MULTI-SPEAKER TRACKING USING MULTIPLE DISTRIBUTED MICROPHONE ARRAYS

Axel Plinge and Gernot A. Fink

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

ABSTRACT

Tracking multiple speakers with microphone arrays is one of the key tasks in smart environments. For good accuracy in reverberant environments, several arrays should be distributed in the room. The method presented is using distributed nodes with microphone arrays that compute local angular speech detections. In an integrating node, these are associated using the spectra and tracks for multiple concurrent speaker are computed. Euclidean coordinates are derived by triangulation, which is improved by a quality based weighting. The method is not only robust against reverberation, but also against transmission errors and jitter. Test with real recordings show that good precision for practical applications can be achieved.

***Index Terms*—** sensor node, speaker tracking, machine hearing

1. INTRODUCTION

A smart environment is defined by its capability of understanding of and interaction with human behavior by means of sensors, actuators and dedicated processing. The implementation of perceptual capabilities that make sense of the sensory data is the basis for this ability. Robust acoustic speaker tracking is required for many practical scenarios like online lectures, video conferencing or meetings [1–3]. It facilitates automated camera control and selection as well as speaker and context identification.

Acoustic localization is often done by the SRP-PHAT or generalized cross-correlation approach [4]. The resulting spatial likelihood can be modeled as mixture of Gaussians (MoG) because reverberant speech is found to produce Gaussian distributed peaks over time [5]. In order to track speakers using a single microphone array, temporal integration can be realized with a maximum likelihood approach [6] or particle filtering [7,8]. Multiple microphone arrays can be used for robust localization in reverberant environments. Their relative and absolute positioning can be derived manually or automatically [9–11]. If the arrays themselves do not have to be strictly synchronized, the system is more robust and easy to realize, especially when using radio connections.

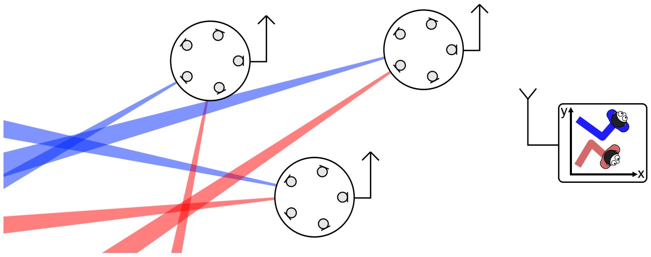

Fig. 1. Proposed system: Multiple distributed nodes localize speech using microphone arrays and transmit to an integration node that calculates speaker tracks in Euclidean space.

The emerging research field of machine hearing tries to incorporate features of the human hearing process into computational processing [12]. One key influence is the study and modeling of the hearing process according to the “Auditory Scene Analysis” (ASA) theory [13]. Recently, several biologically inspired systems were shown to outperform technical approaches to localization [14–16].

A recent application of machine hearing for speaker tracking suggested the use of spectra as cues for associating concurrent estimates, inspired by the observation that human listeners integrate streams from both ears using common cues [17]. It has been successfully applied to recordings of two microphone arrays [18]. Here we extended this approach to the use of multiple distributes nodes. The proposed system is illustrated in Fig. 1. Multiple distributed nodes are equipped with microphone arrays. Each node calculates concurrent angular speaker detections and transmits the spatial and spectral estimates to an integrating unit. There, the spectral information is employed to solve the ambiguity problem for multiple simultaneous detections for multiple concurrent speakers. The angular estimates are used to integrate the localizations over time; Euclidean coordinates are calculated by triangulation.

It is shown that the spectral cues allow to resolve the association ambiguity for multiple nodes. For triangulation, a precision oriented weighting is introduced that both allows the integration of an arbitrary number of nodes and improves the Euclidean localization over unweighted averaging. The system is able to handle transmission errors and is robust against drift and jitter.

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

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. NODE-BASED LOCALIZATION

Each node is equipped with a circular microphone array. The machine hearing approach introduced in [14] is applied. The cochlear and mid-brain model uses $B = 16$ frequency bands. For time windows of $L = 0.6$ s with a time step of $k \cdot L/4$ a set D_k of combined tuples $x = (\theta, \boldsymbol{s})$ with azimuth θ and spectrum $\boldsymbol{s} = (s_0, \ldots, s_{B-1})^T$ is extracted as set of speech detections.

Sources Ψ_i are estimated by calculating clusters $\Psi_i = (\Theta_i, \sigma_i, \boldsymbol{t}_i)$ with mean angle Θ_i, deviation σ_i, and spectrum $\boldsymbol{t}_i$ from the detections in the current and adjacent time frames $x \in D_{k-1} \cup D_k \cup D_{k+1}$ using the EM-algorithm as described in [18]. The probability of x to originate from Ψ_i is

$$p(x|\Psi_i) = p_a(x|\Psi_i) p_s(x|\Psi_i) \tag{1}$$

where the angular probability density for a detection x is calculated using the angular distance $d(\alpha, \beta) = \min\{360 - |\alpha - \beta|, |\alpha - \beta|\}$ as

$$p_a(x|\Psi_i) = p_a(x|\Theta_i, \sigma_i) = \frac{e^{-0.5 d(\theta, \Theta_i)^2 \sigma_i^{-2}}}{\sqrt{2\pi\sigma^2}} \tag{2}$$

and the spectral similarity of a detection $x = (\theta, \boldsymbol{s})$ to a model spectrum $\boldsymbol{t}$ is calculated as normalized scalar product

$$p_s(x|\Psi_i) = \left\langle \frac{\boldsymbol{s}}{||\boldsymbol{s}||}, \frac{\boldsymbol{t}_i}{||\boldsymbol{t}_i||} \right\rangle = \frac{\sum_b s_b t_{i,b}}{\sqrt{\sum_b s_b^2 \sum_b t_{i,b}^2}} . \tag{3}$$

The number of sources can be estimated by observing the typical variance of speaker localizations for the given array geometry. If $\sigma_i > \Gamma_{\text{split}} = 20°$, the source i is split into two sources with $\theta_{i,j} = \theta_i \pm \sigma_i$, when two estimates get closer than a threshold $d(\theta_i, \theta_j) < \Gamma_{\text{join}} = 12°$, the sources $\Psi_{i,j}$ are merged. The EM loop is converging quickly in less than 20 iterations, allowing for real-time calculation. After this step, there are clustered source estimates $E_k^{(m)} = \{\Psi_i\}$ for each time frame at each node m. The resulting sparse set of positions and spectra amounts to below 32 kbps of data which may be transmitted over a wireless connection.

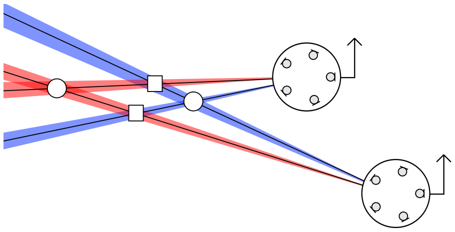

Fig. 2. The problem of ambiguity of multiple concurrent estimates: Without additional information, all four intersections are possible source positions. Using the spectral similarity, the correct intersections (circles) are chosen and the others (squares) are discarded.

3. INTEGRATION AND TRACKING

The problem of ambiguity of multiple concurrent localizations by multiple nodes is illustrated in Fig. 2. To associate the estimates from different nodes, their spectra are correlated using (3), and the pairs with the strongest correlation are computed. By thereafter combining all pairs with common angles, sets of angular estimates over all nodes are derived.

The Euclidean position of the source can be derived by triangulation using these sets. By calculating the intersection of the lines originating at two nodes' center positions $c^{(m,n)}$ with the cluster angles $\Theta^{(m,n)}$ the 2D position $z^{(m,n)}\left(\Theta^{(m)}, \Theta^{(n)}\right)$ is derived. Given two angles α, β the quality of the localization by intersection may be expressed as

$$q(\alpha, \beta) = |\sin(\alpha - \beta)| \tag{4}$$

to reflect the fact that an angular difference of 90° yields the highest precision and an angular difference near 0° or 180° the worst. In order to calculate one point from multiple intersections, the weighted sum

$$z = \frac{\sum_{m,n} q\left(\Theta^{(m)}, \Theta^{(n)}\right) z^{(m,n)}\left(\Theta^{(m)}, \Theta^{(n)}\right)}{\sum_{m,n} q\left(\Theta^{(m)}, \Theta^{(n)}\right)} \tag{5}$$

is used. The effect is illustrated in Fig. 3.

A combined tracking state $\Omega_{j,k} = (\Psi_{j,k}^{(m)}, \ldots, \Psi_{j,k}^{(n)}, z_{j,k})^T$ represents the states of the track with label j for time step k. The probability of a new detection $\Psi_{*,k+1}$ to belong to a track Ψ_j given the cluster angles is calculated for each node as

$$p_a\left(\Psi_{*,k+1}|\Psi_{j,k}\right) = p_a(\Theta_{j,k}|\Theta_{*,k+1}, (\sigma_{*,k+1} + \sigma_{j,k})/2) \tag{6}$$

and then multiplied

$$p(\Psi_{*,k+1}^{(m)}, \ldots, \Psi_{*,k+1}^{(n)}|\Omega_{j,k}) = \prod_o p_a\left(\Psi_{*,k+1}^{(o)}|\Psi_{j,k}^{(o)}\right) \tag{7}$$

to compute the consensus. For each set of new estimates, the track with the highest likelihood above a threshold ϵ_a is

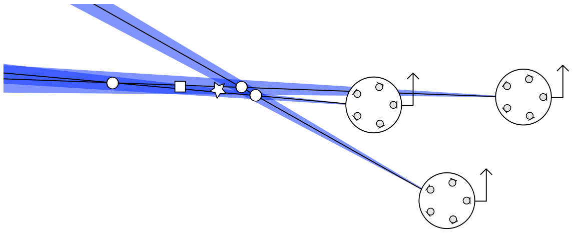

Fig. 3. Incorporating angular intersection quality into the triangulation: Using all pairwise intersections (circles) equally, the center (square) is computed as localization. When weighting by intersection quality, the steeper angles get favored and a more precise localization is achieved (star).

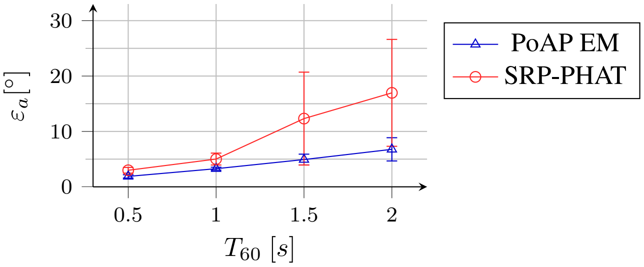

Fig. 4. Localization error for a single node for simulation of a single speaker.

chosen from all tracks not older than a t_{TTL} (e.g. 10 s). The time-to-live covers gaps caused by speech pauses or detection or transmission failure. If no such track exists, a new one is started.

4. EVALUATION

The method was investigated using simulations and then evaluated on recordings in a real conference room. A localization is considered correct if the angles hit the target within an average head width of $0.2\,\text{m}$ or the 2D coordinates are within a typical persons shoulder width of $0.5\,\text{m}$. The precision and recall are calculated in $0.6\,\text{s}$ windows based on the correct localizations. These margins should be sufficient for most practical applications.

4.1. Simulation of a Single Node

A single microphone array was simulated using the image-source method [19] with a shoe-box model. A stationary single speaker at varying positions was simulated with 30 dB SNR and T_{60} times up $2.0\,\text{s}$. The proposed method was compared to the SRP-PHAT approach, Fig. 4 shows the angular RMS (root mean square) error ε_a. The SRP-PHAT shows a higher localization error than the proposed method, especially in cases of strong reverberation. Both an analysis of variance and randomization tests showed the difference to be statistically significant $(p < 0.01)$ for all T_{60}s.

4.2. Simulation of multiple nodes

In order to test the method, a simulation of five nodes located on a table in the center of the room and a single speaker moving around it, speaking from 18 static locations, was done using the image-source method.

4.2.1. Localization

The error and its deviation over the room position for tracking with all five nodes is listed in table 1. The weighting by the intersection quality q consistently improves the results.

$T_{60}[s]$	$\varepsilon_a[°]$		ε_l [m]	pr. [%]	re. [%]
0.00	2.1±1.0	-	0.23±0.19	94±17	92±18
		q	0.15±0.05	100±00	98±05
0.50	4.0±2.5	-	0.93±1.12	75±32	69±31
		q	0.31±0.13	97±09	92±15
1.00	5.0±4.3	-	0.68±0.51	75±29	67±31
		q	0.39±0.43	90±24	73±28

Table 1. Tracking error using five nodes in simulation without (-) and with weighting (q) used in triangulation.

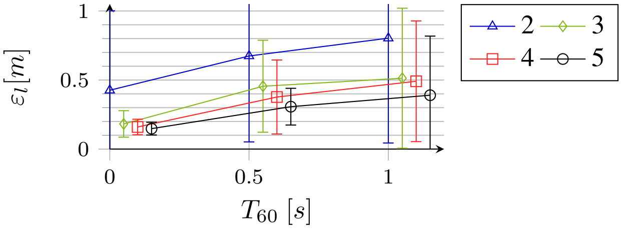

Fig. 5. Tracking error for different node counts.

4.2.2. Node Count

In order to investigate the influence of the node count and the effect of transmission failure, a fixed number of nodes were selected randomly for each time step. Figure 5 shows the RMS 2D localization error ε_l for different counts and reverberation times. When only two nodes can be used, the accuracy is significantly worse because the triangulation is using only bad angles in some cases. The accuracy increases with the number of nodes.

4.2.3. Drift and Jitter

Drift can be avoided by using the integration nodes clock as reference by exchanging information every frame in real time. Exact clock synchronization is not required when using DoA estimates rather than TDoAs [11]. Severe jitter of up to a time step $(L/4 = 0.15\,\text{s})$ may be the result of different clocks at the nodes or the transmission to the integrating node. This was simulated by introducing random jitter to the nodes inputs signals at $T_{60} = 0.5\,\text{s}$. The tracking error increases slightly only for jitter approaching the length of a time window (0.6s) as displayed in Fig. 6. This shows that jitter can be neglected and no compensation is necessary.

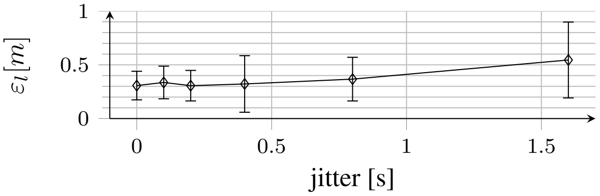

Fig. 6. Tracking error for different intra-array jitter.

	#1 one static speaker			#2 one moving speaker			#3 two static speakers			#4 three speakers		
	precision	recall	RMS	pr.	recall	RMS	pr.	recall	RMS	pr.	recall	RMS
$\theta^{(0)}$	100%	96%	6.4°	100%	99%	7.2°	100%	96%	4.4°	93%	76%	6.3°
$\theta^{(1)}$	100%	96%	4.8°	100%	99%	7.6°	100%	99%	6.5°	100%	99%	5.0°
$\theta^{(2)}$	100%	96%	5.7°	100%	99%	4.6°	72%	69%	10.8°	94%	96%	3.3°
x,y	99%	95%	0.22 m	93%	93%	0.34 m	88%	85%	0.33 m	99%	98%	0.19 m

Table 2. Tracking performance on real recordings.

4.3. Recordings

In order to test the real-world performance, recordings of multiple moving speakers were made in a highly reverberant $3.7 \times 6.8 \times 2.6\,\mathrm{m}^3$ conference room of a smart house installation at our university. Signals from three circular microphone arrays with 5 microphones in a 5 cm radius embedded in a table were recorded at 48 kHz. Each array was captured by a separate sound card. Recordings of coherent white noise showed a jitter of 22 μs between the sound cards. A reverberation time of 670 ± 89 ms over the microphone signals was calculated using a blind estimation algorithm [20]. The linearized ground truth annotations did not reflect slight head movement or speed and position variations.

4.3.1. Speaker Localization

As a general test of localization quality for the intended applications, a single speaker took 15 position sitting at and standing around the table uttering a sentence at each one. As listed in table 2, the angular precision was around 5° and about 95% recall was achieved. The 2D tracking had an error around 0.22 m. Without application of the weighting q, the error was 1.04 m.

In sequence #2, the speaker was moving slowly in front of the table on a linear trajectory. The angular precision was around 7° and 99% recall was achieved. The 2D tracking had an error around 0.34 m (0.54 m without q).

4.3.2. Concurrent Speakers

In sequence #3, speaker starts talking, a second one joins in so that they then both talk concurrently and then only the second one keeps talking. The association by the spectra produced no errors. As listed in table 2, the angular precision was between 5°-10° for the nodes, the higher error may reflect the fact that one speaker was very far from the table. Despite this, the 2D tracking had an error around 0.33 m with 85% recall.

Three speakers were talking around the table in sequence #4. Figure 7 shows the tracking result. The association by the spectra produced no errors. The angular precision was around 5° and about 96% recall was achieved. The 2D tracking had an error around 0.19 m (0.28 m without q).

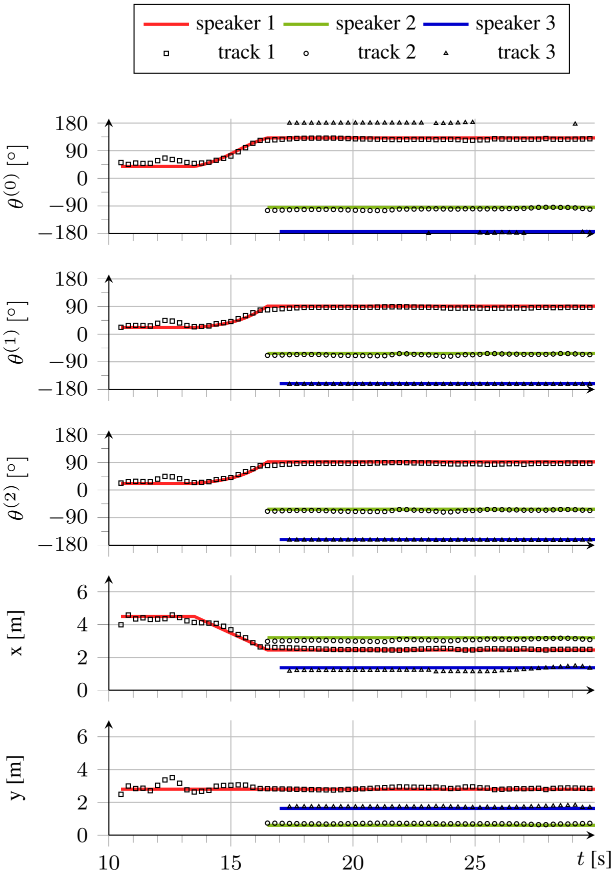

Fig. 7. Tracking three concurrent speakers (sequence #4).

5. CONCLUSION

A method for multi-speaker tracking using distributed nodes with microphone arrays was described. Its capability of accurately tracking static as well as moving speakers was shown with recordings of natural speakers in a reverberant environment. The correct association of the speakers was shown by the ability to handle two or three concurrent speakers. The weighted triangulation is shown to derive improved Euclidean coordinates. The robustness against drift and jitter as well as temporary transmission failure was shown in simulation. Two to five nodes were used, where three nodes already produce good precision for practical applications. The method is real-time capable with a delay of below one second.

6. REFERENCES

[1] Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Guillaume Lathoud, Mark Barnard, and Dong Zhang, "Automatic analysis of multimodal group actions in meetings.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 3, pp. 305–17, Mar. 2005.

[2] Fotios Talantzis, Aristodemos Pnevmatikakis, and Anthony G Constantinides, "Audio-visual active speaker tracking in cluttered indoors environments.," *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 39, pp. 7–15, Feb. 2009.

[3] Damien Kelly, Anil Kokaram, and Frank Boland, "Voxel-Based Viterbi Active Speaker Tracking (V-VAST) with Best View Selection for Video Lecture Post-Production," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, 2011, pp. 2296–2299.

[4] Youssef Oualil, Friedrich Faubel, and Dietrich Klakow, "A Fast Cumulative Steered Response Power for Multiple Speaker Detection and Localization," in *European Signal Processing Conference*, Marrakech, Morocco, 2013.

[5] Maximo Cobos, Jose J. Lopez, and Sascha Spors, "Analysis of Room Reverberation Effects in Source Localization using Small Microphone Arrays," in *International Symposium on Communications, Control and Signal Processing*, Mar. 2010.

[6] Nilesh Madhu and Rainer Martin, "A Scalable Framework for Multiple Speaker Localization and Tracking," in *Int. Workshop on Acoustic Echo and Noise Control*, Seattle, Washington USA, 2008.

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

[8] Axel Plinge, Daniel Hauschildt, Marius H. Hennecke, and Gernot A. Fink, "Multiple Speaker Tracking using a Microphone Array by Combining Auditory Processing and a Gaussian Mixture Cardinalized Probability Hypothesis Density Filter," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, 2011, pp. 2476–2479.

[9] Marc Pollefeys and David Nister, "Direct Computation of Sound and Microphone Locations from Time-Difference-of-Arrival Data," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Las Vegas, NV, 2008, pp. 2445–2448.

[10] Marco Crocco, Alessio Del Bue, Matteo Bustreo, and Vittorio Murino, "A Closed Form Solution to the Microphone Position Self-Calibration Problem," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, Mar. 2012, pp. 2597–2600.

[11] Florian Jacob, Joerg Schmalenstroeer, and Reinhold Haeb-Umbach, "DOA-based Microphone Array Postion Self-Calibration using Circular Statistics," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, 2013.

[12] Richard F. Lyon, "Machine Hearing – An Emerging Field," *IEEE Signal Processing Magazine*, Sept. 2010.

[13] DeLiang Wang and Guy J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, IEEE Press, 2006.

[14] Axel Plinge, Marius H. Hennecke, and Gernot A. Fink, "Robust Neuro-Fuzzy Speaker Localization Using a Circular Microphone Array," in *Int. Workshop on Acoustic Echo and Noise Control*, Tel Aviv, Israel, Aug. 2010.

[15] John Woodruff and DeLiang Wang, "Sequential Organization of Speech in Reverberant Environments by Integrating Monaural Grouping and Binaural Localization," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1856–1866, Sept. 2010.

[16] Tobias May, Steven van de Par, and Armin Kohlrausch, "A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, 2011.

[17] Albert S. Bregman, *Auditory Scene Analysis*, MIT Press, 1990.

[18] Axel Plinge and Gernot A. Fink, "Online Multi-Speaker Tracking Using Multiple Microphone Arrays Informed by Auditory Scene Analysis," in *European Signal Processing Conference*, Marrakesh, Morocco, Sept. 2013.

[19] Jont B. Allen and David A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[20] Heinrich W. Löllmann, Emre Yilmaz, Marco Jeub, and Peter Vary, "An improved algorithm for blind reverberation time estimation," in *Int. Workshop on Acoustic Echo and Noise Control*, Tel Aviv, Israel, Aug. 2010.