A Multimodal Approach to Speaker Diarization on TV Talk-Shows

Félicien Vallet, Slim Essid, and Jean Carrive

***Abstract*—In this article, we propose solutions to the problem of speaker diarization of TV talk-shows, a problem for which adapted multimodal approaches, relying on other streams of data than only audio, remain largely under exploited. Hence we propose an original system that leverages prior knowledge on the structure of this type of content, especially the visual information relating to the active speakers, for an improved diarization performance. The architecture of this system can be decomposed into two main stages. First a reliable training set is created, in an unsupervised fashion, for each participant of the TV program being processed. This data is assembled by the association of visual and audio descriptors carefully selected in a clustering cascade. Then, Support Vector Machines are used for the classification of the speech data (of a given TV program). The performance of this new architecture is assessed on two French talk-show collections: *Le Grand Échiquier* and *On n'a pas tout dit*. The results show that our new system outperforms state-of-the-art methods, thus evidencing the effectiveness of kernel-based methods, as well as visual cues, in multimodal approaches to speaker diarization of challenging contents such as TV talk-shows.**

***Index Terms*—Fusion, joint audiovisual processing, multimodality, speaker diarization, SVM classification, talk-show, unsupervised learning.**

I. Introduction

SPEAKER diarization is the process of partitioning an input audio stream into homogeneous segments according to the speakers' identities. As such, it is the task of determining "who spoke when?" in an audio or video recording that contains an unknown amount of speech as well as an unknown number of speakers. Several studies, such as [1], [2] or [3] highlight the importance of *speaker diarization* in the field of TV content analysis. Indeed, this capability is critical for the structuring of video content. The information conveyed by speakers can allow the retrieval of elementary structural components that are part of the organizational scheme of a particular program. For instance, one such a component is the speech repartition of the various speakers that further allows role recognition as performed in [4] and [5] on audio only data.

Manuscript received February 08, 2012; revised May 18, 2012; accepted July 30, 2012. Date of publication December 12, 2012; date of current version March 13, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nicu Sebe.

F. Vallet and J. Carrive are with the Research Department of Institut national de l'audiovisuel, 94366 Bry-sur-Marne cedex, France.

S. Essid is with Institut Mines-Telecom; Telecom ParisTech; CNRS/LTCI, 75014 Paris, France.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2012.2233724

On this type of content, the speakers cannot be treated as *registered*, in the sense that it is quite difficult to assume that training databases are available for every possible speaker on a TV set. Hence, the *speaker recognition* methods such as *speaker verification* or *speaker identification* turn out to be inapplicable. The focus is thus put here on unsupervised approaches generally categorized as *speaker diarization* methods. Extensive reviews of the research in speaker diarization are proposed in [6] and [7], although the focus is put on the treatment of meetings and broadcast news data.

In contrast with the field of biometrics research, where multimodal datasets have been available since more than a decade (see [8]–[11]), only lately were large enough audiovisual datasets created, that featured real-life situations and allowed for a proper evaluation of tasks such as audiovisual speaker diarization. For instance, the freely available AMI Meeting corpus [12] proposes 100 hours of meeting with several audio and camera recordings (with various microphones and shot angles). Similarly, Canal9 [13] is a database of more than 43 hours of political debates.

Following the creation of the two former datasets, a number of studies have been led on "real-life" audiovisual data. Thus, in [14] the author proposes to detect similar audiovisual scenes—i.e., same speaker, same geographical disposition, same point of view, etc.—and to gather them in order to create a reduced view of the program. In [15] and [16], the authors highlight the importance of computing more robust descriptors than the lips movement traditionally used in biometrics works. Having shown that a speaker's movements are heavily correlated with the audio stream, they estimate the speakers' head and body motion to identify which one of the visible persons is the active speaker. Similarly, entropy of the labial activity for on-screen persons have also been measured to index talk-show participants (see [2] and [17]).

Nevertheless, it is only with the works of Friedland *et al.* (see [18] and [19]) that the first audiovisual speaker diarization systems were proposed. Both exploiting the AMI Meeting corpus, the first of these studies employs multi-camera recordings while only one camera is used in the second. The speaker diarization algorithm proposed is similar to the one presented in [20] (*bottom-up*) but with two GMM models for each cluster (one for each modality).

Thus, it is only very recently that audiovisual diarization has been proposed for *edited content*. In this case, while many cameras are used during the shooting, in general the images of a single one are shown at a time, corresponding to the current shot chosen by the TV director (except for dissolves, surimpressions, etc.). In our work we focus on a particular type of edited TV content, namely *talk-shows*. This is a type of TV shows that

is characterized by a rich and varied content centered at spontaneous and lively conversations between the show participants [1], which makes it particularly challenging for speaker diarization systems. This is further developed in Section II where we show the limitation of state-of-the art approaches when applied to talk-shows.

On the basis of this observation, we propose a completely novel multimodal speaker diarization architecture, well-adapted to talk-shows (and edited video content in general). This original architecture leverages our prior knowledge on the structure of the type of content considered [1], especially the visual information relating to the active speakers, with kernel-based methods, for an improved diarization performance.

Two systems (of increasing complexity and performance) are actually presented. The first, decribed in Section III, uses the visual modality only to perform a pre-clustering of the *visual* frames in order to assemble speaker training data to be used for learning *audio* classifiers. The second is an improvement of the first that exploits *audiovisual* classifiers and an original fusion scheme, described in Section IV.

Fig. 1. Excerpts from the talk-show datasets *Le Grand Échiquier* and *On n'a pas tout dit*.

Fig. 2. Position of the desktop (red) and lapel (green) microphones on the TV set of one of the show of the dataset *Le Grand Échiquier*..

II. Peculiarities of Speaker Diarization on TV Talk-Shows

As introduced above, we deal in this study with a particular type of content, that is TV talk-shows. In this case, in contrast to databases used in most of the speaker diarization works, the audio stream is not the only source of information available. Moreover, the video content is in this case *edited*, i.e., the shot shown on-screen is chosen by the TV director (typically through a video switcher) who generally tries to follow the speaker. The lively nature of TV data makes the task of speaker diarization quite challenging, especially owing to the potentially noisy sound conditions and the spontaneity of the speech. Indeed, talk-shows are known for being structured around the appearance of natural conversation between the host(s) and his/her/their guest(s), and are thus far less predictable than other TV contents such as broadcast news for instance.

A. Corpus Presentation

The main database used in this work is a collection of French programs: *Le Grand Échiquier* (GE). This talk-show, whose duration is between 2 and 3.5 hours, has been broadcasted live each month between 1972 and 1986 and was presented by Jacques Chancel. The program is centered around a main guest and shows a succession of musical performance, TV excerpts, interviews, etc. According to this guest and his/her personnal interests, great variations can be observed from one show to the other. For instance, an important amount of musical passages are generally proposed when the main invitee is a musician and/or a singer, while more film excerpts are shown for a movie director. Six shows have been annotated in speech turns using the software *Transcriber*.[1] For our experiments, the dataset has been split into a development set (over which system parameter tuning is performed) and a test set. The development set is composed by the first three shows (representing a total time of 8 hours 34 minutes) while the last three account for the test set (representing a total time of 8 hours 5 minutes).

In view of a further validation of our system a second dataset has been considered and used as a complementary test set. This is the corpus *On n'a pas tout dit* that consists of four shows annotated in speaker turns by Bendris [21]. *On n'a pas tout dit* is a one-hour long program, daily broadcasted between 2007 and 2008, and presented by Laurent Ruquier (see Fig. 1).

Contrary to *Le Grand Échiquier*, this show is almost exclusively made up of talk. However, instead of being centered on the life and achievements of a guest, it is constituted by humoristic chronicles and the majority of the persons present on the TV set are the hosts and commentators. One of the four shows has been singled out as a development set (55 minutes) while the remaining of the corpus is the actual test set (2 hours 45 minutes).

B. Technical Challenges: Talk-Shows versus Meetings

As mentioned earlier, reference speaker diarization works have been mostly concerned with meeting-conference recordings. Hence we here briefly compare this type of content to the talk-show content in order to highlight the technical challenges relating to the latter.

Meetings are generally recorded using distant wall-mounted or desktop microphones, while in TV talk-shows, the microphones generally used are high-quality boom and/or lapel microphones as shown in Fig. 2. However, audio conditions are potentially very noisy due to the presence of an audience and numerous overlapping-speech passages. Moreover, one of the main features of the talk-show genre is the high spontaneity of the speech between the participants.

Furthermore, while TV talk-shows and meeting conference data present common characteristics, the presence of musical passages, video excerpts imported from other audiovisual productions, applause, laughter, etc. is specific to talk-shows. Besides, silence segments generally observed at the speech turns are extremely short if not negligible. That can be explained by

[1]Transcriber—http://trans.sourceforge.net/

TABLE I
A COMPARISON OF THE NIST RT'09 AND *LE GRAND ÉCHIQUIER* (GE) DATASET CHARACTERISTICS. DURATIONS ARE GIVEN IN MINUTES/SECONDS AND PERCENTAGES REFLECT THE FRACTION OF EACH DURATION OVER THE TOTAL SPEECH DURATION, EXCEPT FOR THE FIRST TWO WHERE THE FRACTION IS COMPUTED WITH RESPECT TO THE TOTAL SHOW DURATION

event	NIST RT'09	GE
Nb. shows	7	6
Avg. duration	25' (100%)	147' (100%)
Avg. speech time	13' (52%)	50' (34%)
Avg. nb. of speech turns	882	1033
Avg. speech segment length	2" (0.2%)	3" (0.1%)
Avg. overlapping-speech time	3' (20%)	5' (10%)
Avg. nb. of speakers	5	16
Avg. speech time of the most active	8'55" (68%)	24'36" (49%)
Avg. speech time of the least active	2'26" (19%)	7" (0.2%)

the fact that TV talk-shows exhibit a more lively speech (almost scripted) due to the fact that TV show guests are prepared in advance by the host [1]. Table I (initially presented in [22]) proposes a more quantitative comparison between meeting data (NIST RT'09 [23]) and the talk-show corpus *Le Grand Échiquier* that is used in this study.

The GE corpus appears to be a challenging speaker diarization dataset, especially as far as the number of possible speakers and the repartition of the speech load are concerned. Indeed, the average number of speakers jumps from 5 for the NIST RT'09 meeting database to 16 for *Le Grand Échiquier*. The speaking time is also much larger, which can be a matter of importance since some state-of-the-art systems are not designed to deal with the computational burden that it can represent. Finally, the fact that the talk repartition is much better balanced in meetings than in talk-show programs (for which the least active speakers talk only a few seconds) is noteworthy.

The detection of speech data (i.e., when the talk show participants are speaking) has been treated quite extensively (see for instance [24]) and is not the main focus in this study. Thus, this detection is here performed semi-automatically by selecting, from the ground-truth annotations, speech parts of more than 30 seconds that are surrounded by music or applause lasting no less than 10 seconds. The speech sections obtained are therefore subject to contain laughter, silence, etc. on top of pure speech. The evaluation presented hereafter is therefore carried out on this set of noisy audio data, ensuring the generalization of the proposed approach since results obtained with fully automatic speech detection are expected to be at least as good.

C. Adapted Evaluation Methods

For speaker diarization campaigns, such as NIST RT [23], standard evaluation methods have been proposed in order for the participants to compare the performances of their systems. The usual evaluation metric is the *diarization error rate* (DER[2]) which is the sum of 3 or 4 errors: $E_{speaker}$, the percentage of scored time that a speaker ID is assigned to the wrong speaker, $E_{false-alarm}$, the percentage of scored time that a hypothesized speaker is labeled as a non-speech in the reference, $E_{missed-speech}$, the percentage of scored time that a hypothesized non-speech segment corresponds to a reference speaker segment and, if specified, $E_{overlap}$, the percentage of scored time that some of the multiple speakers in a segment do not get assigned to any speaker. A mapping function between the result of the automatic and the reference clustering is used so as to minimize $E_{speaker}$.

$$DER = E_{speaker} + E_{false-alarm} + E_{missed-speech} + E_{overlap}. \quad (1)$$

However, as it is explicated in [25], the DER cannot be expected to fully reflect the systems' capacities. Following this idea, it can be observed that the DER is sensitive to the speech repartition. Now, for talk-show programs it can be noticed (as in Table I) that speakers present on a given TV set do not carry the same load of speech. The correct identification of only the most prominent speakers, namely the host(s) and the main guest(s) can ensure low DER values even if the majority of the remaining speakers is incorrectly spotted. Hence, the standard DER poorly describes the diarization systems ability to correctly cluster the non-dominant speakers, which are as important as the main ones in a number of applications, especially when the ultimate goal behind the use of speaker diarization is to obtain a structure of the talk-show (based on speakers interventions as proposed in [1]).

Therefore, we propose two new error rates (the first one has been introduced in [26]). These measures, called *unipond* and *semipond*, compute weighted error rates. As it can be seen in the definitions (2) and (3), *unipond* and *semipond* weight the speech amount of each person so that it is more sensitive to its correct identification than to his/her speech load. These new error rates are again constructed as the sum of $E_{speaker}$, $E_{false-alarm}$ and $E_{missed-speech}$ as in the standard DER, however in this case the overlap errors are not taken into account. In both cases, the mapping function used in $E_{speaker}$ is the same as in the standard DER. So, with $T_{total}(i)$ the total time of the speech produced by the speaker i, $T_{error}(i)$ the total time of speech wrongly attributed for this same speaker, N the total number of speakers and k the number of major speakers (to distinguish from the minor speakers that talk way less) we have:

$$E_{speaker_{unipond}} = \frac{1}{N}\sum_{i=1}^{N}\frac{T_{error}(i)}{T_{total}(i)} \quad (2)$$

$$E_{speaker_{semipond}} = \frac{1}{2k}\sum_{i=1}^{k}\frac{T_{error}(i)}{T_{total}(i)} + \frac{1}{2(N-k)}\sum_{j=k+1}^{N}\frac{T_{error}(j)}{T_{total}(j)}. \quad (3)$$

The *unipond* metric equally weights every talk-show participants, meaning that, independently of their speech load, speakers contribute to an equal portion of the error rate. The *semipond* metric proceeds similarly but distinguishes primary and secondary speakers. Primary speakers include persons speaking the most during the TV show such as the main host(s) and guest(s) and secondary speakers the remaining participants. The error rate is then split in two equal parts between primary and secondary speakers. In the case of the program *Le Grand Échiquier*, k, the number of primary speakers is set to 2: the host Jacques Chancel and the main guest.

[2]NIST RT tools—http://www.itl.nist.gov/iad/mig/tools/index.html

TABLE II
NIST DER (IN PERCENTAGE) FOR THE NIST RT'09 AND *LE GRAND ÉCHIQUIER* (GE) DATASETS WITH AND WITHOUT SCORING ON THE OVERLAPPING-SPEECH PASSAGES

system	corpus	with ovlp-speech	without ovlp-speech
Baseline	**NIST RT'09**	21.1	16.0
	GE devel.	38.7	33.0
	GE test	38.2	34.9

TABLE III
UNIPOND AND *SEMIPOND* DER (IN PERCENTAGE) FOR THE NIST RT'09 AND *LE GRAND ÉCHIQUIER* (GE) DATASETS

system	corpus	*unipond*	*semipond*
Baseline	**NIST RT'09**	28.3	21.3
	GE devel.	68.8	47.1
	GE test	67.0	45.6

D. Evaluation of a State-of-the-Art System

The system proposed in [27] is used in order to situate the results obtained by a state-of-the-art approach for the task of speaker diarization on talk-show content (as previously done in [22]). The latter has been chosen on account of the very good results that it achieved for the NIST RT'09 evaluation campaign for the SDM task (Single Distant Microphone).

Table II displays the results obtained for the NIST RT'09 corpus and the GE corpus (*Le Grand Échiquier*) with the standard DER. For the NIST RT results, the data of the previous campaigns (RT'04, 05, 06 and 07) have been used as a development set so that RT'09 shows could be used entirely in the test set. For the GE dataset the results are given over both the development and test sets.

Table III gives the results obtained on the corpus *Le Grand Échiquier* with the newly introduced metrics *unipond* and *semipond*. As it can be seen in Table II, the results obtained on this database are much worse than those obtained on the NIST one. Moreover Table III emphasizes this trend, showing that on average only one third of the speakers present on a TV set are correctly recognized.

III. A NEW MULTIMODAL SYSTEM

As briefly exposed in Section I, a number of solutions have been proposed for the speaker diarization of TV content. In our previous work [22], we have shown that robust visual features may prove very useful for the initialization of a top-down speaker diarization system. Indeed, the results obtained have shown an improvement of the overall diarization performance in comparison with the original state-of-the-art system. However, the joint use of audio and visual features during the clustering/classification phase have turned out to be inefficient (that is hardly better than an audio-only system) with this GMM-HMM architecture, despite our extensive efforts.

Now the new architecture founded on the use of Support Vector Machines (SVM) we proposed in [26] has proven effective at fusing audio and visual features for an improved speaker classification. However, this system was labeled as a *semi-automatic system*, since it relied on the fact that a user would provide very short training examples (typically between 5 and 15 seconds for each participant that appears in the TV show) to be used to learn SVM classifiers to discriminate the speakers. Thus

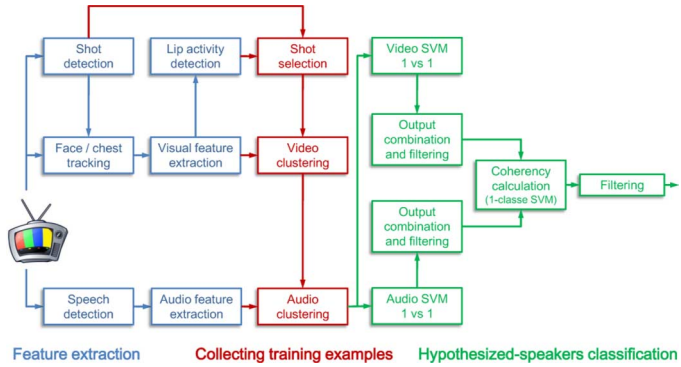

Fig. 3. Description of the speaker diarization system proposed.

inspired by this previous work, the new architecture presented hereafter employs the same type of classifiers, yet this time the training examples are collected in an unsupervised fashion (instead of being provided by a user), and an improved fusion strategy is adopted.

As shown in Fig. 3, the proposed system can be divided in three distinct steps: the extraction of audio and video features, the collection of training data to create a model for each speaker and finally the classification of all speech parts of the talk-show being processed.

A. Feature Extraction

As audio descriptors, we use, as it is usually done, Mel-Frequency Cepstral Coefficients (MFCC, see [28]) along with their first and second derivatives. We also add Line Spectral Frequencies (LSF, see [29]). The extraction is performed at a rate of 100 Hertz using the *YAAFE*[3] software.

As for the visual descriptors, we consider features characterizing the clothing of the TV show-participants, building upon the method initially proposed in [30]. Indeed, though the field size of the shots are varying (from long shots to close-ups), most of the time the person talking is seen on-screen. However, the use of a facial recognition system as in [31] is rather difficult in our task due to the camera movements, the changing field-size and angles of shot, the changing postures of the filmed persons and the varying lighting conditions.

The approach that we choose has the advantage that features relating to on-screen persons' clothing can be extracted even more robustly than the persons' facial features. Another motivation is the fact that in the TV production domain, the clothing of the participants on a TV set is often carefully chosen in order for the viewer not to confuse them. This is asserted in details in specialized TV production publications such as [32] and [33].

Thus, our main assumption being that we often "see who we hear", we suppose that the information carried by the clothing can help us for speaker identification. This assumption is reinforced by Fig. 4 showing the correlation between the clothing dominant color of the person on-screen and the speech turns, the two speakers being those presented just above. The localization of the chest for the extraction of the clothing information is performed using the method described in [26]. In a nutshell, faces are first detected using the Viola-Jones algorithm [34], and the

[3]YAAFE—http://yaafe.sourceforge.net/

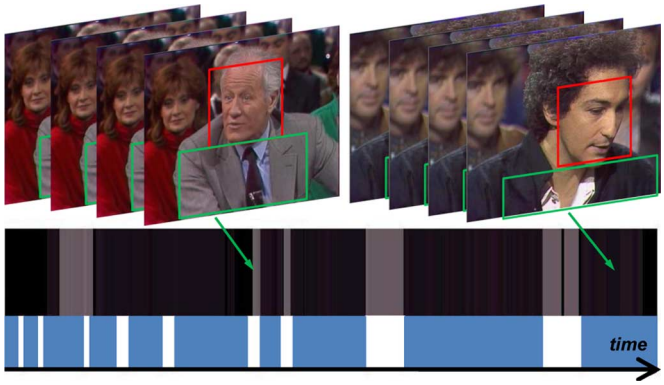

Fig. 4. Clothing dominant color (above) and speech turns (below) for a two-minute speech segment with two participants.

TABLE IV
AUDIO AND VIDEO FEATURES USED IN THE NEW SPEAKER DIARIZATION SYSTEM PROPOSED

modality	features	dimension
audio	*MFCC, Δ, Δ²*	39
	LSF	10
video	*HSV Histogram*	22
	Cumulated HSV Histogram	22

corresponding chest localization is then derived from heuristic rules. Then a tracking is realized within each video shot, as done in [26], using the OpenCV software (see [35]) so that the position of the faces and chests can be determined in the case of non-detection of faces on some frames.

In this work we rely on a more thorough description of the clothing color by extracting HSV (*Hue Saturation Value*) histograms computed for each frame, as well as cumulated histograms for each shot. The former are supposed to reflect the color distribution at each frame while the latter represent a compact description for each shot of the TV show.

Following a preliminary study to tune the histogram extraction parameters (led on the Canal9 political debate dataset [13]), they are computed on respectively 16, 4 and 2 bins, with amplitudes varying from 0 to 180 for the Hue component, and 0 to 255 for the Saturation and the Value. The shot detection is performed using *Shodetect*.[4] The video frame rate is 25 Hertz as it is generally observed for TV productions. Table IV gives an overview of the extracted descriptors and gives their dimension.

B. Collecting Training Examples

The goal of this stage is to gather, in a non-supervised fashion, training examples as pure as possible, in order to create the best training database for the speakers present on the TV set of the considered talk-show. By "examples as pure as possible" we refer to the fact that each speaker model to be created has to be composed by as much data from the same participant as possible.

1) Shot Selection: The major hypothesis used in this selection step is again that "you often see who you hear". Nevertheless, it is important to account for the fact that exceptions to the previous hypothesis are susceptible to arise. Thus, we are interested in keeping only the shots where this assumption is more likely to be verified—performing what is usually called *speaker detection* (see for instance [36] and [37]).

[4]Shotdetect—http://shotdetect.nonutc.fr/

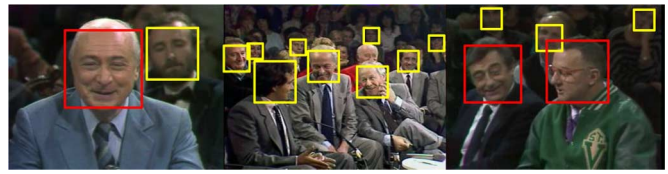

Fig. 5. Face detection and filtering. The faces in red are kept and those in yellow discarded. The shot in the middle is discarded since no face big enough can be found.

2) Shot Filtering: First of all, only shots that are long enough and contain faces in the foreground are kept. This can be explained by the fact that it is much more likely to have an active speaker shown on-screen during a lengthy shot than during a short one. Indeed the latter is generally used to show reactions to the speech currently given (either from the audience or some listening participant).

Therefore, a subset of shots of the show that are considered to be "long enough" are selected by retaining those which are longer than 10 seconds for *Le Grand Échiquier* against 5 for *On n'a pas tout dit*. This length being directly linked with the average duration of a shot (7.5 against 3 seconds).

Another threshold has also to be set to characterize shots with too small faces (usually located in the background). This allows for discarding audience views or long shots that cannot be taken into account for the creation of reliable speaker models. Fig. 5 offers an illustration of this last principle.

3) Lip Activity Detection: Now our goal is to try to make sure that, in the selected shots, the on-screen person in the foreground is effectively the active speaker. Indeed, while we highlighted earlier that most of the time the speaker is shown on-screen, it is necessary, to assemble training sets as pure as possible, to discard shots for which the person visible is not the one doing the talking. Therefore, inspired by [38] a lip activity detector is proposed. However, in our case, owing to the great spontaneity of the participants (that has to be distinguished from sitcom content as proposed in [38] where actors are playing a well-defined role), we need a more robust detector. We proceed as follows: a face detector [34] and a face tracker [39] are combined so that a detected face can be followed in a given shot as explained in [26]. Then, as explained in [38], facial features of the mouth are detected in the face region using a generative model of the feature positions combined with a discriminative model of the feature appearance. The probability distribution over the joint position of the features is modelled using a mixture of Gaussian trees, a Gaussian mixture model in which the covariance of each component is restricted to form a tree structure with each variable dependent on a single parent variable. This model is an extension of the single tree proposed in [40].

Thus, once the mouth corners are located in shot-frames where faces could be detected, a grid of points of interest is placed on a rectangle containing the mouth (with grid points being placed every two pixels). This rectangle is slightly down shifted in order to measure the movements of the lower lip that is the one moving while talking. Optical flow tracking is then performed using [39] across the whole shot duration. This

Fig. 6. Setting of the rectangle on the mouth of the detected face and tracking of the grid points.

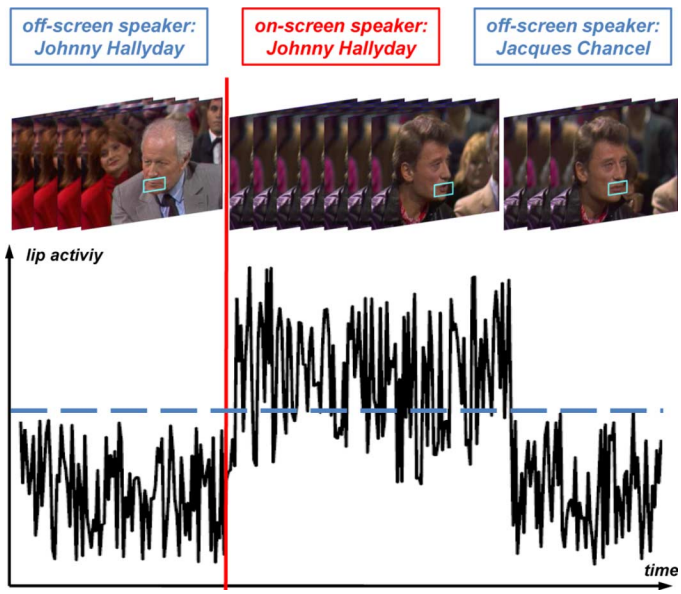

Fig. 7. Lip activity detector for two consecutive shots of the talk-show *Le Grand Échiquier*.

tracking is performed either forward or backward, depending on the location, within the shot, of the frames where mouth corners could be detected.

Fig. 6 shows how the bounding rectangle is set along with the displacement of the grid points between two consecutive frames. The grid points for the frame $f-1$ are displayed in green while the grid points for the current frame f are in blue. The corresponding displacement is visible in red. The lip activity can then be computed for each video frame as the difference of the average head motion and the average mouth motion (with the average taken over all the points of interest of the grid). As explained in [26], the average head motion can be computed during the extraction of visual features since a face tracker is employed.

An illustration of the lip activity detector is provided in Fig. 7. Two persons appear on-screen consecutively: Jacques Chancel, the host of the talk-show *Le Grand Échiquier* and the French singer Johnny Hallyday. During these two shots (separated by the vertical red line), Johnny Hallyday starts talking off-screen and carries on with the camera switching from Jacques Chancel to him. Then, Jacques Chancel interrupts Johnny Hallyday who remains silent for the rest of the second shot, while Chancel is speaking off-screen. The measured lip activity shown in the bottom-half of Fig. 7 testifies that the proposed method is coherent with what is practically observed.

For an easier thresholding, the instantaneous variations of the lip-activity signal are actually smoothed using a 0.5-second median filter (with 50% overlap). Due to the absence of visual annotations in terms of "talking faces", the lip activity detector has not been directly evaluated. However, the validity of this contribution is implicitly verified through the good diarization results obtained in the following. Indeed, on average, a three to four point increase (depending on the evaluation metric used) is observed in the final results with the adjunction of the lip activity detector.

In a nutshell, to sum up the selection process, the shots have to be long enough and to exhibit a significant lip activity to be picked up for the automatic creation of speaker models. Of course, errors are susceptible to happen, for instance, if the on-screen persons' lips are moving while the actual speaker is off-screen. This is precisely the reason for which SVM are chosen as classifiers. Indeed they show a great ability to cope with outliers, as it has been observed in [26].

4) Clustering Cascade: Once obtained a selection of shots with on-screen speakers, the shots corresponding to the same speakers need to be grouped together. For this, two clusterings are combined: the first is done on visual features and the second on audio features. This cascade of monomodal clusterings has been experimentally found to be more efficient than a single clustering based on combined audiovisual features. In particular, a trade-off between performance and complexity is thus achieved as the first visual clustering is quite simple, while the subsequent audio clustering is more computationally intensive. This is further explained in the following.

5) Visual Clustering: Using the cumulated HSV color histograms (computed on the clothing of the on-screen person) for the selected shots, a first (visual) grouping is performed. This one is a hierarchical agglomerative clustering. The distance used to measure the shot similarities is a χ^2 distance computed on the cumulated color histograms. It is defined as follows:

$$d_{\chi^2} = \frac{1}{2}\sum_{i=1}^{b}\frac{(Hx_i - Hy_i)^2}{(Hx_i + Hy_i)}\ ; \tag{4}$$

where Hx and Hy are histograms of b bins each (see [41]).

The grouping stops when a "satisfying" number of visual clusters $\mathcal{C}_\mathcal{V}$ is reached. In order to avoid confusions, that is avoiding the creation of clusters combining two distinct speakers, this number is chosen to be quite large (meaning way above the roughly expected number of speakers in the tested TV show). We set experimentally this number to 40 in the case of *Le Grand Échiquier* and *On n'a pas tout dit* since never more than 20 speakers are expected in such programs.

Fig. 8 displays the principle of the hierarchical grouping. The clustering is stopped when the blue dotted line is reached. As depicted, since this criterion is met pretty quickly, clusters remain rather pure, meaning that the shots collected in a given cluster usually correspond all to the same person. Besides, due to the choice of the stopping criterion, several clusters can actually represent the same speaker (as it is the case for the two at the left and the two at the right of Fig. 8). This over-clustering issue is then solved by the addition of an audio clustering as explained afterwards.

6) Audio Clustering: Given the previously formed 40 visual clusters $\mathcal{C}_\mathcal{V}$, the goal is now to obtain fewer clusters (ideally as many as the number of target speakers), so that each of them can

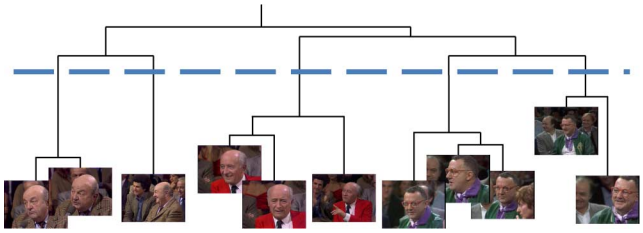

Fig. 8. Dendrogram illustrating the hierarchical agglomerative clustering based on the cumulated HSV histograms of the selected shots.

TABLE V
PURITY OF THE LEARNING DATABASE AUTOMATICALLY ASSEMBLED, ON AVERAGE ON THE DEVELOPMENT SET OF THE TALK-SHOW *LE GRAND ÉCHIQUIER*. THE LAST COLUMN REPRESENTS THE TOTAL TIME OF THE SELECTED SHOTS (IN MINUTES, SECONDS)

shot selection		video clustering		audio clustering		time
nb. clust.	purity	nb. clust.	purity	nb. clust.	purity	
107.3	91.7%	40	88.1%	13.7	87.7%	33'34"

be considered as an appropriate set of training examples (for a particular speaker). Recall that the training data thus collected is to be used to learn classifiers capable of discriminating between the speakers.

Since it was found that the visual features did not allow for further grouping the clusters obtained so far without introducing confusions, this final grouping exploits the audio features (MFCC, Δ, Δ^2 and LSF). It is achieved by applying agglomerative hierachical clustering to the clusters $\mathcal{C}_\mathcal{V}$, yet using a more sophisticated inter-cluster distance, namely Bhattacharryya distances in a Reproducing Kernel Hilbert Space (RKHS) [42], which are kernel-based probabilistic distances. In fact, these distances have been found well-suited for clustering complex audio data in previous work (see [24], [43], [44]). Exploratory experiments have shown that much better results could be obtained than with traditional BIC-related approaches.

Contrary to the initial visual clustering, this time a dynamic threshold is used to stop the clustering at a correct level and thus ideally obtain as many clusters $\mathcal{C}_\mathcal{A}$ as target speakers. This threshold—defined as the best compromise between the duplication/merging of speaker models—is learned on the development set of each dataset (*Le Grand Échiquier* and *On n'a pas tout dit*).

Table V testifies that this clustering cascade allows us to collect acceptable training sets for the different speakers. Indeed, the number of clusters diminishes drastically while their average purity stays almost constant. This purity, sometimes referred to as the precision, is the ratio of the time attributed to the main speaker of the cluster over the total duration of the cluster. Of course, the speaker models created contain outliers. However, their effect during the classification phase are expected to be downplayed by the use of support vector machines.

C. SVM Audio Classification of Hypothesized Speakers

The previous phase allows the collection of then reliable training datasets for speaker classification, as attested by the results presented in Table V. The goal is now to process the remaining parts of the show (that were not selected during the data collection stage and not taken into account in the previous clustering cascade). Note that these remaining parts represent a higher fraction of the content, compared to the shots selected in the previous stage.

As mentioned earlier, we choose to use support vector machines on account of their ability to perform robust classification of the hypothesized speakers. For more information on SVM, we refer the reader to the many good tutorials on this powerful tool (see for instance [45]). The classification is performed over the previously extracted audio features (MFCC, Δ, Δ^2 and LSF).

Practically, we use one-vs-one SVM classifiers, meaning that for the N hypothesized speakers obtained earlier, $N(N-1)/2$ biclass classifiers are trained. Since the training sets are potentially imbalanced we use a different C value[5] for positive and negative training examples, which we will refer to as $C+$ and $C-$ respectively, so that the solution is not biased by the over representation of one class at the expense of another in every bi-class problem (one speaker against another). C_+ and C_- are respectively set as the ratios of positive and negative examples over the total number of examples. Also the kernel used in this work is the usual Gaussian kernel:

$$\kappa(x,y) = \exp\left(-\frac{\|x-y\|^2}{2d\sigma^2}\right) \tag{5}$$

with d the dimension of the feature vector and σ the kernel width parameter. The latter is set in the development phase through cross-validation.

We have chosen to use the *LIBSVM*[6] toolbox [46] to build the various classifiers and obtain probabilized outputs. The obtention of the frame by frame speaker probabilities (for each of the N hypothesized speakers) is then performed using the *minpair* coupling as proposed in [47].

Then, a median filtering is applied using a 0.5-second window, with a 50% overlap. For each of these temporal segments the speaker label that is elected is the one corresponding to the hypothesized speaker with maximum probability.

TABLE VI
DIARIZATION ERROR RATES (IN PERCENTAGE) FOR THE DATASET *LE GRAND ÉCHIQUIER* (GE) AND THE THREE AVAILABLE METRICS. NIST DER IS GIVEN WITH AND WITHOUT SCORING ON THE OVERLAPPING-SPEECH PASSAGES

system	corpus	NIST DER	*unipond*	*semipond*
Baseline	**GE devel.**	38.7 - 33.0	68.8	47.1
	GE test	38.2 - 34.9	67.0	45.6
System I	**GE devel.**	34.5 - 30.5	38.6	33.4
	GE test	33.6 - 30.6	54.2	41.7

D. Experimental Results

We propose here to evaluate the system previously described using the three metrics introduced in Section II: NIST DER, *unipond* and *semipond*. Table VI shows the speaker diarization scores obtained for the corpus *Le Grand Échiquier*.

It can be observed, when comparing with the results from Tables II and III of Section II, that the scores are largely improved. For instance, the improvement is as high as 7.8% (in absolute) with NIST DER or 14% with the *unipond* metrics on

[5]C is the regularization factor penalizing outliers in so-called C-SVM schemes.

[6]LIBSVM—http://www.csie.ntu.edu.tw/~jlin/libsvm/

TABLE VII
DIARIZATION ERROR RATES (IN PERCENTAGE) FOR THE DATASET *ON N'A PAS TOUT DIT* (OAPTD) AND THE THREE AVAILABLE METRICS. NIST DER IS GIVEN WITH AND WITHOUT SCORING ON THE OVERLAPPING-SPEECH PASSAGES

system	corpus	NIST DER	*unipond*	*semipond*
System I	**OAPTD devel.**	35.8 - 27.7	30.8	29.5
	OAPTD test	38.4 - 31.2	45.5	40.3

the test set. Furthermore, it can be noted that while the results are almost the same on the development and test sets for the standard DER, they are much worse for the test set with *unipond* and *semipond*. This can be explained by the number of speakers that is greater in the test set, compared to the development set (on average 17.6 against 15). Still, the performance of our system on the test set is significantly better than the reference system. Besides, one can be surprised by the fact that the NIST DER of the *System I* is lower for the test set than the development set while it is the opposite for *unipond* and *semipond*. That implies that with the weighting introduced in the computation of the new metrics, the importance of each speaker is altered (more for the less active speakers and less for the most active).

To avoid any risk of overfitting to the corpus *Le Grand Échiquier*, the algorithm presented above has also been tested on the TV talk-show dataset *On n'a pas tout dit* (OAPTD). The system run is exactly the same except that two parameters had to be adjusted: the minimum duration for the shot selection and the dynamic threshold for the audio clustering. Their tuning is performed on one show of the corpus that is used as development set while the three remaining ones constitute the test set.

From the results presented in Table VII it can be noticed that the score discrepancy between the standard DER with overlapped speech, compared to the one without overlap is much more important for *On n'a pas tout dit* than for *Le Grand Échiquier*. This can be explained by the speech characteristics of the former show, which are much quicker and concise and for which the participants are more prone to speak at the same time. Otherwise, the results obtained for the three metrics seem in adequation with those presented in Table VI, as they follow the same trend.

IV. AUDIOVISUAL CLASSIFIER FUSION

A shortcoming of the system presented so far is that the visual information is not used during the classification phase. Visual features are however clearly much more stable than their audio counterparts (when the speaker is visible on-screen). Indeed, during a shot, the color distribution is very little subject to change, while audio descriptors exhibit major fluctuations. The exploitation of visual features can thus be of great value, provided that we are able to determine, at every time instant, whether the current speaker is on-screen or not. Therefore, we add a visual classifier to the audio one.

In the following, we first describe the visual classifier, then explain how its output is combined with the audio classifier output, after a confirmation as to whether the speaker is on-screen has been obtained.

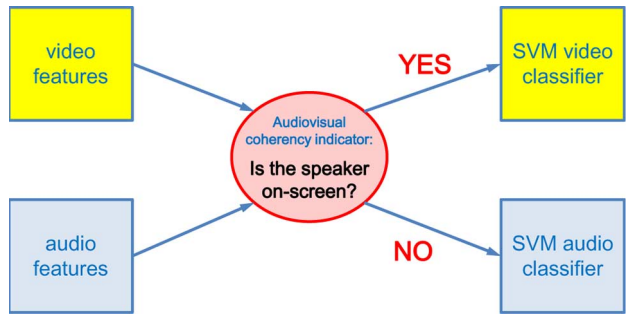

Fig. 9. Principle of the audiovisual coherency indicator.

A. SVM Visual Classification of Hypothesized Speakers

The visual classification is again done using one-vs-one SVM classifiers with probabilistic outputs. They are built using HSV color-histograms from the selected on-screen person clothing region.

The only difference with the SVM audio classification previously proposed is the kernel chosen. In this case, the histogram intersection kernel is preferred to the Gaussian kernel (see [48]). With Hx and Hy two b-bin histograms, this kernel is defined as:

$$\kappa(Hx, Hy) = \sum_{i=1}^{b} \min\{Hx_i, Hy_i\}. \quad (6)$$

B. Audiovisual Coherency Indicator

Since the visual features are extracted on the participants' clothing bounding boxes (which are determined based on face detection), all frames where faces have not been detected (such as those inside cutaway shots) do not allow for extracting those features. Therefore, in this case we can only rely on the results obtained *via* the SVM audio classifiers.

For shots where both audio and visual features are available we however do not know whether the on-screen person is the one talking. As exposed in Fig. 9 we thus need to create an audiovisual coherency indicator, that, by pointing out if the person talking appears on-screen will allow us to use either the audio or the video modality. As stated earlier, the video descriptors are known for being steadier than the audio ones for our task (this has been confirmed experimentally). Thus, when the speaker is shown on-screen the visual classifier's output will be preferred.

Numerous approaches have been proposed to compute the audiovisual correlation between audio and video signals, for instance the Canonical Correlation Analysis (CCA, [49]), the Co-Inertia Analysis (CoIA, [50]) or the Cross Factor Analysis (CFA, [51]). We here follow a different approach that consists in learning an audiovisual distribution "profile" for each speaker.

Let $P(S_i|x_A(f))$ and $P(S_i|x_V(f))$ be the SVM output probabilities for the frame f to belong to the speaker S_i given the audio feature vector $x_A(f)$ or the visual feature vector $x_V(f)$. The key idea is that the audiovisual classification outputs $P(S_i|x_A(f))$ and $P(S_i|x_V(f))$, for all S_i, have different "profiles" when the active speaker is on-screen compared to when they are off-screen. This "profile" can be learned from the training dataset that has been assembled after the shot

selection and the clustering cascade, as will be described in Section IV-B1. Then, the audiovisual coherency indicator can verify for a tested frame f if the speaker probabilities obtained *via* visual classification are coherent with the ones obtained by audio classification. Two actions are then possible: if the indicator says the audio and visual classification outputs are compatible, then the visual label is kept. If not, meaning that the audio and video information are "uncorrelated" (and the person on-screen is not the one talking), the visual information is discarded and the label attributed to the frame f is determined by the audio classifier.

1) Modeling the Joint Distribution of Audiovisual Classifiers Outputs: Our audiovisual coherency detector decides if the audio and visual classifier outputs are compatible by checking whether the meta-feature vector

$$z_{AV}(f) = \left[P\left\{S_1|x_A(f)\right\}, \ldots, P\left\{S_N|x_A(f)\right\}, P\left\{S_1|x_V(f)\right\}, \ldots, P\left\{S_N|x_V(f)\right\}\right]$$

computed on a test frame f can be considered as an observation of the distribution $P(Z_{AV})$ learned on the previously assembled training set.

Here, we propose to use the one-class SVM technique to learn this distribution. This is in fact an effective non-parametric density estimation technique. We refer the reader to [45] for further details on one-class SVM, and merely indicate here how we exploit them. Similarly to the more traditional bi-class SVM setting, in one-class SVM a decision function $g(x)$ is learned whose sign indicates whether or not a tested observation belongs to the modeled density.

In summary, speaker classification based on coherency detection is performed as follows:

i) one-class SVMs are learned on the training dataset (assembled after the shot selection and the clustering cascade) for every hypothesized speaker S_i, yielding decision functions $g_{S_i}(x)$;
ii) during the testing phase, for a frame f, the decision function relating to the speaker S_m maximizing $P\left\{S_i|x_V(f)\right\}$ is selected and applied to $z_{AV}(f)$;
iii) if $g_{S_m}(z_{AV}(f))$ is positive (i.e., the on-screen person is the active speaker), then the label S_m is validated for frame f, otherwise the decision (as to who is the active speaker) is left to the audio classifier by taking $\arg\max_{S_i} P\left\{S_i|x_A(f)\right\}$.

2) Comparison With Existing Techniques: In order to validate the one-class SVM approach proposed here for testing the hypothesis "*the person appearing on-screen is speaking*" we make a comparison with state-of-the-art methods, namely CCA, CoIA and CFA (see [49]–[51]). The correlation between the output probability vectors $x_A(f)$ and $x_V(f)$ is then computed. If the correlation is positive, then the person on-screen is the one talking, otherwise it is not the case. The projection bases are learned on the training dataset created after the shot selection and the clustering cascade similarly to the one-class SVM training.

Since unfortunately no groundtruth is available for the event "the active speaker is on-screen", the efficiency of the audiovisual coherency indicator has not been directly evaluated.

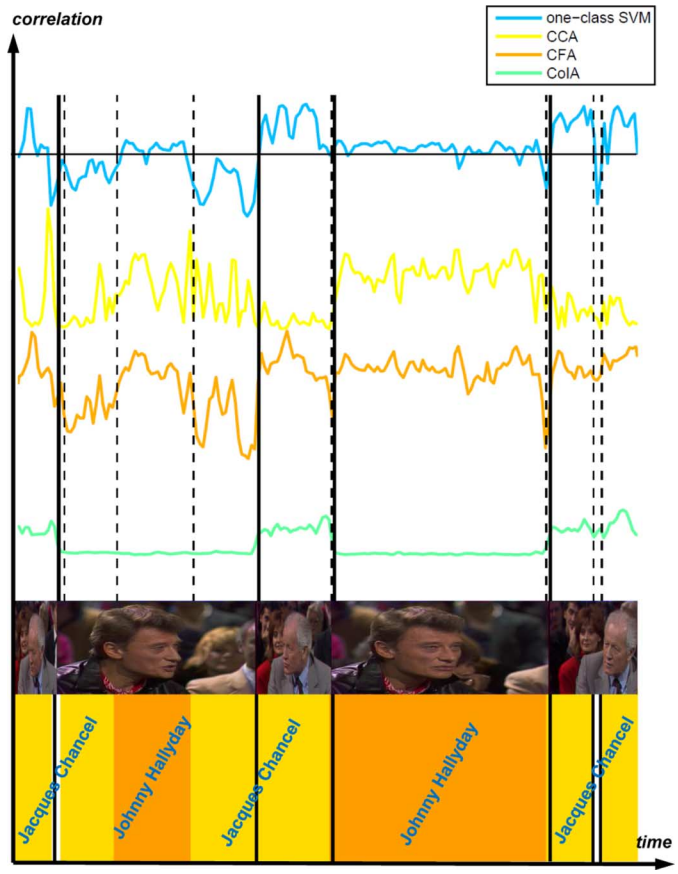

Fig. 10. Comparison of various methods measuring the audiovisual correlation for a few minutes of the talk-show *Le Grand Échiquier*.

However, note that it is implicitly assessed through its positive impact on the speaker diarization results as will be shown in Section IV-C. Fig. 10 proposes a visualization of the correlation/coherency results achieved by our method compared to CCA, CoIA and CFA, over a few minutes of video. The visual cuts are indicated by the black vertical lines while the speaker turns are in dotted lines. The bottom row gives the groundtruth for the speech repartition (with the names of the speakers). Just above a key-frame of each shot is shown. It is therefore possible to check whether the audio and visual information are correctly correlated.

Fig. 10 highlights that the method we propose behaves much better than CCA and CoIA. Indeed, the correlations measures obtained with the latter exhibit great amplitude variations (generally correlated with the shot changes) that make the dynamic thresholding between correlated and uncorrelated parts rather difficult to tune. CFA seems to work fairly well even if here again the tuning of a threshold is problematic (which has been further confirmed on numerous examples during our experiments). In comparison, the new method based on the verification of the coherency of the audio and video streams using one-class SVM allows for a much easier adjustment of the decision threshold.

C. Experimental Results

We now test the impact of the added audiovisual fusion components on the performance of our speaker diarization system.

TABLE VIII
DIARIZATION ERROR RATES (IN PERCENTAGE) FOR THE DATASET *LE GRAND ÉCHIQUIER* (GE) AND THE THREE AVAILABLE METRICS. NIST DER IS GIVEN WITH AND WITHOUT SCORING ON THE OVERLAPPING-SPEECH PASSAGES

system	corpus	NIST DER	*unipond*	*semipond*
Baseline	**GE devel.**	38.7 - 33.0	68.8	47.1
	GE test	38.2 - 34.9	67.0	45.6
System I	**GE devel.**	34.5 - 30.5	38.6	33.4
	GE test	33.6 - 30.6	54.2	41.7
System II	**GE devel.**	31.9 - 27.6	38.5	32.6
	GE test	32.8 - 29.4	53.5	40.8

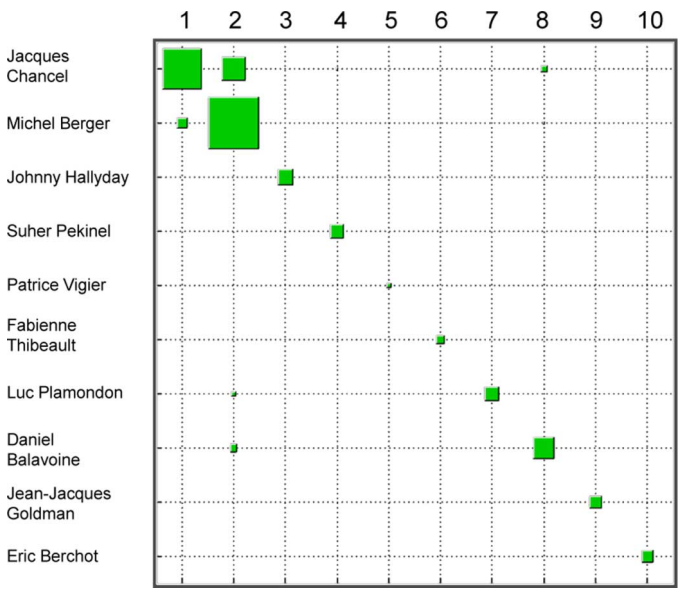

Fig. 11. Confusion matrix showing the actual/hypothesized speakers association for a program of the dataset *Le Grand Échiquier*.

The results are presented in Table VIII. It can be observed that the results are globally better compared to the previous system (that did not exploit the visual classifiers). However, for the *unipond* metric and, to a lesser extent *semipond*, the improvement is not as high as for the standard NIST DER. Audiovisual fusion modules improve the clustering of the dominant speakers. However, as expected, since they do not introduce new clusters, they cannot increase as much the *unipond* and *semipond* metrics.

Fig. 11 shows a confusion matrix for a particular broadcast program of *Le Grand Échiquier*. The columns relate to the hypothesized speakers (deduced after the clustering cascade as detailed in the previous sections) while the rows are the target groundtruth speakers present on the talk-show set. Therefore, the green blocks indicate the proportion of frames classified as cluster i, with $i \in [1, 10]$ that belong to a given speaker (Jacques Chancel, Michel Berger, etc.). The sizes of the blocks clearly indicate that the speech load varies a lot from one person to another. Besides, it allows for the identification of common mistakes such as the fact that main speakers tend to "attract" frames that belong to secondary speakers. Also, it can be generally observed that the host (Jacques Chancel in this example) has a lot of speech frames that are classified in the cluster corresponding to the main guest (here cluster 2 for Michel Berger). This phenomenon can be explained by the type of interventions of the host: very short and concise that correspond to questioning, often interrupting the guest.

TABLE IX
DIARIZATION ERROR RATES (IN PERCENTAGE) FOR THE DATASET *ON N'A PAS TOUT DIT* (OAPTD) AND THE THREE AVAILABLE METRICS. NIST DER IS GIVEN WITH AND WITHOUT SCORING ON THE OVERLAPPING-SPEECH PASSAGES

system	corpus	NIST DER	*unipond*	*semipond*
System I	**OAPTD devel.**	35.8 - 27.7	30.8	29.5
	OAPTD test	38.4 - 31.2	45.5	40.3
System II	**OAPTD devel.**	33.3 - 24.7	27.6	27.1
	OAPTD test	37.8 - 29.2	44.7	38.0

Table IX provides the diarization results of the last system on the validation dataset *On n'a pas tout dit*. Again, improvements are observed, particularly for the standard DER. However, due to the higher frequency of speaker turns, speakers are more likely to talk at the same time, which explains why the difference while taking or not the overlapped speech parts into account is so important.

V. CONCLUSION

In this article we have proposed a new multimodal speaker diarization system exploiting kernel methods. The focus has been put on talk-show programs, as such TV content raises challenging research issues, especially from an indexing and event-retrieval perspective for which speaker diarization is a crucial capability.

Once audio and visual features have been extracted, our system proceeds through two major phases: collecting learning examples for each speaker (in a non-supervised fashion) and classifying audiovisual frames based on SVM. The first phase is accomplished by selecting (long enough) shots of each speaker, before grouping them together in a cascade of visual and audio clusterings. Once obtained these clusters constitute training data for hypothetical speakers, that is used to learn SVM speaker classifiers. Two schemes are then considered: an audio-only classification scheme and a parallel audio/visual classification scheme. In the latter, the classifier output is chosen through an audiovisual coherency analysis which checks if the person talking appears on-screen. If this is the case, the speaker label output by the visual classifier is chosen, otherwise the audio one is kept. Both classification schemes display appreciable improvements in comparison with state-of-the-art systems.

Thus, the exploitation of visual cues (when available) has been shown to be very valuable for the task of speaker diarization, even though the audio modality is the only one which is always reliable. This is owing to the fact that the active speaker is not always seen on-screen.

Also of note is the fact that discriminative methods such as SVM classifiers prove to be very competitive in a field that usually employs quasi-exclusively generative methods such as GMM-HMM.

ACKNOWLEDGMENT

The authors would like to thank Prof. G. Richard from Telecom ParisTech and Dr. N. Evans and Dr. S. Bozonnet from Eurecom for their precious help.

REFERENCES

[1] F. Vallet, S. Essid, J. Carrive, and G. Richard, *TV Content Analysis: Techniques and Applications (Chapter High-Level TV Talk Show Structuring Centered on Speakers' Interventions)*, Y. Kompatsiaris, B. Merialdo, and S. Lian, Eds. Boca Raton, FL, USA: CRC, Taylor Francis, 2012.

[2] M. Bendris, D. Charlet, and G. Chollet, "Talking faces indexing in TV-content," in *Proc. Content-Based Multimedia Indexing*, Grenoble, France, Jun. 2010.

[3] E. E. Khoury, G. Jaffré, J. Pinquier, and C. Senac, "Association of audio and video segmentations for automatic person indexing," in *Proc. Int. Workshop Content-Based Multimedia Indexing*, Bordeaux, France, Jun. 2007.

[4] H. Salamin and A. Vinciarelli, "Automatic role recognition in multiparty conversations: An approach based on turn organization, prosody and conditional random fields," *IEEE Trans. Multimedia*, vol. 14, no. 2, pp. 338–345, 2012.

[5] B. Bigot, I. Ferrané, J. Pinquier, and R. André-Obrecht, "Speaker role recognition to help spontaneous conversational speech detection," in *Proc. ACM Workshop Searching for Spontaneous Conversational Speech*, Firenze, Italy, Oct. 2010.

[6] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1557–1565, 2006.

[7] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, 2012.

[8] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Orlando, FL, USA, May 2002.

[9] Multiple Biometric Grand Challenge, MBGC. [Online]. Available: http://www.nist.gov/itl/iad/ig/mbgc.cfm.

[10] The Extended Multi Modal Verification for Teleservices and Security Applications DataBase, XM2VTSDB. [Online]. Available: http://www.ee.surrey.ac.uk/cvssp/xm2vtsdb/.

[11] Multimedia, Vision and Graphics Laboratory Audio-Visual Database, MVGL-AVD, [Online]. Available: http://mvgl.ku.edu.tr/databases.

[12] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Proc. Machine Learning for Multimodal Interaction: 2nd Int. Workshop*, Edinburgh, U.K., Jul. 2005.

[13] Canal 9 Political Debates Database, Canal 9. [Online]. Available: http://canal9-db.sspnet.eu/.

[14] A. Dielmann, "Unsupervised detection of multimodal clusters in edited recordings," in *Proc. Multimedia Signal Processing*, Saint-Malo, France, Oct. 2010.

[15] H. Vajaria, T. Islam, S. Sarkar, R. Sankar, and R. Kasturi, "Audio segmentation and speaker localization in meeting videos," in *Proc. Int. Conf. Pattern Recognition*, Hong Kong, China, Aug. 2006.

[16] H. Hung and G. Friedland, "Towards audio-visual on-line diarization of participants in group meetings," in *Proc. Workshop Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, Marseille, France, Oct. 2008.

[17] M. Bendris, D. Charlet, and G. Chollet, "Lip activity detection for talking faces classification in TV-content," in *Proc. Int. Conf. Machine Vision*, Hong Kong, China, Dec. 2010.

[18] G. Friedland, H. Hung, and C. Yeo, "Multi-modal speaker diarization of real-world meetings using compressed-domain video features," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Taipei, Taiwan, Apr. 2009.

[19] G. Friedland, C. Yeo, and H. Hung, "Visual speaker localization aided by acoustic models," in *Proc. ACM Int. Conf. Multimedia*, Beijing, China, Apr. 2009.

[20] C. Wooters and M. Huijbregts, "The ICSI RT'07s speaker diarization system," *Multimodal Technol. Percept. Humans*, vol. 4625, pp. 509–519, 2008.

[21] M. Bendris, "Indexation audio-visuelle des personnes dans un contexte de télévision," Ph.D. dissertation, Telecom ParisTech, Paris, France, 2011.

[22] S. Bozonnet, F. Vallet, N. Evans, S. Essid, G. Richard, and J. Carrive, "A multimodal approach to initialisation for top-down speaker diarization of television shows," in *Proc. European Signal Processing Conf.*, Aalborg, Denmark, Aug. 2010.

[23] The NIST Rich Transcription 2009 (RT'09) Evaluation, NIST, 2009. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-evalplan-v2.pdf.

[24] G. Richard, M. Ramona, and S. Essid, "Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Honolulu, HI, USA, Apr. 2007.

[25] J.-F. Bonastre, "La reconnaissance du locuteur: Un problème résolu?," in *J. d'études sur la Parole*, Avignon, France, Jun. 2008.

[26] F. Vallet, S. Essid, J. Carrive, and G. Richard, "Robust visual features for the multimodal identification of unregistered speakers," in *Proc. Int. Conf. Image Processing*, Hong Kong, China, Oct. 2010.

[27] S. Bozonnet, N. Evans, and C. Fredouille, "The LIA-EURECOM RT'09 speaker diarization system: Enhancements in speaker modeling and cluster purification," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Dallas, TX, USA, Mar. 2010.

[28] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1993.

[29] T. Backstrom and C. Magi, "Properties of line spectrum pair polynomials—A review," *Signal Process.*, vol. 86, no. 11, pp. 3286–3298, 2006.

[30] G. Jaffré and P. Joly, "Costume: A new feature for automatic video content indexing," in *Proc. Int. Conf. Adaptivity, Personalization and Fusion of Heterogeneous Information*, Avignon, France, Apr. 2004.

[31] M.-Y. Chen and A. Hauptmann, "Searching for a specific person in broadcast news video," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Montreal, QC, Canada, May 2004.

[32] S. Clements, *Show Runner: Producing Variety and Talk Shows for Television*. Los Angeles, CA, USA: Silman-James Press, 2004.

[33] L. Brown and E. Newman, *Your Public Best: The Complete Guide to Making Successful Public Appearances in the Meeting Room, on the Platform, and on TV*. New York, NY, USA: Newmarket Press, 2002.

[34] P. Viola and M. Jones, "Robust real-time object detection," in *Proc. Int. Workshop Statistical and Computational Theories of Vision-Modeling, Learning, Computing and Sampling*, Vancouver, BC, Canada, Jul. 2001.

[35] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. Sebastopol, CA, USA: O'Reilly Media, 2008.

[36] P. Besson, V. Popovici, J.-M. Vesin, J.-P. Thiran, and M. Kunt, "Extraction of audio features specific to speech production for multimodal speaker detection," *IEEE Trans. Multimedia*, vol. 10, no. 1, pp. 63–73, 2008.

[37] C. Zhang, P. Yin, Y. Rui, R. Cutler, P. Viola, X. Sun, N. Pinto, and Z. Zhang, "Boosting-based multimodal speaker detection for distributed meeting videos," *IEEE Trans. Multimedia*, vol. 10, no. 8, pp. 1541–1552, 2008.

[38] M. Everingham, J. Sivic, and A. Zisserman, "Hello! my name is… Buffy—Automatic naming of characters in TV video," in *Proc. British Machine Vision Conf.*, Edinburgh, U.K., Sep. 2006.

[39] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artificial Intelligence*, Vancouver, BC, Canada, Aug. 1981.

[40] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, no. 1, pp. 57–79, 2005.

[41] G. W. Snedecor and W. G. Cochran, *Statistical Methods*. Ames, IA, USA: Iowa State Univ. Press, 1967.

[42] S. K. Zhou and R. Chellappa, "From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel Hilbert space," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 917–929, 2006.

[43] S. Essid, G. Richard, and B. David, "Instrument recognition in polyphonic music based on automatic taxonomies," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 14, no. 1, pp. 68–80, Jan. 2006.

[44] O. Gillet, S. Essid, and G. Richard, "On the correlation of automatic audio and visual segmentations of music videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 347–355, 2007.

[45] B. Schölkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[46] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
[47] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, 2004.
[48] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. Computer Vision and Pattern Recognition*, Anchorage, AK, USA, Jun. 2008.
[49] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.
[50] H. Bredin and G. Chollet, "Audiovisual speech synchrony measure: Application to biometrics," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 179–185, 2007.
[51] D. Li, N. Dimitrova, M. Li, and I. Sethi, "Multimedia content processing through cross-modal association," in *Proc. ACM Int. Conf. Multimedia*, Berkeley, CA, USA, Nov. 2003.

Félicien Vallet is a research engineer at the Institut National de l'Audiovisuel (INA). He received the state engineering degree and the Ph.D. degree from Telecom ParisTech (the École Nationale Supérieure des Télécommunications), Paris, France, in 2007 and in 2011, respectively. His Ph.D. thesis was completed jointly between The Institut National de l'Audiovisuel and Telecom ParisTech in the field of Automatic TV talk-shows structuring. He has been involved in several national and European research projects. His research interests mainly focus on multimodal analysis and automatic structuring of video content.

Slim Essid is an Associate Professor at the Department of Image and Signal Processing (TSI) of Telecom ParisTech with the Audio, Acoustics and Waves group. He received the State Engineering degree from the École Nationale d'Ingénieurs de Tunis, Tunisia, in 2001, the M.Sc. (D.E.A.) degree in digital communication systems from the École Nationale Supérieure des Télécommunications (ENST), Paris, France, in 2002, and the Ph.D. degree from the Université Pierre et Marie Curie, in 2005, after completing a thesis on automatic audio classification. He has published over 60 peer-reviewed conference and journal papers with more than 50 distinct co-authors, and has served as a program committee member or as a reviewer for various audio and multimedia conferences and journals. His research interests are in machine learning for multimodal signal processing with applications in music and multimedia content analysis, and human activity/behavior analysis. He has been actively involved in various French and European funded research projects.

Jean Carrive received a Ph.D. in Computer Science at Paris 6 University in 2000 in collaboration with INA. The subject of the thesis was Classification of Audiovisual Sequences. Since then, he is in charge of research at the Research Department of INA in the field of Analysis and Documentation of Audiovisual Documents. He has been involved in several national and European projects. The Institut National de l'Audiovisuel or INA is the first audiovisual archive center in the world and the first digital image databank in Europe. The main missions of INA are to preserve the national audiovisual heritage, to make it more available and to keep abreast of changes in the audiovisual sector through its research, production and training activities. The research work at INA is mainly directed towards finalization of digital tools in the field of restoration and indexing of audiovisual documents. INA archives more than 1.110.000 hours of radio and TV broadcasts (60 years of radio and 50 years of TV). Overall, about 144.000 hours are on-line for professional users. Because of cable and satellite TV, the archive will soon grow by 500.000 hours each year.