2D Audio-Visual Localization in Home Environments using a Particle Filter

Stephan Gerlach[1], *Stefan Goetze*[1], *Simon Doclo*[1,2]

[1]Project group Hearing, Speech and Audio Technology, Fraunhofer Institute for Digital Media Technology (IDMT), 26129 Oldenburg, Germany
[2]University of Oldenburg, Institute of Physics, Signal Processing Group, 26111 Oldenburg, Germany
Email: {stephan.gerlach,s.goetze}@idmt.fraunhofer.de, simon.doclo@uni-oldenburg.de
Web: http://www.idmt.fraunhofer.de/hsa

Abstract

Multimodal algorithms benefit from the advantage that they can mutually compensate the weaknesses of the individual modalities. Therefore, we propose a system to localize concurrent speakers in a two dimensional (2D) space jointly using a combined audio-visual localization algorithm. The acoustic source localization is calculated by the multichannel cross-correlation coefficient (MCCC) algorithm and the visual localization is accomplished by the SHORE^{TM,1}, video localization system. The multimodal fusion is performed by a particle filter with adaptations to the particle weighting. An evaluation of the proposed algorithm in an home-environment living lab is performed focussing on possible gains obtained by the complementary localization modalities.

1 Introduction

Source localization is of high interest in various application areas including, e.g., video conferencing or the emerging field of ambient assistive technologies in home environments which gain importance due to demographic changes [1]. Often localization is performed on a single modality using video or audio sensors only. Each modality has its own strengths and weaknesses and it is apparent that a combined usage can compensate the respective weaknesses. However, only a minority of publications address the combined usage. Approaches for joint audio-visual localization algorithms make use of, e.g., Kalman filtering [2], neuronal networks [3], Baysian networks [4] or particle swarm optimization [5] as well as particle filtering [6–9]. In this contribution we examine different weighting methods for a particle filter to estimate the 2D position of speakers. The acoustic modality is supported by visual localization to increase the overall performance. Author Pnevmatikakis et al. proposed in [6] to use separate particle filters for both modalities and fuse the separate position estimates in a geometric approach afterwards. For this approach an estimate from both modalities, audio and video is mandatory. We will use only one single particle filter for both modalities and, by this, inherently data-fusion is also done in the particle filter as in [7–9]. To obtain a fused localization in [8, 9] a multiplication of the complementary probabilities is used. The authors in [8] go beyond the purely combined localization and propose self-calibration of sensors and high-level semantic analysis of the observed scene. In [7] the authors propose an adaptive weighting based on a degree of certainty in the acoustic measurement (c.f. Section 3). Papers [7, 8] use multiple cameras, while this paper focuses on a simple sensor setup with only a single camera.

[1]Sophisticated High-speed Object Recognition Engine (SHORE). Trademark of Fraunhofer IIS, 91058 Erlangen (Germany)

In the following we introduce the individual localization algorithms in Section 2 succeeded by the particle filter and the proposed adjustments to it in Section 3. The contribution is completed with an evaluation of the algorithm in Section. 4 and a conclusion in Section. 5, respectively.

2 Localization

2.1 Audio Localization

For the 2D acoustic source localization we use a time difference of arrival (TDOA) based algorithm, referred to as multichannel cross-correlation (MCCC) algorithm [10, 11],

$$\rho(\theta,l) = 1 - \det(\mathbf{R}(\theta,l)), \qquad 0 \leq \det(\mathbf{R}(\theta,l)) \leq 1, \tag{1}$$

where $\rho(\theta,l)$ can be interpreted as a two dimensional spatial map of correlation values at time block l, $\det(\mathbf{R}(\theta,l))$ is the determinant of the spatial correlation matrix $\mathbf{R}(\theta,l)$ composed of the results of all microphone pairs i,ℓ for the spatial coordinates $\theta = [x,y]^T$. The operator T indicates transposition. With (1) we are able to define a 2D grid with points of interest. Each element $r_{i\ell}(\tau(\theta))$ of the matrix $\mathbf{R}(\theta)$ (neglecting block index l for simplicity) is calculated using the well-known generalized cross-correlation (GCC) [12] algorithm:

$$r_{i\ell}(\tau(\theta)) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} Z_{i\ell}(e^{j\omega\tau})X_i(e^{j\omega\tau})X_l^*(e^{j\omega\tau})e^{j\omega\tau}d\omega. \tag{2}$$

In (2) $X_{i,\ell}$ is a microphone signal in the frequency-domain, $Z_{i\ell}$ is a transformation factor and $\tau(\theta)$ is the theoretical time delay of the impinging sound wave between two microphones for a given sound source at position θ. The relative delay is computed assuming an acoustic far-field with plane wave propagation,

$$\tau_{il}(\theta) = \frac{\|\theta-\theta_i\|_2 - \|\theta-\theta_l\|_2}{c}, \tag{3}$$

where $\theta_{i,l}$ denote the spacial coordinates of microphone i and l and c is the speed of sound. An often chosen weighting for $Z_{i\ell}(e^{j\omega\tau})$ is the phase transformation (PHAT) [12]:

$$Z_{i\ell}(e^{j\omega\tau}) = \frac{1}{|X_i(e^{j\omega\tau})X_l(e^{j\omega\tau})|}. \tag{4}$$

Taking into account that $r_{i\ell}(\tau(\theta)) = r_{\ell i}(\tau(\theta))$ [10] one can write $\mathbf{R}(\theta)$ as,

$$\mathbf{R}(\theta) = \begin{pmatrix} 1 & \hat{r}_{12}(\tau(\theta)) & \vdots & \hat{r}_{1\ell}(\tau(\theta)) \\ \hat{r}_{12}(\tau(\theta)) & 1 & \vdots & \hat{r}_{2\ell}(\tau(\theta)) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{r}_{1\ell}(\tau(\theta)) & \hat{r}_{2\ell}(\tau(\theta)) & \cdots & 1 \end{pmatrix}, \tag{5}$$

$$\hat{r}_{i\ell}(\tau(\theta)) = \frac{r_{i\ell}(\tau(\theta))}{\sqrt{r_{ii}(\tau(\theta))r_{\ell\ell}(\tau(\theta))}}. \tag{6}$$

Peaks in the resulting spatial map $\rho(\theta)$ can be considered as acoustic source candidates. Intuitively speaking, $\sqrt{\rho(\theta)}$ can be understood as an overall correlation of all microphone channels [11]. One advantage of the MCCC algorithm compared to a straightforward mean calculation of the correlation between all microphone pairs is its robustness against microphone malfunctions, because it inherently neglects results witch have no correlation to other microphone channels and oppositely becomes one if two ore more microphone signals are perfectly correlated. For a detailed explanation we refer to [11].

2.2 Video Localization

The visual face detection is performed by means of the *Sophisticated High-Speed Object Recognition Engine* (SHORE) [13] using a single web-cam. Illumination invariant local structure features designated as *modified Census Transformation* are used for the face detection. The interested reader is referred to [13] for more detailed information. SHORE provides the area in pixel where the face is detected in the camera image. This information is translated in a 2D position estimate assuming a proper camera calibration. The size of the face bounding box on the camera image may be interpreted as the distance to the camera. By knowing the distance, the horizontal shift of the face midpoint indicates the lateral displacement. In the following, the resulting face position estimates will be denoted as,

$$\hat{\theta}_k = \begin{bmatrix} x_k \\ y_k \end{bmatrix}, \qquad k = 1 \dots K, \tag{7}$$

K is the number of detected faces in the video frame.

3 Particle Filter

To perform the multimodal fusion and to obtain a combined source position estimate we are using a particle filter. Generally spoken, a particle filter tries to represent a unknown probability function by performing a Monte Carlo simulation of the particle set with discrete probability weights [14]. The particles $\mathbf{s}$ are random samples in a state space. Assuming a first order physical dynamic, each particle is a vector defined as:

$$\mathbf{s}^{(b)}(l) = \left[x_1^{(b)}, y_1^{(b)}, \dot{x}_1^{(b)}, \dot{y}_1^{(b)}, \dots, x_Q^{(b)}, y_Q^{(b)}, \dot{x}_Q^{(b)}, \dot{y}_Q^{(b)} \right]^T, \tag{8}$$

where $\theta_q = [x_q, y_q]^T$ denotes the current coordinates of a source in the considered 2D space and $\dot{x}_q, \dot{y}_q$ are the respective velocities. The index $b = 1 \dots B$ indicates the number of the particle and $q = 1 \dots Q$ the number of the source.

The evolution of this particles over time can be summarized as a two stage process [7]. In the first stage, the physical movement of the particles is calculated by means of a first order *Langevin Model* [14]. This can be understood as predicting the state $\mathbf{s}^{(b)}(l)$ by knowing the prior density $\mathbf{y}(l-1)$, i.e. $p(\mathbf{s}^{(b)}(l)|\mathbf{y}(l-1))$. The second stage implies the link between the particles and the current observation $\mathbf{y}(l)$ by means of computing the likelihood $p(\mathbf{y}(l)|\mathbf{s}^{(b)}(l))$ of the observation, given that the state of particle $\mathbf{s}^{(b)}(l)$ is the true state of the observed system. In this paper the particles $\mathbf{s}^{(b)}$ are associated with two sets of weights $\mathbf{w} = \left[w^{(1)}, \dots, w^{(B)} \right]$, derived separately from the audio observations $\mathbf{w}_a$ and video observations $\mathbf{w}_v$. Each weight is valued by a likelihood function, i.e.,

$$w_a = p(\mathbf{y}_a|\mathbf{s}) = F_a(\mathbf{y}_a, \mathbf{s}), \tag{9}$$

$$w_v = p(\mathbf{y}_v|\mathbf{s}) = F_v(\mathbf{y}_a, \mathbf{s}), \tag{10}$$

where $F(\mathbf{y}, \mathbf{s})$ denotes the likelihood function, neglecting the block index l, and particle index b for simplicity. To build the likelihood functions we use both *pseudo likelihood* and *gaussian likelihood* approaches from [14]. For the acoustic likelihood we simply adapt the spatial map $\rho(\theta)$ given in (1) as a pseudo likelihood function.

$$F_a(\mathbf{y}_a, \mathbf{s}) = \prod_{q=1}^{Q} \max(\rho(\theta_q), \xi) \tag{11}$$

The parameter $\xi \geq 0$ ensures a non-negative likelihood function. The video modality already results in distinct position estimates $\hat{\theta}_k$ (c.f. Section 2.2), therefore we use

$$F_v(\mathbf{y}_v, \mathbf{s}) = \prod_{q=1}^{Q} \sum_{k=1}^{K} \mathcal{N}(\theta_q; \hat{\theta}_k, \sigma^2), \tag{12}$$

where $\mathcal{N}(\cdot)$ is a Gaussian distribution with the face location estimate $\hat{\theta}_k$ as mean value and variance $\sigma^2 = 0.01$ at point θ_q. The position estimate $\hat{\theta}_q = [\hat{x}_q, \hat{y}_q]^T$ for each individual source s is obtained by the weighted sum of all particles $\mathbf{s}^{(b)}$:

$$\begin{bmatrix} \hat{x}_q \\ \hat{y}_q \end{bmatrix} = \sum_{b=1}^{B} w \cdot \begin{bmatrix} x_q^{(b)} \\ y_q^{(b)} \end{bmatrix}. \tag{13}$$

At this point we are able to easily manipulate the influence of the single modalities to the resulting source estimate. The simplest ways to obtain a combined weight is to multiply both weights:

$$\mathbf{w} = \mathbf{w}_a \cdot \mathbf{w}_b \tag{14}$$

as in [8] or to calculate the mean. In [7] both weights are combined by an adaptive factor γ,

$$\mathbf{w} = \gamma \cdot \mathbf{w}_a + (1-\gamma) \cdot \mathbf{w}_v, \quad \gamma = \frac{m}{m_0} \cdot \varepsilon, \tag{15}$$

wherein γ depends on the "acoustic confidence" [7], m_0 is the total number of microphone pairs and m is the number microphone pairs with an observation value $\rho(\delta) > 0$. The maximum factor ε was empirically determined by [7]. Besides the different methods in obtaining the audio and video observations from [7, 8] and our contribution, we propose an exponential weighting approach

$$\mathbf{w} = \mathbf{w}_a^{\alpha} + \mathbf{w}_v^{\beta}, \tag{16}$$

wherein α and β are set individually for each modality. This approach is motivated by the demand to particularly highlight likelihoods with high values compared to poor likelihoods by an non-linear adjustment. In a second specification the weights should be modified independently to be able to adjust the best combination of the weights. A reasonable normalization of the weights has to be considered. The analysis of a proper α and β combination is carried out in Section 4. We exemplary determined the optimal ratio to achieve an optimal detection rate. In the most extreme case when a single modality fails completely (e.g. video camera is occluded or the voice activity detection (VAD) indicates no speech) its influence to the estimation can be neglected $\mathbf{w}_a \vee \mathbf{w}_v := 0$. After all, the combined weights are normalized to again represent a (modified) likelihood of the observations.

$$\sum_{b=1}^{B} w^{(b)} = 1 \tag{17}$$

This in turn can be used in (13) to estimate the source positions. Once the source positions are estimated one have to prevent the particle filter from the degeneracy phenomenon. If this is not considered, after only several blocks the weights of all but one particle will become negligable small. To avoid this, we use the systematic resampling approach as proposed in [15].

4 Evaluation

An evaluation of the proposed algorithm was carried in our home-environment living lab. We recorded the acoustic impulse responses (IRs) of the speaker positions with an 8-channel microphone line array (microphone distance = 20 cm) at a hight of 1.8 m. With this IRs several acoustic signal-to-noise ratios (SNRs) ranging from 20 dB to 0 dB were simulated. The video signal was recorded by a single web-camera positioned on top of the TV at a hight of approx. 1.5 m. A frame rate of 10 fps and a resolution of $1920x1080$ was used. A schematic top view of the setting is shown in Figure 1, including sensor and loudspeaker positions. The faces of the speakers were simulated by the loudspeakers covered with a portrait photo. The red dots in Figure 1 indicate the speaker positions.

$$\mathrm{RMSE} = \sqrt{\frac{1}{L \cdot Q} \sum_{l=1}^{L} \sum_{q=1}^{Q} \left(\hat{\theta}_q(l) - \bar{\theta}_q\right)^2}, \tag{18}$$

The root mean square error (RMSE) of all estimated locations versus the real locations was used as benchmark, with $\bar{\theta}_q$ being the true locations and $l = 1 \cdots L$ the block time. The blocksize was set to 1024 samples at a sampling rate of 48 kHz. Only blocks containing speech were taken into account for the RMSE calculation. Therefore, a VAD [16] was used to indicate time blocks containing speech or non-speech. We evaluated the video and audio localization algorithm only and the combined weighting according to (15) and (16). Table. 1 shows the results. The acoustic localization alone achieves an accuracy of 0.30 m at 20 dB down to 0.95 m at 0 dB. The video localization remained unchanged for all conditions, because only the acoustic SNR was changed. With a RMSE of 0.125 m it is already very accurate compared to the acoustic localization. On average the adaptive combination of both modalities (15) has an RMSE of 0.06 m. Also the other two combinations achieve similar low RMSE values of ≈ 0.06 m. The independence of this results from the SNR indicates the major influence of the visual location estimate. Figure 2 shows the influence of the restricting factor ε in (15) to the adaptive weighting, averaged over all SNRs. In spite of the small differences the optimal value of $\varepsilon \approx 0.45$ is close to the 0.6 mentioned in [7].

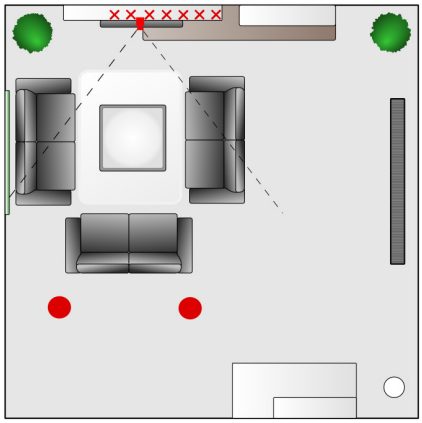

Figure 1: Sketch of the evaluation set up in a living room lab. Crosses indicate the microphone positions and large dots indicate real source positions. Dashed line shows camera perspective.

SNR	video only	audio only	mult weight	adaptive weight	expo. weight
20 dB	0.125 m	0.292 m	0.058 m	0.057 m	0.057 m
15 dB	0.125 m	0.369 m	0.061 m	0.057 m	0.057 m
10 dB	0.125 m	0.361 m	0.060 m	0.058 m	0.059 m
5 dB	0.125 m	0.690 m	0.058 m	0.059 m	0.061 m
0 dB	0.125 m	0.951 m	0.060 m	0.060 m	0.061 m

Table 1: RMSE results: 1st coloumn - only video localization; 2nd col. - only audio localization; 3rd col. - multiply weights (14); 4th col. - adaptive combination using (15); 5th col. - localization using (16)

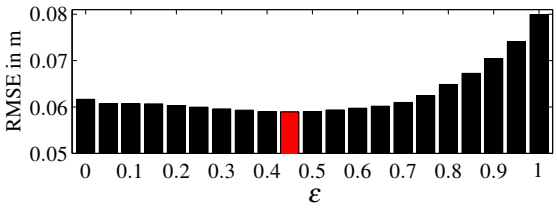

Figure 2: RMSE for adaptive method depending on the restricting ε in (15), averaged over all SNRs

The results of the exponential combination depending on the parameters α and β (c.f. (16)) is presented in Figure 3. Dark areas indicate high RMSE values, while bright coloured areas indicate low RMSE results. It can be seen that the performance of this approach is mainly dominated by the video exponential β. Thus in the examined scenario the video component is the most prominent modality and overrules the acoustic localization. Figure 3b illustrates the results averaged over all α. Best results are achieved by a β of ≈ 0.3. This approach especially tries to reduces the

influence of low weights.

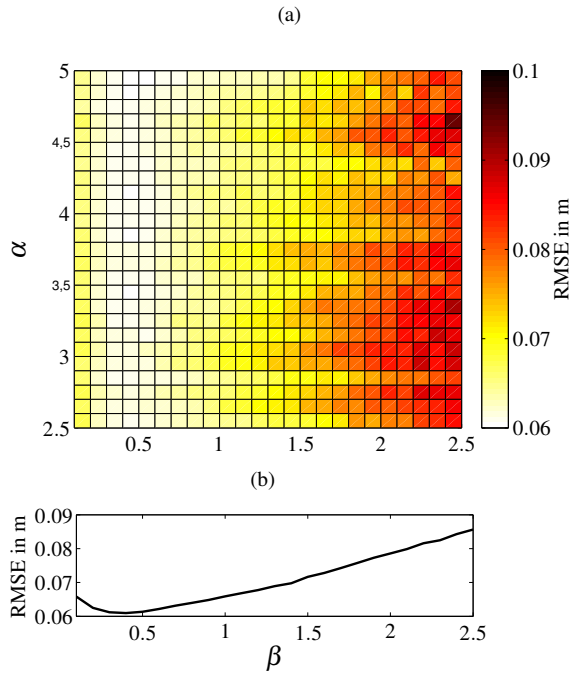

Figure 3: RMSE for weight combination with independent exponential factors α and β (c.f. (16)), mean over all SNR conditions. (b) mean over all α values, indicating the minimum at $\beta = 0.35$

5 Conclusion

As shown in the evaluation the multimodal localization estimation outperforms the single-modality algorithms. The proposed system is well suited to be used as supporting technology for assistive technologies to, e.g. control the home automation depending on the users position and, e.g a subsequent automatic speech recognition system. Particle filter are an appropriate method to combine multimodal localization modalities in a flexible manner. All analyzed combinations of the video and audio weighting could be well adjusted to the specific environmental condition. A simple multiplication of the weights already obtained good results. However, the conditions in the home-environment lab encourage a superior visual location estimate which makes it difficult for an thoroughly examination of the modality fusion. Further investigations with different lighting conditions and covered faces need to be carried out. Also the inherent tracking capabilities of the particle filter have to be analyzed in the future.

References

[1] S. Goetze, J. Schröder, S. Gerlach, D. Hollosi, J.-E. Appell, and F. Wallhoff, "Acoustic Monitoring and Localization for Social Care," *Journal of Computing Science and Engineering (JCSE), SI on uHealthcare*, vol. 6, pp. 40–50, Mar. 2012.

[2] N. Strobel, S. Spors, and R. Rabenstein, "Joint audio-video object localization and tracking," *Signal Processing Magazine, IEEE*, vol. 18, pp. 22 –31, Jan. 2001.

[3] B. Lee, J. Choi, D. Kim, and M. Kim, "Sound source localization in reverberant environment using visual information," in *Intelligent Robots and Systems (IROS)*, pp. 3542 –3547, Oct. 2010.

[4] D. Lo, R. Goubran, and R. Dansereau, "Robust joint audio-video localization in video conferencing using reliability information II: Bayesian network fusion," in *Instrumentation and Measurement Technology Conference (IMTC)*, vol. 2, pp. 1246 – 1249, May 2004.

[5] F. Keyrouz, U. Kirchmaier, and K. Diepold, "Three dimensional object tracking based on audiovisual fusion using particle swarm optimization," in *11th International Conference on Information Fusion*, pp. 1 –5, July 2008.

[6] A. Pnevmatikakis and F. Talantzis, "Person tracking in enhanced cognitive care: A particle filtering approach," in *Proc. of. the 18th European Signal Processing Conference (EUSIPCO)*, Aug. 2010.

[7] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audio-visual speaker tracking," in *Proc. of the 7th International Conference on Multimodal Interfaces*, pp. 61–68, ACM, 2005.

[8] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP Journal on Applied Signal Processing*, vol. 2002, pp. 1154–1164, Jan. 2002.

[9] N. Checka, K. Wilson, V. Rangarajan, and T. Darrell, "A probabilistic framework for multi-modal multi-person tracking," in *Conference on Computer Vision and Pattern Recognition Workshop, 2003. CVPRW '03*, vol. 9, p. 100, june 2003.

[10] J. Chen, J. Benesty, and Y. A. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on Applied Signal Processing*, vol. ID 26503, pp. 1–19, 2006.

[11] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 549–557, Nov. 2003.

[12] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, pp. 320–327, Aug. 1976.

[13] C. Küblbeck and A. Ernst, "Face detection and tracking in video sequences using the modified census transformation," *Image and Vision Computing*, vol. 24, no. 6, pp. 564–572, 2006.

[14] D. Ward, E. Lehmann, and R. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 826–836, Nov. 2003.

[15] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, pp. 174–188, Feb. 2002.

[16] J. Ramırez, J. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3, pp. 271–287, 2004.