Audio-Video Speaker Diarization for Unsupervised Speaker and Face Model Creation

Pavel Campr[1], Marie Kunešová[1], Jan Vaněk[1], Jan Čech[2], and Josef Psutka[1]

[1] University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics
Univerzitni 8, 306 14 Plzen, Czech Republic
{campr,mkunes,vanekyj,psutka}@kky.zcu.cz
[2] Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Cybernetics, Center for Machine Perception
Technicka 2, 166 27 Praha 6, Czech Republic
cechj@cmp.felk.cvut.cz

Abstract. Our goal is to create speaker models in audio domain and face models in video domain from a set of videos in an unsupervised manner. Such models can be used later for speaker identification in audio domain (answering the question "Who was speaking and when") and/or for face recognition ("Who was seen and when") for given videos that contain speaking persons. The proposed system is based on an audio-video diarization system that tries to resolve the disadvantages of the individual modalities. Experiments on broadcasts of Czech parliament meetings show that the proposed combination of individual audio and video diarization systems yields an improvement of the diarization error rate (DER).

Keywords: audio-video speaker diarization, audio speaker recognition, face recognition

1 Introduction

With the increasing amount of multimedia data it is necessary to develop techniques that detect the presence of people and find out their identities. Such information can be used for indexing and searching purposes, for enhancement of automatic speech recognition systems, for building audio or video identification systems or for building audio or video corpora.

Our main goal is to create speaker models in the audio domain and face models in the video domain, in an unsupervised manner, so that the models can be used later for audio-video person identification. Contrary to most other existing systems [3,4] our method produces a slightly larger number of speaker-model candidates than the real number of speakers. The reason is that it is more important that no two different speakers are assigned the same identity than that each speaker is only assigned one. The former error has a negative impact on the performance of the whole system and can never be corrected, while the latter can be discovered and resolved in later processing by automatic or manual verification.

The paper is organized as follows. Section 2 describes the diarization system in audio-only domain, Section 3 in video-only domain. Combined audio-video diarization is described in Section 4 and evaluated in Section 5.

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 465–472, 2014.

2 Audio Speaker Diarization

For audio diarization we use an approach based on Gaussian Mixture Models (GMMs). The system is largely based on the ones proposed in [5] and [6]. However, there are some differences.

Most notably, as this particular application does not require the diarization to be done online, we perform additional offline clustering after all audio files have been processed, to identify and merge speaker-model candidates corresponding to the same speakers, both within a single file and between different ones. Unlike the authors of [5], we also use energy-based speech activity detection as opposed to model-based.

At the beginning, the system starts with only two GMMs, one for each gender, which are trained in advance. Afterwards, for every speech segment, the system decides if the segment corresponds to an already known speaker, or a new one. In the case of a new speaker, a new model is created by copying one of the gender dependent models. Otherwise, one of the existing models is selected. The assigned model is then adapted using the data from the segment.

The system for audio speaker diarization consists of several modules:

1. Feature extraction and voice activity detection
2. Speech segmentation
3. Speaker identification and novelty detection
4. Online GMM learning
5. Offline clustering

2.1 Feature Extraction and Voice Activity Detection

For feature extraction, we used the LFCC with 25 filters in range from 50 Hz to 8 kHz based on 25ms FFT window with 10ms shift. 20 cepstral coefficients were computed without the energy coefficient. No cepstral normalization was performed.

This module also performs an energy-based voice activity detection (VAD), with every frame being labeled as *speech* or *silence* based on a threshold.

2.2 Speech Segmentation

Using the information obtained from the VAD and parameters such as the minimum and maximum segment length and the maximum pause length in a segment, the speech is divided into short segments. Of each segment, only the frames labeled as *speech* are used for the novelty detection and GMM learning modules. In our experiments, this has lead to both reduced computation time and improved performance.

2.3 Speaker Identification and Novelty Detection

For each speech segment the system uses a maximum-likelihood classification to determine both the speaker's gender (using the gender dependent models) and their most likely identity out of the existing speaker-model candidates. Afterwards, a likelihood

ratio test is used to decide whether the segment belongs to the chosen identity, or represents an entirely new speaker.

The likelihood ratio is as follows:

$$L(X) = \frac{P_{sp}}{P_{gen}} , \tag{1}$$

where X is a speech segment and P_{sp} and P_{gen} are the likelihoods of the winning speaker and the appropriate gender dependent model, respectively.

If $L(X) \geq \theta$, the segment X belongs to the old speaker. Otherwise it belongs to an entirely new speaker.

The optimal value of decision threshold θ was found experimentally, and chosen in such a way that the system produces a slightly larger number of speaker models than the real number of speakers. The reason for this is that it is more important to us in this stage that no two different speakers are assigned the same identity than that each speaker is only assigned one. The former error has a negative impact on the performance of the whole system and can never be corrected, while the latter can be discovered and resolved in later stages, either during clustering or once the results from both audio and video are combined.

2.4 Online GMM Learning

For GMM adaptation we use an online variant of the Expectation-Maximization algorithm, as described in [7], with values of the parameters as proposed by [5].

2.5 Audio Clustering

After we have processed all audio recordings, we perform clustering. The main purpose is to find the labels corresponding to the same real speakers between different recordings, although the system also resolves most cases of multiple labels being assigned to the same real speaker within a single recording.

For every pair of speaker models λ_i and λ_j, we calculate the value of the following expression, which is similar to the Cross-Likelihood Ratio [8]:

$$L(i, j) = \frac{1}{N_i} \cdot \log \left(\frac{p(X_i|\lambda_j)}{\max(p(X_i|\lambda_m), p(X_i|\lambda_f))} \right) , \tag{2}$$

where X_i represents all the speech data assigned to the speaker model λ_i, N_i is the number of frames in X_i, and λ_m and λ_f are the gender dependent models.

If both $L(i, j)$ and $L(j, i)$ exceed a certain threshold, we consider λ_i and λ_j to be the same speaker. In order for more than two models to be merged, the condition must be fulfilled for every two of them.

3 Video Speaker Diarization

The goal of this module is to detect and track faces in a video, to extract features from each face image and to perform clustering. The result is a set of clustered face tracks, each cluster representing one identity.

The task of face detection, tracking, and identification has been widely studied. Existing solutions can solve this task with high accuracy [1,2].

3.1 Face Tracking and Feature Extraction

In this paper we use a facial landmarks detector based on the Deformable Part Models [1]. In addition to the detected position of the face, this face detector provides a set of facial landmarks: nose, mouth and canthi corners. Such landmarks are used for the construction of normalized face images from which the face features are computed. The feature computation is based on Local Binary Patterns (LBP).

During the face tracking process, normalized face images are computed for the whole face track. For all normalized face images in the face track, the distances from all the previous normalized images in the same face track are computed. If at least one distance is lower than a threshold θ_1, a similar face appearance was already seen and the image is ignored. As a result, each of the N_T face tracks is represented by a set of key face images and corresponding features λ^V.

3.2 Clustering

Clustering is performed after the processing of all video recordings. The purpose is to find the labels corresponding to the face of the same person between different face tracks, in all videos. The whole process is visualised in Figure 1.

For every pair of face tracks i and j, we compute their distances $D_T(i, j)$, based on the features λ_i^V and λ_j^V. The distance is computed as a min-min distance between the sets of features λ_i^V and λ_j^V. If their minimal distance is lower than a certain threshold

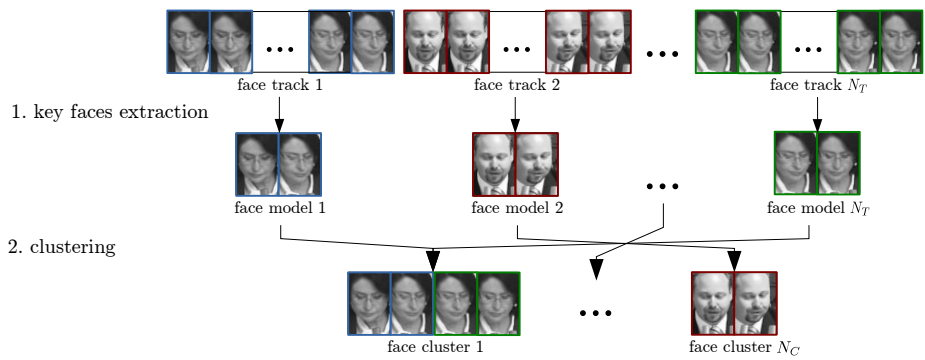

Fig. 1. Video diarization process.

θ_2, then both face tracks are considered to be of the same identity with the same label. In other words, if there are similar faces in the first face track i and the second track j, then the identity of both is considered to be the same.

As a result we have N_C clusters, each representing the face of one person. The threshold θ_2 was experimentally set so that we have multiple clusters with the same identity, but there is no cluster that represents multiple identities. The reason is that the additional merging of clusters is performed in later stage using audio modality.

4 Audio-Video Speaker Diarization

The combined audio-video speaker diarization system tries to resolve the disadvantages of particular modalities. The audio-only diarization requires longer intervals (usually several seconds) to produce a certain decision about the speaker's identity. For the faces in the video domain, the decision can be made in one frame only, but the appearance of one speaker's face can change in time greatly. When compared to audio, the speech style of one speaker changes only a little during the time.

The proposed system is built with the assumption that the speaker's face is present in the video most of the time during his speech. The generated audio and video speakers' models can be used later for any videos where this condition is not fulfilled.

4.1 Audio-Video Speaker Models Clustering

Only the best matching segments (models) were clustered in the previous stages of the processing where the audio and the video modalities were treated independently. Therefore, the number of obtained models was too high number. Further clustering cannot be done accurately according to individual modalities. However, fused audio-video merging is able to reduce the number of models with acceptable error rate. The clustering of the models was done in the following way:

1. The audio similarity matrix was based on the symmetric Kullback-Leibler divergence:

$$L_{KLD}(i,j) = \frac{1}{2}\Big[\mathcal{L}(X_i|\lambda_i) + \mathcal{L}(X_j|\lambda_j) - \mathcal{L}(X_i|\lambda_j) - \mathcal{L}(X_j|\lambda_i)\Big], \quad (3)$$

   where

$$\mathcal{L}(X_i|\lambda_j) = \frac{1}{N_i}\sum_i \log(p(X_i|\lambda_j)). \quad (4)$$

2. The video similarity matrix was based on a transformation of $D(i,j)$ values produced by the face-model distance function. The transformation was chosen to match the audio L_{KLD} similarity:

$$L_V(i,j) = \frac{D(i,j) - \alpha}{\beta}, \quad (5)$$

   where α and β are parameters tuned on a small development data set.

3. The final fused similarity is defined as

$$L_{AV}(i, j) = L_{KLD}(i, j) + L_V(i, j). \quad (6)$$

The evaluated $L_{AV}i, j$ values were compared with a threshold θ and the model pairs with a value higher than the threshold were clustered.

5 Experiments

For testing purposes we used recordings from Czech parliament meetings broadcasted by the Czech Television. A sample image from the broadcasts can be seen in Figure 2. We had 8 recordings with a total of 30 hours of labeled audio.

To evaluate the system we compared the audio-only and audio-video clustering results.

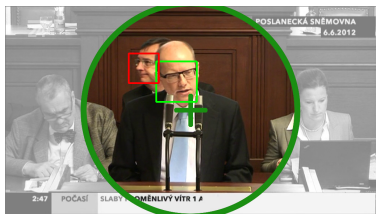

Fig. 2. Sample image from the Czech parliament meetings broadcasts. The two rectangles denote the detected faces.

5.1 Audio Clustering

Gender dependent models were trained using 30 seconds of speech from each of 16 women and 70 men.

In the experiments, we used GMMs with only 8 components and the minimum and maximum segment lengths were set to 1 and 5 seconds respectively.

In order for the diarization results to better correspond to the reference labels, we relabeled any short pauses between two consecutive speech segments which belong to the same speaker as speech as well. The optimal maximum length of such pauses was determined from the reference labels to be 3 seconds.

To evaluate the performance of the system, we used the diarization error rate (DER), which is the sum of three values: the rate of missed speech (i.e. the speech frames that were incorrectly labeled as silence), false alarm (FA, silence incorrectly labeled as speech) and speaker error (SE, the speech labeled as a wrong speaker). It is measured as the fraction of time that is not assigned correctly to a speaker or to non-speech [9].

Additionally, we also computed the speaker error rate and the DER when tolerating the use of multiple models for each real speaker (SE_{MM} and DER_{MM} in the table).

These values essentially represent the ideal error rates we would obtain by performing additional oracle clustering.

Average values from all recordings, obtained both before and after the clustering, are shown in Table 1. The values given were obtained without any forgiveness collar around reference segment boundaries.

Table 1. Audio diarization performance (%)

	miss	FA	SE	DER	SE_{MM}	DER_{MM}
before clustering	2.58	1.09	7.33	11.0	1.68	5.35
after clustering	2.25	1.11	4.15	7.51	1.85	5.22

The audio clustering has lead to a significant decrease of the speaker error, though it does not reach the ideal value represented by SE_{MM}. The slight increase of SE_{MM} suggests that we have also clustered a small number models which did not truly belong to the same speaker, while the decrease of DER_{MM} was caused by the decreased miss rate.

5.2 Audio-Video Clustering

The table 2 presents the results of audio-video clustering method. The previous results of audio-only clustering are compared to the audio-video clustering.

Table 2. Audio-video diarization performance

	DER (%)	number of speakers
audio clustering	7.51	381
audio-video clustering	7.18	292
truth	-	86

The results show that the DER decreased from 7.51% to 7.18% when the video modality is used. The number of resulting clusters decreased from 381 to 292. The merged clusters were either short or contained some noise and the audio-only clustering system was unable to merge these ambiguous clusters. Because the DER is measured as the fraction of time and we clustered mostly short segments, the decrease of DER is 4.5% relatively, although the decrease of number of clusters is 23.4% relatively.

As the face is not present in the video all the time, the proposed algorithm is not able to merge clusters where only audio modality is used and the segments are too short or ambiguous.

6 Conclusion

We addressed the problem of audio-visual speaker diarization. After the description of the baseline audio-only diarization system we presented our proposed method

for association of individual models from the audio and video modalities, all in an unsupervised manner. The method was evaluated on 30 hours of video. The diarization error rate (DER) decreased from 7.51% to 7.18% and the number of clusters decreased from 381 to 292, where the real number of speakers was 86. These results show that the short utterances which the audio-only diarization is unable to associate to a correct identity can be associated more reliably with the addition of video modality and face recognition.

Acknowledgments This research[1] was supported by the Grant Agency of the Czech Republic, project No. GAČR GBP103/12/G084.

References

1. Uřičář, M., Franc, V., Hlaváč, V.: Detector of Facial Landmarks Learned by the Structured Output SVM. In: VISAPP '12: Proceedings of the 7th International Conference on Computer Vision Theory and Applications, pp. 547–556, (2012).
2. Sonnenburg, S. and Franc, V.:COFFIN: A Computational Framework for Linear SVMs. In: Technical Report, Center for Machine Perception, Czech Technical University, Prague, Czech Republic, (2009).
3. Bendris, M., Charlet, D., Chollet, G.: People indexing in TV-content using lip-activity and unsupervised audio-visual identity verification. In: 9th International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 139–144, (2011).
4. El Khoury, E., Sénac, C., Joly, P.: Audiovisual diarization of people in video content. In: Multimedia Tools and Applications, (2012).
5. Markov, K., Nakamura, S.: Never-Ending Learning System for Online Speaker Diarization. In: IEEE Workshop on Automatic Speech Recognition & Understanding, 2007. ASRU, pp. 699–704, (2007).
6. Geiger, J., Wallhoff, F., Rigoll, G.: GMM-UBM based open-set online speaker diarization. In: INTERSPEECH 2010, pp. 2330–2333, (2010).
7. Sato, M., Ishii, S.: On-line EM algorithm for the Normalized Gaussian Network. In: Neural Computation, vol. 12, pp. 407–432, (2000).
8. Reynolds, D., Singer, E., Carlson, B., O'Leary J., McLaughlin, J., and Zissman, M.: Blind clustering of speech utterances based on speaker and language characteristics. In: Proceedings of the 5th International Conference on Spoken Language Processing, vol. 7, pp. 3193–-3196, (1998).
9. National Institute of Standards and Technology, `http://www.itl.nist.gov`

[1] The access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005) is highly appreciated.