathbf{h}_\mathbf{x}^{(i)}(n)\mathbf{h}_{\text{ref}}^{(i)}(n)} \tag{15}$$

where $\mathbf{h}_{\mathbf{x}}^{(i)}(n)$ and $\mathbf{h}_{\text{ref}}^{(i)}(n)$ are the n-th bins of the target and reference histograms of the i-th subregion.

The overall distance $D_{\text{tar,ref}}$ for the multi-region color likelihood is defined as the weighted average of the distances $D^{(i)}$:

$$D_{\text{tar,ref}} = \sum_{i=1}^{n_r} w_h^{(i)} D^{(i)} \tag{16}$$

The weights $w_h^{(i)}$ are chosen based on the importance of each of the subregions. We propose a color model based on four subregions. These are defined based on the face detection: The eyes and lower face subregions are located within the original face detection and are obtained as fixed zones relative to it. The forehead-hair and upper torso subregions are cropped around the original face detection, again with fixed sizes relative to it.

The likelihood for the color measurement $\mathbf{y}^{(\text{color})}$ then is

$$p\left(\mathbf{y}^{(\text{color})}|\mathbf{x}\right) \propto \exp\left(-D_{\text{tar,ref}}/2\sigma^2_{\text{color}}\right). \tag{17}$$

The sensitivity to background color similarities is alleviated by attenuating the effect of colors that appear in the immediate background from the four subregion histograms $\mathbf{h}_{\mathbf{x}}^{(i)}$. To do so the background histogram $\mathbf{h}_{bkg}$ is calculated from the pixels across the face. We calculate the bin values $\mathbf{h}_{\mathbf{x},bkg}^{(i)}(n)$, $n = 1,\ldots,N_h^3$ of the background-aware model histograms as follows: Let $h_{bkg}^{(min)}$ be the minimum non-zero bin value of $\mathbf{h}_{bkg}$. Define:

$$a_n = \min\left(\frac{h_{bkg}^{(min)}}{\mathbf{h}_{bkg}(n)}, 1\right) \tag{18}$$

Then the bin values of $\mathbf{h}_{\mathbf{x},bkg}^{(i)}$ are given by:

$$\mathbf{h}_{\mathbf{x},bkg}^{(i)}(n) = a_n \cdot \mathbf{h}_{\mathbf{x}}^{(i)}(n) \tag{19}$$

Note that (19) yields a non-normalized histogram, whose bins need to be normalized to sum up to unity.

3.3 Partitioned sampling PF tracker

The state-space comprises of the 2D position on the camera plane and the face width. According to the partitioned sampling approach [8], each of the measurement cues is used to update a subspace of the state-space. We utilize face measurements to update position on the camera plane. Subsequently, the position-updated particles are re-assembled with their width dimension and are updated using the color measurement. 50 particles are used.

4. AUDIOVISUAL FUSION

The audio and visual systems described in the previous sections both give location estimates with some uncertainty. Referring to Figure 2, the visual position on the image plane (due to the depth uncertainty) corresponds to any point along the red line connecting the origin of the camera coordinate system $[x_v, y_v, z_v]$ with the depth-normalized coordinates $\mathbf{v}_n$ from the visual track. The image plane coordinates are transformed to $\mathbf{v}_n$ using the intrinsic camera parameters [15]. Similarly, the audio position (with the height uncertainty of the audio tracker) corresponds to any point along the green line connecting the origin of the audio coordinate system $[x_a, y_a, z_a]$ with the depth-normalized coordinates $\mathbf{a}_n$ from the audio track.

Audiovisual fusion utilizes the intersection of the two lines, effectively eliminating the location uncertainty. Ideally the two lines intersect, but in practice audio and visual tracking errors and the different targets (mouth for the audio and center of vaguely frontal face for the visual tracker) result to no intersection. Instead we employ a least squares solution to find the point of minimum distance from both lines, i.e. the center $[x_{av}, y_{av}, z_{av}]$ of the minimum length segment connecting the uncertainty lines. This minimum length is $2e_{av}$, and used as a measure of the quality of match of the audio location with the visual one.

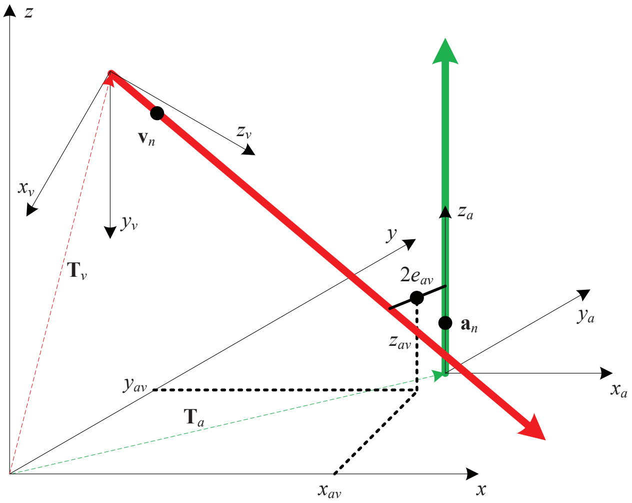

Figure 2: Normalized audio ($\mathbf{a}_n$) and visual ($\mathbf{v}_n$) estimations with their uncertainties (green and red thick lines) on the respective coordinate systems ($[x_a, y_a, z_a]$ and $[x_v, y_v, z_v]$), offset and rotated with respect to the world coordinate system $[x, y, z]$. The audiovisual fusion is the center $[x_{av}, y_{av}, z_{av}]$ of the minimum length ($2e_{av}$) segment connecting the uncertainty lines.

To formulate the problem, we need to relate the audio and visual coordinate systems with the world coordinate system $[x, y, z]$. This is done by finding the respective translation vectors $\mathbf{T}_a$ and $\mathbf{T}_v$ (green and red dashed lines) as well as the rotation matrices $\mathbf{R}_a$ and $\mathbf{R}_v$. For the visual coordinate systems, these are the extrinsic parameters of the camera [15]. For the audio system, the translation vector is simply the location reported by the audio tracker on the floorplan, while the rotation matrix is the identity one, since the orientation of the system does not change with respect to the world one. The normalized audio coordinates then are the z-axis unity vector. Then the least squares solution for $\mathbf{x}_{av}$ is obtained by solving:

$$\begin{bmatrix} \mathbf{I}_3 & -\mathbf{R}_v\mathbf{v}_n & \mathbf{0}_3 \\ \mathbf{I}_3 & \mathbf{0}_3 & -\mathbf{R}_a\mathbf{a}_n \end{bmatrix} \cdot \mathbf{x}_{av} = \begin{bmatrix} \mathbf{T}_v \\ \mathbf{T}_a \end{bmatrix} \tag{20}$$

where $\mathbf{I}_3$ is the 3×3 identity matrix and $\mathbf{0}_3$ is the 3×1 zero vector. The system is solved using the pseudo-inverse of the left-hand matrix.

5. PERFORMANCE DISCUSSION

At every frame, multiple visual targets from the people present and a single audio one from the speaker are reported. The 2D accuracy of the video tracks is quite high, especially with the faces at the camera are approximately frontal, but there is no depth estimate. The audio tracks are on the other hand quite accurate in estimating the angle from the microphones, but cannot give accurate depth estimates. Also, when there is no speech the audio track is quite erratic, jumping around the space.

In order to fuse the two modalities, both the camera and the microphones have to be related to the world coordinate system. For the camera, this is done by calibrating it [2] to extract its intrinsic and extrinsic parameters. For the microphones, their position is simply measured. The fused audiovisual estimate of the location is obtained by attempting to associate the audio track with each of the visual tracks and solving (20). Excessive approximation errors e_{av}, heights z_{av} and face widths (that are estimated as in [6], given the depth from the camera and the tracked face bounding box) are used to discount associations. If no association survived, then the system assumes there is no speech, resulting to the audiovisual Voice Activity Detection. If there is at least one surviving association, then the fused location of the speaker is returned.

5.1 Performance Measures

We use a single metric to evaluate the different systems. For its calculation we test the source location estimates provided by the Audio and the Audio-Visual systems at each time frame against the ground truth (this is a result of manual annotation), for the total duration of the test signals.

The squared error for time frame t is given as $\varepsilon_t = \|\mathbf{s}_t - \bar{\mathbf{s}}_t\|^2$, where $\bar{\mathbf{s}}_t$ denotes the actual (manually annotated) source location. The metric used for comparison, the *Root Mean Square Error* (RMSE) is defined as the square root of the average value of ε_t over the total number of frames. In the following results, the above metric is presented in meters. The lower the values, the better the performance of the corresponding system.

5.2 Experiments

To demonstrate the effect of video tracking upon active speaker localization we measured the performance of the audio tracker and the multi-modal one upon the corpus created for the HERMES project [1]. The corpus contains A/V recordings in typical reverberant rooms equipped with three microphones and one camera. Collection of audio data is performed using a total of three microphones and one camera. The microphones are facing the expected location of the speakers, while the camera is a bit off-center, at one of the corners of the room. The recordings are conducted in presence of ambient noise from both air-conditioning and personal computers. Each recording consists of a discussion between 3 people sitting in armchairs in front of a television set that hosts the 3 co-linear microphones being 0.2 m apart.

There exists significant interaction between the people with discussions that have movements of the speakers, interchanging speakers and numerous acoustic events e.g. interruptions of the discussion due to ringing mobile phones, people coughing and laughing. The A/V data were manually annotated to provide the Cartesian location and the speech activity of each participant at every frame of video. These annotations are considered to be the ground truth for the measurement of our system performance.

The audio system used $N = 50$ particles, $L = 0.27$ s and $f_s = 44.1$ Khz. Also for the external PF, $d_e = 1$ m and $T_e = 1$ sec. The reverberation time of the room was measured to be approximately $T_{60} = 0.5$ s. The camera was recording at 10 frames per second and 1600×1200 resolution. The faces are typically 60 to 65 pixels wide. The associations of the audio with some visual track were accepted is $e_{av} < 600$mm, $z_{av} \in [700, 2100]$mm and the estimated face widths are within $[111, 189] mm$. These result to an equal rate of correctly detected speech and silence at 83.5%. For the frames correctly identified as containing speech, the RMSE of the audio-only tracks is 771mm parallel to the microphone plane (x-axis) and 676mm perpendicular to it (y-axis). The equivalent RMSE for the audiovisual tracks are 391mm and 354mm respectively, while for height it is 146mm. All RMSE and numbers are less than half for the audiovisual system leading to a more precise localization.

6. CONCLUSION

Performing acoustic source tracking to detect the active speaker in the realistic environment of a moderately reverberant office room is severely limited by reverberation and/or background noises. Under these conditions, the use of the video modality can prove to be of advantage compared to more traditional algorithms.

In this paper, we have presented a framework that integrated an audio and a visual tracking system in order to extends the ability of the system in order to track the active speaker or the one that last spoke. The system was designed keeping in mind the minimum setup of 2 cameras and 3 microphones in order to be appropriate for easy installation in homes of elderly users.

Using recordings in a cluttered meeting room, we have demonstrated that the multi-modal framework outperforms the audio only system in all scenarios. Thus, we have a system that remains adequately robust, easy to employ and can serve as the necessary first step in offering complex services in smart-homes.

REFERENCES

[1] HERMES (cognitive care and guidance for active aging) EU FP7 STREP. http://www.fp7-hermes.eu.

[2] J.-Y. Bouguet. Camera calibration toolbox for matlab. www.vision.caltech.edu/bouguetj/calib_doc/htmls/ parameters.html, 2008.

[3] G. Bradski, A. Kaehler, and V. Pisarevsky. Learning-based computer vision with intel's open source computer vision library. *Intel Technology Journal*, 9, 2005.

[4] M. Brandstein, J. Adcock, and H. Silverman. A closed-form location estimator for use with room environment microphone arrays. *IEEE Trans. on Acoust. Speech and Sig. Proc.*, 5:45–50, 1997.

[5] H. Istance, A. Hyrskykari, D. Koskinen, and R. Bates. Gaze-based attentive user interfaces auis to support disabled users: towards a research agenda. *Proceedings of the 2nd Conference on Communication by Gaze Interaction: COGAIN 2006: Gazing into the Future*, 1:56–62, 2006.

[6] N. Katsarakis and A. Pnevmatikakis. Face validation using 3d information from single calibrated camera. In *DSP'09: Proceedings of the 16th international conference on Digital Signal Processing*, pages 972–977, Santorini, Greece, 2009.

[7] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transaction on Acoustics Speech and Signal Processing*, 24(4):320–327, 1976.

[8] P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proc. of IEEE*, 92(3):495–513, 2004.

[9] L. Portoni, C. Combi, and F. Pinciroli. User-oriented views in health care information systems. *IEEE Transactions on Biomedical Engineering*, 49(12):1387–1398, 2002.

[10] J. Vermaak and A. Blake. Nonlinear filtering for speaker tracking in noisy and reverberant environments. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 5:30213024, 2001.

[11] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, pages 511–518, Kauai, HI, USA, December 2001.

[12] D. Ward, E. Lehman, and R. Williamson. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Trans. on Acoust. Speech and Sig. Proc.*, 11(6):826–836, 2003.

[13] M.-H. Yang. Recent advances in face detection. In *IEEE International Conference on Pattern Recognition (ICPR 2004)*, United Kingdom, August 2004.

[14] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(1):34–58, 2002.

[15] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.