Speaker Tracking Based on Face Detection and Voice Activity Detection Using a Microphone Array

Trond F. Bergh
University of Oslo/Squarehead Technology
Oslo, Norway

***Abstract*—We present a framework for multi-speaker localisation and tracking, as well as identification of the active speaker using a camera-assisted microphone array. The proposed algorithm is a hybrid solution, where a face-detection algorithm followed by a voice activity detection algorithm does speaker detection and tracking. Our working thesis is that this combination leads to a more robust solution than either method on its own. We will carry out experiments with a square microphone array with a 40 cm by 40 cm aperture and 256 elements.**

Keywords*—*video conferencing; active speaker localisation; face detection; microphone array; voice activity detection

I. Introduction

A system for robust speaker localisation and tracking is a sought-after solution in many applications. The challenge in general is perhaps best illustrated through the example of video-conferencing: a multi-speaker situation in which it is desirable to automatically detect and track the active speaker. This makes it possible to automatically direct the camera; and with one or more microphone arrays, steer the array beam(s), towards the active speaker. Thus speech is enhanced while interfering noise from elsewhere in the room is attenuated.

Positioning of the speaker can be done by different means, either using video-based methods such as face detection, or using audio-based direction of arrival (DOA) methods [1], or by combined audio-visual methods [2]–[4]. Visual methods by themselves tend to be unreliable, as the speaker is not necessarily facing a camera, or the speakers face may be partially or fully occluded. As microphone arrays are regaining some popularity in recent years, microphone array based DOA estimators used to direct cameras towards the active speaker have been suggested as a more robust solution than multiple single microphones. However, in multi-speaker situations and in reverberant environments, most DOA estimators suffer from correlation problems. Furthermore, the DOA alone might not suffice to accurately locate the active speaker, as non-speech noise might interfere. A solution to this is to use voice activity detection (VAD) [5]–[7] to discriminate human speech from other sound sources. However, spatial sweeping with the array beam combined with VAD quickly leads to a problem too CPU-intensive for realtime use.

Our proposed method is the opposite of the array-guided camera, we use a camera to guide the array beam, similarly to [8], [9], and use VAD to discriminate active speakers from non-speech audio sources. We investigate this method in the hopes that it will lead to a more robust and computationally tractable solution than systems using either audio or video on its own. Specifically, our method avoids beamformer sweeping, which is a computationally expensive process. The array-camera system architecture and calibration is further elaborated in [10].

II. Method

A schematic overview of the proposed method is shown in Fig. 1, with a step-by-step description of the algorithm following.

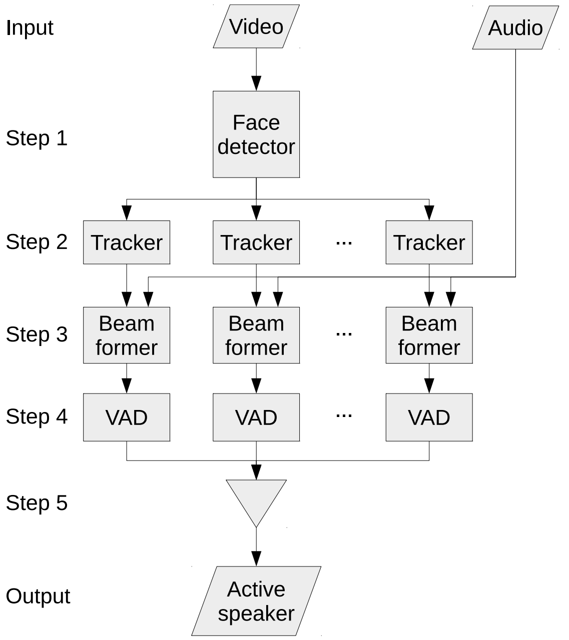

Fig. 1. Overview of the proposed method for automatic speaker identification and tracking.

Step 1: At this step a single video frame is analysed to detect faces. We use the OpenCV implementation of a cascaded classifier for object detection using Haar-like features [11]. The output of this step is a (possibly empty) set of coordinates at which faces have been detected in the frame, which we denote $\mathcal{D}$.

Step 2: We track the face positions across frames with multiple independent particle filters using the common sequential importance resampling (SIR) algorithm [12]. We assign the detections in from step 1 to individual particle filter trackers using a version of the Munkres algorithm for non-square matrices [13], since the number of detections do, in general, not match the number of trackers (initially, the set of trackers, $\mathcal{T}$, is empty). We use the squared distance between detection point and tracker expectation value as the cost function in the modified Munkres algorithm. Since, in general, $|\mathcal{D}| \neq |\mathcal{T}|$, we might have detections that have not been assigned to an existing tracker. In this case we initialise a new tracker with all particle coordinates equal to those of the unassociated detection, and hence all particle weights equal. If, on the other hand, we have unassociated trackers, a counter is incremented on each, and if the counter reaches a certain threshold, the tracker is deleted. We utilise a Langevin type particle motion model, as in [4].

Step 3: At this stage the tracker expectation values from step 2 is transformed from pixel coordinates to 3D cartesian coordinates using camera lens parameters. The transformed coordinates are used to steer the beamformers associated with each active tracker.

Step 4: At this step the output from each beamformer is fed into a VAD, based on cepstral distance, similar to [6]. Each VAD makes a binary decision of voiced/unvoiced audio frames, and keeps a running tally of the decision for the last 50 frames. As the audio sample rate is $44.1\ \mathrm{kHz}$, and the VAD operates with a frame size of 512 samples this means that the last $0.6\ \mathrm{s}$ worth of decisions are remembered by the VAD.

Step 5: In this step the location of the active speaker is determined based on expectation values of the particle filters and the voice activity of the output audio stream from the beamformers. The active speaker is identified by the VAD with the highest number of voice frames during the last $0.6\ \mathrm{s}$ (If this number is zero, no speaker is detected).

III. First Results

To evaluate the proposed method we have compared its accuracy (in terms of locating the current speaker correctly) and processing time (as a proxy for actual processing intensity) to a simple beam sweep method. The beamformers used in the beam sweep method are focused on points of a rectangular grid with a fixed distance ($z = 3\ \mathrm{m}$) from the array plane ($z = 0$). The particle filter trackers used in the proposed algorithm are similarly confined to the same plane. Preliminary results for a scenario of $170\ \mathrm{s}$, with two sitting persons taking turns to speak, indicate that the proposed algorithm predicts a speaker position within $\pm 50\ \mathrm{cm}$ of the correct one about $60\ \%$ of the time. Furthermore, steps 1 and 2 of the proposed algorithm is equivalent to about 13 beams in terms of computational intensity (for an array of 256 elements), thus for two speakers the proposed algorithm is equivalent to about 15 beams. Beam sweeping with 15 beams (3 vertical, 5 horizontal), using the same accuracy measure, is only correct about $10\ \%$ of the time, for the same computational effort (though a more fair comparison would use fewer array elements and more beams).

Acknowledgment

The author would like to thank Professor Sverre Holm (main supervisor) and Ines Hafizovic (co-supervisor); as well as Andreas Holm, Johannes Daleng, Jon Petter Åsen, Frode Olsen and Bjørn Braathen at Squarehead Technology for both helpful advice and contributions to computer code.

References

[1] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[2] N. Strobel, S. Spors, and R. Rabenstein, "Joint audio-video object localization and tracking," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 22–31, Jan. 2001.

[3] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 1, pp. 1154–1164, Jan. 2002. [Online]. Available: http://dx.doi.org/10.1155/S1110865702206058

[4] C. Voges, P. Bauer, and T. Fingscheidt, "A particle filtering algorithm for audiovisual speaker localisation," in *4th Workshop on Positioning, Navigation and Communication, 2007. WPNC '07*, Mar. 2007, pp. 103–108.

[5] R. Tucker, "Voice activity detection using a periodicity measure," *Communications, Speech and Vision, IEE Proceedings I*, vol. 139, no. 4, pp. 377–380, Aug. 1992.

[6] J. A. Haigh and J. Mason, "Robust voice activity detection using cepstral features," in *1993 IEEE Region 10 Conference on TENCON '93. Proceedings. Computer, Communication, Control and Power Engineering*, Oct. 1993, pp. 321–324 vol.3.

[7] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 412–424, Mar. 2006.

[8] J. Wang, Y. Liang, and J. Wilder, "Visual-information-assisted microphone array processing in a high-noise environment," in *Proc. SPIE 3521, Machine Vision Systems for Inspection and Metrology VII, 198*, vol. 3521, Oct. 1998, pp. 198–203. [Online]. Available: http://dx.doi.org/10.1117/12.326960

[9] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato, "A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization," in *Proceedings of the 10th International Conference on Multimodal Interfaces*, ser. ICMI '08. New York, NY, USA: ACM, 2008, pp. 257–264. [Online]. Available: http://doi.acm.org/10.1145/1452392.1452446

[10] I. Hafizovic, C.-I. C. Nilsen, M. Kjølerbakken, and V. Jahr, "Design and implementation of a MEMS microphone array system for real-time speech acquisition," *Applied Acoustics*, vol. 73, no. 2, pp. 132–143, Feb. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003682X11002192

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001*, vol. 1, 2001, pp. I–511–I–518 vol.1.

[12] N. Gordon, D. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-gaussian bayesian state estimation," *Radar and Signal Processing, IEE Proceedings F*, vol. 140, no. 2, pp. 107–113, Apr. 1993.

[13] F. Bourgeois and J.-C. Lassalle, "An extension of the munkres algorithm for the assignment problem to rectangular matrices," *Commun. ACM*, vol. 14, no. 12, pp. 802–804, Dec. 1971. [Online]. Available: http://doi.acm.org/10.1145/362919.362945

Trond Flisnes Bergh is 31 years old, and graduated from NTNU in 2008 with an MSc in Engineering Physics. From 2008 to 2014 he worked as a researcher in DNV GL. Since March 2014 he has been pursuing a PhD degree in Digital Signal Processing under the supervision of Professor Sverre Holm (University of Oslo, main supervisor) and Ines Hafizovic (Squarehead Technology, co-supervisor). He is expected to finish his PhD in spring 2017.