Multimodal speaker diarization utilizing face clustering information

Ioannis Kapsouras, Anastasios Tefas, Nikos Nikolaidis, and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki
54124 Thessaloniki, Greece
jkapsouras@aiia.csd.auth.gr

Abstract. Multimodal clustering/diarization tries to answer the question "who spoke when" by using audio and visual information. Diarization consists of two steps, at first segmentation of the audio information and detection of the speech segments and then clustering of the speech segments to group the speakers. This task has been mainly studied on audiovisual data from meetings, news broadcasts or talk shows. In this paper, we use visual information to aid speaker clustering. We tested the proposed method in three full length movies, i.e. a scenario much more difficult than the ones used so far, where there is no certainty that speech segments and video appearances of actors will always overlap. The results proved that the visual information can improve the speaker clustering accuracy and hence the diarization process.

Keywords: Multiomodal · Diarization · Clustering · Movies

1 Introduction

Speaker diarization/clustering tries to detect speech segments and cluster similar segments in order to group together segments of the same speakers. Diarization can automatically answer the question "who spoke when" when used together with speaker recognition systems, by providing the speakers true identity. Alternatively the same question can be answered in a semi-automatic way by combining diarization with the manual labelling of the speaker clusters with their true identity. Speaker diarization is a process that automatically produces semantic information from audio data.

Movies contain both audio and video information. Usually, video and audio are considered as different modalities and are analysed separately. In this paper, the combination of the two modalities (audio and video) for the task of speaker clustering/diarization is investigated. The intuition behind the modalities fusion is that one can perform a similar to speaker diarization analysis upon the visual data: face clustering. In more detail, assume that faces are detected in the frames of a movie and then the detected faces are tracked over time, resulting in a number of video facial trajectories [13], [2], [5]. A representative face is selected to represent a facial trajectory. The selected faces can then be clustered into clusters, each ideally corresponding to a single actor/person. The face clustering results and/or video information can be used in the audio based diarization process in order to improve the speaker clustering accuracy.

In the proposed method, audio speech segments and video facial trajectories were used as high level features to improve speaker clustering. In more detail, the similarity of two speech segments was increased when these segments have overlap with visual appearances of the same actor and was decreased otherwise.

The proposed multimodal approach was tested in 3D feature length movies. Multimodal analysis of movies content has certain inherent difficulties since unlike meetings or talk shows audio (speech) and video are often not coherent in movies, for example the person depicted in the video might not be the one that is speaking.

2 Previous Work

Multimodal speaker diarization, which is closely related to multimodal person clustering has already been studied in the literature, but mainly on audiovisual data from meetings or talk shows, which impose far less difficulties than movie content. The video information can enhance the audio information during the speaker diarization, hence a multimodal approach to diarization (audio + video) can improve the answer in the diarization question "who spoke when". In [4], Khoury et al. proposed a framework for audio-visual diarization. The authors combined audiovisual information using co-occurrence matrices. Moreover, they used information, such as face size and lip activity rates to improve the audiovisual association. Their method improves all audio, video and audiovisual diarization. The authors evaluated their method in a number of news videos, meetings videos and movies. In [8], Noulas et al. proposed a probabilistic framework to perform multimodal speaker diarization. The proposed method uses a Dynamic Bayesian Network (DBN) to model the people as multimodal entities that are involved in audio and video streams and also in audiovisual space. The model is generated by using the Expectation Maximization algorithm. The proposed DBN, also called factorial HMM, can be treated as an audiovisual framework. The factorial HMM arises by forming a dynamic Bayesian belief network composed of several layers. Each of the layers has independent dynamics, but the final observation vector depends upon the state in each layer. Their method was tested in meetings and news videos. Multimodal speaker diarization is also addressed by Friedland et al. in [6]. The method combines audio and video low level features, by using agglomerative clustering, where GMMs are used to model the clusters. The method proposed by [6] was tested on meetings video.

To our knowledge, there are no methods that use multimodal information for speaker diarization on 3D video and multichannel audio data. Moreover, the task of diarization is much easier, when the input data are from meetings or talk shows. In such setups, the visual appearance of a speaker (i.e., its clothing or facial appearance) does not change within the duration of the meeting/show. The composition of the group of participating persons typically does not change either. For talk shows, one can further assume that the speaker is in a closeup view. Moreover, the possibility that the speaker is the person that is shown (actor) is very high. These observations do not apply in the case of 3D films or films in general. In this case, the speaker/actor visual appearance may change over the duration of the movie. Furthermore, the group of people may change over time. Finally, the coherence between visual and audio scene is not guaranteed, since, for example, 3DTV video and audio scenes often capture only a part of the real scene (there may be people speaking that are not displayed or displayed people may speak but one may hear the voice of somebody else). Due to the above, the situation is much less constrained in the case of 3DTV content and person identification is more difficult than in the previously discussed setups.

3 Method Description

The proposed methods use information derived from video to improve the speaker clustering. In order to combine video and audio information, video facial trajectories and speech segments were used. Video trajectories are series of facial images in consecutive frames (usually depicting the same person) and speech segments are segments where speech has been detected in the audio channel of a movie.

3.1 Audio processing

The first step in speaker diarization is speech detection. Speech segments are detected and subsequently segmented in the audio channel of a video. Finally speaker clustering is performed in order to group together speech segments in clusters that are homogeneous. Each cluster should ideally correspond to a single speaker. In more detail the three steps of the speaker diarization approach used in this paper are:

- **Speech detection**: using Mel Frequency Cepstral Coefficients (MFCC) features and SVM classifiers
- **Change point detection**: in order to further segment the speech segments to homogeneous parts.
- **Spectral clustering**: to group speech segments that belong to the same speaker.

Features are extracted from the segmented speech segments. The audio features used were the Mel Frequency Cepstral Coefficients (MFCC) and the Spectral Flatness Measures (SFM). In speaker diarization, the standard score is based on the Bayesian Information Criterion (BIC) using Gaussian models [3]. A distance matrix $\mathbf{D}$ of dimensions $N \times N$ is derived using the MFCC features and the BIC criterion where N is the number of the audio segments. A novel variant [10] of the spectral clustering proposed in [7] was used for clustering.

3.2 Video processing

The first step in video processing is face detection and tracking. Faces are detected using [11] and tracked in the video channel or channels (in the case of 3D videos) of a movie using the algorithm proposed in [14]. In more detail, each detected face is tracked for K frames. A series of tracked images of a detected face form a facial trajectory (figure 1). Each facial trajectory is represented by any of the images included in it and these trajectories are clustered by using their representative images. It is obvious that all faces included in a trajectory belong to the same actor unless tracking error occur. Local Binary Patterns (LBP) [9] were used as features to represent the facial images.

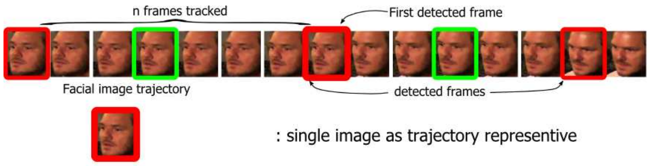

Fig. 1. Face trajectory computed by using face detection and tracking.

Calculating LBPs for all pixels of an image is not the best solution neither in terms of effectiveness nor in terms of calculation time. In our case we have chosen to calculate LBPs only in pixels that carry important information (i.e. mouth, eyes, etc.), thus two passes of fiducial points detectors were used. The first one is for the calculation of 66 fiducial points, such as outline of eyes, eyebrows, mouth etc, [1] and the second one [12] for better localization of these points. Moreover, these fiducial points are used in order to scale and align the detected images. LBPs are calculated upon patches around these 66 aligned points. Final, a histogram with K bins is calculated for each of these features. By this way a descriptor of dimension $66 \times K$ is calculated for each image.

In order to perform face clustering, similarities between each pair of images (each image representing a facial trajectory) have to be computed. The χ^2 distance was used to calculate the distances between two corresponding LBP histograms on a pair of images i, j and the final d_{ij} distance value was computed as the sum of the 66 distances (one per histogram). The similarity between the two images was calculated as $1/d_{ij}$ and a similarity matrix $\mathbf{V}$ between facial images (or more precisely facial image trajectories) was computed. Finally the clustering method in [10] is used to perform face clustering by utilizing $\mathbf{V}$. The result of face clustering can be used to improve speaker clustering (Section 3.3).

3.3 Multimodal approach

As can be seen in Sections 3.1 and 3.2, speaker clustering and face clustering group the speakers and the actors in the audio and visual data of a movie respectively. The speakers and the (visible) actors of a movie are in general the same people (people that speak in a movie, usually appear in it also), thus face clustering can improve most probably the speaker clustering i.e. the diarization process.

The input of the algorithm used for multimodal speaker clustering is the similarity matrix of the audio segments. The main idea is to a) increase the similarity of two speech segments, when these segments overlap with visual appearances of the same actor or b) decrease the similarity value, if no such overlap exists. The matrix derived by audio features (Section 3.1) is actually a speech segments distance/dissimilarity matrix, i.e., has small values when two speech segments are similar and high values otherwise. Therefore, the first step towards combining audio and video information was to transform this matrix to a similarity matrix **S**. This was done by using a sigmoid function:

$$S_{i,j} = \frac{1}{1 + exp(4 * (D_{i,j} - \bar{D})/\sigma)}, \tag{1}$$

where $D_{i,j}$ is an element of the distance matrix $\mathbf{D}$, $\bar{D}$ the mean value of $\mathbf{D}$ and σ the standard deviation of $\mathbf{D}$. To combine information from video and audio, in order to enhance speaker clustering using video, a new matrix $\mathbf{Q}$ is created with dimensions equal to those of the speech similarity matrix $\mathbf{S}$. The next step is to find, for each element (i, j) of the matrix $\mathbf{Q}$, the video trajectories that overlap in time with the speech segments that correspond to this element. Then, if the same actor appears in the corresponding video trajectories, the corresponding element of $\mathbf{Q}$ is increased, otherwise it is decreased. The final similarity matrix $\mathbf{F}$ is formed by combining the speech similarity matrix $\mathbf{S}$ and matrix $\mathbf{Q}$:

$$\mathbf{F} = \mathbf{S} + \alpha\mathbf{Q},\ 0 \leq \alpha \leq 1. \tag{2}$$

Two different approaches were implemented and tested (see Section 4), in order to create the matrix $\mathbf{Q}$, i.e., to change the elements of $\mathbf{Q}$ that correspond to speech segments which overlap with video trajectories. In the first approach, the ground truth for the actors depicted in the video trajectories was used, in order to check performance when the face clustering is perfect, i.e., it contains no errors. In more detail, for each pair of audio segments, the overlapping facial trajectories are found and, if the same actor appears in these trajectories according to the ground truth information, then the value in the corresponding element in matrix $\mathbf{Q}$ is multiplied with q where $q > 1$, otherwise it is multiplied with $1/q$. In the second more realistic approach, the same procedure is used, but instead of using the ground truth for the actors, the results of the face clustering algorithm are used. In other words, the results of face clustering described in Section 3.2 are used to check if the same actor appears in the overlapping facial trajectories.

Finally, after the calculation of matrix $\mathbf{F}$ using (2), the clustering algorithm in [10] is used for speaker clustering.

4 Experimental Results

The evaluation of the proposed multimodal speaker clustering approach was made by using a modified F-measure. F-measure punishes the erroneous split of a class into 2 parts quite strictly. In the modified version of F-measure used in this paper, overclustering is performed by creating more than the needed clusters and then the clusters that correspond to the same speaker are merged. The final F-measure is evaluated upon this merged clustering result. By this way, F-measure becomes less strict in the evaluation of splitted classes and evaluates more the purity of clusters.

The proposed approach was tested in three full length 3D feature films of different duration, size of cast and genre. These movies were selected in order to test the proposed approach in a difficult and realistic scenario. Stereo information of the video channels was exploited in two ways. Face detection [11] was applied on both channels (left and right), mismatches between the two channels were rejected and a stereo tracking algorithm [14] was applied in both channels. By using the above approaches we end up with a number of facial trajectories, namely series of consecutive facial images. As stated in previous section, each of these trajectories is represented by a single facial image for each channel (Left-Right). For 2 trajectories represented each one by 1 (in case of a mismatch) or 2 images (in case of left-right channel) the following similarity is calculated:

$$Simlarity = max_{ij} LBP(x_i, x_j) x_i \in T_k, x_j \in T_m \tag{3}$$

where x_i is an image belonging to the T_k trajectory and x_j is an image belonging to T_m trajectory and $1 \leq k, m \leq 2$.

It should be noted that only a relatively small number of speech segments overlap with facial trajectories, which is a usual phenomenon in movies and makes multimodal diarization difficult in such content.

Experiments have been conducted to verify the performance of the proposed method. The results can be seen in Table 1 alongside with the performance of the clustering when only audio modality was taken into account. As can be seen in this Table, the use of video ground truth information for the actors depicted in each facial trajectory (Multimodal 1 column) improves the clustering performance by approximately 8% in every movie, in terms of the modified F-measure compared to audio only diarization. Since the ground truth was used, it can be deducted that this is the best possible improvement for speaker clustering by using the video information with the proposed approach. When using information derived from actual facial image clustering in video the improvement, as can be seen in Table 1 (column Multimodal 2), is approximately 5%.

Table 1. Speaker Clustering F-measure, when video information is incorporated.

	Audio only	**Multimodal 1**	**Multimodal 2**
Movie 1	0.51	0.59	0.56
Movie 2	0.48	0.57	0.51
Movie 3	0.45	0.53	0.5

As can be seen from the experimental results, information derived from video data can help the audio-based speaker diarization. The increase in performance is higher when ground truth information is used, which leads to the obvious conclusion that the better the face clustering in the video, the better the effect of multimodal speaker clustering in the speaker diarization. It should be noted that, face clustering is not the only way to cluster the actors facial images in a video. Face recognition or label propagation can also be used to cluster the actors to groups and use this information for multimodal speaker clustering.

5 Conclusions

In this paper we proposed a method to improve speaker diarization through a multimodal approach. The improvement of speaker clustering can be done by using video information derived from video data through face clustering. Experiments in three full stereo movies have shown that multimodal speaker clustering achieves better results that single modality speaker clustering.

Acknowledgment

The authors wish to thank A. Roebel, L. Benaroya and G. Peeters (IRCAM, France) for providing speech analysis data (speech segments similarity matrices) used in this work.
The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 287674 (3DTVS). This publication reflects only the authors views. The European Union is not liable for any use that may be made of the information contained therein.

References

1. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3444–3451 (2013)
2. Baltzakis, H., Argyros, A., Lourakis, M., Trahanias, P.: Tracking of human hands and faces through probabilistic fusion of multiple visual cues. In: Proceedings of the 6th International Conference on Computer Vision Systems, ICVS'08, pp. 33–42. Springer-Verlag, Berlin, Heidelberg (2008)
3. Chen, S., Gopalakrishnan, P.: Speaker, environment and channel change detection and clustering via the bayesian information criterion. In: Proceedings of DARPA Broadcast News Transcription and Understanding Workshop (1998)
4. El Khoury, E., Snac, C., Joly, P.: Audiovisual diarization of people in video content. Multimedia Tools and Applications **68**(3), 747–775 (2014)
5. Elmansori, M.M., Omar, K.: An enhanced face detection method using skin color and back-propagation neural network. European Journal of Scientific Research **55**(1), 80 (2011)
6. Friedland, G., Hung, H., Yeo, C.: Multi-modal speaker diarization of real-world meetings using compressed-domain video features. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009., pp. 4069–4072 (2009)

7. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Proceedings of NIPS, pp. 849–856. MIT Press (2001)
8. Noulas, A., Englebienne, G., Krose, B.: Multimodal speaker diarization. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(1), 79–93 (2012)
9. Ojala, T., Pietikainen, M., Harwood, D.: Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In: Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International Conference on, vol. 1, pp. 582–585 vol.1 (1994)
10. Orfanidis, G., Tefas, A., Nikolaidis, N., Pitas, I.: Facial image clustering in stereo videos using local binary patterns and double spectral analysis. In: IEEE Symposium Series on Computational Intelligence (SSCI) (2014)
11. Stamou, G., Krinidis, M., Nikolaidis, N., Pitas, I.: A monocular system for person tracking: Implementation and testing. Journal on Multimodal User Interfaces **1**(2), 31–47 (2007)
12. Uricar, M., Franc, V., Hlavc, V.: Detector of facial landmarks learned by the structured output svm. In: Proceedings of VISAPP 2012, pp. 547–556 (2012)
13. Zoidi, O., Nikolaidis N.and Tefas, A., Pitas, I.: Stereo object tracking with fusion of texture, color and disparity information. Signal Processing: Image Communication **29**(5), 573 – 589 (2014)
14. Zoidi, O., Nikolaidis, N., Pitas, I.: Appearance based object tracking in stereo sequences. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)., pp. 2434–2438 (2013)