GTM-UVigo Systems for Person Discovery Task at MediaEval 2015

Paula Lopez-Otero, Rosalía Barros, Laura Docio-Fernandez,
Elisardo González-Agulla, José Luis Alba-Castro, Carmen Garcia-Mateo
AtlantTIC Research Center
{plopez,rbarros,ldocio,eli,jalba,carmen}@gts.uvigo.es

ABSTRACT

In this paper, we present the systems developed by GTM-UVigo team for the Multimedia Person Discovery in Broadcast TV task at MediaEval 2015. The systems propose two different strategies for person discovery in audio through speaker diarization (one based on an online clustering strategy with error correction using OCR information and the other based on agglomerative hierarchical clustering) as well as intrashot and intershot strategies for face clustering.

1. INTRODUCTION

The Person Discovery in Broadcast TV task at MediaEval 2015 aims at finding out the names of people who can be both seen as well as heard in every shot of a collection of videos [10]. This paper describes the audio, video and multimodal approaches developed by GTM-UVigo team to address this task[1].

2. AUDIO-BASED PERSON DISCOVERY

The audio approaches can be divided in three stages: speech activity detection, division of speech regions in speaker turns and, lastly, speaker clustering.

2.1 Speech Activity Detection

A Deep Neural Network (DNN) based speech activity detector (SAD) was used. The acoustic features used were 26 log-mel-filterbank outputs, and a window of 31 frames was used to predict the label of the central frame. The DNN has the following architecture: 806 unit input layer, 4 hidden layers, each containing 32 tanh activation units, and an output layer consisting of two softmax units. The output layer generates a posterior probability for the presence or non-presence of speech, and the ratio of both output posteriors is used as a confidence measure about speech activity over time. This confidence is median filtered to produce a smoothed estimate of speech presence and, finally, a frame is classified as speech if this smoothed value is greater than a threshold.

2.2 Speaker Segmentation

After performing speech activity detection, the speech segments are further divided into speaker turns following the approach described in [7]. First, Mel-frequency cepstral coefficients (MFCCs) plus energy are extracted from the waveform. After this, the Bayesian Information Criterion (BIC) based segmentation approach described in [2] is employed, performing a coarse segmentation to find candidates followed by a refinement step. A false alarm rejection strategy is applied in the latter step so as to reject change-points that are suspicious of being false alarms [6].

2.3 Speaker Clustering

Two different approaches for speaker diarization were assessed, one working in online mode, used in the primary system, and another working in offline mode. A feature they have in common is the use of the iVector paradigm [3] for speaker turn representation.

2.3.1 Online approach

This clustering strategy consists in comparing the iVectors of the speaker models with the iVector of a given speaker turn by computing its dot product and, if the maximum dot product exceeds a predefined threshold, the speaker turn is assigned to the speaker model; else, it is considered as a new speaker. Every time a new segment is assigned to a speaker, its model is refined by computing the mean of all the iVectors assigned to that speaker model.

A novel feature introduced in this online clustering scheme is the use of written names obtained from OCR [9] for automatic error correction. To that end, the speaker assignment using these written names is considered as more reliable than the clustering assignment, so anytime the clustering and the written name approach make a different decision, the written name will prevail over the clustering decision.

2.3.2 Offline approach

The proposed offline clustering strategy relies on an agglomerative hierarchical clustering scheme. First, a similarity matrix was obtained by computing the dot product between all the pairwise combinations of the iVectors of each speaker turn, and this matrix was used to obtain a dendrogram.The C-score stopping criterion described in [8] was used to select the number of clusters.

3. VIDEO-BASED PERSON DISCOVERY

The video-based strategies encompass three different steps: face detection and tracking, visual speech activity detection and face clustering.

[1]The code of GTM-UVigo systems will be released at `https://github.com/gtm-uvigo/Mediaeval_PersonDiscovery`

MediaEval 2015 Workshop, Sept. 14–15, 2015, Wurzen, Germany.

3.1 Face detection and Tracking

Face detection is based on histogram of oriented gradient features (HOG) and a linear SVM classifier implemented in dlib library [5]. For each detected person, a face tracking and landmark detection method based on CLNF models are used [1]; every time a person stops being visible on screen, a model that has information about presence, speech intervals and the highest quality face templates is stored in a database. To reduce the false alarm rate, face tracks that have a short time duration and a low quality score are rejected; this score is calculated with a weighted sum of face symmetry and sharpness values.

3.2 Visual Speech Activity Detection

The proposed visual speech activity detection method is based on the relative mouth movements which are generally small in silence sections, whereas variations of lip shape are usually stronger during speech [12]. Using face landmarks obtained from the previous step, mouth openness and lips height variance over time are computed. A variable threshold based on face size is applied in order to make the decision at each frame and a low-pass filter is used to smooth results.

3.3 Face clustering

The face clustering strategies consist in a face recognition system so that every time a face track is going to be inserted in the database, a score is computed in order to add it as a new person or to merge it with an existing one. First, Gabor features are extracted from the highest-quality templates of a person and matching scores are obtained using the hyper cosine distance [4]. Second, the final score to compare with the merging threshold is computed as the maximum of all the matching scores obtained from the two sets of face images. In the intrashot strategy, only models that appear within the same shot are compared, aiming at correcting presence intervals when the tracking method fails. The intershot strategy allows to merge all the person appearances in a video.

4. MULTIMODAL PERSON DISCOVERY

Multimodal person discovery was performed using four different sources of information: speaker diarization (SD) using the techniques described in Section 2; face detection (FD) and video-based speech activity detection (VVAD) as described in Section 3; and written names (WN) extracted using the strategy described in [9]. First, the set of evidences is defined as proposed in the baseline fusion strategy provided by the organizers. Given a shot, a person is considered to appear in it if the same name is present in SD, FD and VVAD within the time interval that defines the shot. A late naming strategy was used to assign names to the different sources of information [11]. For each hypothesized name, a confidence is computed as proposed in the baseline strategy, but those hypotheses with confidence lower than 1 are discarded, as they correspond to situations of non-overlap between the evidence and the hypothesized name.

5. RESULTS AND DISCUSSION

Table 2 shows the results achieved by the submitted systems both in REPERE (partition 'test2') and INA datasets; these systems are combinations of the two proposed speaker diarization and face clustering strategies as summarized in Table 1. The results achieved using the baseline metadata (b) are also shown for comparison.

Table 1: Summary of the submitted systems

System	Spk. clustering	Face clustering
Primary (p)	online	intrashot
Contrastive1 (c1)	online	intershot
Contrastive2 (c2)	offline	intrashot
Contrastive3 (c3)	offline	intershot

Table 2 shows that the two speaker diarization strategies are almost equally suitable for this task as they achieve very similar results; however, the online strategy shows a better performance, probably due to the use of the OCR information for error correction. With respect to the face clustering strategies, the intrashot method obtained better results, probably because the intershot combination led to an excessive combination of faces, making the system miss speakers by erroneously combining them with others.

Table 2: Results on development and test datasets corresponding to July 1st deadline.

	REPERE			INA		
	EwMAP	MAP	C	EwMAP	MAP	C
p	**75.76%**	**77.10%**	**78.03%**	**80.34%**	**80.61%**	92.42%
c1	74.90%	75.80%	77.58%	75.42%	75.69%	85.99%
c2	**75.76%**	**77.10%**	77.58%	80.21%	80.49%	92.32%
c3	75.54%	76.43%	77.58%	75.26%	75.54%	85.89%
b	63.58%	63.93%	71.75%	78.35%	78.64%	**92.71%**

The development of the audio-based person discovery approaches showed us that a lower speaker diarization error rate do not lead to a higher EwMAP, as overclustering results in incorrect person detections. Also, we have to increase our efforts in TV programmes featuring challenging acoustic conditions, which are the ones who had a more degraded performance. Lastly, we realised that adding written names obtained from OCR information to the speaker diarization algorithm led to an improvement of the performance, so this type of fusion will be studied in more depth.

The proposed video-based person discovery approaches showed us that the intrashot strategy performed better than the intershot strategy, probably because of the overclustering issue mentioned above. The most challenging aspects, that will have to be addressed in the future, were the variations in pose, scale and illumination, as they made it difficult to develop a robust face matching strategy.

GTM-UVigo team got into this task by developing audio and face modules and combining them through a simple decision-level fusion but, in future work, audiovisual fusion in earlier stages of the system will be researched in order to exploit all the potential of multimodal person discovery.

6. ACKNOWLEDGEMENTS

This research was funded by the Spanish Government ('SpeechTech4All Project' TEC2012-38939-C03-01), the Galician Government through the research contract GRC2014/024 (Modalidade: Grupos de Referencia Competitiva 2014) and 'AtlantTIC Project' CN2012/160, and also by the Spanish Government and the European Regional Development Fund (ERDF) under project TACTICA.

7. REFERENCES

[1] T. Baltrusaitis, P. Robinson, and L. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 354–361, 2013.

[2] M. Cettolo and M. Vescovi. Efficient audio segmentation algorithms based on the BIC. In *Proceedings of ICASSP*, volume VI, pages 537–540, 2003.

[3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[4] E. González-Agulla, E. Argones-Rua, J. Alba-Castro, D. González-Jiménez, and L. Anido-Rifón. Multimodal biometrics-based student attendance measurement in learning management systems. In *IEEE International Symposium on Multimedia (ISM)*, pages 699–704, 2009.

[5] D. King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

[6] P. Lopez-Otero. *Improved Strategies for Speaker Segmentation and Emotional State Detection*. PhD thesis, Universidade de Vigo, 2015.

[7] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo. GTM-UVigo system for Albayzin 2014 audio segmentation evaluation. In *Iberspeech 2014: VIII Jornadas en Tecnología del Habla and IV SLTech Workshop*, 2014.

[8] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo. A novel method for selecting the number of clusters in a speaker diarization system. In *Proceedings of EUSIPCO*, pages 656–660, 2014.

[9] J. Poignant, L. Besacier, G. Quénot, and F. Thollard. From text detection in videos to person identification. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2012.

[10] J. Poignant, H. Bredin, and C. Barras. Multimodal Person Discovery in Broadcast TV at MediaEval 2015. In *Proceedings of the MediaEval 2015 Workshop*, 2015.

[11] J. Poignant, H. Bredin, V. Le, L. Besacier, C. Barras, and G. Quénot. Unsupervised speaker identification using overlaid texts in TV broadcast. In *Proceedings of Interspeech*, 2012.

[12] B. Rivet, L. Girin, and C. Jutten. Visual voice activity detection as a help for speech source separation from convolutive mixtures. *Speech Communication*, 49(7):667–677, 2007.