published in: *European Conference on Signal Processing (EUSIPCO) 2013*

ONLINE MULTI-SPEAKER TRACKING USING MULTIPLE MICROPHONE ARRAYS INFORMED BY AUDITORY SCENE ANALYSIS

Axel Plinge and Gernot A. Fink

Department of Computer Science, TU Dortmund University, Dortmund, Germany

ABSTRACT

Tracking multiple speakers with microphone arrays is used for practical applications such as video conferencing. An important task is the integration of multiple arrays with correct associations of multiple concurrent speakers. A single-array tracking approach based on CASA is extended here to probabilistic tracking with multiple arrays in order to handle a varying number of moving speakers over time and assign the concurrent localizations of multiple sensors to the speakers. Tracking is done simultaneously in angular and Euclidean space. The effectiveness of the method is shown with recordings of real speakers in a reverberant conference room by evaluation on the publicly available AV16.3 corpus.

Index Terms— microphone array, auditory scene analysis, multi-sensor, speaker tracking

1. INTRODUCTION

The influential "Auditory Scene Analysis" (ASA) theory of human hearing is based on psychoacoustic experiments as well as biological and neurological research. Successful computational (CASA) models were developed, cf. [1]. These use only up to two sensors of an artificial human head [2], while technical tracking and beamforming approaches employ microphone arrays with eight or more sensors [3]. Hybrid methods applying neurobiologically inspired processing to microphone arrays were introduced recently [4].

Tracking human speakers with microphone arrays is an important task for many practical applications such as speech separation and enhancement as well as camera control for online lectures and smart conferencing. For unconstraintly moving speakers in a larger room and multi-modal integration, tracking in Euclidean space using distributed microphone arrays is appropriate. The major challenge beyond the reverberation and the natural sparsity of the speech signals is the handling and correct association of concurrent speakers over all microphone arrays.

This work was supported by the German Research Foundation (DFG) under contract number Fi 799/5-1.

We would like to thank Daniel Hauschildt for many fruitful discussions. We also would like to thank the reviewers for their helpful suggestions.

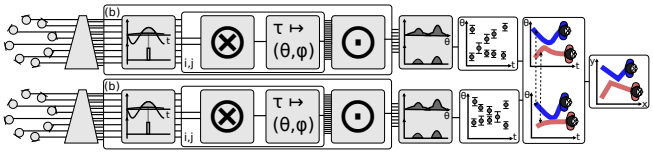

Fig. 1. Processing structure (f.l.t.r.): mic. arrays, filter bank, spike generation, correlation, backprojection, combination, PoA, clustering and tracking with multi-array integration.

Common tracking strategies for a single microphone array are the clustering of localizations, e.g. with a maximum likelihood approach [5] or particle filtering [6]. The fact that not only the location, but also the spectra of the speakers are different is used in source separation techniques [3]. CASA models of both monaural and binaural human hearing use similarities of multiple cues such as location, spectrum, and pitch for grouping and separation of speakers [7,8].

An existing hybrid method [9] applies a model of the inner ear and auditory midbrain to signals from a single microphone array. It computes time-difference-of-arrival (TDOA) based spatial information in several frequency bands separately and uses the spatial and spectral information for clustering localizations. Distinctly localized "glimpses" in reverberant and noisy environments are integrated to continuous speaker tracks. The density based clustering requires multiple thresholds and discards inherent probabilistic information.

This paper presents an extension of that method overcoming these shortcomings in a new simultaneous integration step. It applies maximum likelihood clustering while introducing a methodical sound estimation of the number of speakers. In a new sequential and model based integration, the probabilistic information is used in computing a consensus over multiple microphone arrays, which increases the robustness. The ambiguity of multiple simultaneous detections is resolved using spectral similarity. At the same time, speaker tracks in angular and Euclidean coordinates are calculated with the angular localizations from individual arrays.

2. METHOD

The processing steps of the speaker tracking are illustrated in Figure 1 and described consecutively in this section.

2.1. Cochlear and Midbrain Model

The cochlear and midbrain model described in [9] is used. It filters the microphones' signals with a gammatone filterbank composed of B bands. The peak-over-average-position (PoAP) cochlear model applies onset dominance and glimpses events with high modulation. Correlations are calculated in short time frames as TDOA estimates. These are backprojected to spherical far field source positions and combined for all microphone pairs using a fuzzy t-norm. For the fuzzy operation the signals have to be adjusted into the range $[0, 1]$. Rather than presetting the required gain manually as done in [9], the gain is automatically determined with a short-term histogram of the spike amplitude. A short moving average over 0.3 s is calculated over all data points with a shift of 0.075 s to accommodate for very fast moving speakers. Speakers can be separated by azimuth in most practical scenarios, so the maximum value over all elevations is used. A peak-over-average (PoA) filtering step in analogy to the difference-of-Gaussian processing found in human perception is applied by subtracting an $45°$ average from an $5°$ average and using only positive values.

For each frame index k the resulting sparse spatial likelihood values $e_{k,\theta,b}$ are collected over all bands b for each azimuth θ, yielding an estimate

$$\boldsymbol{s} = (e_{k,\theta,0}, e_{k,\theta,1}, \ldots, e_{k,\theta,B-1})^T \tag{1}$$

of the spectral distribution. The spectral energy summed over all bands is a measure of the correlation strength. It reflects the source probability and position accuracy and is interpreted as likelihood

$$l\big((\theta, \boldsymbol{s})\big) = \sum_b e_{k,\theta,b} \tag{2}$$

for a source at the given angle. Detections with less than $B/4$ nonzero spectral components or a likelihood l below ϵ_s are excluded as non-speech sounds. The tuples in the remaining set

$$D_k = \big\{x = (\theta, \boldsymbol{s}) \,\big|\, l(x) > \epsilon_s \wedge ||\{b|e_{k,\theta,b} > 0\}|| \geq B/4\big\} \tag{3}$$

of combined azimuth-spectrum tuples for each time frame k are considered speech energy detections. The resulting spatial likelihood is less susceptible to reverberation and noise than the SRP-PHAT as is illustrated in Figure 2a,b.

2.2. Simultaneous Grouping

According to the ASA theory location as well as spectral cues are used for grouping the auditory information coming from a certain source. The process of "simultaneous grouping" is modeled by clustering over azimuth and spectral similarity.

Since reverberant speech is found to produce Gaussian distributed peaks over time [10], the spatial likelihood is modeled as mixture of Gaussians (MoG) [5]. The probability density for a detection $x = (\theta, \boldsymbol{s}) \in D_k$ can be calculated as

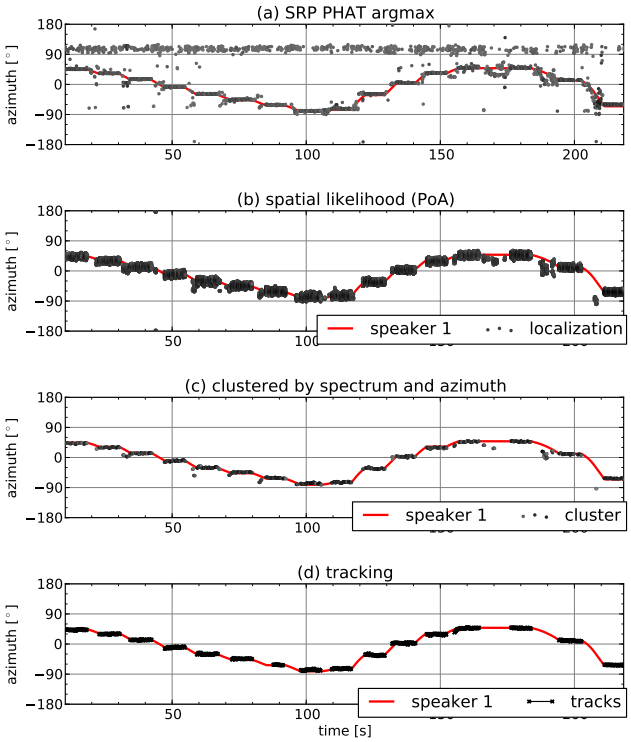

Fig. 2. Localizations for sequence 01 of the AV16.3 corpus. (a) Results from SRP-PHAT argmax processing for 0.3 s time windows for comparison, (b) PoA, (c) grouping and (d) tracking. Note that the noise at about $100°$ apparent in the SRP-PHAT localization is excluded by the preprocessing.

$$p_a(x|\Theta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-0.5\frac{d(\theta, \Theta)^2}{\sigma^2}\right) \tag{4}$$

with the average angle Θ and standard deviation σ using the angular distance

$$d(\alpha, \beta) = \min\{360 - |\alpha - \beta|, |\alpha - \beta|\}\,. \tag{5}$$

Due to the nature of human speech production, spectral magnitudes are dependent across frequency for natural speech, which is still apparent in reverberant conditions. The spectra from different speech sources are dissimilar with a high probability in most practical scenarios. Noise and time domain aliasing artifacts are assumed independent across frequency. The spectral similarity of a detection $x = (\theta, \boldsymbol{s})$ to a model spectrum $\boldsymbol{t}$ is calculated as normalized scalar product

$$p_s(x|t) = \left\langle \frac{\boldsymbol{s}}{||\boldsymbol{s}||}, \frac{\boldsymbol{t}}{||\boldsymbol{t}||} \right\rangle = \frac{\sum_b s_b t_b}{\sqrt{\sum_b s_b^2 \sum_b t_b^2}}\,. \tag{6}$$

The probability of x to originate from $\Psi_i = (\Theta_i, \sigma_i, \boldsymbol{t}_i)$ with average angle Θ_i, standard deviation σ_i and spectrum $\boldsymbol{t}_i$ is

$$p(x|\Psi_i) = p_s(x|\boldsymbol{t}_i) p_a(x|\theta_i, \sigma_i)\,. \tag{7}$$

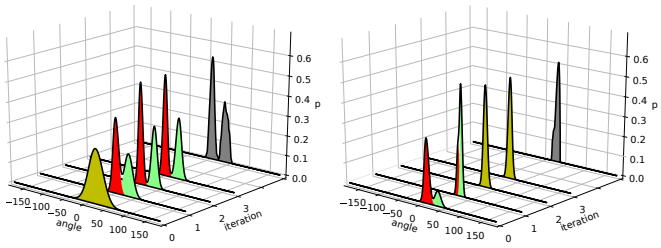

Fig. 3. Split (left) and join (right) within the EM estimation. Mixture components are plotted in color and the mixture as contour line for each iteration. The spatial likelihood histogram $\sum_{x=(\theta,\boldsymbol{s})} l(x)$ for the angles θ to be estimated is shown in gray at the back. Iteration 0 displays the estimate from the previous time frame.

Sources Ψ_i are estimate by the EM-algorithm with the maximum likelihood estimate for all N detections in the current and adjacent time frames $x \in D_{k-1} \cup D_k \cup D_{k+1}$ as

$$p\left(\Psi_i|x\right) = \frac{c_i p(x|\Psi_i)}{\sum_{i'} c_{i'} p\left(x|\Psi_{i'}\right)} \tag{8}$$

using mixture weights c_i. The maximization step is give by

$$\gamma_i(x) = \frac{p(\Psi_i|x) l(x)}{\sum_{x'} p(\Psi_i|x') l(x')} \tag{9}$$

$$\hat{\Theta}_i = \sum_{x=(\theta,\boldsymbol{s})} \gamma_i(x)\theta \tag{10}$$

$$\hat{\sigma_i}^2 = \sum_{x=(\theta,\boldsymbol{s})} \gamma_i(x) d(\theta, \hat{\Theta}_i)^2 \tag{11}$$

$$\hat{\boldsymbol{t}}_i = \sum_{x=(\theta,\boldsymbol{s})} \gamma_i(x)\boldsymbol{s} \tag{12}$$

$$\hat{c_i} = \frac{1}{N} \sum_x p\left(\Psi_i|x\right), \tag{13}$$

where the weighted average of angles in (10) has to be calculated on the circle. (Note that the weighting $\gamma_i(x)$ takes the spatial sum likelihood into account. In relation to the original unweighted EM-implementation, $l(x = (\theta, \boldsymbol{s}))$ can be interpreted as the number of measurements at position θ so that the maximization step equals the original one for a discrete number of measurements.)

The number of sources can be estimated by observing the typical variance of speaker localizations. When two estimates get closer than a threshold $d(\theta_i, \theta_j) < \Gamma_{\text{join}} = 8°$, the sources i, j are merged. If $\sigma_i > \Gamma_{\text{split}} = 12°$, the source i is split into two sources with $\theta_{i,j} = \theta_i \pm \sigma_i$, see Figure 3.

The estimation loop is terminated when the likelihood does no longer change significantly. This typically happens after two to ten iterations, allowing for real-time calculation. After this step, there are clustered source estimates $E_k = \{\Psi_i\}$ for each time frame.

2.3. Sequential / Model based Integration

By calculating the intersection of the lines of length $h^{(1,2)}$ from each array originating at position $m^{(1,2)}$ with the cluster angle $\theta^{(1,2)}$ the 2D position

$$z = m^{(1)} + h^{(1)} \begin{pmatrix} \cos\theta^{(1)} \\ \sin\theta^{(1)} \end{pmatrix} = m^{(2)} + h^{(2)} \begin{pmatrix} \cos\theta^{(2)} \\ \sin\theta^{(2)} \end{pmatrix} \tag{14}$$

of the speaker can be derived. The arrays have to be synchronous up to around one frame shift (75 ms). With only two arrays, the expected error is high for a small intersection angle. A combined tracking state $\Omega_j = (\Psi_j^{(1)}, \Psi_j^{(2)}, z_j)^T$ represents the states of the track with label j. The probability of a new detection Ψ_i to belong to a track j given the cluster angles for one microphone array is calculated with the average deviation and (4) as

$$p_a\left(\Psi_j|\Psi_i\right) = p_a\left(\Psi_i|\Psi_j\right) = p_a(\Theta_j|\Theta_i, (\sigma_i + \sigma_j)/2). \tag{15}$$

As array consensus, we obtain the joint probability

$$p\left(\Psi_i^{(1,2)}|\Omega_j\right) = p_a\left(\Psi_i^{(1)}|\Psi_j^{(1)}\right) p_a\left(\Psi_i^{(2)}|\Psi_j^{(2)}\right). \tag{16}$$

Note that the 2D distance is intently omitted here to allow for tracking with two arrays and small intersecting angles.

When assuming the signals from multiple speakers have different spectra at the same time, the spectral similarity can be used to find out which localizations of multiple microphone arrays originated form the same source. The likelihood based on the spectra of clusters from two arrays originating from the same speaker can be expressed similarly to (6) as

$$p_s\left(\Psi_i^{(1)}, \Psi_i^{(2)}\right) = p_s\left(\Psi_i^{(2)}, \Psi_i^{(1)}\right) = \left\langle \frac{\boldsymbol{t_i^{(1)}}}{||\boldsymbol{t_i^{(1)}}||}, \frac{\boldsymbol{t_i^{(2)}}}{||\boldsymbol{t_i^{(2)}}||} \right\rangle. \tag{17}$$

The pairs of clusters for the arrays are chosen by spectral similarity, then tracks are formed by probabilistic spatial association. A time-to-live rule is added to handle small speech pauses [5]. Any track j whose newest detection is older than $t_{\text{TTL}} = 2\,\text{s}$ is discarded. Smaller gaps are filled by linear interpolation. For each set of clusters $\Psi_i^{(1,2)} \in E_k^{(1)} \times E_k^{(2)}$ the following algorithm is applied:

1. Assign all pairs $\Psi_i^{(1)}, \Psi_i^{(2)}$ their spectral likelihood $w = p_s$. Add a small bias δ_t for all pairs near an existing track.
2. Choose the pair with the highest likelihood $> \epsilon_b$.
3. Calculate $p_j = p(\Psi_i^{(1,2)}|\Omega_j)$ for all tracks i and choose the likeliest track with $p_j > \epsilon_a$ not older than t_{TTL}. If there is a gap, fill it by linear interpolation. If no such track exists, start a new one.
4. Discount w for used angles by δ_b and continue at 2.

Since typically only a few speakers are active, the number of clusters per array is small, leading to a small number of potential combinations for association over the arrays and to the existing tracks. This allows for fast online computation.

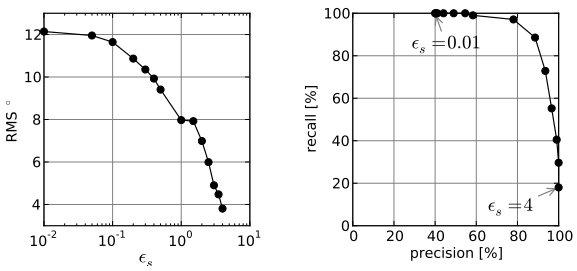

Fig. 4. RMS error (left) and precision-recall curve for the threshold ϵ_s on seq01.

3. EVALUATION

First basic evaluations illustrate some of this papers assumptions. Then the proposed method is tested with a number of recordings from the freely available AV16.3 corpus. It consists of recordings in a reverberant meeting room with a T_{60} of about 0.7 s and two circular eight channel microphone arrays with a radius of 10 cm recorded at 16 kHz [11]. $B = 16$ bands in the range between 300 Hz and 3 kHz were used. A localization is considered correct if the angles hit the target within an average head width of 0.2 m or the 2D coordinates are within a typical persons shoulder width of 0.5 m. The precision and recall are calculated in 0.6 s windows based on the correct localizations. These margins should be sufficient for most practical applications.

3.1. Spatial Likelihood

It was stated that the spatial likelihood value $l(x)$ is correlated with the actual source probability. Figure 4 shows a precision and recall curve and the angular root mean square (RMS) error as function of the threshold ϵ_s. It can be seen that the precision of the estimates increases with the likelihood, while the recall decreases for increasing threshold values.

3.2. Tracking a single speaker

The first AV16.3 sequence consists of a single speaker that is static while speaking at 16 positions in the room. Figure 2 shows the steps of our pipeline and the angular tracking result. The utterances at all 16 positions are tracked correctly.

In sequence 11 one speaker moves his head fast in a short time. Figure 5 shows the tracking result. The track is split when the localizations clusters have a gap of $20°$ leading to a small $p_i < \epsilon_a$ (15). The 2D tracking has good accuracy where the intersection angle $\theta^{(1)} - \theta^{(2)}$ is larger than $20°$.

3.3. Tracking multiple speakers

In sequence 18, two concurrent speakers repeatedly put their heads together and apart. Figure 6 shows the tracking result. When the speakers are very close, both spatial and spectral information overlap so only one track is continued. The tracks

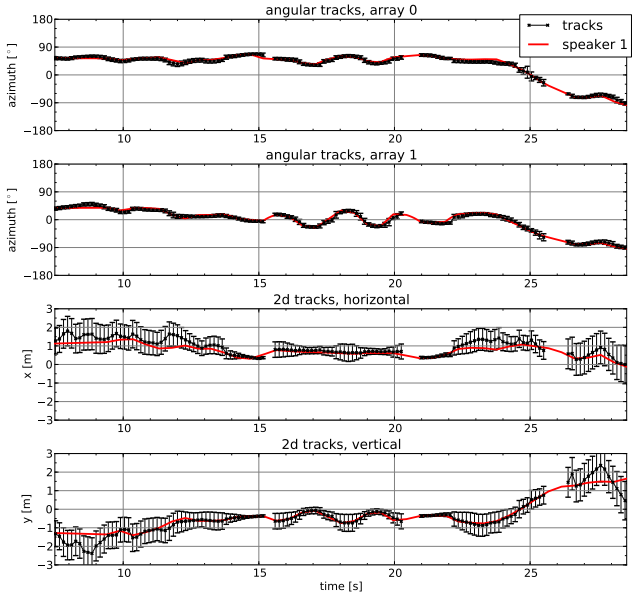

Fig. 5. Tracking for seq11, one fast moving speaker. The standard deviations of the tracks are plotted as errorbars.

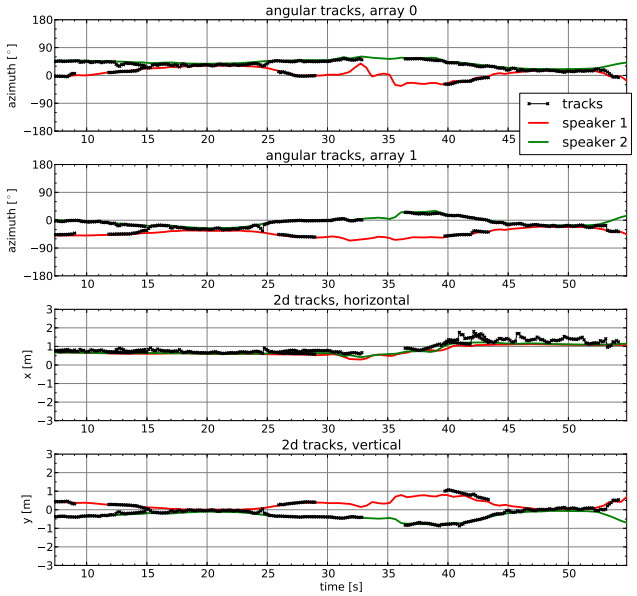

Fig. 6. Tracking for seq18, two concurrent speakers.

follow the speakers within about $4°$ and show gaps where the clusters are sparse. The spectrum-based resolving of the ambiguity can be seen by the 2D tracks. Again, the 2D accuracy deteriorates when the intersecting angle is very small.

Table 1 summarizes the results for all sequences used in the evaluation. In sequence 40, three speakers talk concurrently while two sit still and one moves around them. It can be seen that the consecutive steps of EM-Clustering and tracking increase the accuracy while the consensus leads to a smaller recall. Offline methods with tuned post-processing are reported to achieve an RMS of $2.5°$ on the dataset [10]. The

sequence	method	precision	recall	RMS
seq01	PoA	34%, 29%	90%, 91%	7.2°, 9.3°
	[4] BS	90%, 83%	95%, 93%	3.7°, 4.0°
loc.	[9] DB	90%, 86%	91%, 91%	3.3°, 3.4°
	[*] EM	96%, 94%	99%, 98%	2.5°, 2.4°
	[9]	93%, 87%	98%, 97%	3.1°, 3.0°
trk.	[*] ang.	95%, 98%	89%, 89%	2.5°, 2.0°
	[*] 2D	91%	80%	0.320 m
seq11	PoA	26%, 30%	94%, 97%	8.8°, 9.0°
	[4] BS	67%, 57%	94%, 97%	6.2°, 7.9°
loc.	[9] DB	82%, 81%	90%, 93%	5.5°, 6.5°
	[*] EM	70%, 69%	97%,100%	5.6°, 7.0°
	[9]	67%, 73%	97%, 97%	5.9°, 6.9°
trk.	[*] ang.	66%, 72%	92%, 94%	5.6°, 6.8°
	[*] 2D	83%	97%	0.408 m
seq18	PoA	88%, 88%	100%,100%	10.5°,17.9°
	[4] BS	91%, 88%	88%, 86%	5.9°, 5.8°
loc.	[9] DB	95%, 96%	88%, 83%	13.1°, 8.4°
	[*] EM	92%, 95%	89%, 85%	6.3°, 4.7°
	[9]	93%, 89%	87%, 83%	6.2°, 5.0°
trk.	[*] ang.	97%, 99%	80%, 78%	4.1°, 3.2°
	[*] 2D	99%	93%	0.202 m
seq40	PoA	56%, 44%	88%, 88%	6.4°, 8.2°
	[4] BS	92%, 83%	84%, 87%	3.8°, 4.1°
loc.	[9] DB	90%, 86%	91%, 91%	3.3°, 3.4°
	[*] EM	94%, 93%	83%, 85%	3.4°, 3.5°
	[9]	91%, 87%	77%, 80%	4.5°, 4.3°
trk.	[*] ang.	97%, 94%	78%, 78%	2.9°, 3.2°
	[*] 2D	98%	93%	0.631 m

Table 1. PoA, localization and tracking results for the proposed method [*] and previous [4], [9]. Values for both microphone arrays are given in pairs.

2D tracking results show a rather large RMS of around half a meter due to the occurrence of small itersection angles, however, the precision and recall with respect to practical applications is still reasonable. In comparison, the localizations based on the sum spatial likelihood over all bands [4] shows inferior performance. The EM clustering works better than the density based clustering [9], and the multi-array tracking performs better than the single array tracking [9], especially in the multi-speaker scenarios.

4. CONCLUSION

Based on insights from studies of human perception, common signal processing techniques were reformulated for online tracking of multiple speakers with multiple microphone arrays. The system is real-time capable with a latency of 0.5 s. The results show that the proposed method improves over the previous hybrid approaches and is able to handle the tracking of multiple concurrent speakers in real reverberant conditions. The concurrent speakers are successfully separated and robust postion estimates with good precision for practical applications are derived. For precise Euclidean coordinates, three arrays and/or different geometries avoiding small intersecting angles should be investigated. The coordinate estimates with their individual variance can be used for multi-modal integration.

5. REFERENCES

[1] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley, 2006.

[2] T. May, *Binaural Scene Analysis: Localization, Detection and Recogntion of Speakers in Complex Acoustic Scenes*, Technische Universiteit Eindhoven, 2012.

[3] R. Martin, U. Heute, and C. Antweiler, *Advances in Digital Speech Transmission*, Wiley, 2008.

[4] A. Plinge, M. H. Hennecke, and G. A. Fink, "Robust Neuro-Fuzzy Speaker Localization Using a Circular Microphone Array," in *Proc. Int. Workshop on Acoustic Echo and Noise Control*, Tel Aviv, Israel, 2010.

[5] N. Madhu and R. Martin, "A Scalable Framework for Multiple Speaker Localization and Tracking," in *Proc. Int. Workshop on Acoustic Echo and Noise Control*, Seattle, WA, USA, September 2008.

[6] M. F. Fallon and S. J. Godsill, "Acoustic Source Localization and Tracking of a Time-Varying Number of Speakers," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1409–1415, 2012.

[7] H. Christensen and J. Barker, "Speaker Turn Tracking with Mobile Microphones: Combining Location and Pitch Information.," in *EUSIPCO 2010, Aalborg, Denmark, August, 2010*, 2010, pp. 954–958.

[8] K. Hu and D. Wang, "An Unsupervised Approach to Cochannel Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 120–129, 2012.

[9] A. Plinge, M. H. Hennecke, and G. A. Fink, "Reverberation-Robust Online Multi-Speaker Tracking by using a Microphone Array and CASA Processing," in *Proc. Int. Workshop on Acoustic Echo and Noise Control*, Aachen, Germany, 2012.

[10] G. Lathoud and J.-M. Odobez, "Short-Term Spatio-Temporal Clustering applied to Multiple Moving Speakers," *IEEE Trans. Audio Speech & Language Process.*, 2007.

[11] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: An Audio-Visual Corpus for Speaker Localization and Tracking," in *Proc. MLMI '04 Workshop; LNCS*, 2005, vol. 3361, pp. 182–195.