published in: *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC) 2012*

REVERBERATION-ROBUST ONLINE MULTI-SPEAKER TRACKING BY USING A MICROPHONE ARRAY AND CASA PROCESSING

Axel Plinge, Marius H. Hennecke, and Gernot A. Fink

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

ABSTRACT

Online tracking of speakers is an important task for applications in smart environments such as camera control, meeting annotation and speech separation. Challenges for an audio-only system are small-room reverberation, noise, the unknown number of speakers, and gaps occurring in natural speech. Combining models from neurobiology and cognitive psychology with many-channel signal processing and pattern recognition techniques, a hybrid method was developed. By employing online CASA processing to signals from a microphone array, the real-time capable method is able to track an arbitrary number of concurrent moving speakers in highly reverberant environments.

Index Terms— microphone array, auditory scene analysis, glimpsing model, speaker tracking

1. INTRODUCTION

Human listeners show an impressive ability to locate and separate concurrent speakers by hearing in everyday situations. A popular theory, based on psychoacoustic experiments as well as biological and neurological research, is the "Auditory Scene Analysis" (ASA). Computational ASA localization and tracking applications use two sensors of an artificial human head and evaluate IID and ITD [1], while technical solutions often employ circular or t-shaped microphone arrays with eight or more sensors and evaluate only the time delay of arrival [2]. Biologically inspired systems were shown to outperform technical approaches to localization such as the GCC-PHAT [3,4]. Recently, hybrid approaches applying neuralbiologically inspired processing to microphone arrays were introduced [3,5]. According to ASA, the auditory information is clustered in a step called simultaneous grouping and then combined over time in sequential integration by common features such as location, spectrum and pitch. Thereafter model-based integration is done top-down using *a priori* knowledge such as speaker movement and speech models. Common tracking strategies are the clustering of localizations with probability-based methods such as the EM-algorithm followed by rule-based combination [6] or particle filtering. Difficulties are the movement and varying number of speakers over time, and the fact that human speech contains gaps and pauses. In reverberant and noisy environments, only a few clearly localized time-frequency events can be found. The "glimpsing model" suggests that human speech perception in adverse conditions is based upon sparse clear events with high signal-to-noise ratio [7]. To localize and track multiple speakers in reverberant environments, clearly localized "glimpses" can be integrated to continuous speaker tracks.

We would like to thank Szilárd Vajda for his suggestion of the DBScan algorithm and Heinrich W. Löllmann for his cooperation. This work was in part supported by the German Research Foundation (DFG) under contract number Fi 799/5-1.

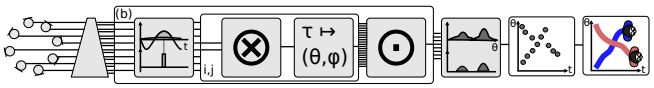

Fig. 1. Processing structure

2. METHOD

The processing steps of the many-channel speaker tracking are shown in figure 1. First, the signals of a circular microphone array are sampled and each signal is processed by a cochlear model, then microphone pairs are correlated by a midbrain model, mapped into the spatial domain and combined by a fuzzy t-norm. Speech localizations are calculated by a simultaneous grouping. Then continuous tracks are derived by using both sequential and model-based integration.

2.1. Cochlear and Midbrain Model

The implementation of the cochlear and midbrain model is described in detail in [3]. The microphones' signals are filtered by a gammatone filterbank composed of $n_B = 16$ bands with center frequencies equidistantly spaced on the ERB scale between 250 Hz and 3 kHz. The cochlear model makes use of onset dominance and glimpses events with high modulation ($t_g = 6\,\text{dB}$). Correlations between microphones are calculated in time frames of above 12 ms advanced in 6 ms steps. Spherical far field source positions (θ, ϕ) for backprojection are discretized with azimuth $\theta \in \{0°, 1°.., 359°\}$, and elevation $\phi \in \{0°, 5°.., 35°\}$. By fuzzy combination, robust but sparse spatial likelihoods are calculated.

2.2. Peak Localization

In the presence of strong reverberation, the number of clearly modulated events per second is small and secondary peaks are hard to avoid. In a larger time context, reverberant speech is found to produce Gaussian distributed peaks. Thus, when considering larger time segments, the correlation results can be modeled as "true" peaks plus noise [6].

To accommodate for fast moving speakers, a short moving average over $0.5s$ is calculated over all data points with a shift of $0.125s$. To find angular peaks, a peak-over-average filtering step in analogy to the difference-of-Gaussian processing found in human perception is applied: A $45°$ average, spanning the reverberation induced artifacts, is subtracted from a $5°$ average. Since in most practical scenarios, speakers can be separated by azimuth, the maximum value over all elevations ϕ can be used:

$$\hat{e}_{l,\theta,b} = \max_{\phi} \{\breve{e}(5)_{l,\theta,\phi,b} - \breve{e}(45)_{l,\theta,\phi,b}\} \ . \tag{1}$$

Then the positive values for each frame index l and azimuth θ are collected over all bands b giving an estimate of the spectral distribution

$$s_{l,\theta,b} = \max\{\hat{e}_{l,\theta,b}, 0\} \tag{2}$$

$$\boldsymbol{s}_{l,\theta} = (s_{l,\theta,0}, s_{l,\theta,1}, \ldots, s_{l,\theta,n_B-1})^T \ . \tag{3}$$

The energy values comprise a set of azimuth-spectrum tuples

$$D_l = \{(\theta, \boldsymbol{s}_{l,\theta}) \mid \textstyle\sum_b s_{l,\theta,b} > t_e\} \tag{4}$$

for each time frame considered speech energy detections where the sum exceed a threshold of $t_e = -40\,\mathrm{dB}$.

2.3. Simultaneous Grouping

According to the ASA theory, location as well as spectral cues are used to group the auditory information as coming from a certain source. Under the sparsity assumption no or only few values originating from different speech sources collide in frequency and angle. Coinciding energy in multiple frequency bands at a similar azimuth provides independent speech indication.

The process of simultaneous grouping is here emulated by clustering of the detections by azimuth and spectral similarity. The spectral similarity between two detections $x = (\alpha, \boldsymbol{s})$ and $y = (\beta, \boldsymbol{t})$ is calculated as normalized scalar product

$$cs(x, y) = \frac{\sum_b s_b t_b}{\sqrt{\sum_b s_b^2 \sum_b t_b^2}} \tag{5}$$

and the angular similarity is represented the angular distance

$$da(x, y) = (\alpha - \beta) \mod 360 \,. \tag{6}$$

Detections D_l over three consecutive frames $l-1, l, l+1$ are clustered for each index l. The clusters are computed with a density-based clustering approach inspired by the DBScan algorithm [8] that iteratively expands regions with sufficient density. For a detection x the set of neighbors is defined as

$$N(x) = \{y | da(x, y) < \Delta\theta_1 \wedge cs(x, y) > \Delta S\}, \tag{7}$$

with a spectral correlation above $\Delta S = 0.7$ and azimuths closer than $\Delta\theta_1 = 12$. The following steps are executed:

1. Mark all detections unvisited.
2. Find next unvisited detection $x \in \{D_{l-1}, D_l, D_{l+1}\}$ and its neighbors $N(x)$.
   (a) If $|N(x)|$ is less than $\varepsilon_n = 12$, all unvisited detections are discarded as noise and marked visited.
   (b) Otherwise, the unvisited detections in $N(x)$ form the next cluster C_κ and are marked visited. The cluster is iteratively expanded by all unvisited neighbors in $N(y)$ to any $y \in C_i$ until no such neighbors are left.
3. If any unvisited detections are left, continue at 2.

By precalculation of a distance matrices, the neighborhood query can be executed in $O(\log n)$ where $n = |D_l|$. Since each detection is visited only once, the clustering is done in $O(n \log n)$. From the elements $(\theta, \boldsymbol{s})$ in each cluster C_κ at a given frame l, the average spectrum

$$\boldsymbol{S}_{\kappa,l} = \frac{1}{n_b} \sum_{(\theta,\boldsymbol{s}) \in C_{\kappa,l}} \boldsymbol{s} \tag{8}$$

is calculated as well as the clusters' centroid azimuth weighted by spectral energy

$$A_{\kappa,l} = \frac{\sum_{(\theta,\boldsymbol{s}) \in C_{\kappa,l}} \sum_b \theta s_b}{\sum_{(\theta,\boldsymbol{s}) \in C_{\kappa,l}} \sum_b s_b} \tag{9}$$

as estimate of the source angle. Considering the typical spectral spread of natural speech sounds, clusters spreading less than a third of the n_b frequency bands are discarded as non-speech such as machine noise or reverberation. The remaining grouping results for each time frame l are represented by the set of time-azimuth-spectrum tuples

$$R = \left\{ (l, A_{\kappa,l}, \boldsymbol{S}_{\kappa,l}) \,\middle|\, |\{s_{\kappa,l,b} > 0\}| > \lfloor n_b/3 \rfloor \right\} \tag{10}$$

representing speaker localizations.

2.4. Sequential and Model-Based Integration

The next step of the human auditory processing according to ASA is the sequential integration of groups into simultaneous streams. It is not practical to compute these over larger time periods by clustering for two reasons: First searching for cluster points over long time periods would lead to long

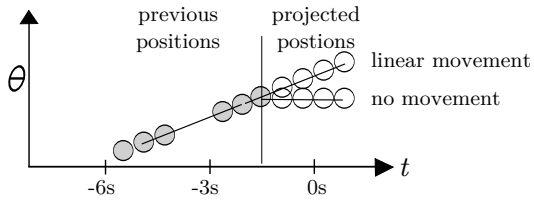

Fig. 2. Projected future speaker positions.

search times. Second, iterative updates are necessary for online processing. Therefore, the sequential integration is done by an online algorithm connecting clusters with similar spectral and location cues.

The final ASA step of model-based integration is implemented by using top-down knowledge in form of a simple model of speaker movement. The tracks of a natural speaker contain gaps due to speech pauses and omissions due to the glimpsing rule. We use a time-to-live (TTL) rule for speaker assignment over these gaps [6]. Adjacent localizations are assigned the same track if they are similar in both spectrum and location. In order to model both fixed and moving speakers, points at the same position as well as points on the same trajectory are assigned the same track if not older than t_{TTL} of up to 12 s and closer than $\Delta\theta_2 = 40°$. The trajectories are derived by linear interpolation of the last two 3 s stretches of the previous elements within t_{TTL} of each track, as illustrated in figure 2. The trajectory can not be steeper than $\Delta\theta/s = 40°$ for normal moving speakers.

Tracks are calculated implementing both steps by applying a number of rules successively for each time frame l.

1. *Sequential Integration* Localizations $x = (l, \cdot, \cdot) \in R$ for which a track element in the previous frame $y = (l-1, \cdot, \cdot) \in T_i$ with similar spectra $cs(x, y) > \Delta S$ and close azimuth $da(x, y) < \Delta\theta_2$ exists are added to the same track, $T_i = T_i \cup x$.
2. *Linear Movement* Localizations are added to the nearest track $T_j = T_i \cup x$ if they are on the same trajectory if joined by either a gap or having similar spectra, $cs(x, y) > \Delta S$.
3. *No Movement* Localizations x are added the nearest track $T_j = T_i \cup x$ containing an element $y \in T_j$ no farther than $\Delta\theta_2$ and t_{TTL} away.
4. *Birth Rule* For any remaining points, a new track k is started, $T_k = \{x\}$.

Tracks consisting of one or two isolated localizations are discarded as noise, gaps shorter than 2 s in the remaining tracks are closed by linear interpolation.

3. EVALUATION

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

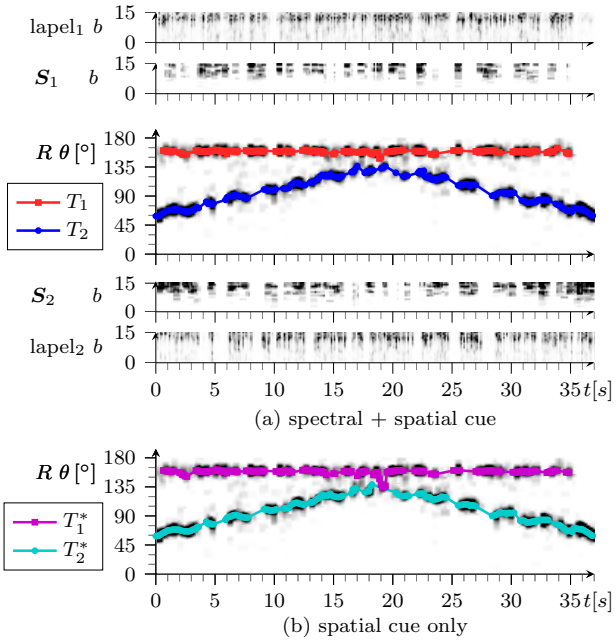

Fig. 3. Scenario #2: Close speakers tracked using spectral and spatial information (a) and spatial information only (b).

our university. Signals from a circular microphone array with 8 microphones in a 5 cm radius were recorded at 48 kHz. The speakers were wearing lapel microphones for most recordings. They were ordered to follow a preassigned trajectory that was matched to a linear ground truth. The optimal subpattern assignment (OSPA) distance was used to evaluate the tracking quality, it is defined on the space of finite sets of tracks and reflects labeling errors [9]. Precision and recall were calculated allowing an error corresponding to an average width of a human head of 0.2 m in the target distance.

In a first recording, one speaker was talking for one minute towards the array on a fixed position. A reverberation time of 624 ± 54 ms over all microphone signals was calculated using a blind estimation algorithm [10]. The speaker was localized with an OSPA distance of 0.65°, 100% precision and recall were achieved.

In a second recording, the effect of spatial proximity was tested. One speaker was sitting at a fixed position and talking, while a second speaker was talking overlapping while moving towards the first till up close and away again. In figure 3a, speaker localizations R are shown with the calculated tracks before linear interpolation plotted on top, connected by lines. The sum over all bands is plotted with black representing high values, white zero values or non-elements. Above and below, the spectra for each track as well as the lapel microphone signals are plotted. The OSPA distance was 5.48°, the precision 100% and recall 90.9%. When only the location cue is used (ΔS=0), the overlap in the spatial likelihood leads to distortion of the tracks and deviation of the tracks into each other, cp. figure 3b. The OSPA distance increased to 6.61°.

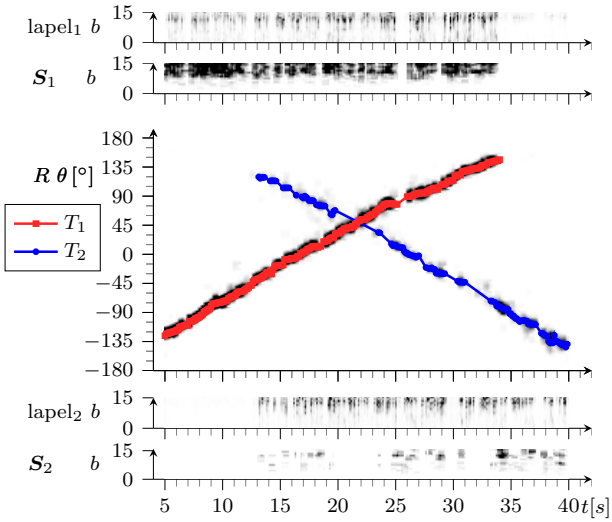

Fig. 4. Scenario #3: Tracks for crossing trajectories.

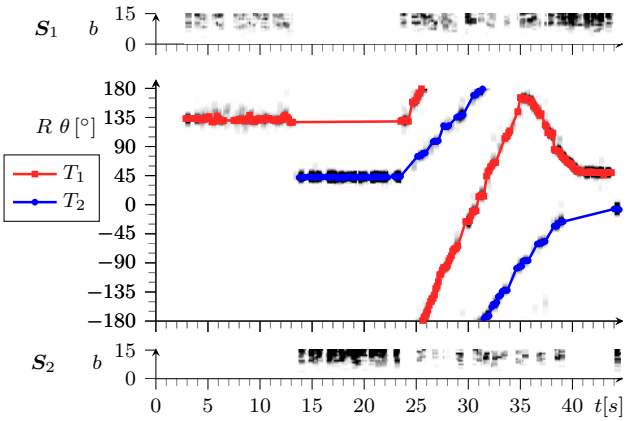

Fig. 5. Scenario #4: Two moving and talking speakers.

In a third scenario, crossing trajectories were investigated. Two speakers were walking in concentric circles around the microphone array in different directions, therefore crossing their angular trajectory. As shown in figure 4, the tracks computed accurately represent the crossing trajectories. Speaker one was louder, almost drowning out the other at the crossing point. However, due to the moving speaker rule, the trajectory of speaker two is continued correctly after the crossing. The OSPA distance was $6.68°$, the precision 100% and recall 86.7%. When comparing the spectra of the lapel microphones to the clustered ones, it becomes apparent that the latter contain more gaps, which is due to the glimpsing and onset dominance.

In a fourth scenario, two speakers were first talking in turns, then walking around the array while talking simultaneously. Results are depicted in figure 5. The OSPA distance was $4.43°$, the precision 100% and recall 81.2%. The association of the speakers over speech pauses via the large t_{TTL} is correct, but should be verified by integrating a speaker model.

4. CONCLUSIONS

The application of a cochlear and midbrain model to a microphone array derives reverberation-robust spatial likelihoods. It was combined with simultaneous grouping by density-based clustering employing both spectral and spatial cues, which efficiently provides sound speaker localizations. Sequential and model-based integration using spectral and spatial cues is able to calculate accurate tracks for concurrent speakers in a highly reverberant conference room. The implementation copes well with difficult scenarios such as fast movement and close or crossing trajectories. It is realtime-capable with a latency below 1 s.

5. REFERENCES

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

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

[3] 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.

[4] T. May, S. van De Par, and A. Kohlrausch, "A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 1, pp. 1–13, 2011.

[5] R. M. Stern, E. B. Gouvea, and G. Thattai, "Polyaural Array Processing for Automatic Speech Recognition in Degraded Environments," in *Proc. INTERSPEECH*, 2007, pp. 926–929.

[6] 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.

[7] M. P. Cooke, "A Glimpsing Model of Speech Perception in Noise," *J. Acoust. Soc. Am.*, vol. 119, pp. 1562–1573, 2006.

[8] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proc. Conf. on Knowledge Discovery and Data Mining*, Portland, OR, USA, 1996, pp. 226–231.

[9] Branko Ristic, Ba-Ngu Vo, Daniel Clark, and Ba-Tuong Vo, "A Metric for Performance Evaluation of Multi-Target Tracking Algorithms," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3452–3457, 2011.

[10] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *12th Int. Workshop on Acoustic Echo and Noise Control*, Tel Aviv, Israel, 2010.