Advances in Radio Science (2003) 1: 113–117

Advances in Radio Science

Speaker tracking with a microphone array using Kalman filtering

D. Bechler, M. Grimm, and K. Kroschel

Institut für Nachrichtentechnik, Universität Karlsruhe, Kaiserstr. 12, 76128 Karlsruhe, Germany

Abstract. In this publication a method for tracking a speaker with acoustical information by means of a microphone array is presented. A sound source localization algorithm based on the time delays of arrival of sound waves in microphone pairs provides initial position estimates. These significantly varying estimates are spatially filtered by an adaptive Kalman filter to obtain a smoothed trajectory of the speaker's movement. Experimental results with real data are presented for a variety of scenarios recorded in a noisy and reverberant office room environment.

1 Introduction

Integrating acoustic perception by means of a microphone array into autonomous humanoid robots is nowadays an important area of research. The aim of the robot's hearing system is not only to be able to interact with a human operator but also to create an acoustic map of the sound environment and to perform an acoustic scene analysis, i.e. the localization, separation and classification of sound sources present in the acoustic environment. Especially the task of acoustically localizing a speaker in a room environment is of great interest not only in robotics but also for teleconferencing or acoustic surveillance systems.

The technique of choice in most recent acoustic localization systems using microphone arrays is a two-step procedure. First, the time delay of arrival (TDOA) of sound signals in a pair of spatially separated microphones is estimated. In a second step the estimated TDOAs of different microphone pairs are used in combination with the microphone array geometry to localize the sound source. To avoid the computational demanding solution of a set of non-linear equations for the exact sound source position, sub-optimal closed-form localization estimators with satisfactory precision can be applied (DiBiase et al., 2001; Huang et al., 2000).

Correspondence to: D. Bechler
(bechler@int.uni-karlsruhe.de)

Due to the one-sample-precision of the TDOA estimation algorithm and due to noise and reverberation influences, the TDOA estimates and the real TDOA values are not identical which leads to relatively high variances in consecutive position estimates. In this publication a method to smooth the speaker trajectory and to assure robustness by means of an adaptive Kalman filter is evaluated for a real data single speaker scenario in a noisy and reverberant office environment.

The paper is organized as follows: In Sect. 2 the localization of sound sources based on time delay estimates is presented. Filtering the initial estimates of the sound source position by an adaptive Kalman filter is discussed in Sect. 3. Section 4 describes the experimental setup. In Sect. 5 we present and evaluate the obtained results. Finally, some conclusions are drawn and an outlook for future work is given.

2 Time- delay- based localization of sound sources

2.1 Signal model

For a given pair of spatially separated microphones M_i and M_j, the recorded sensor signals $x_i(t)$ and $x_j(t)$ for a signal $s(t)$, emanated from a remote sound source in a reverberant and noisy environment, can be modeled mathematically as

$$\begin{aligned} x_i(t) &= h_i(t) * s(t) + n_i(t) \\ x_j(t) &= h_j(t - \tau_{ij}) * s(t) + n_j(t), \end{aligned} \tag{1}$$

where τ_{ij} represents the relative signal delay of interest, $*$ signifies the convolution operator, $h_i(t)$ is the acoustic impulse response between the sound source and the i^{th} microphone and the additive term $n_i(t)$ summarizes the channel noise in the microphone system as well as environmental noise for the i^{th} sensor. This noise $n_i(t)$ is assumed to be uncorrelated with $s(t)$.

2.2 TDOA estimation with GCC method

The most popular approach for determining the TDOAs is called the Generalized Cross-Correlation (GCC) method Knapp and Carter (1976). The relative time delay τ_{ij} is estimated as the time lag with the global maximum peak in the GCC function $R_{ij}^{(g)}(\tau)$

$$\hat{\tau}_{ij} = \underset{\tau}{\operatorname{argmax}} R_{ij}^{(g)}(\tau). \tag{2}$$

This GCC function $R_{ij}^{(g)}(\tau)$ is defined as

$$R_{ij}^{(g)}(\tau) = \int_{-\infty}^{+\infty} \psi_{ij}(\omega) X_i(\omega) X_j(\omega)^* e^{j\omega\tau} d\omega. \tag{3}$$

The weighting function $\psi_{ij}(\omega)$ intends to decrease noise and reverberation influences and tries to emphasize the GCC value at the true TDOA value τ_{ij}. For real environments the Phase Transform (PHAT) technique (Knapp and Carter, 1976) has shown the best performance. This PHAT weighting function is defined as

$$\psi_{ij}^{PHAT}(\omega) = \frac{1}{|X_i(\omega) X_j(\omega)^*|}. \tag{4}$$

2.3 Confidence criteria for TDOA estimates

To detect outliers in TDOA estimates, two confidence criteria can be used: the value of the maximum peak and the ratio between the 1^{st} and the 2^{nd} peak in the GCC function (Bechler and Kroschel, 2002a,b). These criteria allow a reliability scoring of individual estimates and can be used to reject erroneous measurements. The higher the value of these properties of the GCC function the higher the probability that the TDOA was estimated correctly.

2.4 Data association and clustering of TDOA estimates

With the confidence criteria described in Sect. 2.3, a tradeoff has to be made between a high number of estimates necessary for a continuous source tracking, and a high percentage of correct TDOA estimates, which is crucial for robust source localization. Additionally, the one-sample-precision of the GCC algorithm can be too imprecise, which is problematic for the localization algorithm. As a solution we propose data association and clustering techniques for the TDOA estimates. To initialize an acoustic track, a high reliability is demanded and therefore a high threshold for the value of the confidence criteria is used. Once initialized, this decision threshold can be lowered, as now it is possible to search in a region of interest around the initial TDOA estimate value. Thus, erroneous estimates outside this region of interest can be rejected. In addition, several consecutive TDOA estimates are averaged, which solves the problem of the one-sample-precision. With these data association and clustering techniques, the TDOA estimates become sufficiently accurate for the following localization algorithm to produce robust sound position estimates.

2.5 Localization algorithm

To derive from the TDOAs and the microphone array geometry the source position, the exact localization necessitates solving a set of non-linear equations, which can be computationally demanding. To accelerate the sound source position determination, the One-Step Least-Squares (OSLS) algorithm (Huang et al., 2000) is used. This closed-form location estimator approximates precisely enough accurate the exact solution to the non-linear problem.

3 Spatial filtering

For a continuous source trajectory these initial, significantly varying position estimates obtained by the localization algorithm described in Sect. 2 are spatially smoothed by a Kalman Filter (KF) as post-processing unit.

3.1 Kalman filtering

For the motion of a speaker the time-discrete state space description by means of a state and an observation equation is used. The state equation is defined as

$$\boldsymbol{x}(k) = \mathbf{A}\boldsymbol{x}(k-1) + \mathbf{B}u(k) \tag{5}$$

where $\boldsymbol{x}(k)$ is a state vector at time instant k containing the position, the velocity and the acceleration of the sound source in Cartesian coordinates. The time-invariant state transition matrix $\mathbf{A}$ as well as the variance σ_U^2 of the white noise $u(k)$ representing the system error are assumed to be known. $\mathbf{B}$ is the time-invariant system noise coupling matrix mapping the system error to the elements of the state vector $\boldsymbol{x}(k)$. In the observation equation

$$\boldsymbol{y}(k) = \mathbf{C}\boldsymbol{x}(k) + \boldsymbol{n}(k) \tag{6}$$

$\boldsymbol{y}(k)$ is the observation vector containing the source position, $\mathbf{C}$ is the time-invariant measurement matrix mapping the state of the system to the observation vector and $\boldsymbol{n}(k)$ is white noise with the covariance matrix $\mathbf{R}_{NN}(k)$ representing the measurement error. For our scenario, in which a speaker can move arbitrarily in the acoustical environment, three motion models are reasonable: a static, a constant velocity and a constant acceleration model (for the definitions of the according matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ cf. Grimm, 2001).

3.2 Filter algorithm

One cycle of the KF is done with the following filter algorithm. For a detailed derivation cf. Brown (1983).

In a first step, the updated estimation error covariance matrix $\mathbf{R}_{EE}(k+1|k+1)$ can be derived from the previous covariance matrix $\mathbf{R}_{EE}(k|k)$. Therefore, the *a priori* estimation error covariance matrix $\mathbf{R}_{EE}(k+1|k)$ at the $(k+1)^{th}$ iteration given k observations is calculated with

$$\mathbf{R}_{EE}(k+1|k) = \mathbf{A}\mathbf{R}_{EE}(k|k)\mathbf{A}^T + \mathbf{B}\sigma_U^2\mathbf{B}^T. \tag{7}$$

With Eq. (7), the covariance matrix $\mathbf{R}_{\nu\nu}(k+1)$ of the innovation $\nu(k+1)$ can be calculated, where $\nu(k)$ is defined as the difference between the actual observation vector $\boldsymbol{y}(k)$ and the estimated observation vector at the k^{th} iteration given $k-1$ observations $\hat{\boldsymbol{y}}(k|k-1)$:

$$\nu(k) = \boldsymbol{y}(k) - \hat{\boldsymbol{y}}(k|k-1). \tag{8}$$

$\mathbf{R}_{\nu\nu}(k+1)$ is given by

$$\mathbf{R}_{\nu\nu}(k+1) = \mathbf{C}\mathbf{R}_{EE}(k+1|k)\mathbf{C}^T + \mathbf{R}_{NN}(k). \tag{9}$$

The *a posteriori* estimation error covariance matrix $\mathbf{R}_{EE}(k+1|k+1)$ is then found from

$$\begin{aligned}\mathbf{R}_{EE}(k+1|k+1) &= \mathbf{R}_{EE}(k+1|k) \\ &\quad -\mathbf{W}(k+1)\mathbf{R}_{\nu\nu}(k+1)\mathbf{W}^T(k+1),\end{aligned} \tag{10}$$

where $\mathbf{W}(k+1)$ is the filter gain at instant $k+1$ defined as

$$\mathbf{W}(k+1) = \mathbf{R}_{EE}(k+1|k)\mathbf{C}^T(\mathbf{R}_{\nu\nu}(k+1))^{-1}. \tag{11}$$

Similarly, the estimate of the system state can be iteratively derived from the previous estimate. First, the *a priori* system state estimate

$$\hat{\boldsymbol{x}}(k+1|k) = \mathbf{A}\hat{\boldsymbol{x}}(k|k) \tag{12}$$

is calculated and out of it the source position is estimated with

$$\hat{\boldsymbol{y}}(k+1|k) = \mathbf{C}\hat{\boldsymbol{x}}(k+1|k). \tag{13}$$

By means of innovation and filter gain the actual state estimation results from

$$\hat{\boldsymbol{x}}(k+1|k+1) = \hat{\boldsymbol{x}}(k+1|k) + \mathbf{W}(k+1)\nu(k+1) \tag{14}$$

and the updated source position is found from

$$\hat{\boldsymbol{y}}(k+1|k+1) = \mathbf{C}\hat{\boldsymbol{x}}(k+1|k+1). \tag{15}$$

3.3 Adaptive Kalman Filter

Up to now the different motion model approaches were regarded independently. As the motion dynamics of a speaker in an office environment are variable, a single model for the KF is not suitable for all situations. Thus, a multiple model system would be advantageous. By means of the definition of a model probability described in this section, several models are used simultaneously for the determination of the final position estimate.

3.3.1 Model probability

For L models α_i with $1 \le i \le L$ the conditional probability $f(\alpha_i|\mathbf{y}(k))$ is the probability that the i^{th} model based on the measurements $\mathbf{y}(1), \cdots, \mathbf{y}(k)$ is the one describing the speaker's movement correctly. With the Bayes formula we get

$$f\left(\alpha_i|\boldsymbol{y}(k)\right) = \frac{f\left(\boldsymbol{y}(k)|\alpha_i\right) f\left(\alpha_i\right)}{f(\boldsymbol{y}(k))}. \tag{16}$$

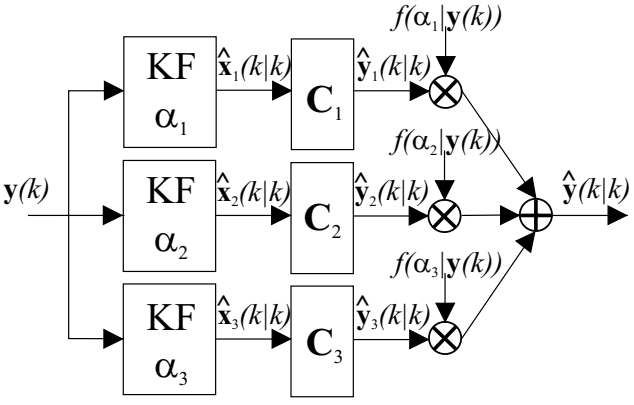

Fig. 1. Block diagram of the adaptive Kalman filter

For this conditional probability a recursive formula can be given:

$$f\left(\alpha_i|\boldsymbol{y}(k)\right) = \frac{f\left(\boldsymbol{y}(k)|\boldsymbol{y}(k-1),\alpha_i\right) f\left(\alpha_i|\boldsymbol{y}(k-1)\right)}{\sum_{j=1}^{L} f\left(\boldsymbol{y}(k)|\boldsymbol{y}(k-1),\alpha_j\right) f\left(\alpha_j|\boldsymbol{y}(k-1)\right)}. \tag{17}$$

The terms $f\left(\boldsymbol{y}(k)|\boldsymbol{y}(k-1),\alpha_i\right)$ can be calculated in every recursive step with

$$\begin{aligned}f\left(\boldsymbol{y}(k)|\boldsymbol{y}(k-1),\alpha_i\right) &= \frac{1}{(2\pi)^{\frac{3}{2}}\sqrt{|\mathbf{R}_{\nu\nu,i}(k)|}} \\ &\quad \cdot \exp\left\{-\tfrac{1}{2}\nu_i^T(k)\mathbf{R}_{\nu\nu,i}^{-1}(k)\nu_i(k)\right\}.\end{aligned} \tag{18}$$

3.3.2 Multiple Model Adaptive Estimator (MMAE)

The approach using several models for the final estimate simultaneously is called adaptive KF or Multiple Model Adaptive Estimator (MMAE) (Bar-Shalom, 1988). In Fig. 1 the block diagram of our adaptive KF is shown. The measurement $\boldsymbol{y}(k)$ is used in three KFs in parallel to determine one estimate of the system state $\hat{\boldsymbol{x}}_1(k|k)$, $\hat{\boldsymbol{x}}_2(k|k)$ and $\hat{\boldsymbol{x}}_3(k|k)$ for every of the three models. As mentioned, we use a static, a constant velocity and a constant acceleration model to describe the motion of a speaker. By means of the observation matrices $\mathbf{C}_1, \mathbf{C}_2$ and $\mathbf{C}_3$ the system state estimates are mapped to the observation estimates $\hat{\boldsymbol{y}}_1(k|k)$, $\hat{\boldsymbol{y}}_2(k|k)$ and $\hat{\boldsymbol{y}}_3(k|k)$. These position estimates are weighted with the corresponding model probabilities $f(\alpha_i|\boldsymbol{y}(k))$ according to (Eq. 17) with $1 \le i \le 3$ and finally summed up. Hence, the actual position estimate for the adaptive KF is determined with

$$\hat{\boldsymbol{y}}_{MMAE}(k|k) = \sum_{i=1}^{3} f(\alpha_i|\boldsymbol{y}(k)\mathbf{C}_i\hat{\boldsymbol{x}}_i(k|k). \tag{19}$$

4 Experimental setup

Real data experiments have been carried out in a (5 m × 5 m × 3 m) office room with typical environmental noise (fans, ...) and relatively strong reverberation. For

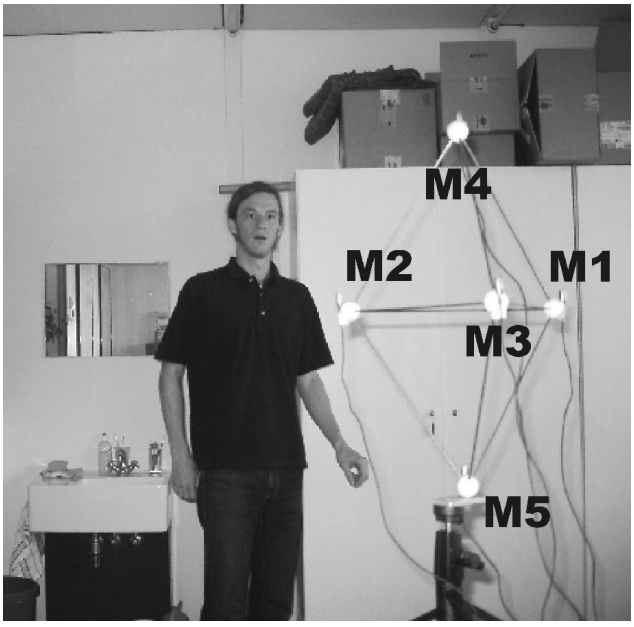

Fig. 2. Experimental setup

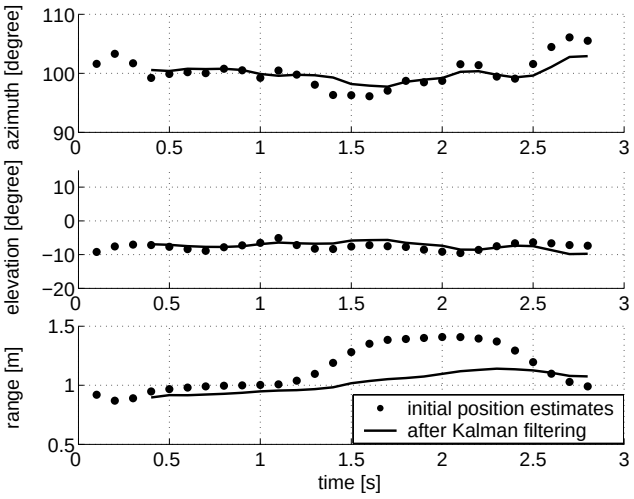

Fig. 3. Stationary speaker at $azimuth = 95°$, $elevation = -7.9°$ and $range = 1.5$ m

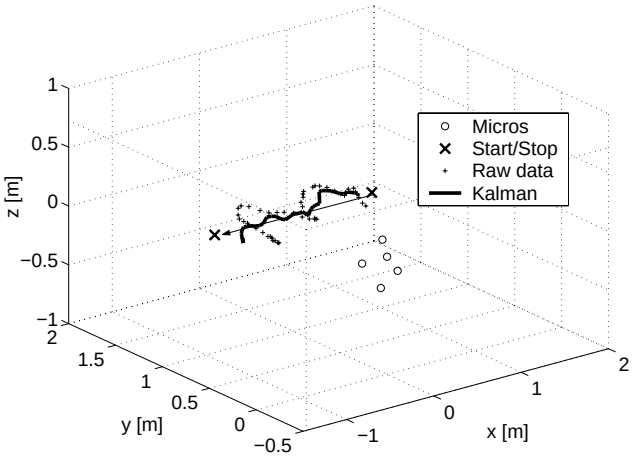

Fig. 4. Speaker moving from Cartesian coordinates $(-1m, 1m, 0.13m)$ to $(1m, 1m, 0.13m)$

data recording we used a 5-microphone array in an equilateral double-tetrahedron geometry with a side length of $D = 28$ cm as shown in Fig. 2. 25 scenarios were recorded in which a speaker is uttering German sentences. In addition to stationary positions for a sitting or standing speaker, motions along given trajectories with constant velocity and transitions from stationary positions to constant walking were studied. The sampling frequency was $f_s = 16$ kHz. The recorded data were analyzed in frames of 32 ms to assure quasi-stationarity. For this data segmentation, a Hamming window with a 50% overlap was applied. For the data association and clustering techniques described in Sect. 2.4, the average of the TDOA estimates in a block out of 12 consecutive frames is calculated where blocks are overlapped by 50%. This generates a new position estimate for the sound source every 96 ms. With this averaging, robust speaker tracking is still possible since at least four estimates per second are required (Silverman and Kirtman, 1992) for a moving speaker.

5 Results

The proposed system shows robust speaker localization capabilities in a noisy and reverberant environment. The adaptive KF performs its task as post-processing unit in guaranteeing a smoothed continuous trajectory compared with the initial position estimates.

Exemplarily, Fig. 3 shows a comparison of the localization of a stationary speaker at the spherical coordinates $azimuth = 95°$, $elevation = -7.9°$ and $range = 1.5$ m before and after Kalman filtering. The temporal runs of all spherical coordinates are significantly smoothed. Note that the variance in range is reduced, but the overall range error increases by applying a KF in this case. Until 1.2 s the localization algorithm underestimates the range by about 50 cm and the KF evidently cannot remedy this systematic error.

An example for tracking a speaker uttering a German sentence while moving with constant velocity from (−1 m,1 m,0.13 m) to (1 m,1 m,0.13 m) is presented in Fig. 4. In this 3D-plot the true trajectory (arrow) and the positional estimates before and after applying the adaptive KF for a walking speaker are shown. Also in this example the KF functions reliably and delivers a smoothed speaker trajectory. As for this scenario there is no systematic error in the initial position estimates the overall error in all coordinates is reduced.

6 Conclusions and outlook

The good results for the determination of a smoothed continuous trajectory from the initial sound source positions jus-

tify the complexity of our adaptive Kalman filter as post-processing unit. But the reduction of the overall error of position estimates with an adaptive KF highly depends on the quality of the initial estimates used as KF input. If these estimates show low systematic errors, the KF delivers an efficient smoothing of the initial trajectory. Even if there are erroneous TDOA estimates in only one of the 4 microphone pairs, this can lead to wrong initial position estimates by the localization algorithm, which renders the operation of the filter unreliable. Therefore future work will be devoted to the improvement of the TDOA reliability by considering psychoacoustics, e.g. the precedence effect (Litovsky et al., 1999) to suppress reverberation influences. An additional advantage of the KF lies in its ability to predict source positions in case of missing current position estimates due to speech pauses or non-reliable TDOA estimates.

In this work only one speaker was present in the acoustic scene. Current investigations are concerned with the simultaneous tracking of multiple active sound sources requiring an according number of KFs working in parallel. Furthermore, it is aimed to extend this tracking system by cameras. With the fusion of acoustic and visual data strong improvements should be achieved compared with a pure acoustic tracker.

Acknowledgements. This work is part of the Sonderforschungsbereich (SFB) No. 588 *"Humanoide Roboter - Lernende und kooperierende multimodale Roboter"* at the University of Karlsruhe. The SFB is supported by the Deutsche Forschungsgemeinschaft (DFG).

References

Bechler, D. and Kroschel, K.: Confidence scoring of time difference of arrival estimation for speaker localization with microphone arrays, 13. Konferenz Elektronische Sprachsignalverarbeitung ESSV, September 2002a.

Bechler, D. and Kroschel, K.: Reliability measurement of time difference of arrival estimations for multiple sound source localization, 17th Annual Meeting of the IAR, November 2002b.

Brown, R. G.: Introduction to Random Signal Analysis and Kalman Filtering, Wiley, 1983.

Bar-Shalom Y.: Tracking and data association, Academic Press, 1988.

DiBiase, J. H., Silverman, H. F., and Brandstein, M. S.: Microphone Arrays, chapter Robust Localization in Reverberant Rooms, Springer, 2001.

Grimm, M.: Passives Audio-Tracking sich bewegender Geräuschquellen, Studienarbeit, Institut für Nachrichtentechnik, Universität Karlsruhe, 2001.

Huang, Y., Benesty, J., and Elko G. W.: Passive acoustic source localization for video camera steering, IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 909–912, June 2000.

Knapp, C. H. and Carter, G. C.: The generalized correlation method for estimation of time delay, IEEE Trans. on Acoustics, Speech and Signal Processing, 24(4):320–327, August 1976.

Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J.: The precedence effect, Journal of the Acoustical Society of America, 106(4):1633–1654, October 1999.

Silverman, H. F. and Kirtman, S. E.: A two-stage algorithm for determining talker location from linear microphone array data, Computer, Speech and Language, 6(2):129–152, 1992.