Audio-Video detection of the active speaker inmeetings

Francisco Madrigal, Frédéric Lerasle, Lionel Pibre, Isabelle Ferrané
LAAS-CNRS, Université de Toulouse UPS, IRIT, CNRS, Toulouse, France

Abstract:

Meetings are a common activity that provides challenges when creating systems that assist them. Such is the case of the Active speaker detection, which can provide useful information for human interaction modeling, or human-robot interaction. This is mostly done using speech, however, certain visual and contextual information can provide additional insights. In this paper we propose an active speaker detection framework that integrates audiovisual features with social information, from the meeting context.

Fig. 1: Example of active speaker detection

Proposal

We propose to reinforce audio estimate by including (1) features directly extracted from videoclips and (2) contextual information from the social interactionof participants in a meeting.
Visual cue is processed using a Convolutional Neural Network (CNN) that captures the spatio-temporal relationships with both cues: raw pixels (RGB images) and motion (estimated with optical flow).

Contextual reasoning is done with an original methodology, based on the gaze of all participants.

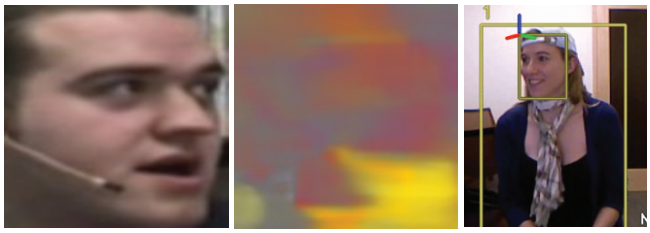

Fig. 2: Example of input images used for training. Left: RGB image. Middle: Magnitude of optical flow image. Right: Example of HyperFace. RGB lines define the orientation of the head.

Multi-modal speaker detection

Audio feature. We take as a baseline the audio recognition framework of Xie et al. VGG-Speaker-Recognition framework. The spectrograms are computed from audio clipsof 2.5 sec, with 256 frequency components. The spectrogramis normalized by subtracting the mean and dividing by the standard deviation.

Visual features. We evaluate different networks that extract spatio-temporal information from video clips: one 2D and three 3D CNN.

***ResNet2D**: Basic 2D CNN, ResNet50, that we use as a baseline.

***C3D:** 3D CNN performs the convolution and pooling spatio-temporally. It has 5 convolution layers, each followed by a pooling layer, 2fully connected layers, and at the end a softmax loss layer.

***ResNet3D**:This architecture takes up the idea of ResNet residual blocks but using the 3D convolution instead of 2D.

Social information. Computed from the head orientation, which represents the gaze of the person, with HyperFace. Since we knowa priori the camera location, the configuration can be projectedto a topological space.

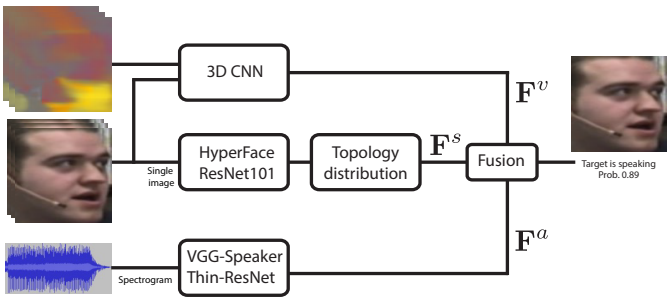

Fig. 3: Pipeline of our multi-modal speaker detection. Inputs are on the left, in the center the independent detection modules and in the end the fusion of all estimates F.

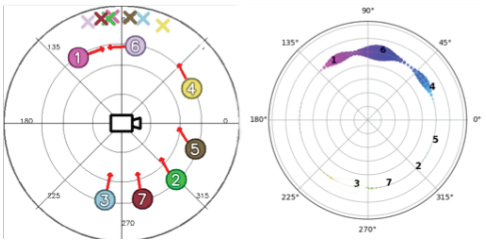

Fig. 4: Left: Example of the topological space of a meeting. Each circle represents a participant and its angular position is the position of the participant from the point of view of the camera, i.e. the left side is 0deg and the right end 360deg. Right: Speaker probability distribution.

Results

Dataset: We evaluate our speaker estimation frameworkusing the AMI Corpus (Augmented Multi-party Interaction). This dataset consists of over 100 hours of meetingsequences with 4 participants each.
First, we use this dataset to train and evaluate the performance of the visual-based networks following across-validation methodology. Meeting sequences are randomly divided into 5 groups, 4 are used to train the visual-based networks, and one group for testing. We call each groups CV.

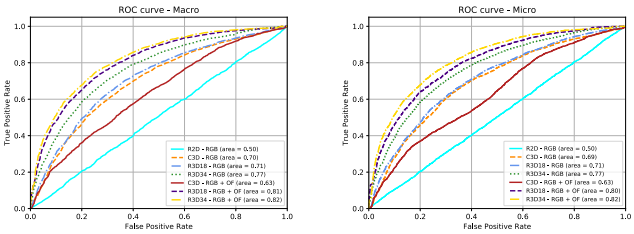

Fig. 5: Macro and micro ROC curves of the fold CV2.

	C3D		ResNet3D-18		ResNet3D-34	
Fold	RGB	RGB-OF	RGB	RGB-OF	RGB	RGB-OF
CV1	0.5	0.55	0.7	0.75	0.7	0.78
CV2	0.7	0.63	0.71	0.81	0.77	0.82
CV3	0.65	0.5	0.79	0.82	0.79	0.85
CV4	0.5	0.77	0.76	0.85	0.78	0.84
CV5	0.67	0.5	0.68	0.76	0.68	0.76
Mean	0.60	0.59	0.73	0.80	0.74	0.81

Table I. RESULTS OF MACRO AREA UNDER CURVE (AUC) FOR ALL FOLDS OF AMI CORPS USING THE 3D CNNS.

	C3D		ResNet3D-18		ResNet3D-34	
Fold	RGB	RGB-OF	RGB	RGB-OF	RGB	RGB-OF
CV1	0.48	0.55	0.69	0.75	0.7	0.78
CV2	0.69	0.63	0.71	0.8	0.77	0.82
CV3	0.64	0.5	0.76	0.82	0.79	0.85
CV4	0.5	0.73	0.76	0.84	0.77	0.84
CV5	0.64	0.5	0.68	0.75	0.68	0.76
Mean	0.59	0.59	0.72	0.79	0.74	0.81

Table I. RESULTS OF MICRO AREA UNDER CURVE (AUC) FOR ALL FOLDS OF AMI CORPS USING THE 3D CNNS.

In C3D, the use of optical flow only improves in some groups. The results with ResNet3D always improve with OF, reaching in certain cases up to 10%. These results have been obtained using the test samples.
We evaluate the sequences in real-time, i.e., frame by frame. Since the sequences have a long duration, we take a sub-part of each video. More precisely, we begin the analysis 15 minutes starting from minute 3 of each sequence.

	Audio Feature	Social Featur	Visual Feature ResNet3D-34	Joint prob.
Macro	0.79	0.66	0.7	0.84
Micro	0.78	0.68	0.67	0.84

Table I. EVALUATION OF THE FRAMEWORK USING AMI CORPUS SEQUENCES DIRECTLY.

References

[1] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterancelevel aggregation for speaker recognition in the wild," in International Conference on Acoustics, Speech, and Signal Processing, 2019.
[2] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multitask learning framework for face detection, landmark localization, pose estimation, and gender recognition," IEEE Trans. on PAMI, pp. 121–135, Jan 2019.
[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in IEEE Int. Conf. on Computer Vision (ICCV), December 2015.
[4] G. Liu, Y. Yu, K. A. Funes Mora, and J. Odobez, "A differential approach for gaze estimation," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2019.