USING THE VOICE SPECTRUM FOR IMPROVED TRACKING OF PEOPLE IN A JOINT AUDIO-VIDEO SCHEME

Eleonora D'Arca, Neil M. Robertson and James Hopgood

Joint Research Institute for Signal and Image Processing,
Heriot-Watt University & University of Edinburgh, UK
visionlab.eps.hw.ac.uk

ABSTRACT

In this paper we present a new solution to the problem of speaker tracking among people where occlusions occur (disappearance and non-speaking). In a normal conversation between two or more people, we learn speaker mel-cepstral coefficients (MFCC) and incorporate this information into a sequential Bayesian audio-video position tracker. The joint video-to-audio data association step is thus improved and we achieve robust person recognition which in turn aids tracking performance. We provide comprehensive evaluation via simulations and real data quoting tracking accuracy, precision and diarisation error rate (DER) compared to ground truth. For simulate and real experiments in an open space the trajectory tracking performance increases by 20% measured against ground truth using our approach. As a further enhancement versus the state-of-the-art, speaker identity recognition at a distance is improved by 20% by exploiting audio-video localisation cues.

Index Terms— Distant Speaker Recognition, Speaker Tracking, Multimodal tracking, MFCC, EKF

1. INTRODUCTION

Bayesian multi-modal speaker tracking based on audio and video positional data has been shown to effectively address most of the common problems that audio-only and video-only tracker normally faces in meeting-room applications (e.g. diarisation) [1, 2, 3, 4, 5, 6]. There are difficulties scaling these approaches up to larger spaces. In fact, from structured and extensive sensor networks (lapel/arrays of microphones and multi/panoramic cameras strategically positioned) and small area of interests ($4 - 5\ m^2$) focus is moved towards sparely displaced sensors and analysed spaces which are double or three times the size of the previous ones. In such larger dynamic scenarios, when people occlude each other, as in normal social interactions, systems cannot distinguish the actual speaker location and identity. This is principally because video tracks merge [7, 8, 9].

NR and JH are supported by EC FP7 LOCOBOT (Grant EC/260101).

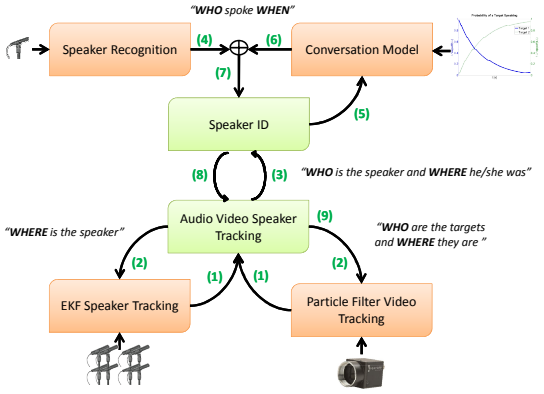

Fig. 1: A schematic of the system presented in this paper. Constituent parts of this diagram are referred to explicitly in the text (e.g. "arrow 1").

In confined meeting rooms audio-video (AV) speaker segmentation, i.e. correctly inferring over time who is the actual speaker, is done via lip detection and head movements plus GCC or SRP-PHAT audio localisation cues and voiceprint recognition techniques [6, 5, 10, 11, 12]. However, in more general scenarios (larger spaces) people are often recorded from a distance, meaning they: (a) do not face the cameras, and (b) their voices are captured from microphones widely scattered (rather than in an array or worn on the person). If targets get too close to one another speaker position is no longer resolvable by the AV estimator using time-difference of arrival [13]. Proximity-based association algorithms [7, 8, 9] fail to correctly match noisy audio and video data streams. One approach to solve this would be to make recourse to full 3-D tracking but the system complexity may become very high.

The grand aim of our work is to recognise person's "role" in a larger social interaction (e.g. a party). This task normally involves non-trivial data association [14] and blind source separation [15]. We first tackle, in this paper, a slimmed-down version of this problem. The novel system we present here can localise and recognise speaker identity among two people in difficult circumstances. We integrate speaker recognition

(SR) to make an audio-video tracking (AVT) system (based on Kalman filtering) robust to close audio sources (conversation). In summary, our work results in the following contributions: *a*) detection and tracking of speaker identity through occlusion; *b*) a novel smoothing prior (conversation model (CM)) increases the speaker recognition detection rate at a distance; *c*) exploitation of a small sensor network and fast visual tracking algorithms (using a single camera and 8 microphones) which works in a 12 m^2 area where people freely move; *d*) evaluation vs. the state-of-the-art using recognised metrics for multi object tracking (2006 and 2007 CLEAR dataset and speech recognition 2006 NIST dataset), in contrast to the closest papers to ours in the literature [6, 5, 8, 9].

2. THEORY

A schematic diagram of our system is shown in Figure 1. We now describe in more detail the components of this system.

2.1. Audio-Video Tracking

A description of the basic speaker trajectory 2-D audio-video tracking (AV) unit can been found in our previous work [16]. We repeat the salient details for clarity. The posteriors of an extended Kalman filter (EKF) audio tracker and of an independent video tracker based on a GPU-accelerated particle filter with ellipsoid models for people [17], $\mathbf{x}_a$ and $\mathbf{x}_v$, are fused in a common Kalman filter node (arrow 1, Figure 1). Hence, its joint output $\mathbf{x}_{av} = \mathbf{P}_a^{-1}\mathbf{x}_a + \mathbf{P}_v^{-1}\mathbf{x}_v$, is fed back into the individual audio and video trackers to improve the single modality estimation (arrow 2).Assuming people speak alternatively, as in a normal conversational mode, to a single audio signal $\mathbf{z}_a$, corresponds several video measurements at a time, $\mathbf{z}_{v_i}$, one for each of the N detected targets. By basing the audio-to-video data association step on spatial proximity, i.e. nearest neighbour (NN), speaker segmentation and recognition can also be obtained (arrow 3) as long as people *are resolved* by the AVT and its measurements can be considered robust with respect to the speaker motion model. (This is not possible where speakers are close and is the prime motivation for introducing MFCC to the audio track). In particular, the speaker identity inferred by the AVT is equal to the one of the i-th target if $S_{AV} = \arg\max_i \left\{ \mathbf{p}(\mathbf{z}_a, \mathbf{z}_{v_i} \mid \mathbf{x}) \right\}, i = 1, ..., N$.

2.2. Text-Independent Speaker Recognition

When more people are detected in an image and occlusions happen, appearance-based video trackers can no longer support the audio track hypothesis. This makes it for different information to be integrated. As microphones already gather audio information for AVT purposes, the frequency content of at least one of them can be used to segment over time the speaker and recognise their identity (ID).

The speaker recognition (SR) module performs text-independent speaker recognition based on Gaussian Mixture Models (GMM) [18] (arrow 4), under the assumptions that there exist $M = N$ possible speaker ID, whose voiceprints models $\mathbf{p}(s_j)$ are learned before the experiment is performed. We also assume that no false measurements nor missed detections are present. In particular, the speaker models are calculated on the base of 60 s training signal for each speaker. From every voice sequence 12 sets of mel-cepstral coefficients (MFCC) are extracted. Each model is represented by a 16-mixture GMM whose parameters are estimated by the extracted acoustic MFCC vectors. Particularly, an EM algorithm iteratively estimates them to monotonically increase the likelihood of the proposed GMM. It converges when the model likelihood reaches a local maximum. The test conversation sequence, not recorded in matching conditions, is framed in small speech-only sub-segments which are considered to be long enough to detect a speaker change. For each speech sub-segment its MFCCs are extracted and compared to the available database of speaker models to determine the likelihood of the particular speaker S_{SR} to be the one who uttered the considered speech segment s_j i.e. $S_{SR} = \arg\max_j \left\{ \mathbf{p}(S \mid s_j) \right\}, j = 1, ..., M = N$. Speaker DER results are presented in Tables 1 and 2 in column *'Experiment'*.

2.3. Speaker Conversation Model

In our experiments, speakers can be at most 2 m distant from the sensors. Distance has been proved to be a critical factor to diarisation error rate (DER) accuracy. In [19] a 2 m distant microphone shows a $\approx 20\%$ DER. Thus, we further introduce a new speaker switching probability to model the amount of time that has to be elapsed before a person stop talking once they have started (arrow 6) i.e. $S_{CM} = \arg\max_j \left\{ \mathbf{p}(S \mid u_j) \right\}$. This acts as a smoothing prior on speaker ID recognition. In particular, we assume that the amount of time we have to wait before a speaker u_j finishes talking is proportional to the elapsed time t. We define this as an exponential probability density function i.e. $\boldsymbol{p}(u_j) = exppdf(\lambda, t)$. The remaining potential speakers $M - 1$ are characterised by a probability of starting the conversation which is given by the complementary speaking probability scaled by $M - 1$. The CM is triggered by the $j - th$ speaker ID detection which is given by an averaged decision fusion of the AVT and the SR modules i.e. $S_{CM}(0) = \mathrm{w}_{AV}\ \mathrm{S}_{AV} + \mathrm{w}_{CM}\ \mathrm{S}_{CM}$ (arrow 5), where w_{AV} and w_{CM} are evaluated on the base of the module confidence in inferring its decision.

2.4. Combining Tracking and Recognition

Once an identity i has been assigned to every target in a image frame, the person recognition decision derived from SR+CM

Experiment	System	MOTP (m)	MOTA (%)	DER (%)
'InOut' $SR_{DER} = 8.3\%$	AVT	0.34	79	50.61
	AVT + SR	0.27	80	17
	AVT + SR + CM	0.27	80	4.8
'CloselySpaced' $SR_{DER} = 10.7\%$	AVT	0.25	89	47
	AVT + SR	0.15	96	16.9
	AVT + SR + CM	0.11	96	7.3
'Crossing(Incorrect Split)' $SR_{DER} = 9.5\%$	AVT	0.08	100	44.6
	AVT + SR	0.08	100	9.7
	AVT + SR + CM	0.08	100	3.6
'Crossing(Merge on Listener)' $SR_{DER} = 9.5\%$	AVT	0.10	100	26.5
	AVT + SR	0.09	100	12.1
	AVT + SR + CM	0.07	100	2.4

Table 1: Simulations results.

Experiment	System	MOTP (m)	MOTA (%)	DER (%)
'Single Speaker' $SR_{DER} = 17.68\%$	Audio only	0.69	27	–
	Video only	1.30	45	–
	AVT	0.25	94	4.70
	AVT+SR	0.25	94	4.70
'Abandoning' $SR_{DER} = 23.5\%$	AVT	0.25	91	43.7
	AVT + SR	0.30	84	23.5
	AVT + SR + CM	0.30	84	2.2
'CrossingReal' $SR_{DER} = 20.6\%$	AVT	0.47	72	14.8
	AVT + SR	0.56	55	9.63
	AVT + SR + CM	0.55	55	2.2

Table 2: Real experiment results.

may be used in order to recover ID tracking when occlusions occur. When competitive association hypotheses exist for the AVT (i.e. the AVT confidence drops below a certain threshold), the third part voiceprint decision is averaged, according to their confidence value defined as in Jin *et al.* [19]. The CM decision is added in (arrow 7) so as to discriminate among the different video data i.e. $S = \mathrm{w}_{SR}\,\mathrm{S}_{SR} + \mathrm{w}_{CM}\,\mathrm{S}_{CM}$. The speaker ID is first fed back into the AVT to aid the resolving of the NN association ($S = S_{AV} = i \implies \mathrm{z}_v = \mathbf{z}_{v_i}$) and hence the speaker recognition plus the tracking (arrow 8). Secondly, it is sent to the video tracking unit to indirectly re-assign the correct targets appearance models z_{v_i}(t+1) = $\mathbf{x}_{av}$ (t) , thus resolving the occlusion (arrow 9).

3. EXPERIMENTATION AND RESULTS

We now present comprehensive results on simulated and real data. No comparison is made for them other than against our original system since, as far as we are aware, no other works exist with same experiment setups. In order to validate the proposed concept we simulate a $10 \times 10 \times 6.5\ m^3$ open room, characterised by a reverberation time ($T_{60} = 0.3\ s$), in which two people talk alternatively for $60\ s$ within an area of interest of $3 \times 4\ m^2$. In the room there are two speakers, only one of whom is active at any given time. Only 1 camera and 4 pairs of directional microphones are used. Full details can be found in [16] as well as all the details about the synchronisation and calibration of the sensors and the filters parameters. Specifically, filters were initialised using the video detected position of their correspondent targets and static matrices Q and R [20], whose values were chosen on the basis of an optimisation step. At last the CM parameters were learned from similar sequences, testing different values. Performance is evaluated using position tracking precision (MOTP) and accuracy (MOTA) [21]. The tracker is considered to have correctly hit the target if the distance between is output and the ground truth is within $50\ cm$. We also compute the diarisation error rate (DER) expressing the speaker error only [22]. The initial AVT results [16] are compared against a half-way AVT+SR solution which does not include the smoothing prior and to the final AVT+SR+CM solution to firstly shows benefit of introducing the SR module only and, secondly, of the smoothing prior (CM).

Simulated Experiment *'InOut'* considers a person speaking *Speaker*1 inside the FOV and leaving it from a side where at the same time another speaker *Speaker*2 is entering. Then, the last also leaves the FOV. The difficulty here is given by the fact that the speakers look alike so that the video tracker cannot distinguish between the targets. Our objective is met as it is shown the AVT cannot correctly identify the speaker ID whereas the AVT+SR+CM can.

Simulated Experiment *'Crossing'* simulates two people walking along crossing diagonal trajectories. *Speaker*1 speaks for the first half of their trajectories, while *Speaker*2 does it for their second half. The crossing points represents a potential occlusion zone for the camera, meaning the detected target trajectories could at worse either merge on the listener as in the *'Crossing(Merge on Listener)'* experiment, or diverge incorrectly in that area as in *'Crossing(Incorrect Split)'*. In both cases, the AVT+SR+CM solution can still track and recognise the actual speaker whereas the AVT is mistaken.

Simulated Experiment *'CloselySpaced'* reproduces a long term occlusion for the camera. In particular, both targets trajectories are close ($50\ cm$) and parallel to the camera image plane. The conversation is segmented as for the previous experiment. Being one of the target invisible to the video trackers for almost the whole experiment, the AVT tracking error is always quite large as the NN data association is compromised along the whole speaker trajectory. In turn, the AVT + SR + CM system, thanks to the external voiceprint information, can resolve the association notably decreasing the tracking error as well as the speaker error.

Results of tracking and speaker ID recognition are presented, averaged over 100 Montecarlo runs, in Table 1. Performance improvements are shown in Figure 2.

Having proved the concept with simulations, we move to a real indoor room where people can freely move. Audio and video data was gathered in a typical open office room, whose size is $111.44\ m^2$, where the area considered of interest is $12\ m^2$ (as seen in Figure 3**(a)**). Also we made no attempt to reduce normal background noise (desk fans, footsteps, talking etc.). A significant reverberation time ($T_{60} \approx 0.5\ s$) was measured. Ground-truth data was hand labelled considering feet position to $10\ cm$ of accuracy on a ground plane common

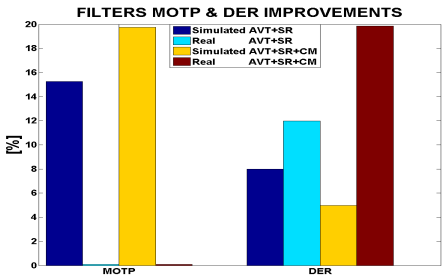

Fig. 2: Averaged filters performances for simulated and real experiments. MOTP improvement reaches 20% for the AVT+SR+CM algo as the simulations were specifically chosen to have low AVT tracking and recognition errors. Nevertheless, it must be noted that this is not quite the same for the real results as they replicate different scenarios were the AVT on its own is incorrect on ID recognition only. As for the DER, AVT+SR+CM outperforms the AVT+SR because of the conversation smoothing prior. Moreover real results are better than the simulated ones as the real experiment SR accuracy is lower than the SR accuracy in the simulations (f.e. 90.5% vs 79.4%).

to the cameras and the microphones. Synchrony of data was obtained by processing audio and video signals accordingly to the cameras frame rate 7.5 Hz. The point of this section is actually proving that this algorithm can maintain and recover tracking ID. Therefore we describe the results in terms of ID recognition rather than the precision which is obviously not high in such a challenging scenario if any further signal processing is used.

Real Experiment *'Single Speaker'* considers a person speaking along a rectangular trajectory for two times its perimeter, appearing and disappearing from behind an occlusion. Results as presented in Figure 3.

Real Experiment *'Abandoning'* shows a person walking and talking along a rectangular trajectory disappearing behind an occlusion. Then a second person, who looks alike the first one and who is speaking as well, reappears from behind the occlusion and walks along the rectangular trajectory till the point he disappears again. Visual results are presented in Figure 4.

Real Experiment *'CrossingReal'* shows two people with very similar appearance walking while having a conversation. They meet along a diagonal where they keep on walking past each other causing an occlusion in the resulting image. Results are presented in Figure 5.

Results of tracking and speaker ID recognition are presented in Table 2. Improvements are shown in Figure 2.

4. CONCLUSION AND FUTURE WORK

AV position-based speaker tracking and recognition at a distance is insufficient when speakers are close because of ID mismatches. We have shown that by further integrating voiceprint information and a conversation model the accuracy of speaker localisation increases by 38%. Also using localisation cues, speaker ID recognition improves on average by 20% in real room scenarios. This results in better scene understanding, which was our stated goal, and also AV diarisation. Given the high correlation between speech and body gestures, we are currently working on learning correlations parameters by observing speakers *gestures* to further improve speaker localisation and ID recognition.

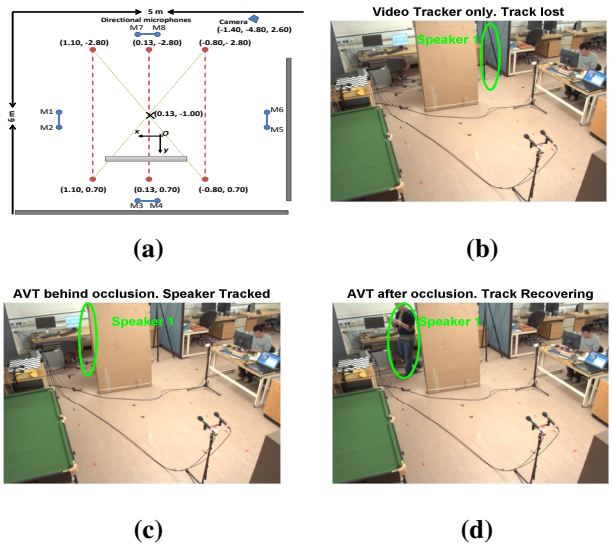

Fig. 3: Real experiments layout **(a)** and *'Single Speaker'* tracking results. In **(b)** the video tracker only loses the speaker track when a long occlusion occurs. In turn, **(c)** shows the AVT correctly locating the speaker through the occlusion. Finally **(d)** shows speaker track recovering - the video tracker alone is not capable of doing this.

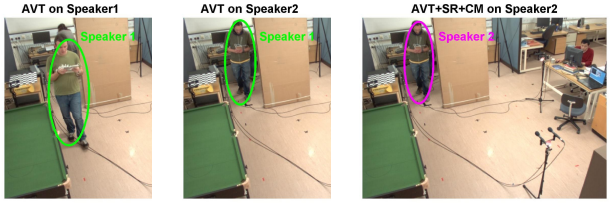

Fig. 4: *'Abandoning'* Tracking Results. **(a)** Shows the AVT locked onto *Speaker1*. In **(b)** *Speaker2* appears while *Speaker1* has left the scene. The tracking ellipse is still green coloured meaning the AVT cannot make a distinction between IDs as person appearance models are very similar. In **(c)** instead, the magenta ellipse indicates the AVT+SR+CM solution can correctly infer the person ID.

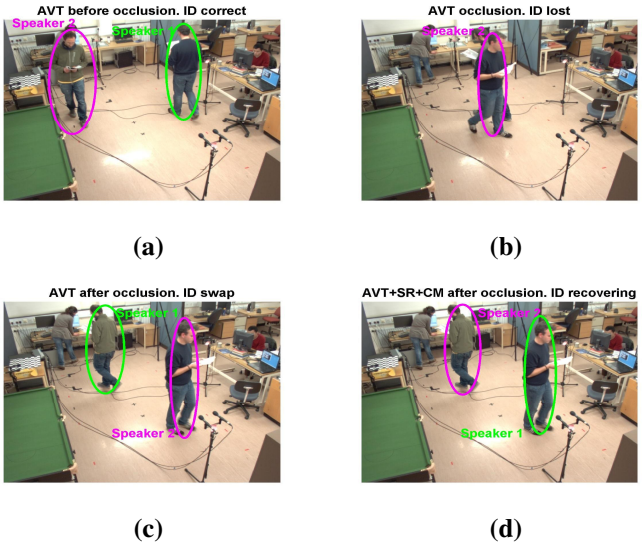

Fig. 5: *'CrossingReal'* Tracking Results. In **(a)** the AVT correctly identifies both people ID. In **(b)** a short term occlusion leads track to merge. This results in **(c)** in an ID swap as the ellipses colors have exchanged. On the other hand **(d)** presents the AVT+SR+CM result for the same situation i.e. correct ID recovering after the occlusion.

5. REFERENCES

[1] Dmitry N. Zotkin, Ramani Duraiswami, and Larry S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 1, pp. 1154–1164, 2002.

[2] Guillaume Lathoud, Jean-Marc Odobez, and Daniel Gatica-Perez, "Av16.3: An audio-visual corpus for speaker localization and tracking," in *MLMI*, 2004, pp. 182–195.

[3] M. Lincoln, I. McCowan, J. Vepa, and H.K. Maganti, "The multi-channel wall street journal audio visual corpus (mc-wsj-av): specification and initial experiments," in *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, nov. 2005, pp. 357 –362.

[4] Kai Nickel, Tobias Gehrig, Hazim Kemal Ekenel, John W. McDonough, and Rainer Stiefelhagen, "An audio-visual particle filter for speaker tracking on the clear'06 evaluation dataset," in *CLEAR*, 2006, pp. 69–80.

[5] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 601–616, Feb. 2007.

[6] Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi, "Audio-visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation," *J. Sel. Topics Signal Processing*, vol. 4, no. 5, pp. 882–894, 2010.

[7] N. Checka, K.W. Wilson, M.R. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," vol. 5, pp. V–881–4 vol.5, May 2004.

[8] Yeongseon Lee and R. Mersereau, "Data association for people tracking using multiple cameras," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 312008-april4 2008, pp. 2585 –2588.

[9] Huiyu Zhou, M. Taj, and A. Cavallaro, "Target detection and tracking with heterogeneous sensors," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 2, no. 4, pp. 503–513, Aug. 2008.

[10] H.K. Ekenel, M. Fischer, Qin Jin, and R. Stiefelhagen, "Multi-modal person identification in a smart environment," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, june 2007, pp. 1 –8.

[11] Gerald Friedland, Chuohao Yeo, and Hayley Hung, "Visual speaker localization aided by acoustic models," in *Proceedings of the 17th ACM international conference on Multimedia*, New York, NY, USA, 2009, MM '09, pp. 195–202, ACM.

[12] Xavier Alameda-Pineda, Vasil Khalidov, Radu Horaud, and Florence Forbes, "Finding audio-visual events in informal social gatherings," in *Proceedings of the 13th international conference on multimodal interfaces*, New York, NY, USA, 2011, ICMI '11, pp. 247–254, ACM.

[13] John McDonough Matthias Wolfel, *Distant Speech Recognition*, Wiley, 2009.

[14] C.-Y. Chong, "Tracking and data fusion: A handbook of algorithms (bar-shalom, y. et al; 2011) [bookshelf]," *Control Systems, IEEE*, vol. 32, no. 5, pp. 114 –116, oct. 2012.

[15] Shoji Makino, Te-Won Lee, and Hiroshi Sawada, *Blind Speech Separation*, Springer, 2007.

[16] Eleonora D'Arca, Neil Robertson, and James Hopgood, "Audio-video tracking of active speakers through occlusion," in *In Proc. of the 9th IET Data Fusion and Target Tracking Conference*, 2012.

[17] Wasit Limprasert, Andrew M. Wallace, and Greg Michaelson, "Accelerated people tracking using texture in a camera network," in *VISAPP (2)'12*, 2012, pp. 225–234.

[18] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72 –83, jan 1995.

[19] Qin Jin, Yue Pan, and Tanja Schultz, "Far-field speaker recognition," in *International Conference on Acoustic, Speech, and Signal Processing (ICASSP*, 2006.

[20] T. Gehrig, K. Nickel, H.K. Ekenel, U. Klee, and J. McDonough, "Kalman filters for audio-video source localization," pp. 118–121, Oct. 2005.

[21] Keni Bernardin and Rainer Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *J. Image Video Process.*, vol. 2008, pp. 1:1–1:10, January 2008.

[22] Jonathan G. Fiscus, Jerome Ajot, Martial Michel, and John S. Garofolo, *The Rich Transcription 2006 Spring Meeting Recognition Evaluation*, NIST, 2006.