KALMAN FILTERS FOR AUDIO-VIDEO SOURCE LOCALIZATION

Tobias Gehrig, Kai Nickel, Hazim Kemal Ekenel, Ulrich Klee, John McDonough

Institut für Logik, Komplexität, und Deduktionssysteme
Universität Karlsruhe
Am Fasanengarten 5
D-76131 Karlsruhe, Germany
{tgehrig,nickel,ekenel,klee,jmcd}@ira.uka.de

ABSTRACT

In prior work, we proposed using an extended Kalman filter to directly update position estimates in a speaker localization system based on time delays of arrival. We found that such a scheme provided superior tracking quality as compared with the conventional closed-form approximation methods. In this work, we enhance our audio localizer with video information. We propose an algorithm to incorporate detected face positions in different camera views into the Kalman filter without doing any explicit triangulation. This approach yields a robust source localizer that functions reliably both for segments wherein the speaker is silent, which would be detrimental for an audio only tracker, and wherein many faces appear, which would confuse a video only tracker. We tested our algorithm on a data set consisting of seminars held by actual speakers. Our experiments revealed that the audio-video localizer functioned better than a localizer based solely on audio or solely on video features.

1. INTRODUCTION

Most practical acoustic source localization schemes are based on *time delay of arrival estimation* (TDOA) for the following reasons: Such systems are conceptually simple. They are reasonably effective in moderately reverberant environments. Moreover, their low computational complexity makes them well-suited to real-time implementation with several sensors.

Time delay of arrival-based source localization is based on a two-step procedure:

1. The TDOA between all pairs of microphones is estimated, typically by finding the peak in a cross correlation or *generalized cross correlation* function [1].
2. For a given source location, the squared-error is calculated between the estimated TDOAs and those determined from the source location. The estimated source location then corresponds to that position which minimizes this squared error.

If the TDOA estimates are assumed to have a Gaussian-distributed error term, it can be shown that the least squares metric used in Step 2 provides the maximum likelihood (ML) estimate of the speaker location [2]. In prior work [3], we proposed using a variation of a Kalman filter to directly update the speaker position estimate based on the observed TDOAs. In particular, the TDOAs comprise the observation associated with an extended Kalman filter whose state corresponds to the speaker position. Hence, the new position estimate comes directly from the update formulae associated with the Kalman filter. It is worth noting that similar approaches have been proposed by Gannot *et al* [4] for an acoustic source localizer, as well as by Duraiswami *et al* for a combined audio-video source localization algorithm based on a particle filter [5].

In this work, we enhance our audio localizer with video information. We propose an algorithm to incorporate detected face positions in different camera views into the Kalman filter without doing any triangulation. Our algorithm differs from that proposed by Strobel *et al* [6] in that no explicit position estimates are made by the individual sensors. Rather, as in the work of Welch and Bishop [7], the observations of the individual sensors are used to *incrementally* update the state of a Kalman filter. This combined approach yields a robust source localizer that functions reliably both for segments wherein the speaker is silent, which would be detrimental for an audio only tracker, and wherein many faces appear, which would confuse a video only tracker. We tested our algorithm on a data set consisting of seminars held by actual speakers. Our experiments revealed that the audio-video localizer functioned better than a localizer based solely on audio or solely on video features.

The rest of this work is organized as follows. In Section 2, we review the acoustic source localization based on time-delay of arrival estimation. We also discuss how the three-dimensional position of a speaker can be back projected onto the two-dimensional image plane of a camera. We show in both cases that a squared-error criterion can be used as a basis for speaker localization. Section 3 summarizes a variant of the Kalman filter, known as the iterated extended Kalman filter. Section 4 presents a simple model for speaker motion, then discusses how the development in the preceding sections can be combined to develop a localization algorithm capable of tracking a moving speaker. Section 5 presents the results of our initial experiments comparing our combined audio-video localization scheme with localizers based only on audio or only on video features.

2. SOURCE LOCALIZATION

Here we present the audio and video features to be used in our source localization algorithm.

Audio Features

Consider the i-th pair of microphones, and let $\mathbf{m}_{i1}$ and $\mathbf{m}_{i2}$ respectively be the positions of the first and second microphones in the pair. Let $\mathbf{x}$ denote the position of the speaker in $\mathcal{R}^3$. Then the *time delay of arrival* (TDOA) between the two microphones of the

pair can be expressed as

$$T_i(\mathbf{x}) = T(\mathbf{m}_{i1}, \mathbf{m}_{i2}, \mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{m}_{i1}\| - \|\mathbf{x} - \mathbf{m}_{i2}\|}{s} \quad (1)$$

where s is the speed of sound.

For N microphone pairs, audio source localization based on a maximum likelihood (ML) criterion [2] proceeds by minimizing the error function

$$\epsilon(\mathbf{x}) = \sum_{i=0}^{N-1} \frac{1}{\sigma_i^2} \left[\hat{\tau}_i - T_i(\mathbf{x})\right]^2 \quad (2)$$

where $\hat{\tau}_i$ is the observed TDOA for the i-th microphone pair and σ_i^2 is the error covariance associated with this observation. The TDOAs can be estimated with a variety of well-known techniques [1, 8]. Perhaps the most popular method involves the generalized cross correlation (GCC), which can be expressed as

$$R_{12}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{X_1(e^{j\omega\tau}) X_2^*(e^{j\omega\tau})}{|X_1(e^{j\omega\tau}) X_2^*(e^{j\omega\tau})|} e^{j\omega\tau} \, d\omega \quad (3)$$

Typically $R_{12}(\tau)$ is calculated with an inverse FFT. Thereafter, an interpolation is performed to overcome the granularity in the estimate corresponding to the sampling interval [1].

Let us denote the sensor observation error covariance matrix as

$$\mathbf{\Sigma} = \text{diag}\begin{bmatrix}\sigma_0^2 & \sigma_1^2 & \cdots & \sigma_{N-1}^2\end{bmatrix} \quad (4)$$

In [3], it is shown that (2) can be linearized about the prior position estimate $\mathbf{x}(t-1)$ as

$$\epsilon(\mathbf{x}; t) = \left[\bar{\boldsymbol{\tau}}(t) - \mathbf{C}(t)\mathbf{x}\right]^T \mathbf{\Sigma}^{-1} \left[\bar{\boldsymbol{\tau}}(t) - \mathbf{C}(t)\mathbf{x}\right] \quad (5)$$

where the rows of $\mathbf{C}(t)$ are given by

$$\mathbf{c}_i(t) = \left[\nabla_{\mathbf{x}} T_i(\mathbf{x})\right]^T = \frac{1}{s} \cdot \left[\frac{\mathbf{x} - \mathbf{m}_{i1}}{d_{i1}} - \frac{\mathbf{x} - \mathbf{m}_{i2}}{d_{i2}}\right]^T \quad (6)$$

for $d_{ij} = \|\mathbf{x} - \mathbf{m}_{ij}\|$ and the components of $\bar{\boldsymbol{\tau}}(t)$ are

$$\bar{\tau}_i(t) = \hat{\tau}_i(t) - T_i(\mathbf{x}(t-1)) + \mathbf{c}_i(t)\hat{\mathbf{x}}(t-1) \quad (7)$$

Video Features

In order to localize the speaker visually, we minimize a squared-error criterion very much like that in (2). In this case, we seek to minimize the difference between the output of a face detector described in Section 4 and the speaker's predicted position. This *two-dimensional* difference is calculated in the camera's image plane. As shown in Figure 1, the predicted speaker position $\mathbf{x}$ is projected onto the image plane $\mathbf{I}$ of the camera at position $\mathbf{t}$ with focal length $\mathbf{f}$. This results in the image point $\hat{\mathbf{x}}$. We see to minimize the difference between $\hat{\mathbf{x}}$ and the position $\mathbf{y}$ returned by the face detector.

The extrinsic parameters $\mathbf{t}$ and $\mathbf{R}$ define a camera's translation and rotation with respect to the global 3D coordinate frame. To project a point onto the image plane, we also need information about the camera's intrinsic parameters: The camera matrix $\mathbf{P}$ is made up of the focal length f, the sensor pixel size p_x and p_y, and the principle point $\begin{bmatrix}c_x & c_y & 1\end{bmatrix}^T$; see [9]:

$$\mathbf{P} = \begin{pmatrix} \frac{f}{p_x} & 0 & c_x \\ 0 & \frac{f}{p_y} & c_y \\ 0 & 0 & 1 \end{pmatrix} \quad (8)$$

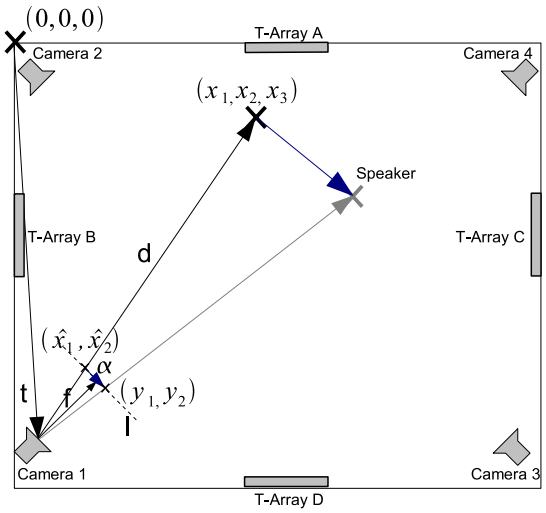

Figure 1: Back projection of the speaker's position onto the image plane of a camera.

Assuming a simple pinhole camera model, we can project the position estimate $\mathbf{x}$ onto the image plane using the projection equation

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{pmatrix} = \mathbf{P}\mathbf{R}^T(\mathbf{x} - \mathbf{t}) \quad (9)$$

and obtain the 2D point

$$f(\mathbf{x}) = \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix} = \begin{pmatrix} \frac{\bar{x}_1}{\bar{x}_3} \\ \frac{\bar{x}_2}{\bar{x}_3} \end{pmatrix} \quad (10)$$

For efficiency, we can precalculate

$$\mathbf{A} = \mathbf{P}\mathbf{R}^T \quad (11)$$

as this term does not change.

As in (6) for the audio features, we need a linearization for this nonlinear projection function. Hence, we take the partial derivative of $f(\mathbf{x})$ with respect to $\mathbf{x}$

$$\mathbf{C} = \nabla_{\mathbf{x}} f(\mathbf{x}) \quad (12)$$

where

$$c_{ij} = \frac{a_{ij} - a_{3j}\hat{x}_i}{\bar{x}_3} \quad (13)$$

for $1 \le i \le 2$, $1 \le j \le 3$ and $\{a_{ij}\}$ are the elements of A.

We now wish to apply the standard Kalman filter update formula directly to recursively estimate a speaker's position based on our audio and video features. Moreover, we want to avoid any closed-form approximation for the speaker position on the audio side, as well as any triangulation on the video side. To see more clearly how such an approach can be implemented, we briefly review a variant of the Kalman filter in Section 3.

3. ITERATED EXTENDED KALMAN FILTER

Let $\mathbf{x}(t)$ denote the *state* of a Kalman filter at time t, and let $\mathbf{y}(t)$ denote the associated observation. Moreover, define a transition matrix $\mathbf{F}(t+1,t)$ which specifies how the state evolves in time, and functional $\mathbf{C}(t,\mathbf{x}(t))$ which specifies how the state is related to the current observation. The Kalman filter is then described by the *process* and *observation* equations:

$$\mathbf{x}(t+1) = \mathbf{F}(t+1,t)\,\mathbf{x}(t) + \boldsymbol{\nu}_1(t) \tag{14}$$

$$\mathbf{y}(t) = \mathbf{C}(t,\mathbf{x}(t)) + \boldsymbol{\nu}_2(t) \tag{15}$$

where $\boldsymbol{\nu}_1(t)$ and $\boldsymbol{\nu}_2(t)$ are the *process* and *observation noise* respectively, which by assumption are zero mean with covariances matrices $\mathbf{Q}_1(t)$ and $\mathbf{Q}_2(t)$.

Let $\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})$ denote the predicted estimated state of a Kalman filter using all the observations $\mathcal{Y}_{t-1} = \{\mathbf{y}(n)\}_{n=0}^{t-1}$ up to time $t-1$. The *innovation* is defined as the difference between the observation $\mathbf{y}(t)$ and the *prediction* $\mathbf{C}(t,\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}))$ at time t:

$$\boldsymbol{\alpha}(t,\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) = \mathbf{y}(t) - \mathbf{C}(t,\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) \tag{16}$$

Using the innovation and the *Kalman gain* $\mathbf{G}_f(t,\mathbf{x}(t|\mathcal{Y}_{t-1}))$, the state estimate can be updated according to [10, §10]

$$\mathbf{x}(t|\mathcal{Y}_t) = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) + \mathbf{G}_f(t,\mathbf{x}(t|\mathcal{Y}_{t-1}))\,\boldsymbol{\alpha}(t,\mathbf{x}(t|\mathcal{Y}_{t-1})) \tag{17}$$

Recursively updating the Kalman gain [10, §10] requires knowledge of the predicted $\mathbf{K}(t,t-1)$ and filtered $\mathbf{K}(t)$ state errors, which are calculated through the *Riccati equation*:

$$\mathbf{K}(t+1,t) = \mathbf{F}(t+1,t)\mathbf{K}(t)\mathbf{F}^T(t+1,t) + \mathbf{Q}_1(t) \tag{18}$$

$$\mathbf{K}(t) = [\mathbf{I} - \mathbf{F}(t,t+1)\mathbf{G}_f(t)\mathbf{C}(\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}))]\,\mathbf{K}(t,t-1)$$

where $\mathbf{C}(\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}))$ is linearization of $\mathbf{C}(t,\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}))$ about $\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})$.

Jazwinski [11, §8.3] describes an *iterated extended Kalman filter* (IEKF), in which (16–17) are replaced with the *local iteration*,

$$\boldsymbol{\alpha}(t,\boldsymbol{\eta}_i) = \mathbf{y}(t) - \mathbf{C}(t,\boldsymbol{\eta}_i) \tag{19}$$

$$\boldsymbol{\zeta}(t,\boldsymbol{\eta}_i) = \boldsymbol{\alpha}(t,\boldsymbol{\eta}_i) - \mathbf{C}(\boldsymbol{\eta}_i)\,[\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) - \boldsymbol{\eta}_i] \tag{20}$$

$$\boldsymbol{\eta}_{i+1} = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) + \mathbf{G}_f(t,\boldsymbol{\eta}_i)\boldsymbol{\zeta}(t,\boldsymbol{\eta}_i) \tag{21}$$

The local iteration is initialized by setting

$$\boldsymbol{\eta}_1 = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) = \mathbf{F}(t,\hat{\mathbf{x}}(t-1|\mathcal{Y}_{t-1}))$$

Note that $\boldsymbol{\eta}_2 = \hat{\mathbf{x}}(t|\mathcal{Y})$ as defined in (17). Hence, if the local iteration is run only once, the IEKF reduces to an extended Kalman filter. Normally (19–20) are repeated, however, until there are no substantial changes between $\boldsymbol{\eta}_i$ and $\boldsymbol{\eta}_{i+1}$. Both $\mathbf{G}_f(t,\boldsymbol{\eta}_i)$ and $\mathbf{C}(\boldsymbol{\eta}_i)$ are updated for each local iteration. After the last iteration, we set $\hat{\mathbf{x}}(t|\mathcal{Y}_t) = \boldsymbol{\eta}_f$. Jazwinski [11, §8.3] reports that the IEKF provides faster convergence in the presence of significant nonlinearities in the observation equation, especially when the initial state $\mathbf{x}(1|\mathcal{Y}_0)$ is far from the optimal value.

4. SPEAKER TRACKING

In this section, we discuss the specifics of how the linearized least squares position estimation criterion (5) can be recursively minimized with the iterated extended Kalman filter presented in the prior section. We begin by associating the observation $\mathbf{y}(t)$ with the TDOA estimate $\boldsymbol{\tau}(t)$ for the audio features, and with the detected face position for the video features. Moreover, we recognize that the linearized observation functional $\mathbf{C}(t)$ required for the Kalman filter is given by (6) and (12) for the audio and video features respectively. Furthermore, we can equate the TDOA error covariance matrix $\boldsymbol{\Sigma}$ in (4) with the observation noise covariance $\mathbf{Q}_2(t)$ and define a similar matrix for the video features. Hence, we have all relations needed on the observation side of the Kalman filter. We need only supplement these with an appropriate model of the speaker's dynamics to develop an algorithm capable of tracking a moving speaker, as opposed to finding his position at a single time instant.

Consider the simplest model of speaker dynamics, wherein the speaker is "stationary" inasmuch as he moves only under the influence of the process noise $\boldsymbol{\nu}_1(t)$. The transition matrix is then $\mathbf{F}(t+1|t) = \mathbf{I}$. Assuming the process noise components in the three directions are statistically independent, we can write

$$\mathbf{Q}_1(t) = \sigma^2 T^2 \mathbf{I} \tag{22}$$

where T is the time since the last state update. Although the audio sampling is synchronous for all sensors, it cannot be assumed that the speaker constantly speaks, nor that all microphones receive the direct signal from the speaker's mouth; i.e., the speaker sometimes turns so that he is no longer facing the microphone array. As only the direct signal is useful for localization [12], the TDOA estimates returned by those sensors receiving only the indirect signal reflected from the walls should not be used for position updates. This is most easily done by setting a threshold on the GCC (3), and using for source localization only those microphone pairs returning a peak in the GCC above the threshold [12]. This implies that no update at all is made if the speaker is not speaking.

The face detector used for visual speaker localization is based on the concept of boosted classifier cascades presented in [13, 14]. In order to be able to detect faces from different views, two separate cascades—one for frontal and one for profile faces—were trained, thus covering a range of $\pm 90^\circ$ horizontal head rotation. In order to reduce the rate of false detections, we maintain an adaptive background model of the scene and ignore detections that are not supported by the foreground-background segmentation.

Performing face detection on an entire video image is very costly. Hence, to keep computational expense within reason, each camera first receives the most recent position estimate from the Kalman filter. This three-dimensional position estimate is then projected onto the camera's image plane, and the face detector searches for a face within a relatively small neighborhood about this point. If a face is discovered within this region, the innovation vector, given by the difference between the projected position estimate and the location of the detected face, is calculated and returned to the Kalman filter. This two-dimensional innovation vector is then used to update the three-dimensional speaker position.

As all video data arrives asynchronously, there is at most a two-dimensional innovation vector available from a single camera to update the three-dimensional speaker position at any given time instant. As Welch and Bishop note [7], this implies the state of the Kalman filter is not *observable* based on the data obtained from any single video sensor. Nonetheless, the state of the Kalman filter can be updated based on a single observation. Moreover, the true state of the Kalman filter is observable when estimates from all sensors, both audio and video, are sequentially combined, subject only to very mild restrictions on the positions of the sensors with respect to the speaker and on the update rate [15].

Tracking Mode	RMS Error (cm)				
	X	Y	Z	2D	3D
audio only	46.7	43.5	22.8	65.1	69.4
video only	101.5	119.3	24.4	162.6	164.6
audio-video	41.4	36.9	12.5	56.0	58.6

Table 1: Results of the source localization experiments: The columns X, Y, Z show the average error (cm) in speaker position for each dimension. 2D and 3D represent the RMS error on the floor plane (X,Y) and the entire space (X,Y,Z) respectively.

5. EXPERIMENTS

The test set used to evaluate the algorithms proposed here contains approximately 2.5 hours of audio and video data recorded during five seminars by students and faculty at the University of Karlsruhe (UKA) in Karlsruhe, Germany. Prior to the start of the seminars, four video cameras in the corners of the room had been calibrated with the technique of Zhang [16].

The location of the centroid of the speaker's head in the images from the four calibrated video cameras was manually marked every 0.7 second. Using these hand-marked labels and the calibration information, the true position of the speaker's head in three dimensions was calculated using the technique described in [17]. These "ground truth" speaker positions are accurate to within 10 cm.

As the seminars took place in an open lab area $5\ \mathrm{m} \times 7\ \mathrm{m}$ used both by seminar participants as well as students and staff engaged in other activities, the recordings are optimally-suited for evaluating source localization and other technologies in a realistic, natural setting. In addition to speech from the seminar speaker, the far field recordings contain noise from fans, computers, and doors, in addition to cross-talk from other people present in the room. For these initial experiments, the seminars were recorded with four T-shaped microphone arrays with four elements each, located on the four walls of the room.

Table 1 shows the results of a set of experiments that were made to compare the accuracy of source localizers running in different modes. The audio-video experiment used the same parameters that were used to run the experiments on a single modality. To initialize the localization algorithm, we used a fixed starting position for all seminars so that the Kalman filter was forced converge to the true position. We filtered the innovation sequence of the Kalman filter, using twice the standard deviation of the innovation covariance matrix as a threshold, in order to remove outliers. The IEKF was iterated at most fives times. As process noise we used (154.62, 184.13, 34.24). We also restricted the position estimates returned by the Kalman filter to be within the physical room and the time delays to be within the bounds determined by the dimensions of the room. Moreover, we set a threshold of 0.18 on the maximum peak of the GCC for each microphone pair and used only those pairs that had correlation values that exceeded that threshold for further position estimation. As source for the audio source localization we used all combinations of microphone pairs of the T-Arrays B and D in Figure 1. The measurement noise for for microphone pairs was held between 0.11 ms to 0.54 ms.

The size of the face detector's search window was determined by the projection of a cube with an edge size of 50cm. Additionally, the state error of the Kalman filter projected into the camera space was added to get a dynamic search window. The measurement noise for the cameras was approximately 20 pixels.

6. ACKNOWLEDGEMENTS

This work was sponsored by the European Union under the integrated project CHIL, *Computers in the Human Interaction Loop*, contract number 506909.

7. REFERENCES

[1] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proc. ICASSP*, vol. II, 1994, pp. 273–6.

[2] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[3] U. Klee, T. Gehrig, and J. McDonough, "Kalman filters for time delay of arrival-based source localization," *EURASIP Special Issue on Multichannel Speech Processing*, submitted for publication.

[4] T. Dvorkin and S. Gannot, "Speaker localization exploiting spatial-temporal information," in *The International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Sep. 2003, pp. 295–298.

[5] R. Duraiswami, D. Zotkin, and L. Davis, "Multimodal 3-d tracking and event detection via the particle filter," in *Workshop on Event Detection in Video, International Conference on Computer Vision*, 2001, pp. 20–27.

[6] N. Strobel, S. Spors, and R. Rabenstein, "Joint audio-video signal processing for object localization and tracking," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Heidelberg, Germany: Springer Verlag, 2001, ch. 10.

[7] G. Welch and G. Bishop, "SCAAT: Incremental tracking with incomplete information," in *Proc. Computer Graphics and Interactive Techniques*, August 1997.

[8] J. Chen, J. Benesty, and Y. A. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech Audio Proc.*, vol. 11, no. 6, pp. 549–57, November 2003.

[9] M. Pollefeys, *Tutorial on 3D Modeling from Images*. Katholieke Universiteit Leuven, 2000.

[10] S. Haykin, *Adaptive Filter Theory*, 4th ed. New York: Prentice Hall, 2002.

[11] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. New York: Academic Press, 1970.

[12] L. Armani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of a CSP-based voice activity detector for distant-talking ASR," in *Proc. Eurospeech*, vol. II, 2003, pp. 501–4.

[13] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in *IEEE ICIP*, 2002.

[14] M. Jones and P. Viola, "Fast multi-view face detection," in *IEEE Conference on Computer Vison and Pattern Recognition*, 2003.

[15] G. F. Welch, "SCAAT: Incremental tracking with incomplete information," Ph.D. dissertation, University of North Carolina, Chapel Hill, NC, 1996.

[16] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Analysis Machine Intel.*, vol. 22, pp. 1330–1334, 2000.

[17] D. Focken and R. Stiefelhagen, "Towards vision-based 3-D people tracking in a smart room," in *IEEE Int. Conf. Multimodal Interfaces*, October 2002.