# Audio-Video detection of the active speaker in meetings

Jorge Francisco Madrigal Diaz, Frédéric Lerasle, Lionel Pibre, Isabelle Ferrané

# Audio-Video detection of the active speaker in meetings

Francisco Madrigal
LAAS-CNRS
Toulouse, France
Email: jfmadrig@laas.fr

Frédéric Lerasle
LAAS-CNRS
Université Paul Sabatier
Toulouse, France
Email: lerasle@laas.fr

Lionel Pibre
IRIT, Université de Toulouse,
Toulouse, France
Email: pibre@laas.fr

Isabelle Ferrané
IRIT, Université de Toulouse
Université Paul Sabatier
Toulouse, France
Email: ferrane@laas.fr

***Abstract*—Meetings are a common activity that provide certain challenges when creating systems that assist them. Such is the case of the Speaker recognition, which can provide useful information for human interaction modeling, or human-robot interaction. Speaker recognition is mostly done using speech, however, certain visual and contextual information can provide additional insights. In this paper we propose a speaker detection framework that integrates audiovisual features with social information, from the meeting context. Visual cue is processed using a Convolutional Neural Network (CNN) that captures the spatio-temporal relationships. We analyse several CNN architectures 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. We evaluate our proposal with a public benchmarks in state-of-art: AMI corpus. We show how the addition of visual and context information improves the performance of the speaker recognition.**

KEYWORDS

Speaker recognition, Convolutional Networks, Audiovisual modeling, Feature fusion.

## I. INTRODUCTION

Meetings are essential in our society, from universities and industries, where they are commonly held to coordinate professional matters such as projects, researches, funds, etc., to informal meetings where people discuss everyday issues. In all cases, the person of interest is the one who is speaking because she/he is (usually) the center of attention of all participants. Moreover, communication takes place not only through voice, but also with certain behaviors or gestures that we can observe. Interpreting this audiovisual information in order to estimate the person of interest or active speaker can be useful in several scenarios, such as human-robot interaction [1], multi-media, among others, so a system capable of interpreting both cues is desirable.

In the literature there are different proposals to carry this out in a general way without considering the context. Most of them focus on analyzing the audio only, which truly is a source that provides rich information. For example, multiple works explore ways to recognize the speaker, others propose diarization techniques that allow partitioning the incoming audio into homogeneous streams, i.e. audio sections with a single speaker. Likewise, there are approaches that interpret the audio turning it into text. Although, the audio detection cannot be associated with a visual detection of the person without considering some assumptions and different people talking at the same time can lead to misdetection or recognition errors.


Fig. 1. Example of active speaker recognition using the Talking face video made available from the Face and Gesture Recognition Working group [2]

On the other hand, visual-only approaches for speaker detection has not been widely explored in recent years due to they are confused by occlusions or face movements such as facial expressions or yawning. Nevertheless visual information has been exploited to reinforce audio-based estimation, see Fig. 1. Here, the movement of the lips is encoded by spatio-temporal features, commonly estimated through Convolutional Neural Networks (CNN). Nonetheless, there is still relevant information that can be extracted in this application context.

In this work we focus on identifying speakers at meetings based on video and audio cues. For this goal, our proposal focuses on three aspects: (i) analyze the meeting sequences with a state of the art audio-based recognition method and our own handcrafted video-based speaker detector using spatio-temporal features, (ii) extracting contextual information in an original way from the video meetings, such as the fact that most of the participants usually gaze towards the active speaker; therefore the correlation of head orientations inferred by vision would reinforce the speaker detection, (iii) merge visual and audio estimations such that we obtain a robust final detection.

Regarding (ii), our approach is based on the prior image detection of the face using techniques in the state of the

art. The principle is analyzing frames to characterize the orientation of the face in a continues space in terms of pitch, yaw and roll. Therefore, the speaker / non-speaker status can also be inferred at once with pure visual percept.

However, in literature there are few multi-person public audio-video datasets in our application context (meeting), which limits the scope of the evaluations and comparisons that we can perform. Nevertheless, the evaluations on the available datasets with increasing complexity and mono / multi-people, observed within the field of view, are encouraging for the different analysis and experimentation that we have carried out.

The main contributions of this paper are:

1) A pure visual speaker classifier, based on 3D CNNs, applied in this original context. It takes as input video frames (i.e. clips) with several features: optical flow and RGB images.
2) A new model that exploits the classic topology of a meeting room and the social behavior of the participants.
3) A fusion methodology which combines these 2 visual percepts with a state of the art audio recognition.
4) A robust evaluation in datasets well adapted to our problem of speaker detection in meetings. The results show considerable improvement by merging both video and audio percepts because the fusion allows to disambiguate certain situations.

This paper has the following structure: Section 2 presents the related work and Section 3 provides the formulation of our proposal for speaker detection. Quantitative and qualitative results including a discussion are given in Section 4. Last, Section 5 describes conclusions and future work.

## II. Related work

Communication is the key for two entities to interact. When a human is involved it is, commonly, performed by the voice and gestures, which the computers process them as audio and video respectfully. Our work focuses on exploiting these signals to estimate the person (s) who speak in an instant of time, i.e. active speaker. From both signals, the audio has been widely explorer because it provides natural information.

**Video** Pure visual information is an important source to consider in speaker recognition, especially if the audio is not available, is corrupted or unintelligible. Zhou et al. [2] give an in-depth review of advances in visual speaker recognition until 2014. The authors provide a list of datasets aimed for this purpose. Also, the presented methods are grouped according to the type of feature used, therefore we can assume 4 groups: (1) Image-based. Here raw pixels are transformed directly as visual features with the aid of methods such as Principal Component Analysis (PCA) [3], [4]. (2) Motion-based. Features describe the observed movement during speaking. Instead of creating handcrafted features, several proposals estimate the motion directly from the video with techniques such as Optical Flow (OF) [4]. (3) Geometric-based. The feature considers the geometric information of a moving mouth. In [5], the movement is computed by measuring the distance between points detected over the mouth. Then, the difference of distances between successive images represents the motion. The main limitation of these methods is that any movement of the mouth is considered as speaking. So, the number of false positives is commonly high.

**Audio** In literature there is a large body of work on Speaker recognition [6]. One way to achieve it is with diarization techniques where the audio stream is partitioned in segments according to the identity of the speaker. In [7], Bonastre et al. propose a diarization method based on the Binary key (BK) modelling, which transform the audio into a feature representing the speaker within the binary space. Then, the diarization is performed by an iterative agglomerative clustering algorithm which forms segments of the same speaker. Patino et al. [8] improves the method by considering spectral clusterization. One of the main challenges is to be able to recognize the same person regardless the intensity of speaker voice, e.g. whispering, or that background noise alters the identification. Vestman et al. [6] do a deep taxonomy on different features that address these issues and propose a sound time-varying feature which gives state of the art results. With the rise of the Convolutional Neural Network (CNN), some proposals [9] have exploit this end-to-end solution for speaker diarization. The most recurrent architectures are those based on Long Short Term Memory (LSTM) Networks ([10], [11]) since they capture the variations in the voice of the announcer. One of the limitations of CNNs is they are computationally expensive, generally requiring powerful GPUs to produce good results. In [12], the authors propose an end-to-end utterance-level diarization framework. This method proposes a new 'thinResNet' trunk architecture, which incorporates a GhostVLAD layer allowing to aggregate features across time. It is trained and evaluated using the VoxCeleb dataset [13], an audio-visual dataset of short clips, extracted from YouTube, of interviews. In this dataset, with over of 2000 hours of recording and more than 7000 persons, [12] has demonstrated the effectiveness of this compact network by providing state of the art performance.

However, murmurs, background noises or interspersed audio cause a bad estimate. To overcome these difficulties, several approaches include visual features since those are not affected by these phenomena.

**Audio-Video** Audiovisual methods are robust to background noise and different speak modes (whisper). Recently deep learning has provided advances in audiovisual speaker estimation by capturing the temporal relationships of visual and acoustic cues. The use of recurrent neural network (RNN) is explored in [14] to extract video features using 2D CNN. Then, in a similar way as [5], they train a Long Short-Term Memory layer for each feature. The output of both layers is concatenated and the result used to train a final LSTM layer. This process is known as early fusion, on the contrary it is called the fusion when the results are merged at the end without any other training layer pursuing them. [15] compares between both fusion methods in the context of speaker recognition. Afouras et al. [16] explore the impact

of training an audio-visual lip reading network with different loss function.

In the literature, the application of LSTM layers has been widely studied because it captures well the spatio-temporal information of a speaker. In addition, there are other CNN architectures that learn this aspect but used in different classification contexts. Such is the case of the deep 3-dimensional convolutional networks (C3D) [17], here the objective is to classify actions such as biking, running, among others. This 3D CNN model has been extended to other architectures, in [18] they study the use of 3D residual neural network (ResNet3D). In both cases, the goal is to obtain a 3D feature that encodes appearance and motion simultaneously. We aim to study these networks in our context of speaker detection because the proposals show good performance to state-of-the-art methods. The classic CNN-based image classificators learn from the raw image, some works have shown that the use of additional features like optical flow [4], [19], [18] or depth cue [19] to improve the estimation. In this work we explore the use of optical flow as a second visual cue.

The biggest drawback of video-based applications (whether pure or in conjunction with audio) is that their evaluations are conducted in constrained situations where there are usually only one (or few) people looking in the straight at the camera. This scenario is not realistic in terms of a classical meeting. Therefore, we propose using a classifier, which considers spatio-temporal features, trained from realistic video sequences in a meeting situation. Additionally, the video can provide more information if we consider the social aspect among the participants of a meeting.

If we assume the gaze of almost all participants observes the main speaker, then we can infer the latter if we can estimate the gaze of everyone. One way to achieve it is using algorithms such as [20]. It predicts the gaze direction by training a differential convolutional neural network that take into account the gaze differences between two images of the eye. However, robust and accurate results require good image resolution of the eye which cannot be guaranteed in situations such as meetings since the camera(s) observes all people together at a given distance, reducing the image details. Another way is assuming that the orientation of the head and the gaze share the same direction. In the literature, HyperFace [21] is a state of the art framework based on ResNet101. Its architecture allows it to perform, at the same time, Face Detection, Head Pose Estimation, Landmarks Localization and Gender Recognition. Fig. 2 shows an example of the head pose estimated by HyperFace, head pose is depicted by the RGB lines.

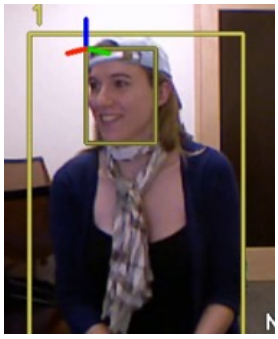

Fig. 2. Example of HyperFace output using the ICT-3DHP dataset [22]. RGB lines define the orientation of the head in terms of roll, yaw and pitch respectively.

## III. Multi-modal speaker detection

Since audio is generally a very discriminating cue in the speaker recognition problem, we propose to reinforce this estimate by including (1) features directly extracted from video clips and (2) contextual information from the social interaction of participants in a meeting. Fig. 3 shows the pipeline of our proposal. We take as baseline the audio recognition framework of Xie et al. [12] because it has shown a good performance in the challenging dataset VoxCeleb. In the case of visual features, we evaluate different 3D networks that extract spatio-temporal information from video clips. Social information is computed from the head orientation, which represents the gaze of the person. This orientation is estimated using HyperFace [21] which is a framework with results in the state of the art and has more than 500 cites.

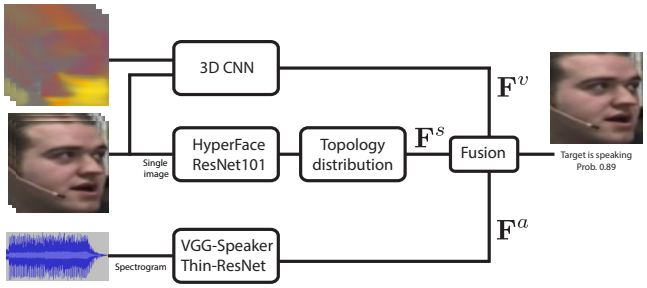


Fig. 3. Pipeline of our multi-modal speaker detection. Inputs are on the left, in the center the independent detection modules and in the end the fusion of all estimates $\mathbf{F}^*$.

### A. Audio features

For the audio representation we use author's implementation of [12], named VGG-Speaker-Recognition framework [3]. The spectrograms are computed from audio clips of 2.5 sec, with 256 frequency components. The spectrogram is normalized by subtracting the mean and dividing by the standard deviation. The thin-ResNet network was trained with Adam optimizer and an initial learning rate of $1 - e3$. In our case, the evaluation is done with an audio at a 40kHz sample rate. This network outputs an audio-based feature $f^a$ that is compared against a bank of features ($f_i^a$), previously calculated, of each participant $i$. Thus, we obtain a vector $\mathbf{F}^a$ that indicates the probability that the recorded audio belongs to a person.

[3]https://github.com/WeidiXie/VGG-Speaker-Recognition

Fig. 4. Example of input images used for training the CNN. Left: RGB image. Right: magnitude of optical flow image.

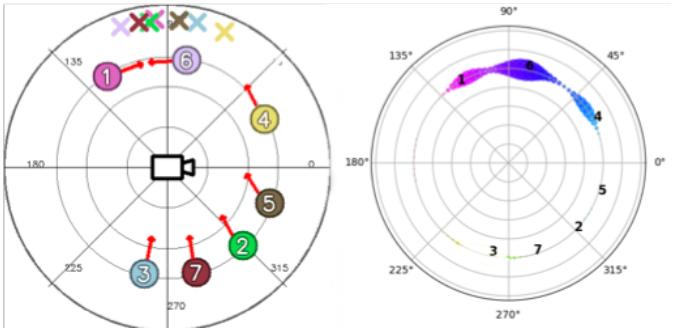

Fig. 5. Left: Example of the topological space of a meeting. Each circle represents a participant and its angular position is the position of the participant from the point of view of the camera, i.e. the left side is $0°$ and the right end $360°$. Right: Speaker probability distribution.

### B. Visual features

**Settings** We consider 4 network architectures: one 2D and three 3D CNN. Since we want to capture the temporal relationship that exists between the images, we group several consecutive images to form a non-overlapped video clip of 16 frames. The input clip has a size $3 \times L \times H \times W$, where H and W are the height and width of the frame in the 3 RGB channels and $L = 16$ is the size of the clip, i.e., the number of frame. Thus, the clips are input to the networks.

**ResNet2D** The first network tested is a basic 2D CNN, more precisely a ResNet50, that we use as baseline. In this kind of architecture, the spatio-temporal information can be handle by treating the L frames as channels of the same image. Therefore, the input dimension of ResNet2D is $3L \times H \times W$, forming a 3-dimensional tensor. In this case, image size is set to $H = 224$ and $W = 224$. In this network, the convolution is performed over the spatial dimension and the first layer collapses the temporal information into 2D features maps. Thus, any subsequent layers do not consider the temporal meaning.

**C3D** Unlike 2D networks, 3D CNN performs the convolution and pooling spatio-temporally. This network [17] has 5 convolution layers, each followed by a pooling layer, 2 fully connected layers and at the end a softmax loss layer. Th input size is $3 \times 16 \times 112 \times 112$. In comparison with ResNet2D, here we have a 4 dimensional tensor. We can observe the architecture follows a classic CNN with the different the convolution is performed 3D, allowing the feature maps to learn temporal context. Thus, the convolution layers create a 3D feature where the initial part focuses on the appearance of the first images and the rest considers the salient motion [17].

**ResNet3D** This architecture takes up the idea of ResNet residual blocks but using the 3D convolution instead of 2D. This preserve and propagate the temporal reasoning through network's layers. Input is a 4D tensor as C3D but image size is the same as ResNet2D: $3 \times 16 \times 224 \times 224$. We evaluate two versions of ResNet3D, one considering 18 blocks and another with 34.

All networks are trained with RGB images that show only the face. Inspired by other works [4], [19], [18], we use optical flow as an additional cue since it allows to encode the movement of the face. Fig. 4 shows an example of the input images. In this case, we follow a Bi-CNN architecture, where we have two networks each trained by one cue, i.e., RGB images or optical flow image. We remove the top or each network and concatenate both visual features. Finally, we train 2 fully connected layers and a softmax loss layer.

The output of these networks indicates the probability that the participant $i$ is speaking or not. If the input is a clip vector, then the network gives a visual-based vector $\mathbf{F}^v$ which represent the probability that each participant is speaking.

### C. Social feature

In Fig. 6 we can see a basic scenario of the type of meetings from we want to detect if someone(s) is speaking. Here, 4 cameras are set in the center of the table observing each one a participant. We obtain a $360°$ view of the room by concatenating the 4 camera images, see Fig. 7. Since we know the camera location, the configuration can be projected to a topological space similar to the one shown in Fig. 5-left. In this example, 7 participants are represented as circles and their position corresponds to the relative position of the participant with respect to the camera in the $360°$ view, i.e., the left border depicts $0°$ and the right border is $360°$. The arrows depict the head orientation of each target and the crosses are the projection of the estimated orientation. Then, we use these projections to compute a probability distribution representing the speaker detection based on the gaze of all the participants.

To estimate the orientation, we use the tool HyperFace [21], based on ResNet-101 and trained with the AFLW dataset [23]. HyperFace computes the head pose of a target $i$ as $\theta_i^{HF} = \{\theta_i^x, \theta_i^y, \theta_i^z\}$, where each angle is in the range $\theta_i^* = [-1, 1]$. Lets assume $c_i$ as the central position, on the image, of the $i-th$ person considering only the horizontal axis. This can be projected to the topological space by :

$$\mathbf{C}_i = 2\pi * \frac{c_i}{I_w},$$

where $I_w$ is the width of the $360°$ image. Then, we turn it around according to the orientation of the estimated pose but considering only $\theta_i^z$. This is because it represents the rotations left-right of the head. Therefore, the projected gaze is defined as follows:


Fig. 6. Example of a classical meeting from the AMI corpus [24]. There are 4 cameras where each one records a single participant.

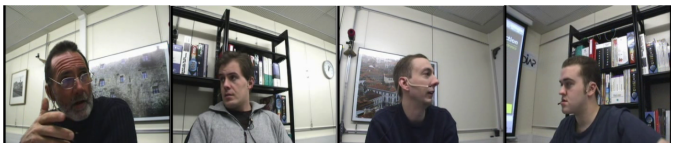

Fig. 7. Example of a 360° image generated by concatenating the single view of each camera.

$$\mathbf{G}_i = \mathbf{C}_i + \pi * (1 - 2 * \theta_i^z).$$

We can observe that the rotation is done proportionally to the head pose. If $\theta_i^z = 0$ then the target is regarding straight to the camera. This means the gaze is at the opposite position of $\mathbf{C}_i$. Even if $\theta_i^z = [-1, 1]$, the real value will never overpass $[-0.5, 0.5]$ due to physical constrains of the human head movement. Then, we double the value of $\theta_i^z$ allowing to cover all possible orientations. We transform the gaze point $\mathbf{G}_i$ into probability using the von Mises distribution :

$$f_i(x|\mathbf{G}_i, \kappa) = \frac{\exp^{\kappa \cos(x - \mathbf{G})}}{2 * \pi \mathbf{I}_0(\kappa)},$$

where $\mathbf{I}_0(\kappa)$ is the Bessel function of order 0 and $\kappa = 10$ measures the concentration. The final distribution is then calculated as the average of each participant distribution. After normalization, we obtain a probability similar as shown in Fig. 5-Right. We can observe the distribution is concentrated mostly at targets 1, 4 and 6 being the later the most probable speaker. By extracting the exact probability of each participant, we obtain the vector $\mathbf{F}^s$.

### *D. Feature fusion*

The previous steps have calculated vectors that represent the probability that a participant is a speaker by analyzing different cues. The simplest fusion strategy would be to average all vectors. However, the failure of one of them can lead to a degradation of the other two estimates. Therefore, we merge the probabilities based on a simple heuristic as follows:

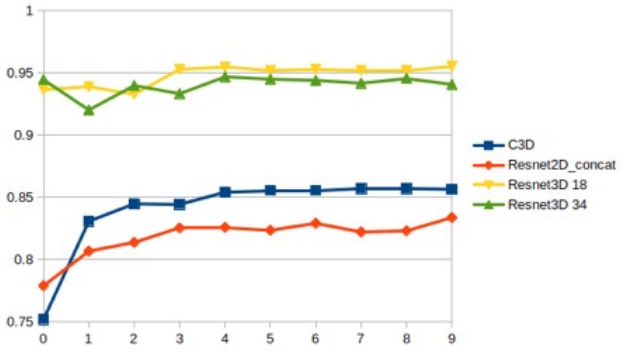


Fig. 8. Speaker detection training: accuracy vs. number of epochs.

$$\mathbf{F}_i = \begin{cases} (\mathbf{F}_i^a + \mathbf{F}_i^v + \mathbf{F}_i^s)/3, & \text{if } \forall \mathbf{F}_i^* > 0.5 \text{ or} \\ & \forall \mathbf{F}_i^* < 0.5 \\ (\mathbf{F}_i^\alpha + \mathbf{F}_i^\beta)/2, & \text{if } \forall \mathbf{F}_i^\alpha > 0.5 \text{ and } \mathbf{F}_i^\beta > 0.5| \\ & \alpha = \{a, v, s\}, \beta = \{a, v, s\} \\ & \text{and } \alpha \neq \beta \end{cases}$$

## IV. Evaluations

**Dataset** We evaluate our speaker estimation framework using the AMI Corpus (Augmented Multi-party Interaction) [24]. This dataset consists of over 100 hours of meeting sequences. In order to limit the evaluation, we select the IDAIP scenario that has 38 recorded meetings with 4 participants each, see Fig. 6. There are 4 cameras and each one records a single target. Fig. 7 shows the 4 cameras views. AMI corpus has a detailed ground-truth (GT) at the audio level but has no information regarding the video.

First, we use this dataset to train and evaluate the performance of the visual-based networks following a cross-validation methodology. Thus, meeting sequences are divided in 5 groups, 4 are used to train the visual-based networks and one group for testing. We call these groups *CV*, we add a number at the end to denote the group that is used for testing. Then CV1 uses group 1 for testing and the rest for training, the same applies for CV2 through CV5.

The CNNs are trained considering only the target face using RGB and Optical Flow (OF). The face is detected with CAFFE's ResNetSSD FaceDetector and OF is computed with Farneback's algorithm, both implemented in OpenCV. The samples are grouped in clips of 16 frames and labeled as Speaker or non-Speaker according to the audio-based GT. The images are re-scaled to 224x224 pixels which is the size required for ResNet-based network. For C3D, images are set to 11x112 pixels. Since all sequences last several hours, we limit the sample to 100 clips per person per class. Thus, each class has of over 15000 clip samples.

**Training** All visual-based CNNs are trained with batches of 20 clips, Adam optimizer and an initial learning rate of 0.003, as proposed in [17], and it is divided by 10 after every 4 epochs. The training is stopped after 10 epochs.

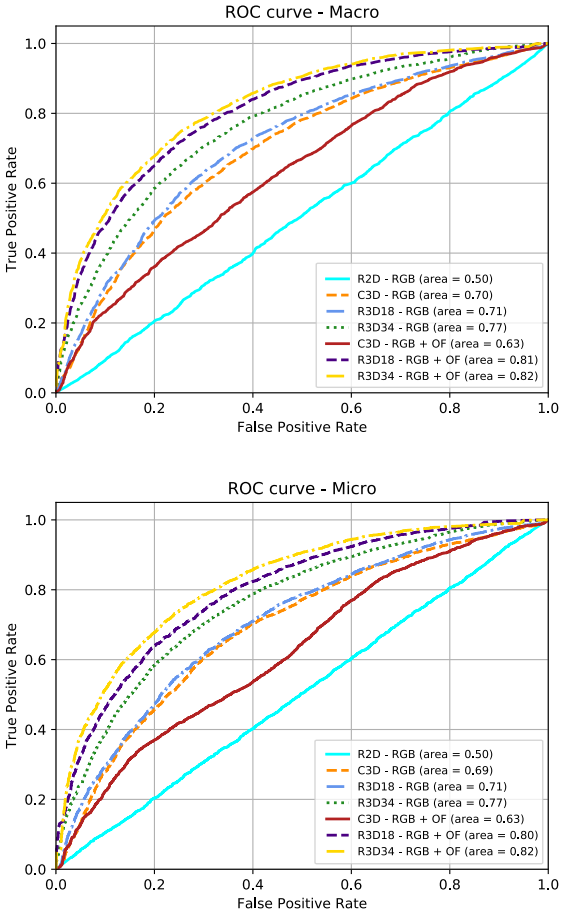


Fig. 9. Macro and micro ROC curves of the fold CV2.

**Results** First, we evaluate the performance of the CNN architectures. In Fig. 8 we can see the results of the training. it compares the number of epochs (horizontal axis) against precision (vertical axis). This first evaluation is done with cross-validation using CV2, where all groups but number two is used for training. This group represents well all sequences. We can see that the ResNet 2D network, which has collapsed the clip as additional channels of an image, brings the smallest performance. On the contrary, its 3D versions have a better performance reaching around the $95\%$.

In Figure 9 we compare the results of all networks using the test part of the CV2 group. Here we use the ROC curves as an evaluation metric, which measures the power of classification of the network. The higher the true positive rate, the better the classification. The evaluations confirm a significant improvement between the base network C3D and deeper network ResNet3D, in both versions (18 and 34 blocks). Although C3D is not benefited by including optical flow information, we can see how ResNet3D increases its performance by 5 percent. The results of macro and micro versions are similar since the number of samples per class is well balanced.

Tables I and II show the results considering all the groups, using the area under the ROC curve as a metric. This metric summarizes the measure of the ROC curve characterizing the performance the binary classification. In the case of 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. To evaluate the performance of the entire system, 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 from minute 3 of each sequence. Then, we analyze 15 minutes, stopping at minute 18, or at the end of the video in certain cases.

We show the results in Table III, each column indicates the feature used for the classification, i.e., column 2 shows the results using only the audio to estimate the speaker.

It is clearly observed that of the three features, audio is the most discriminating to recognize the speaker. Both visual features give similar results. The last column shows the results obtained by merging all the features. We can see that this fusion improves the estimate, this is because the video allows to provide supplementary information that pure audio cannot capture.

We have prepared a video that shows the qualitative results of our framework in one AMI corpus sequence. The video is available at the following link:

https://drive.google.com/file/d/11B-1f6EhvH3lPVV080KoY_ITdW483emZ/view?usp=sharing

## V. Conclusion

This paper has presented a framework for estimating speakers in a meeting context using 3 features. The first feature is audio, which naturally captures whether a person is speaking. In this paper we show that audio-based results can be improved by including additional information.

Such is the case of visual-based information, we have presented an analysis of how 3D CNNs, which have been used in other classification contexts, can encode the space-time aspect of a speaker. In parallel, we have studied how the use of optical flow incorporates the movement knowledge that allows to improve the estimation in most cases.

Finally, an original methodology was presented to model social aspects that aids to estimate the speaker in the context of the meeting. The analysis based on participant gaze, whose orientation we assume is the same as the orientation of the head. Additional research on social behavior could provide new information that could help to improve the results.

## Acknowledgment

This work was carried at LAAS-CNRS and supported by the project LinTO, funded by Bpifrance as part of the French project Program d'Investissements d'Avenir 3.

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 | 0.5 | 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 II
RESULTS OF MICRO 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 III
EVALUATION OF THE FRAMEWORK USING AMI CORPUS SEQUENCES DIRECTLY.

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

## REFERENCES


[1] W. He, P. Motlicek, and J. Odobez, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE Int. Conf. on Robotics and Automation (ICRA)*, May 2018, pp. 74–79.

[2] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A review of recent advances in visual speech decoding," *Image and Vision Computing*, vol. 32, no. 9, pp. 590 – 605, 2014.

[3] X. Hong, H. Yao, Y. Wan, and R. Chen, "A pca based visual dct feature extraction method for lip-reading," in *2006 Int. Conf. on Intelligent Information Hiding and Multimedia*, Dec 2006, pp. 321–326.

[4] N. Le and J.-M. Odobez, "Learning multimodal temporal representation for dubbing detection in broadcast media," in *Proceedings of the 24th ACM International Conference on Multimedia*, ser. MM '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 202–206.

[5] D. C. M. G. M. M. B. B. A. L. P. Korshunov, M. Halstead and S. Marcel, "Tampered speaker inconsistency detection with phonetically aware audio-visual features," in *Proceedings of the International Conference on Machine Learning*, 2019.

[6] V. Vestman, D. Gowda, M. Sahidullah, P. Alku, and T. Kinnunen, "Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction," *Speech Communication*, vol. 99, pp. 62 – 79, 2018.

[7] J. Bonastre, X. Anguera, G. Sierra, and P. Bousquet, "Speaker modeling using local binary decisions," in *Interspeech*, 2011.

[8] J. Patino, H. Delgado, and N. Evans, "The EURECOM submission to the first DIHARD challenge," in *Conf. of the International Speech Communication Association, September 2-6, 2018, Hyderabad, India*, Hyderabad, INDIA, 09 2018.

[9] M. Hrúz and Z. Zajíc, "Convolutional neural network for speaker change detection in telephone speaker diarization system," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 4945–4949.

[10] A. Sarkar, S. Dasgupta, S. K. Naskar, and S. Bandyopadhyay, "Says who? deep learning models for joint speech recognition, segmentation and diarization," in *2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5229–5233.

[11] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with lstm," in *2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5239–5243.

[12] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *International Conference on Acoustics, Speech, and Signal Processing*, 2019.

[13] W. X. A. Z. A. Nagrani, J. S. Chung, "Voxceleb: Large-scale speaker verification in the wild." *Computer Speech  Language*, 2019.

[14] F. Tao and C. Busso, "End-to-end audiovisual speech activity detection with bimodal recurrent neural models," *Speech Communication*, vol. 113, 2019.

[15] D. C. M. P. S. Petridis, J. Shen, "Visual-only recognition of normal, whispered and silent speech," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[16] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.

[17] 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.

[18] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Computer Vision and Pattern Recognition*, 2018.

[19] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "Poseidon: Face-from-depth for driver pose estimation," in *Procs. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[20] 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.

[21] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, Jan 2019.

[22] T. Baltrušaitis, P. Robinson, and L. P. Morency, "3d constrained local model for rigid and non-rigid facial tracking," in *Conf. on Computer Vision and Pattern Recognition*, June 2012, pp. 2610–2617.

[23] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Nov 2011, pp. 2144–2151.

[24] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 2, pp. 181–190, 2007.