Tracking sound sources by means of HMM

Antonio Rodà, Christian Micheloni
University of Udine
Via Margreth 3, 33100 Udine, Italy
{antonio.roda;christian.micheloni}@uniud.it

Abstract

Video-based surveillance systems may benefit from the integration with microphone arrays for the localization of sound events. Applying the sound localization techniques to the surveillance of large areas requires addressing some open issues, such as the non uniform resolution of the microphones-based localization systems. This paper presents a new method for tracking moving sound events based on an Hidden Markov Model (HMM), which exploits a priori information derived from medium and long-term observations of the monitored area. The results obtained with simulated trajectories show that the HMM-based tracker is able to significantly reduce the localization error. Applications can be found in surveillance systems for large areas, such as square, streets, or parking lots, where it is of interest the monitoring of moving vehicles and people.

1. Introduction

Video-based surveillance systems may benefit from the integration with other types of sensors. In particular, audio sensors can provide a major improvement because, unlike a normal camera, they are omnidirectional and do not require direct line-of-sight with the sound source. Such capabilities can nicely complement vision in order to help to localize interesting or dangerous events in the monitored area. Examples of integrated audio-video frameworks for the recognition of the surrounding scene can be found in robotics [16] [9] [6] and in the human computer interfaces field [12]. These systems are able to localize a sound source in a room using microphone arrays. In such a context, the use of signal processing techniques, generally based on Time Delay Of Arrival estimation (TDOA, i.e. the time delay with which the waveform arrives to the different sensors of the array) and delay-and-sum beamforming allow to operate at short distance (e.g. few meters). Recently, microphone arrays have been tested for the localization of sound events [15] [7] [3] with application to the surveillance of large areas. The use of the localization techniques to monitoring large areas, such as squares or parks, introduces difficulties not yet fully addressed. The localization error of sound events by means of microphone arrays generally depends on several factors: the distance and the angle of the sound source with respect to the array, the shape of the array, its size (i.e. number of microphones), the distance between the microphones, the sampling frequency, the acoustic response of the environment and the presence of competing sound sources. The localization errors can be reduced using arrays with a large aperture [13]. This approach, however, needs a large number of microphones, thus requiring greater computational resources and a larger space, not always available in a real scenario, to install the array.Another approach involves the use of algorithms based on the Kalman filter [8] or particle filter [11] for the tracking of moving sound events. In general, these algorithms exploit a priori information given by the previous positions of the event in motion.

This paper presents a new system for tracking sound events based on an Hidden Markov Model (HMM), which exploits a priori information derived from medium and long-term observations of the monitored area. This approach can be adopted in all those contexts where the events of interest do not move randomly, but rather follow more or less stable paths. Some examples are squares, streets, or parking lots, where vehicles and people go preferably through some paths rather than others. The HMM-based algorithm aims to reduce the tracking error of the sound events especially in areas where the resolution of the microphone array is low. At the same time, this system allows to adapt the resolution of the audio and video sensors, yielding to an easier integration between these kind of sensors in a single surveillance system. Let be the Map of the monitored area divided into N rows and M columns, which define NxM equally spaced cells (see Section 3), and these values can be chosen so that the obtained grid corresponds to the resolution of the video analysis subsystem. It can be noted that this approach can be seen as a particular case of

*This work is partially supported by the Interreg IV Italy-Austria project n. 4697 "SRSNet - Intelligent Audio/Video Sensor Networks".

Cartesian Hidden Markov Model, defined in [17].

Hidden Markov Models have already been used for the tracking of sound parameters (e.g. [14], [4]), but their application to localize sound sources has not yet been explored. Other related works, that however do not concern sound aspects, are [1], [2], [5].

The assumption behind our system is that in a real space, such as a square or a street, people and objects move according to certain preferred trajectories which can be described by a Markov process. Observing these trajectories, it is possible to estimate the probability with which an object moves from one cell to the adjacent ones, thus obtaining the transition probabilities of the HMM. Out of all the possible sources of error in localizing a sound event by means of microphone arrays, we will consider only the spatial sampling effect, resulting from the sampling over time of the audio signal. Indeed, it can become the predominant source of error in the monitoring of large and low reverberating areas. The presented approach, however, remains valid even in the presence of other sources of error.

Another requirement is the continuity of the sound source. While this condition can be considered reasonably satisfied in the case of motor vehicles, some problems may be represented by moving people who talk. In this case the sound source is characterized by more or less short silences, separating syllables and words. Our approach remain applicable if the duration of this silence is negligible compared to the transit time within a cell (a condition usually occurred in the case of walking people). If pauses are longer, however, the movement can be segmented into multiple continuous trajectories, applying the HMM-based tracker to each segment individually.

The rest of the paper is organized as follows. In Section 2 we will detail the problem of spatial sampling in the case of uniform linear array, giving some equations to describe the spatial resolution. In Section 3 we will present an implementation of the tracking system based on HMM. The results of the system evaluation, carried out with simulated trajectories, will be showed in Section 4.

2. Spatial resolution of a ULA

Consider for convenience a uniform linear array (ULA) in a reference system (x, y, z) in which the xy plane is the plane where the sound sources lie, the center of the array is located at the coordinates (0, 0, h) where $h > 0$ is the distance between the array and the plane of interest, and the axis of the array has the same direction of the axis x. Assume that the array has been designed to capture the acoustic waves coming from ahead only (the microphone capsules point in that direction), i.e the array works in the half-space $y > 0$. By convention, α is the angle between the direction of arrival of the sound and the perpendicular to the array axis, defined between -90^o (equivalent to a sound that comes from extreme left) to $+90^o$ (corresponding to a sound coming from the far right). In this reference system, the Time Difference Of Arrival (TDOA) denotes a conical surface (see Figure 1) represented by the parametric equation:

$$\begin{cases} x = t \\ y = sink \cdot \cot\alpha \cdot t \\ z = cosk \cdot \cot\alpha \cdot t + h \end{cases} \tag{1}$$

where $\alpha = \arcsin(TDOA \cdot c/d)$ (far field condition), c is the sound propagation velocity, d is the distance between the microphones of the array, $k \in [0, \pi]$ is the independent parameter of the equation, $t \in \mathbb{R}^+$ if $\alpha > 0$ and $t \in \mathbb{R}^-$ if $\alpha < 0$. For $\alpha = 0$ the cone degenerates into the half-plane $\{x = 0,\ y > 0\}$. The condition that the sound sources lie on the plane $z = 0$ is equivalent to put $cosk \cdot cot\alpha \cdot t + h = 0$, which gives

$$\sin k = \sqrt{1 - \frac{h^2}{\cot^2\alpha \cdot t^2}} \tag{2}$$

with $cot\alpha \neq 0$ and $1 - h^2/(cot^2\alpha \cdot t^2) > 0$, from which it follows the condition $t \geq h/\cot\alpha$ if $\alpha > 0$ and $t \leq h/\cot\alpha$ if $\alpha < 0$. Substituting Eq. 2 in Eq. 1 we obtain

$$cot^2\alpha \cdot x^2 - y^2 - h^2 = 0 \tag{3}$$

that is the equation of a hyperbola (Figure 1). Therefore, if the array measures a certain value of TDOA, it means that the sound source is located in one of the points described by the Eq. 3. The strategy we used to estimate the two-dimensional coordinates of the source is to use a second array, placed in a different location than the first. The source location will be determined by the intersection between the hyperbolas detected by the two arrays. Solving the equation system between the two hyperbolas, respectively related to the left and right array, we obtain the sound source position

$$\begin{aligned} x &= \tfrac{d_a}{2} \cdot \tfrac{tg(\alpha_{dx}) + tg(\alpha_{sx})}{tg(\alpha_{dx}) - tg(\alpha_{sx})} \\ y &= \sqrt{d_a^2 \cdot \left(\tfrac{tg(\alpha_{dx}) \cdot tg(\alpha_{sx})}{tg(\alpha_{dx}) - tg(\alpha_{sx})}\right)^2 - h^2} \end{aligned} \tag{4}$$

where α_{sx} and α_{dx} are the DOAs estimated respectively by the left and right array, d_a is the distance between the two arrays.

Due to the sampling over time of the audio signal, the TDOA can only assume values that are integer multiples of the sampling period. It follows that $\alpha_s x$ and $\alpha_d x$ assume discrete values and the space of the solutions of the system given by the Eq. 4 is a discrete set of points (see Fig. 2 for a graphical representation of this set).

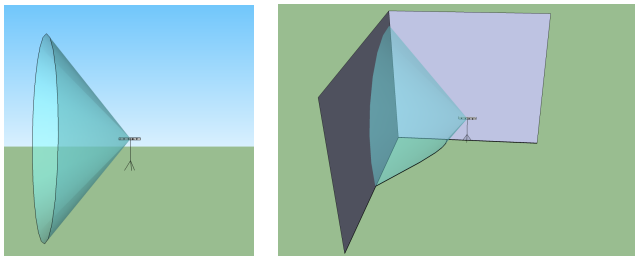

Figure 1. *Conical surface corresponding to a given TDOA and its intersection with the plane z=0*

3. HMM-based tracking

3.1. Model definition

Given a sound source placed at cell coordinates (i, j), we can estimate its position calculating the TDOA among the microphones and solving the equation Eq. 4. In general, the estimated position (k, l) will be different from the actual one (i, j). We define the function

$$m : (i, j) \rightarrow (k, l) \tag{5}$$

that maps the actual position into the estimated position, where i, j, k, and l are integer numbers, $i, k = 1, \ldots, M$ and $j, l = 1, \ldots, N$. We define an HMM with MxN states, denoted as $S_{(i,j)}$, each one associated to a cell of the monitored plane, and observations, denoted as $\nu_{(k,l)}$, associated to the elements of the set

$$O = \{(k, l) \,|m(i, j) = (k, l) \forall i = 1, \ldots, M, j = 1, \ldots, N\}. \tag{6}$$

that contains all the positions estimated by the microphone arrays.

We denote the transition probabilities as

$$a_{(i,j)(m,n)} = P[q_t = S_{(m,n)} \,|q_{t-1} = S_{(i,j)}] \tag{7}$$

and the emission probabilities as

$$b_{(i,j)}(k, l) = P[\nu_{(k,l)} \text{ at t} \,|q_t = S_{(i,j)}] \tag{8}$$

where $t = 1, 2, \ldots$ are the time instants associated with the state changes and q_t is the actual state at time t.

In order to test the tracking system based on HMM, a rectangular area of size 80x60 meters, representing the space on which the events of interest lie, has been simulated in MATLAB environment. The area is divided in 3072 cells, uniformly spaced along 64 rows and 48 columns. Two ULAs are placed at cell coordinates (1,9) and (1,18), that corresponds to a distance between the arrays of about $11.3m$, at a height from the plane of interest $h = 12m$. The distance between the microphones of each array has been set at $d = 0.25m$ and the sampling frequency has been set at $48000Hz$. Given these parameters, the cells observable

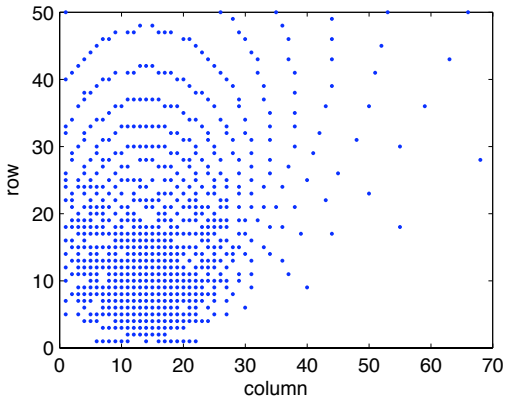

Figure 2. *Cells observable by the microphone arrays.*

by the audio arrays (i.e. the space of solutions of the Eq. 3 for the two arrays) are shown in the Figure 2; they represent the possible observations of the HMM.

The transition probabilities among the 3072 hidden states of the HMM are calculated by comparing a set of known trajectories. For this purpose, a procedure was implemented for the generation of stochastic trajectories.

3.2. Trajectory generator

The trajectory generator has been constructed starting from an activity map [10] c(i,j) $\rightarrow$ [0,1], i.e. a function that associates to each cell at the $i - th$ row and $j - th$ column the probability that an event of interest is in that position. Figure 4 shows an example of activity map: the red cells represent the locations where the presence of an event of interest is more probable. From the activity map, we define four Markov models for the generation of four types of trajectories, named up-down (U-D), down-up (D-U), left-right (L-R), right-left (R-L). Regarding the first type (U-D), the initial state of the Markov model is randomly chosen among the cells in the upper end of the Map, following the probability distribution defined by the activity map: $c(48, j)$, where $j = 1, ..., 64$. The probability of transition to the other cells are all null except the three neighbor cells along the upper-down direction (see Fig. 3). These transition probabilities are calculated by the equation

$$\begin{aligned} \varphi(x_{i,j}, x_{i-1,j-1}) &= c(i-1, j-1)/k \\ \varphi(x_{i,j}, x_{i-1,j}) &= c(i-1, j)/k \\ \varphi(x_{i,j}, x_{i-1,j+1}) &= c(i-1, j+1)/k \end{aligned} \tag{9}$$

where $k = c(i-1, j-1) + c(i-1, j) + c(i-1, j+1)$.

The other models are defined in a similar manner, starting from the left, right and bottom end of the Map. The trajectory generator is capable of generating a rich variety of trajectories (in all the four directions). For each point of the trajectories, the equations of Section 2 are used to calculate

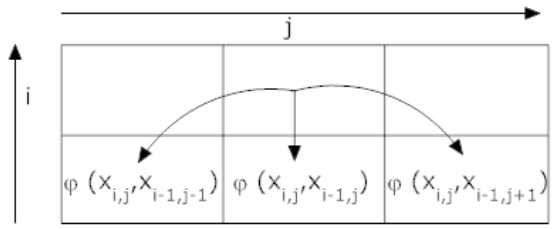

Figure 3. *Transition probabilities of the U-D Markov model.*

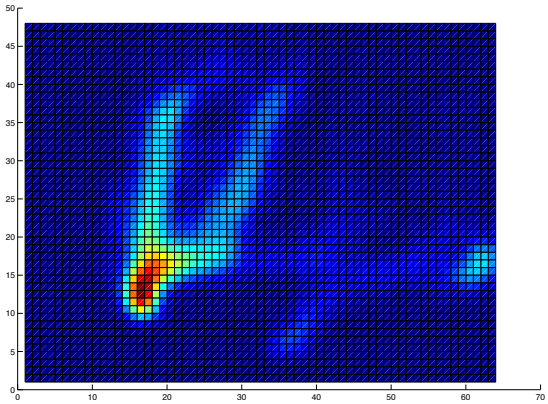

Figure 4. *Activity map used to train the HMM.*

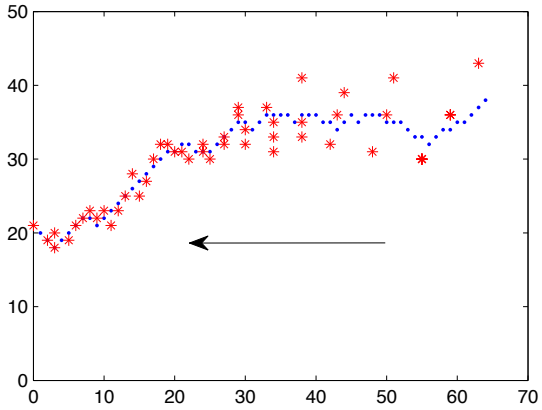

Figure 5. *A trajectory R-L generated by means of the Markov model. Points are the real trajectory and stars represent the corresponding positions estimated by the microphone arrays.*

the corresponding position estimated by the microphone arrays.

Figure 5 shows an example of trajectory directed from right to left, generated by the Markov model: the blue dots represent the actual location of the event, while the red asterisks are the corresponding positions estimated by the arrays. Note that the localization error depends on the position of the event, following the distribution in Figure 2.

A number of trajectories have been generated by the model and used to compute the transition probabilities between the states of the HMM. At the same time, the emission probabilities have been computed by taking into account the positions estimated by the microphone arrays. The training was repeated six times, with a number of different trajectories: 250, 500, 1000, 2000, 4000, and 8000. After the training phase, other 400 trajectories (100 for each directions) have been generated in order to form the test set to evaluate the performance of the tracking system.

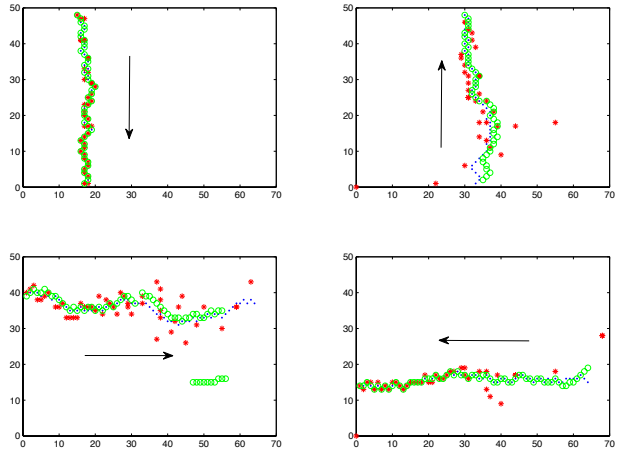

Figure 6. *Some examples of trajectory computed by the HMM. Points are the testing trajectory, stars represent the corresponding positions estimated by the microphone arrays without the HMM tracker, circle are the positions estimated using the HMM tracker.*

4. Results

After the training phase, the HMM receives as input a sequence of positions estimated by the arrays and generates as output the more likely sequence of visited cells, given those observations. Figure 6 shows four trajectories of test (represented by the blue dots), the positions estimated by the microphone arrays (red asterisks), and the trajectories estimated by the HMM (green circles). It can be noted that, apart from some exceptions, the tracking system follows the unknown trajectory with good approximation even in areas where the location resolution of the array is low.

For a quantitative assessment of the performance, the deviation between the actual position (i, j) and the estimated one $(\tilde{i}, \tilde{j})$ has been calculated for each step t of the trajectory. The sum of this deviations provides a measure of the capability to follow the actual trajectory:

$$err = \sum_{t=1}^{N} \sqrt{(i(t) - \tilde{i}(t))^2 + (j(t) - \tilde{j}(t))^2} \qquad (10)$$

Table 1 summarizes the performance by varying the size of the training set (from 250 to 8000 trajectories). Statistics have been computed on the 100 test trajectories; the *outperf* column shows the percentage of test trajectories where the

training	outperf	min	max	mean	dev	minHMM	maxHMM	meanHMM	devHMM
250	57.8	4.5	22.1	9.4	4.3	1.0	28.7	9.0	7.1
500	56.8	4.5	23.3	9.4	4.7	0.2	25.6	8.5	6.7
1000	67.3	4.6	22.8	9.8	4.5	0.3	21.9	6.8	5.7
2000	71.0	4.7	20.5	9.1	4.2	0.2	19.3	5.6	4.4
4000	80.3	4.4	21.1	9.6	4.5	0.3	17.1	4.7	4.0
8000	76.5	4.7	21.1	9.2	4.4	0.2	18.3	4.5	3.7

Table 1. Results depending on the size of the test set.

direction	outperf	min	max	mean	dev	minHMM	maxHMM	meanHMM	devHMM
U-D	94	1.1	23.5	7.5	7.3	0.0	17.1	3.3	4.5
D-U	97	9.1	13.3	10.7	0.9	0.3	15.3	1.7	3.0
L-R	74	3.7	15.1	8.3	2.1	0.3	21.1	6.2	4.6
R-L	56	3.8	32.6	11.8	7.6	0.4	14.9	7.7	3.8

Table 2. Results depending on the trajectory type: upper-down, down-up, left-right, and right-left.

HMM has outperformed the array-only system (i.e., the system without the HMM-based tracker). Overall, the system with HMM has better behavior in up to 80% of the trajectories. In particular, the minimum error (in the case of the most favorable trajectory) goes from 4.5 cells (array-only approach) to 0 (with HMM). This means that, unlike the array-only system, the HMM is able to follow the unknown trajectory without error. On average, the error goes from about 21 cells (array-only) to 4.5 cells (with HMM). As for the size of the training set, we see a significant performance increase from 250 to 4000 trajectories, while from 4000 to 8000 trajectories the performance difference is minimal.

Table 2 shows the performance achieved with a training set of 4000 trajectories, by varying the trajectory direction. While an excellent performance is achieved with the U-D and D-U trajectories ($outperf > 94\%$), in the case of R-L trajectories the results of the HMM are worst, though still better than the array-only system.

This disparity may depend on the different properties of the trajectories generated by the Markov model described in Section 3. In fact, Figure 7 shows that the trajectories generated by the U-D model (on the left) follow more defined paths than those generated by the R-L model (on the right), where the trajectories are distributed more uniformly across the Map. Therefore, if the sound source moves along a trajectory that does not follow a pre-defined path, the tracker can make mistakes as in the case showed on the lower left panel of Figure 6. This result confirms that, as expected, the performance of HMM-based tracking system depends on the predictability of the trajectories; the system is therefore suitable for all those context where the movement of the interesting events follows stable paths.

Table 3 resumes the average results of the evaluation test. For comparison, we add the performances of a Kalman-based tracker (see [8] for more details), that receives as input the positions estimated by the microphone arrays. It can be noted that the HMM-based tracker outperforms the other approaches in the 79% of the test trajectories.

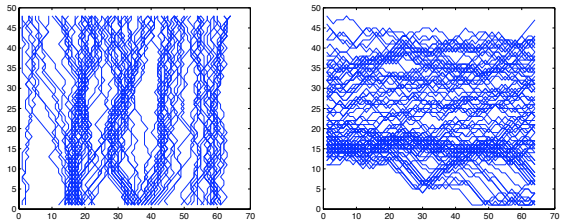

Figure 7. *The 100 test trajectories: U-D direction on the left, R-L direction on the right.*

	outperf	min	max	mean	std
array-only	0	1.1	32.6	9.6	4.5
Kalman	21	0.7	28.3	6.1	4.1
HMM	79	0	21.1	4.7	4

Table 3. Average results of the HMM-based tracker compared with an array-only approach and a Kalman-based tracker.

5. Conclusions

An HMM-based tracker for the localization of sound sources was presented. The results of the validation test show that the tracker is able to reduce the localization error when the movement of the events of interest can be statistically modeled after mid- and long-term observations. This makes possible to apply the HMM-based tracker to surveillance systems for large areas such as squares or streets, where people and vehicles are used to move along preferential paths.

References

[1] H. H. Bui, S. Venkatesh, and G. West. *Tracking and surveillance in wide-area spatial environments using the abstract hidden Markov model*, pages 177–196. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2002.

[2] H. Buxton and S. Gong. Advanced Visual Surveillance using Bayesian Networks. In *International Conference on Computer Vision*, June 1995.

[3] P. Dostalek, V. Vasek, V. Kresalek, and M. Navratil. Utilization of audio source localization in security systems. In *43rd International Carnahan Conference on Security Technology*, pages 305–311, Zurich, 5-8 Oct 2009.

[4] V. Emiya, R. Badeau, and B. David. Automatic transcription of piano music based on HMM tracking of jointly-estimated pitches. In *Proc. Eur. Conf. Sig. Proces. (EUSIPCO)*, Lausanne Suisse, 2008. FP6-027026-K-SPACE.

[5] R. Fraile and S. Maybank. Vehicle trajectory approximation and classification. In *British Machine Vision Conference*, 1998.

[6] J.-S. Hu, C.-Y. Chan, C.-K. Wang, and C.-C. Wang. Simultaneous localization of mobile robot and multiple sound sources using microphone array. In *ICRA'09: Proceedings of the 2009 IEEE international conference on Robotics and Automation*, pages 4004–4009, Piscataway, NJ, USA, 2009. IEEE Press.

[7] K. K. Jung, H. S. Shin, S. H. Kang, and K. H. Eom. Object tracking for security monitoring system using microphone

array. In *17-20 Oct.*, pages 2351–2354, Seoul, 17-20 Oct 2007.

[8] U. Klee, T. Gehrig, and J. McDonough. Kalman filters for time delay of arrival-based source localization. *EURASIP Journal on Applied Signal Processing*, pages 167–167, 2006.

[9] E. Menegatti, M. Cavasin, E. Pagello, E. Mumolo, and M. Nolich. Combining audio and video surveillance with a mobile robot. *International Journal on Artificial Intelligence Tools*, 16(2):377–398, 2007.

[10] C. Piciarelli, C. Micheloni, and G. L. Foresti. Occlusion-aware multiple camera reconfiguration. In *International Conference on Distributed Smart Cameras (ICDSC 2010)*, Atlanta, GA, USA, Aug 31 - Sep 4 2010.

[11] D. Riva, D. Saiu, A. Sarti, M. Tagliasacchi, S. ubaro, and F. Antonacci. Tracking multiple acoustic sources using particle filtering. In *Proc. of the European Signal Processing Conference*, Florence, Italy, September 4-8, 2006.

[12] S. Shivappa, M. Trivedi, and B. Rao. Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey. *Proceedings of the IEEE*, 98(10):1692–1715, 2010.

[13] H. F. Silverman, Y. Yu, J. M. Sachar, and W. R. Patterson. Performance of real-time source-location estimators for a large-aperture microphone array. *IEEE Trans. Speech, Audio Process*, pages 593–606, 2005.

[14] R. Streit and R. Barrett. Frequency line tracking using hidden markov models. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 38(4):586–598, Apr 1990.

[15] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection and localization for audio-surveillance systems. In *AVSS '07: Proceedings of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 21–26, Washington, DC, USA, 2007. IEEE Computer Society.

[16] J.-M. Valin, F. Michaud, J. Rouat, and D. Letourneau. Robust sound source localization using a microphone array on a mobile robot. In *Intelligent Robots and Systems, 2003*, volume 2, pages 1228 – 1233, 27-31 Oct. 2003 2003.

[17] L. B. White. Cartesian hidden markov models with applications. *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, 40(6):1601–1604, June 1992.