IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

UPC Multimodal Speaker Diarization System for the 2018 Albayzin Challenge

Miquel India, Itziar Sagastiberri, Ponç Palau, Elisa Sayrol, Josep Ramon Morros, Javier Hernando

Universitat Politècnica de Catalunya (UPC)

miquel.india@tsc.upc.edu, itziar.sf@opendeusto.es, ppalaupuigdevall@gmail.com,
[elisa.sayrol, ramon.morros, javier.hernando]@upc.edu

Abstract

This paper presents the UPC system proposed for the Multimodal Speaker Diarization task of the 2018 Albayzin Challenge. This approach works by processing individually the speech and the image signal. In the speech domain, speaker diarization is performed using identity embeddings created by a triplet loss DNN[1] that uses i-vectors as input. The triplet DNN is trained with an additional regularization loss that minimizes the variance of both positive and negative distances. A sliding window is then used to compare speech segments with enrollment speaker targets using cosine distance between the embeddings. To detect identities from the face modality, a face detector followed by a face tracker has been used on the videos. For each cropped face a feature vector is obtained using a Deep Neural Network based on the ResNet 34 architecture, trained using a metric learning triplet loss (available from dlib library). For each track the face feature vector is obtained by averaging the features obtained for each one of the frames of that track. Then, this feature vector is compared with the features extracted from the images of the enrollment identities. The proposed system is evaluated on the RTVE2018 database.

Index Terms: Speaker Diarization, Face Diarization, Multimodal System

1. Introduction

Video broadcasting generates a massive amount of multimedia information that once archived generates a need to access its contents. In particular, it is essential to develop tools that are able to automatically search and detect the presence of people. Challenges such as REPERE [1] or the MediaEval Multimodal Person Discovery in Broadcast TV [2], [3] addressed the identification of people appearing and speaking.

Two main approaches can be found in the literature for retrieval of person identification in videos: Perhaps the most popular is based on clustering face tracks, speech segments or both [4, 5]. This provides multiple clusters (aggrupations of signal segments) that correspond to the identities in the video. Then, an assignment of names to clusters is performed. The main problem of this approach is that a large number of non important identities can appear, and that the clusters are highly non-homogeneous. Combining speech and face modalities is also a challenge in these systems. The second usual approach is verification on the signal segments [4, 6]. Here, two stages are defined: enrollment and search. Each segment is compared against the enrollment data and a decision is made for each segment. In several cases, both approaches (clustering and verification) use of some kind of metric learning to improve the discriminativeness of the feature vectors.

This paper presents the UPC proposal to the Albayzin Evaluation: IberSPEECH-RTVE 2018 Speaker Diarization Challenge. In this challenge a list of people occurrences within the RTVE2018 database should be provided as a result. This list must contain people either if they are talking, or their faces appear in the video, or both at the same time. The two modalities should therefore be provided. Identities and enrollment data are given, so the identification might be considered supervised. Nevertheless, in the evaluation data unknown persons may appear and should be distinguished from those that are known.

This paper is organized as follows: Section 2 describes the system that has been developed to detect identities from the image and the speech modalities. Section 3 gives the technical details of the setup and the experimental results on the provided database. Finally, conclusions and final remarks are given in section 4.

[1]Deep Neural Network

2. System Description

The UPC Multimodal Speaker Diarization System consists on two monomodal systems and a fusion block that combines the outputs of the previous systems so as to obtain a more refined speaker labelling. The speech and the image are processed in an independent manner with a speaker and a face diarization system. The face tracks and the speaker segments are then fused with an algorithm that combines the intersection of these sources according to a set of assumptions made on the data. The next section describes in detail the monomodal approaches and the fusion system to combine them.

2.1. Video System

The video system is responsible of localizing the faces of the individuals appearing in the scene and to determine if these faces belong to one of the N given identities.

Our approach is based on performing face tracking to identify the intervals where a given person is appearing in the video. A face track consists of the location of the faces of an individual in the consecutive frames where he/she appears in the video. Thus, the face track determines the spatial location of the faces and the temporal interval in the video where this person appears. Then, each face track is forwarded through a classifier with $N + 1$ classes, namely the identities of the *known* persons (this is, the set of persons in the enrollment set) plus the *unknown* class. The approach is summarized in Figure 1.

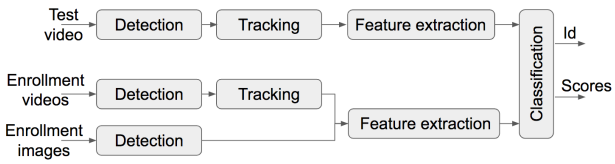

Figure 1: *Block diagram for the face modality.*

10.21437/IberSPEECH.2018-40

Our approach uses a tracking by detection approach: First, all the faces in the video sequence are detected. For this, we use a detector based on HOG+SVM[2] [7] from the dlib [8] library.

Once the faces have been detected, a KLT[3] tracker [9, 10, 11] is used to relate the detections in successive frames. We used the implementation[4] provided for the baseline system of the Multimodal Person Discovery in Broadcast TV task in MediaEval 2015 [2].

As mentioned previously, a face track provides the spatial location of a set of faces of a given individual, which are used for feature extraction, and the temporal interval where this person is visible in the video.

In the video system, the track is the basic unit of recognition: we will output a result for each track that is classified as belonging to one of the *known* persons. Tracks classified as *unknown* are discarded and no output is provided.

To characterize each track, we follow a two step process: first, a feature vector is extracted for each detected face in the track. Then, the final feature vector for the track is obtained by averaging all the track's feature vectors.

These feature vectors are obtained using the last fully connected layer from a Deep Neural Network based on the ResNet 34 architecture [12], trained using the metric learning triplet loss process described in FaceNet [13]. This learns a mapping from the detected faces to a compact space where the feature vectors (i.e. 128 dimensional FaceNet embeddings) originating from the faces of a given individual are located in a separate and compact region of the space. Thus, the vectors are highly discriminative, allowing to use standard techniques to perform classification/verification. We have used the off-the shelf *dlib* [8] implementation, without any adaptation nor fine-tuning to the task identities.

A similar method is used to extract the feature vectors for the images and videos of the enrollment set. For each person, 10 still images and one short clip were provided. For each still image, we detect the face and we extract a feature vector. The short video is processed similarly to the test video: scene detection, face detection and face tracking. A feature vector is extracted for each resulting track. This results in a variable number of enrollment vectors for each person, depending on the number of tracks in the short video. These vectors are associated with the name of the corresponding person and used as a person model.

To decide the track identity, we used a k-NN classifier with a cosine distance metric. A global threshold is applied to determine if the track belongs to any of the persons in the database. If this is the case, the identity corresponding to the nearest vector in the database is used as the track identity. This simple method is possible because the highly discriminative properties of the FaceNet 128 dimensional embeddings.

2.2. Speaker System

The speaker system works as a tracking algorithm that uses speaker embeddings to compare speech signal segments with the speech utterances of the enrollment identities. These representations are created with a DNN which is feed with i-vectors and is trainned with a multi-objective loss (Figure 2). This loss is based on a triplet margin loss and a regularitzation function which minimizes the variance of both positive and negative tuple distances.

[2]Histogram of Oriented Gradients + Support Vector Machine

[3]Kanade-Lucas-Tomasi tracker

[4]https://github.com/MediaevalPersonDiscoveryTask/Baseline2015

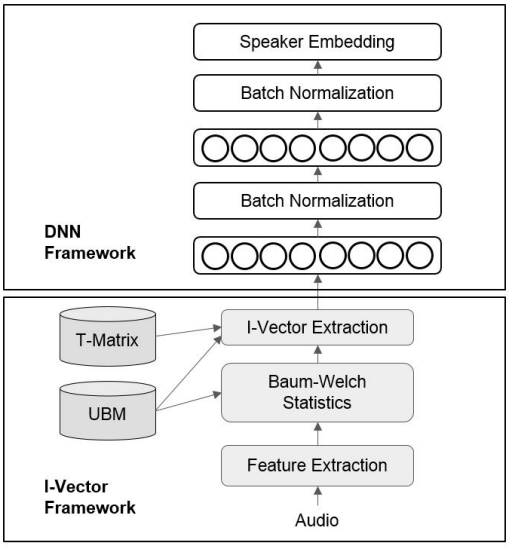

Figure 2: *Speaker front-end diagram.*

I-vectors are low rank vectors, typically between 400 and 600, representing a speech utterance. Given a speech signal, acoustic features like Mel Frequency Cepstral Coefficients (MFCC) are extracted. These feature vectors are modeled by a set of Gaussian Mixtures (GMM) adapted from a Universal Background Model (UBM). The mean vectors of the adapted GMM are stacked to build the M supervector, which can be written as:

$$M = \mu + T\omega \tag{1}$$

where μ is the speaker- and session-independent mean supervector from UBM, T is the total variability matrix, and ω is a hidden variable. The mean of the posterior distribution of ω is referred to as i-vector. This posterior distribution is conditioned on the Baum-Welch statistics of the given speech utterance. The T matrix is trained using the Expectation-Maximization (EM) algorithm given the centralized Baum-Welch statistics from background speech utterances. More details can be found in [14].

Given an i-vector, a DNN is used to extract a more discriminative speaker vector. This DNN is composed by 2 hidden layers of 400 nodes, where the activations of the second layer are used as a speaker embedding. This neural network is fed with i-vectors and an initial L2 normalization is applied to these inputs before the first hidden layer. After each hidden layer, a batch normalization layer is used as regularizer. Initially, the DNN is pretrainned as a speaker classifier. Therefore, a softmax layer is added in the output of the network and the DNN is trainned minimizing the cross-entropy loss. Following to this pretrainning, the softmax layer is removed and the DNN is trainned with the following multiple objective loss:

$$Loss = \frac{1}{N}\sum_{i=1}^{N} TLoss_i + \frac{\lambda}{2} RLoss_i \tag{2}$$

$$TLoss_i = \max(0, d(A_i, P_i) - d(A_i, N_i) + margin) \tag{3}$$

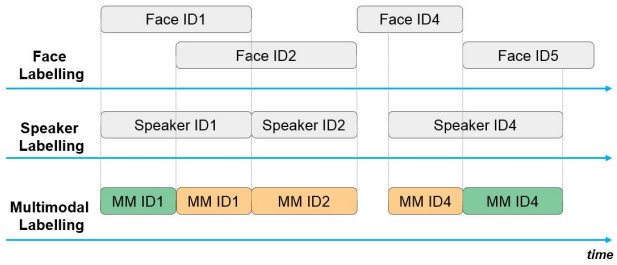

Figure 3: *Fusion scheme. Green boxes refer to the segments where id assignation has been directly propagated. Orange boxes refers to segments which id ask has been assigned after the score combination.*

$$RLoss_i = (|d(A_i, P_i) - \frac{1}{N}\sum_{j=1}^{N}|d(A_j, P_j)|)+$$
$$(|d(A_i, N_i) - \frac{1}{N}\sum_{j=1}^{N}|d(A_j, N_j)|) \quad (4)$$

where $TLoss$ corresponds to the triplet margin loss [13] and $RLoss$ corresponds to our proposed regularization function. $TLoss$ is computed with hinge loss and $d(x, y)$ refers to the 2_{nd} order euclidean distance between a pair of vectors. On the other hand, $RLoss$ is a function that forces the DNN to minimize the variance of both positive and negative tuples distances. Hence, in each batch we estimate the means of the positive and negative scores. These means are then used to minimize the distance between each positive or negative pair distance and its corresponding mean. We add a λ penalty term so as to balance the magnitude of the regularization function in the global loss.

The i-vector framework combined with the DNN is used as a front-end block in the speaker tracking system. This front-end allows to extract features of the speech signal and compare it with the signals of the speaker targets. In our approach, a sliding window strategy have been used to extract speaker embeddings from 3 seconds length speech segments with a 0.25 seconds shift size. For the enrollment identities, we have used the whole signal to extract an embedding for each target. Cosine distance metric is then used to evaluate the similarity of speech segment embeddings for each target. The target with the biggest similarity is then assigned to the corresponding speech segment. In order to classify the non-interest or unknown speakers, a threshold is imposed to determine the assignation between the best candidate and the speech segment. If the most similar target distance is below the threshold, the speech segment is automatically tagged as an unknown identity.

2.3. Fusion System

A fusion system has been considered in order to combine the previous information sources. Speaker and video diarization are performed first in an individual manner. The results of both modalities are then fused so as to obtain a better speaker assignation. In order to combine both outputs properly, we made the following assumptions:

- Speaker segments and face tracks of the same person are temporarily correlated. Hence, it is very likely that the person who appears in the video is the one who is speaking.
- Some speakers do not come into view any time in the show and there are other people who are shown in the screen but do not speak. These faces and speakers correspond in major part to the unknown identities.

According to these assumptions, an algorithm has been designed based in weighting temporal overlaps between the tracks of the face system and the speech segments of the speaker system (Figure 3). As its shown in the figure, the intersection between face tracks and speaker segments produces a new multimodal segmentation. The temporary segments where face and speech are not overlapped are discarded. We use this new segmentation to combine the assignations of both modalities:

- The segments where the corresponding face/speaker segments have the same target assigned are automatically tagged with that identity.
- When the speaker and face assignations are not the same, we produce a new scoring combining both modalities distances between the segment and the enrollment targets. First we extract the scores of the multimodal segment for each modality. The range of these scores are different for each source, hence it is needed to normalize them. This normalization is produced with a softmax activation which has a different temperature τ parameter for each modality. A new set of scores is then produced with the average of both modalities scores. Given these new multimodal scores, a new threshold is used to determine whether the segment correspond to the most similar target or to an unknown identity.

3. Optimization and Experimental Results

In the following section we describe the setup of the proposed approaches and we present the results of these systems for the Multimodal Speaker Diarization task of the 2018 Albayzin Challenge.

3.1. Speaker System

The speaker front-end block has been trained on the *VoxCeleb2* [15] database. Feature extraction is performed with 20 size MFCC plus delta features. The UBM has been trained with a 1024 mixtures GMM and the T Matrix size is 400. For the whole i-vector framework we have used the *Alize* [16, 17] toolkit and we have only used the first 1000 speakers of *VoxCeleb2* development partition. The DNN used is composed by two 400 size hidden layers. The pretraining has been performed using the same data used for the i-vector framework. For the triplet based DNN training, the whole *VoxCeleb2* development partition have been used. In order to obtain a good estimation of the positive and negative pair means, batch size have been set to 1024. The λ for the $RLoss$ have been set to 1. Both network trainnings have been performed with Adam optimizer. Learning rate have been set to 0.01 and the pretraining has been regularized with an additional 0.001 weight decay. For the target assignation, the decision threshold has been tuned to improve DER results on the RTVE2018 development set. A final value of 0.08 threshold over the the cosine distance (in range [-1,1]) has been obtained.

3.2. Video system

The method described in Section 2.1 has been used to obtain the results. We have filtered short tracks (tracks shorter than 1s) because they are likely to belong to non-important faces.

System	Miss	FA	SER	DER
Monomodal Speaker	3.5%	5.7%	31.9%	41.13%
Monomodal Face	37.9%	0.5%	1.9%	40.24%
Fusion (Spk. Eval.)	26.6%	2.3%	38.2%	66.99%
Fusion (Face Eval.)	51.7%	0.3%	26.9%	78.92%

Table 1: *DER results on the development partition.*

This also allows to reduce the computational load of the system. For each track, a 128D feature vector has been generated. The final identity decision is determined by a k-NN classifier. As the number of enrollment vectors is low, a value of $k = 1$ has been used. By looking at the small Speaker Error Rate value in Table 1, this approach is effective, thanks to the discriminating power of the embeddings. The principal challenge in this task was the high number of tracks belonging to persons that are not in the enrollment set. To reject these tracks, a global threshold th has been used. This threshold has been determined as the value providing the highest DER measure in the development set. A final value of $th = 0.47$ over the cosine distance (in range $0 - 1$) with the nearest neighbor has been obtained.

3.3. Fusion system

Given the scores between signal speaker/face segments and the target vectors, a softmax activation have been used to normalize the scores of each modality. In order to obtain similar scores, the softmax of each modality has been applied with a different temperature τ parameter. For speech $\tau = 3$ has been used and for the face modality τ has been set to 2. For the fusion system the target/non-target distance threshold have been set to 0.03.

3.4. Results

The proposed systems have been evaluated in the RTVE2018 database for the Multimodal Speaker Diarization task of the 2018 Albayzin Challenge. The development partition is composed of one video, with a duration of around 2 hours. Enrollment data (10 still images and a short video) is provided for a total of 34 identities. The test partition is composed of three test videos, with a total duration of around 4 hours with enrollment data for 39 identities. The metric used to evaluate the systems is the Diarization Error Rate (DER), which is the sum of three different errors: Miss Speech (MISS), False Alarm (FA) and Speaker Error Rate (SER). In this challenge, the presented approeaches are evaluated individually in each modality. Hence, it is needed to produce a diarization result for both speaker and face sources.

Table 1 shows the results of the presented approaches on the development partition. The first two rows correspond to the face and speaker system evaluated with their corresponding face/speaker groundtruth. Fusion system corresponds to the combination approach described in *Section* 2.3. Therefore, the third and fourth row of the table correspond to the fusion system evaluated with the speaker and the face groundtruth.

Speaker system shows a 41.13% DER, where the main source of error is the SER with a 31.9%. The threshold used to decide whether a segment corresponds to a target or to an unknown identity produces a low MISS but leads to a higher FA and SER. We noticed that our system failed in segments where music was included in the background and with these targets whose enrolment signal was very different to the show in terms of channel variability. Adapting the model to the RTVE corpus could have improved the rate of error caused by these factors. On the other hand, using an initial speaker segmentation on the signal instead of a sliding windows strategy could also lead to a better system performance.

For the face modality, the main source of error is the high number of missed face time (37.9%). On the other side, the FA and the SER are very low. The missed face time error could be originated from two different motives. For one side, a threshold too low could cause many false rejections of valid tracks (i.e. tracks belonging to valid enrollment identities). On the other side, this error could be originated because the face detection/tracking failed to extract valid tracks. To determine which one of this errors is predominant we have set the rejection threshold at its maximum value ($th = 1$), meaning that all tracks should be accepted. After doing that, we found that the missed face error was still very high (37.2%). This indicates that the errors are mainly produced by the tracking step.

The fusion system presents worst results than the the speaker and the face systems used individually. Both fusion systems evaluated with the speaker and the video groundtruths present higher MISS and SER in comparison with the monomodal systems. We have noticed that the multimodal segmentation does not improve the results because it automatically discards a lot of speaker segments and face tracks. In one hand, it discards the segments where there is no overlapping between speaker segments and face tracks. On the other hand, when more than one face appears in the video, the system automatically discards the face track of the person who is not speaking. Therefore, our assumptions would work better with the aim of looking for who is shown and speaking at the same time but not for this kind of multimodal evaluation.

4. Conclusions

We have presented two monomodal and one multimodal technologies to perform person identification in broadcast videos. A quantitative analisis has been performed on the RTVE 2018 dataset as provided in the Albayzin challenge. From the experiments it can be seen that the monomodal systems should be improved. For the speaker approach, it would be interesting to explore transfer learning methods to adapt our generic model in a smaller scenario and to include a speaker segmentation algorithm in the system. For the face modality, we plan to improve the face detection and tracking step as it has been proven that is the main source of error for the face modality. There is also a large room for improvement for the multimodal fusion system. Instead of fusing the systems from the output of the monomodal systems, an end to end multimodal system could work better if a big amount of data is available.

5. Acknowledgements

This work was supported in part by the Spanish Project DeepVoice (TEC2015-69266-P) and the project MALEGRA (TEC2016-75976-R), financed by the Spanish Ministerio de Economa, Industria y Competitividad and the European Regional Development Fund (ERDF).

6. References

[1] O. G. G. Bernard and J. Kahn, "The first official repere evaluation," in *SLAM-Interspeech*, 2013.

[2] J. Poignant, H. Bredin, and C. Barras, "Person discovery in broadcast tv at mediaeval 2015," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.

[3] H. Bredin, C. Barras, and C. Guinaudeau, "Person discovery in broadcast tv at mediaeval 2016," in *Working Notes Proceedings of the MediaEval 2016 Workshop*, 2016.

[4] N. Le, H. Bredin, G. Sargent, M. India, P. Lopez-Otero, C. Barras, C. Guinaudeau, G. Gravier, G. Barbosa da Fonseca, I. Lyon Freire, J. Z. Patrocinio, S. J. F. Guimares, M. Gerard, J. Morros, J. Hernando, L. Docio-Fernandez, C. Garcia-Mateo, S. Meignier, and J. Odobez, "Towards large scale multimedia indexing: A case study on person discovery in broadcast news," in *CBMI 2017*.

[5] M. India, D. Varas, , V. Vilaplana, J. Morros, and J. Hernando, "Upc system for the 2015 mediaeval multimodal person discovery in broadcast tv task," in *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.

[6] M. India, G. Marti, , E. Sayrol, J. Morros, J. Hernando, C. Cortillas, and G. Bouritsas, "Upc system for the 2016 mediaeval multimodal person discovery in broadcast tv task," in *Working Notes Proceedings of the MediaEval 2016 Workshop*, 2016, pp. 1–3.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR'05*, vol. 1, June 2005, pp. 886–893.

[8] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[9] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI'81*, 1981, pp. 674–679.

[10] C. Tomasi and T. Kanade, "Detection and tracking of point features," International Journal of Computer Vision, Tech. Rep., 1991.

[11] J. Shi and Tomasi, "Good features to track," in *CVPR 1994*, June 1994, pp. 593–600.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR 2016*, June 2016, pp. 770–778.

[13] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR 2015*, 06 2015, pp. 815–823.

[14] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[15] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[16] J.-F. Bonastre, F. Wils, and S. Meignier, "Alize, a free toolkit for speaker recognition," in *ICASSP'05*, vol. 1, 2005, pp. I–737.

[17] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. W. Evans, B. G. Fauve, and J. S. Mason, "Alize/spkdet: a state-of-the-art open source software for speaker recognition." in *Odyssey*, 2008, p. 20.