SIMULTANEOUS LOCALIZATION OF MULTIPLE SOUND SOURCES USING BLIND ADAPTIVE MIMO FILTERING

Herbert Buchner, Robert Aichner, Jochen Stenglein, Heinz Teutsch, and Walter Kellermann

Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen, Germany
{buchner,aichner,teutsch,wk}@LNT.de

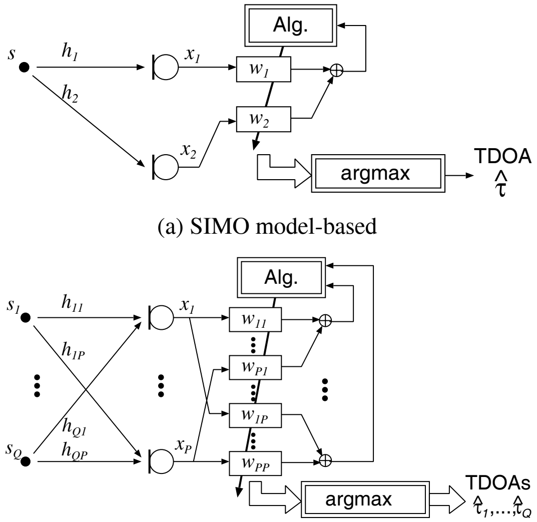
ABSTRACT

Blind adaptive filtering for time delay of arrival (TDOA) estimation is a very powerful method for acoustic source localization in reverberant environments with broadband signals like speech. Based on a recently presented generic framework for blind signal processing for convolutive mixtures, called TRINICON, we present a TDOA estimation method for simultaneous multidimensional localization of multiple sources. Moreover, an interesting link to the known single-input multiple-output (SIMO)-based adaptive eigenvalue decomposition (AED) method is shown. We evaluate the novel multiple-input multiple-output (MIMO)-based approach and compare it with the known SIMO-based method in a reverberant acoustic environment using reference data of the positions obtained from infrared sensors. The results show that the new approach is very robust against reverberation and background noise.

1. INTRODUCTION

A widely used approach to estimate multidimensional source positions in a flexible way is to employ a two-step procedure [1]: In the first step, a set of TDOAs are estimated using measurements across various combinations of microphones. The second step then determines the source positions by geometric considerations. The key to the effectiveness of many localizers is thus an accurate and robust TDOA estimator. The most widely used and conceptually simple method for this is to use the generalized cross-correlation function (GCC) [2]. However, in real acoustic environments, the room reverberation and the background noise are two major sources of signal degradation severely affecting the estimation results. In particular, since most methods, such as GCC, are inherently based on a free-field propagation model, they can lead to good results if the room reverberation is not very long but their performance may break down in highly reverberant environments [3, 4].

To address this reverberation problem, a completely different approach of TDOA estimation based on blind adaptive filtering was proposed in [5]. This so-called adaptive eigenvalue decomposition (AED) algorithm may be seen as a major advance in localization since it focuses directly on the impulse responses h_1 and h_2 (assumed to be FIR) between a source s and the microphones, and thus, this approach is inherently based on the real reverberant propagation model (Fig. 1 (a)). TDOA estimation with AED is based on blind single-input multiple-output (SIMO) system identification by exploiting the linear relation, $x_1(n) * h_2(n) - x_2(n) * h_1(n) = 0$, between the microphone signals due to the common sound source s, as first proposed in [6, 7]. By minimizing the mean square of the output signal



(a) SIMO model-based

(b) MIMO model-based

Fig. 1. TDOA estimation via (a) SIMO and (b) MIMO model-based blind adaptive filtering.

$$e(n) = s(n) * (h_1(n) * w_1(n) + h_2(n) * w_2(n)) \tag{1}$$

by an adaptive algorithm, we ideally obtain independently of $s(n)$

$$h_1(n) * w_1(n) = -h_2(n) * w_2(n) \tag{2}$$

so that with proper initialization [5] and filter length the estimated filters are (see also Fig. 1 (a)) $w_1(n) = c \cdot \hat{h}_2(n)$ and $w_2(n) = -c \cdot \hat{h}_1(n)$ with a scaling constant c. The TDOA can then be calculated from these filters by

$$\begin{aligned} \hat{\tau} &= \arg\max_n |\hat{h}_2(n)| - \arg\max_n |\hat{h}_1(n)| \\ &= \arg\max_n |w_1(n)| - \arg\max_n |w_2(n)|. \end{aligned} \tag{3}$$

Note that we consider here in this paper only the use of microphone pairs. However, there are generalizations of GCC [8], AED [7, 8] (and also of the proposed approach in Sect. 2) to more sensors to further improve the robustness by this redundancy.

Motivated by the high accuracy of the above-mentioned adaptive SIMO filtering approach for single source localization, the

0-7803-8874-7/05/$20.00 ©2005 IEEE

objective of this paper is to study blind adaptive MIMO filtering for simultaneous localization of *multiple sources* in a similar way in order to maintain the realistic reverberant propagation model. Fig. 1 (b) shows the corresponding MIMO-based structure. As with AED, we would like to calculate the Q TDOAs for the Q sources from the estimated FIR filters w_{pq}. Thereby, we assume in this paper that we have at least as many microphones as sources, i.e., $P \geq Q$. We further assume that the sources are mutually uncorrelated. In general this assumption holds for speech and audio signals.

Several different algorithms for blind adaptive MIMO filtering have been proposed in the literature. In the following, we consider in this paper the TRINICON framework, first proposed in [10] as a very general and versatile tool for blind adaptive MIMO filtering, suitable for broadband applications such as speech and audio. Many known and novel algorithms follow from this information-theoretic framework as special cases, and so far, promising results have been obtained for blind source separation (BSS), blind deconvolution, and also blind dereverberation of speech signals [10, 11, 12]. For BSS, a real-time broadband implementation for reverberant environments has been presented in [13]. Moreover, it has been shown that this system exhibits a high robustness against diffuse background noise [14]. These properties make the approach attractive for localization, as shown in the next section. The close relation between BSS and source localization can be motivated by considering BSS as a set of 'blind adaptive beamformers' which inherently include automatic beam steering to the sources without any prior information. As in case of AED, the MIMO-based approach yields robust solutions to the TDOA estimation problem when the channels do not share common zeros. Using multichannel techniques this is well fulfilled in acoustic environments [15]. Moreover, we will see in Sect. 2.3 that there is indeed a very close relation to the SIMO-based TDOA approach.

2. TRINICON-BASED TDOA ESTIMATION FOR PASSIVE ACOUSTIC SOURCE LOCALIZATION

In the context of independent component analysis (ICA), different approaches exist to blindly estimate the MIMO filtering matrix $\mathbf{W}$ according to Fig. 1 (b) for the above mentioned tasks by criteria other than traditional MSE. In general, the following source signal properties can be utilized [9] which are all jointly taken into account in TRINICON:

(i) Nongaussianity is exploited by using higher-order statistics. A fundamental and versatile criterion is to minimize the Kullback-Leibler distance [16] between the joint pdf of the estimated output signals and a certain *desired* joint pdf (essentially a hypothesized stochastic source model [10]). By factorizing the desired pdf w.r.t. the different sources and maintaining the temporal dependencies within individual channels, the minimum mutual information (MMI) criterion follows. MMI is known as the most general approach for BSS based on ICA. MMI also appears to be a suitable MIMO generalization of the SIMO-based approach in Fig. 1 (a). In both cases, the filters are designed so that, ideally, the contributions of (different) sources are (mutually) cancelled out in the output signal(s). In the case of AED, this is expressed by (2).

(ii) Nonwhiteness is exploited by simultaneous minimization of output cross-relations over multiple time-lags. We therefore consider multivariate pdfs, i.e., 'densities including D time-lags'.

(iii) Nonstationarity is exploited by simultaneous minimization of output cross-relations at different time-instants. We assume ergodicity within blocks of length N so that the ensemble average is replaced by time averages over these blocks.

2.1. Matrix notation for convolutive mixtures

To express the algorithm for block processing of convolutive mixtures in a general way, we first need to formulate the convolution of the FIR demixing system of length L in the following convenient matrix form [11]:

$$\mathbf{y}(m,j) = \mathbf{x}(m,j)\mathbf{W}(m), \tag{4}$$

where m denotes the block index, and $j = 0, \cdots, N-1$ is a time-shift index within a block of length N, and

$$\begin{aligned}
\mathbf{x}(m,j) &= [\mathbf{x}_1(m,j), \ldots, \mathbf{x}_P(m,j)], & (5)\\
\mathbf{y}(m,j) &= [\mathbf{y}_1(m,j), \ldots, \mathbf{y}_P(m,j)], & (6)\\
\mathbf{W}(m) &= \begin{bmatrix} \mathbf{W}_{11}(m) & \cdots & \mathbf{W}_{1P}(m) \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{P1}(m) & \cdots & \mathbf{W}_{PP}(m) \end{bmatrix}, & (7)\\
\mathbf{x}_p(m,j) &= [x_p(mL+j), \ldots, x_p(mL-2L+1+j)], & (8)\\
\mathbf{y}_q(m,j) &= [y_q(mL+j), \ldots, y_q(mL-D+1+j)] & (9)\\
&= \sum_{p=1}^{P} \mathbf{x}_p(m,j)\mathbf{W}_{pq}(m). & (10)
\end{aligned}$$

In (9), D denotes the number of time lags taken into account to exploit the nonwhiteness of the source signals as shown below. $\mathbf{W}_{pq}(m)$ denotes a $2L \times D$ Sylvester matrix that contains all coefficients of the respective filter:

$$\mathbf{W}_{pq}(m) = \begin{bmatrix}
w_{pq,0} & 0 & \cdots & 0 \\
w_{pq,1} & w_{pq,0} & \ddots & \vdots \\
\vdots & w_{pq,1} & \ddots & 0 \\
w_{pq,L-1} & \vdots & \ddots & w_{pq,0} \\
0 & w_{pq,L-1} & \ddots & w_{pq,1} \\
\vdots & & \ddots & \vdots \\
0 & \cdots & 0 & w_{pq,L-1} \\
0 & \cdots & 0 & 0 \\
\vdots & & \vdots & \vdots \\
0 & \cdots & 0 & 0
\end{bmatrix}. \tag{11}$$

2.2. Coefficient optimization

2.2.1. General Coefficient Update

Based on the *natural gradient* [9] of the cost function in [10], the generic TRINICON-based update rule reads:

$$\begin{aligned}
\mathbf{W}(m) &= \mathbf{W}(m-1) - \mu\Delta\mathbf{W}(m), & (12)\\
\Delta\mathbf{W}(m) &= \frac{2}{N}\sum_{i=0}^{\infty}\beta(i,m)\sum_{j=0}^{N-1}\mathbf{W}(i) \\
&\quad \cdot \left\{\mathbf{y}^H(i,j)\mathbf{\Phi}_{s,PD}(\mathbf{y}(i,j)) - \mathbf{I}\right\}, & (13)
\end{aligned}$$

with the score function

$$\mathbf{\Phi}_{s,PD}(\mathbf{y}(i,j)) = -\frac{\partial \log \hat{p}_{s,PD}(\mathbf{y}(i,j))}{\partial \mathbf{y}(i,j)} \tag{14}$$

resulting from the hypothesized source model (a factorization of $\hat{p}_{s,PD}(\cdot)$ among the sources yields BSS, while further factorizations lead to efficient dereverberation algorithms [10]). In (13), β is a window function with finite support that is normalized according to $\sum_{i=0}^{m} \beta(i,m) = 1$ allowing for online, offline, and block-online algorithms [11, 13].

2.2.2. *Sylvester constraint and coefficient initialization*

When implementing the update rule (13) the Sylvester structure of $\Delta\mathbf{W}$ has to be ensured by enforcing a Sylvester constraint $\mathcal{SC}$. Two possible ways to do so are to select the L filter taps in the first column or to select the L-th row of $\Delta\mathbf{W}$ as a reference and generate the Sylvester structure (11) from it. We denote these two special types of constraints as $\mathcal{SC}_C$ and $\mathcal{SC}_R$, respectively. As shown in [12], the choice of $\mathcal{SC}$ affects the way how the natural gradient is implemented, and selects certain subclasses of TRINICON which can be related after certain approximations to some known algorithms from the literature. In particular, it can be shown that $\mathcal{SC}_C$ leads to very robust algorithms if the different sources in Fig. 1 (b) are located in different half planes w.r.t. the microphone array (as, e.g., in car environments). However, in general, if we do not have such a prior knowledge, $\mathcal{SC}_R$ as also used for the experiments in Sect. 3 is preferable.

The choice of $\mathcal{SC}$ also affects the applicable methods for coefficient initialization. For $\mathcal{SC}_R$ a proper way is to set all coefficients, except for $w_{pp,L/2} = 1$ to zero, whereas for $\mathcal{SC}_C$, we have to set $w_{pp,0} = 1$ instead.

2.2.3. *Realizations based on Second-Order Statistics (SOS)*

Here, the source models are simplified to sequences of multivariate Gaussian pdfs described by $PD \times PD$ correlation matrices $\hat{\mathbf{R}}_{\cdot\cdot}$ within the length-N signal blocks. This leads to the coefficient update

$$\Delta\mathbf{W}(m) = 2\sum_{i=0}^{\infty} \beta(i,m)\mathbf{W}(i)\left\{\hat{\mathbf{R}}_{\mathbf{yy}}(i) - \hat{\mathbf{R}}_{\mathbf{ss}}(i)\right\}\hat{\mathbf{R}}_{\mathbf{ss}}^{-1}(i). \tag{15}$$

The BSS variant of this generic SOS natural gradient update also used for multiple TDOA estimation in Sect. 3 follows immediately by setting

$$\hat{\mathbf{R}}_{\mathbf{ss}}(i) = \mathrm{bdiag}_D\, \hat{\mathbf{R}}_{\mathbf{yy}}(i). \tag{16}$$

This algorithm leads to very robust practical solutions even for a large number of filter taps due to an inherent normalization by the auto-correlation matrices, reflected by the inverse of $\mathrm{bdiag}_D\, \hat{\mathbf{R}}_{\mathbf{yy}}$ in (15). Note that there are also efficient approximations of this algorithm with a reduced computational complexity allowing already real-time operation on a regular PC platform [13]. Similarly to the choice of $\mathcal{SC}$ in Sect. 2.2.2, the definition of the estimation method for the correlation matrices (covariance vs. correlation method) is an important practical aspect. For complexity reasons, we choose the correlation method for the evaluation in Sect. 3. Further details on this implementation can be found in [13].

2.3. TDOA estimation and relation to SIMO-based approach

To obtain TDOA estimates for the Q sources from the MIMO filter matrix $\mathbf{W}$ according to Fig. 1 (b), several different methods are conceivable. Since $\mathbf{W}$ can be considered as an estimate of the inverse of the MIMO system of room impulse responses (at least in case of deconvolution), one could try to invert the result again, followed by a detection of the peaks in a similar way as outlined in Sect. 1. However, from a complexity point of view using an inversion may not be the desired method.

Here, we consider another way, based on a very interesting relationship to the SIMO-based method (and blind system identification in general [17]). Let us define a mixing matrix $\mathbf{H}$ with Sylvester structure in the same way as shown in (7), (11) for $\mathbf{W}$. With compatible dimensions, the corresponding matrix for the overall system $\mathbf{C}$ can then be expressed as $\mathbf{C} = \mathbf{HW}$ [11]. It can be shown by simple considerations (or directly from the TRINICON update for all types of applications) that ideally, upon convergence, we have [11]

$$\mathrm{boff}\{\mathbf{C}\} = \mathbf{0}, \tag{17}$$

except a channel-wise (external) permutation ambiguity which can be easily resolved in the localization application (not to be confused with the internal permutation ambiguity, known from narrowband approaches for BSS). boff denotes the blockwise off-diagonal elements of a matrix. For simplicity, we rewrite this set of equations (17) for the case of two sources and two microphones. Expressed by the convolution operator, we obtain

$$h_{11} * w_{12} \quad = \quad -h_{12} * w_{22} \tag{18}$$

$$h_{21} * w_{11} \quad = \quad -h_{22} * w_{21} \tag{19}$$

By comparing Fig. 1 (a) and (b), we see that (18) and (19) can in fact directly be considered as the generalization of the AED condition (2) to multiple sources. Using a proper coefficient initialization, (18) is the corresponding equation to estimate the TDOA of source 1, while (19) gives the TDOA of source 2. Moreover, since the coefficient initialization, described in Sect. 2.2.2, also corresponds to the one recommended for the AED in [5], we can expect similar steady-state performances due to this close link. This is verified in Sect. 3. From these findings, we can express the TDOA estimates immediately in the same way as in (3) as

$$\hat{\tau}_1 \quad = \quad \arg\max_n |w_{12,n}| - \arg\max_n |w_{22,n}|, \tag{20}$$

$$\hat{\tau}_2 \quad = \quad \arg\max_n |w_{11,n}| - \arg\max_n |w_{21,n}|. \tag{21}$$

3. EXPERIMENTAL RESULTS

The audio data used for the evaluation have been recorded at a sampling rate of 48 kHz in a TV studio with a reverberation time of $T_{60} \approx 700$ ms. These data are made available as part of an audio-visual database [18]. This database also includes reference data of the speaker positions measured using infrared sensors. From the reference positions reference TDOAs are calculated by geometric considerations. This allows us to consider both, fixed and moving speakers in a real acoustic environment. From the database, we chose two scenes in the same environment with one fixed and one moving source, respectively. Those are used separately for the SIMO-based approaches, and a superposition (Fig. 2) is used for the MIMO-based approach. The distance between the two microphones was 16 cm. For the adaptation algorithms, the filter lengths were optimized leading to 1024 for the SIMO case (AED) and to 256 for the MIMO case (BSS, (15)). The block length for the GCC (using a phase-transform (PHAT) weighting rule [2]) has been set to 1024. GCC and AED have been complemented by a voice-activity detector. Fig. 3 (a) and (b) show the reference and estimated TDOAs for the fixed and the moving speakers, respectively. In these first experiments, only one speaker was active (also in case

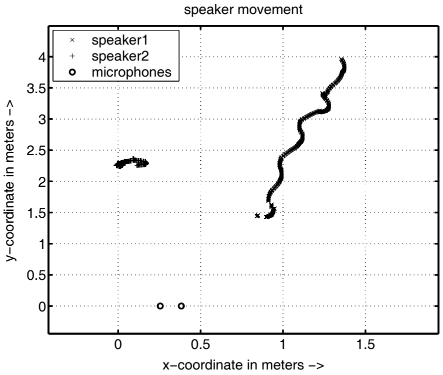
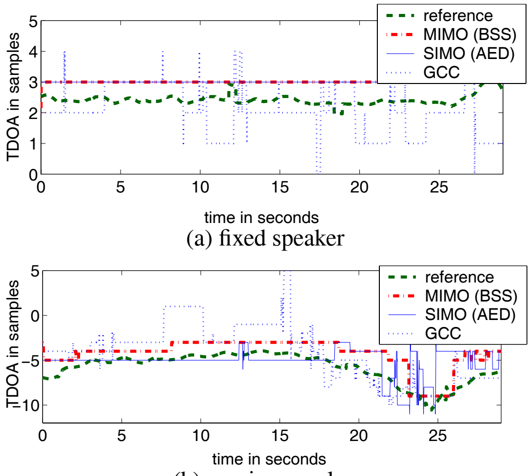


Fig. 2. Scenario used for the simulations.

of the MIMO-based approach). Subplot (a) confirms that both of the blind adaptation algorithms lead to the same accurate TDOA estimates in this static case, as expected from the considerations in Sect. 2.3. Note that the TDOA estimates can only attain integer values. In Fig. 4 we consider the simultaneous estimation of two TDOAs by the proposed MIMO approach. Due to the scenario in Fig. 2 the two TDOAs exhibit different signs. The estimates deviate only slightly from the corresponding results of the MIMO-based approach in Fig. 3 (a) and (b) during some very short time intervals. This may be explained by the different speech activity of the two sources which is inevitable. However, the short peaks in Fig. 4 may be easily removed by appropriate post-processing.



(a) fixed speaker

(b) moving speaker

Fig. 3. TDOA estimation for one source.

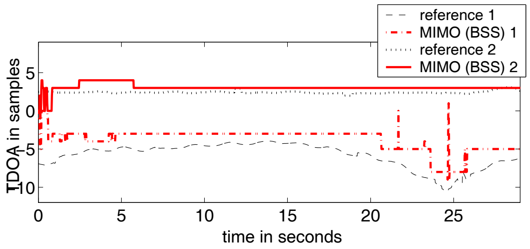


Fig. 4. Simultaneous TDOA estimation for two sources.

4. CONCLUSIONS

We presented a novel approach for simultaneous estimation of multiple TDOAs based on blind adaptive MIMO filtering. For the adaptation, no voice activity detection is required. Some similarities to the known SIMO-based approach have been identified. The experimental results show a robust performance in reverberant environments. Based on the experience with the closely related broadband BSS application, a robust localization in environments with diffuse noise can also be expected.

5. REFERENCES

[1] M.S. Brandstein and D.B. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, Berlin, 2001.

[2] C.H. Knapp and G.C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 320-327, Aug. 1976.

[3] B. Champagne, S. Bedard, and A. Stephene,“Performance of time-delay estimation in the presence of room reverberation,” *IEEE Trans. Speech Audio Processing,* vol. 4, pp. 148-152, Mar. 1996.

[4] J.P. Ianniello, “Time delay estimation via cross-correlation in the presence of large estimation errors,” *IEEE Trans Acoust Speech Signal Processing,* vol. ASSP-30, no. 6, pp. 998-1003, Dec. 1982.

[5] J. Benesty, “Adaptive Eigenvalue Decomposition Algorithm for passive acoustic source localization,” *J. Acoust. Soc. Am.,* vol. 107, pp. 384-391, Jan. 2000.

[6] H. Liu, G. Xu, and L. Tong, “A deterministic approach to blind identification of multi-channel FIR systems,” *Proc. IEEE ICASSP*, Adelaide, Australia, Apr. 1994.

[7] M.I. Gürelli and C.L. Nikias, “EVAM: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals,” *IEEE Trans Signal Processing,* vol. 43, no. 1, pp. 134-149, Jan. 1995.

[8] J. Chen, Y. Huang, and J. Benesty, “Time delay estimation” in Y. Huang and J. Benesty (eds.), *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Kluwer, 2004.

[9] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley & Sons, Inc., New York, 2001.

[10] H. Buchner, R. Aichner, and W. Kellermann, “TRINICON: A versatile framework for multichannel blind signal processing,” in *Proc. IEEE ICASSP*, Montreal, Canada, May 2004, vol. 3, pp. 889-892.

[11] H. Buchner, R. Aichner, and W. Kellermann, “Blind Source Separation for Convolutive Mixtures: A Unified Treatment,” in Y. Huang and J. Benesty (eds.), *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Kluwer, Boston, 2004.

[12] R. Aichner, H. Buchner, and W. Kellermann, “On the causality problem in time-domain blind source separation and deconvolution algorithms,” in *Proc. IEEE ICASSP*, Philadelphia, PA, USA, Mar. 2005.

[13] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, “Real-time convolutive blind source separation based on a broadband approach,” in *Proc. Int. Symp. ICA*, Granada, Spain, Sept. 2004, pp. 833-840.

[14] R. Aichner, H. Buchner, and W. Kellermann, “Convolutive blind source separation for noisy mixtures,” *Proc. Joint Meeting of the German and the French Acoustical Societies (CFA/DAGA)*, March 2004.

[15] M. Miyoshi and Y. Kaneda, “Inverse filtering of room acoustics,” *IEEE Trans. on Acoust., Speech and Audio Processing*, vol 36, no. 2, pp. 145-152, Feb. 1988.

[16] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley & Sons, New York, 1991.

[17] H. Buchner, R. Aichner, and W. Kellermann, “Relation between blind system identification and convolutive blind source separation,” in *Proc. Joint Workshop on Hands-Free Communication and Microphone Arrays,* Piscataway, NJ, Mar. 2005.

[18] M. Krinidis, G. Stamou, H. Teutsch, S. Spors, N. Nikolaidis, R. Rabenstein, I. Pitas, ”An audio-visual database for evaluating person tracking algorithms,” in *Proc. IEEE ICASSP*, 2005.