Combining Audio Features and Visual I-Vector @ MediaEval 2015 Multimodal Person Discovery in Broadcast TV

Fumito Nishi, Nakamasa Inoue, Koichi Shinoda
Tokyo Institute of Technology, Tokyo, Japan
{nishi, inoue, shinoda}@ks.cs.titech.ac.jp

ABSTRACT

This paper describes our diarization system for the Multimodal Person Discovery in Broadcast TV task of the MediaEval 2015 Benchmark evaluation campaign [1]. The goal of this task is naming speakers, who are appearing and speaking simultaneously in the video, without prior knowledge. Our diarization system is based on multimodal approach to combine audio and visual informations. We extract features from a face in each shot to make visual i-vectors [2], and introduce them to the provided baseline system. In the case of faces are extracted correctly, the performance becomes better, but based on the test run, clear improvement could not be observed.

1. INTRODUCTION

The Multimodal Person Discovery in Broadcast TV task can be split into following three subtasks: speaker diarization, face detection and tracking, and name detection. We focused on the speaker diarization.

For speaker diarization, combining audio and visual features are often effective. For example, in previous work, visual features extracted from faces [3], cloths [4], and whole images [5], are used to decrease influence of background noise. In this paper, we decide to use facial features because they represent personal characteristics directly.

In the field of speaker recognition and diarization, using i-vector [6-7] is one of the state-of-the-art methods. The main idea of the audio i-vector is to find a subspace, which represents speaker- and channel-variabilities simultaneously. We apply this method to images. HOG features, which are used in face recognition [8]. To extract visual i-vectors, we use HOG features instead of MFCCs or PLPs used in the audio i-vector framework [2].

Here, we assume that a visual i-vector represents face- and channel-variabilities. We expect it works as face recognition and is easy to introduce to a diarization system.

2. APPROACH

Figure 1 and Figure 2 show the overview of visual i-vector extraction, and the whole system, respectively. First, we extract HOG features from a face in each shot to make a visual i-vector. Second, we estimate the distance between each pair of visual i-vectors. Third, we combine audio score and visual i-vector's score to apply hierarchical clustering. Finally, we combine diarization results and other results using the baseline system [9-10].

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

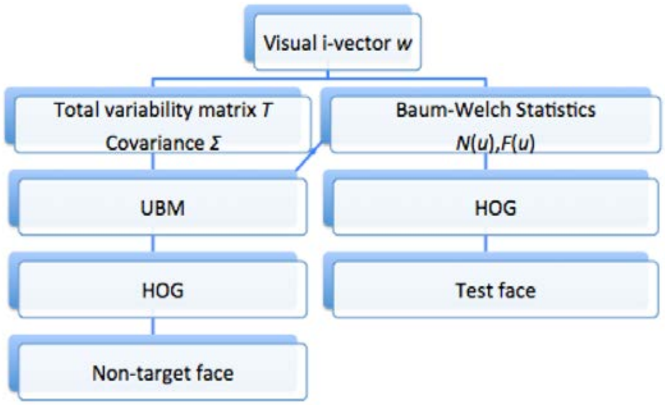

Figure 1: Overview of visual i-vector extraction

2.1 i-vectors for a face

Let M be a GMM super-vector, which is the concatenation of normalized mean vectors of an estimated GMM for a targeted video shot. An i-vector w is extracted from it, by assuming that M is modeled as

$$M = m + Tw,$$

where m is a face- and channel- independent super-vector, and T is a low rank matrix representing total variability. The EM algorithm is used to estimate the total variability as proposed in [11]. Note that w is associated with face tracks. I-vector w_u for utterance u is calculated by the following equation.

$$w_u = \ (I + T^t \Sigma\ N(u)T)^{-1} T^t \Sigma^{-1} F(u),$$

where N(u), and F(u) are the zero, and first order Baum-Welch statistics on the UBM for the current utterance u, and Σ is the covariance matrix of the UBM.

After visual i-vectors are extracted, we calculate distance between each pair of vectors in cosine distance. The distance D_{ij} between i-vectors w_i and w_j is calculated by

$$D_{ij} = 0.5 - \frac{w_i w_j}{||w_i||_2\ ||w_j||_2}.$$

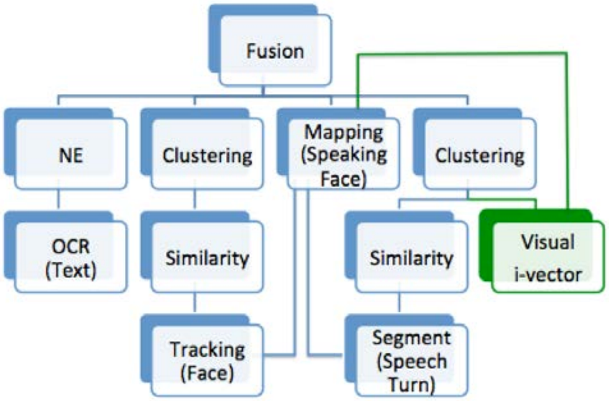

Figure 2: System overview. The rightmost visual i-vector is our proposal method.

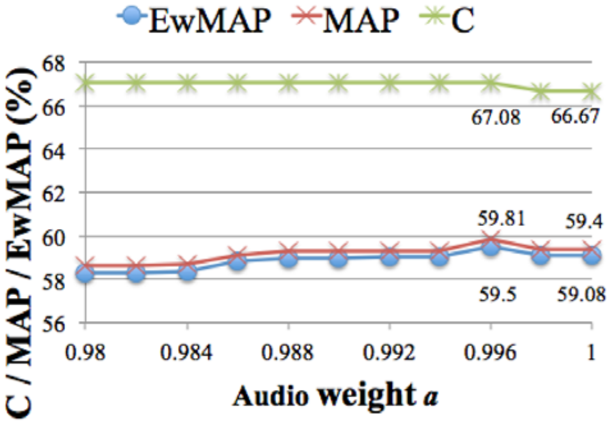

Figure3: EwMAP, MAP, C by audio weight *a* in development set.

2.2 Late Fusion

To combine our system and the baseline diarization system, we use late fusion. The final score *F* is given by

$$F = aA + (1 - a)V \ (0 \leq a \leq 1),$$

where *A* is a score from the baseline system (normalized BIC between each utterance), *V* is a score from the visual i-vector, and *a* is a weighting parameter.

2.3 Clustering

After two scores combined, we apply hierarchal clustering according to fusion score *F*.

3. EXPERIMENTS AND RESULTS

3.1 Experimental Settings

HOG Features are extracted from the face in each shot detected by using the baseline speaking-face system. If there is no candidate, the previous shot's candidate is used instead of it. A HOG feature is 34 dimensions, which consists of 32 dimensional histograms of oriented gradients and its x-y coordinates. The number of Gaussian mixture components of the UBM is 32. In the development set, we select the most effective audio weight by grid search. The results are the first deadline version.

Baseline	Dev.: Test2	Test
Correctness (C)	66.67	92.71
Mean Average Precision (MAP)	59.40	78.64
Evidence-weighted MAP (EwMAP)	59.08	78.35
run #1 visual i-vector (*a* = 0.996)	Dev.: Test2	Test
Correctness (C)	67.08	92.71
Mean Average Precision (MAP)	59.81	78.67
Evidence-weighted MAP (EwMAP)	59.90	78.38

Table 1: Correctness, Mean Average Precision, Evidence-weighted mean average precision of the baseline system and our visual i-vector system.

3.2 Experimental Results

Figure 3 shows the Evidence-weighted Mean Average Precision (EwMAP), Mean Average Precision (MAP), Correctness (C) of the each audio weight. Table 1 shows the C, MAP, and EwMAP for the development and test set. As we can see, *a*=0.996 is the best.

In the development and test videos, direction of speaker face is frontal for the most part. This is good condition for our system. However, the results do not show significant improvement compared with the baseline system. One of the reasons is the insufficient face detection accuracy. We observed that 63% of detected bounding boxes capture a face or part of a face in our analysis on randomly sampled 10 videos in the test set. However, since only 82% of them capture an entire face precisely, our system did not improve the overall performance significantly. We conclude that normalization of the face area for more sophisticated face detection is needed.

4. CONCLUSION

We presented an audio-visual based speaker diarization system, which uses visual i-vectors. The results did not show obvious advantage of our method due to the accuracy of face detection. In our future work, normalizing the face area and combining with other visual features extracted from whole image would help the improvement of the accuracy.

5. REFERENCES

[1] J. Poignant *et al.* “Multimodal Person Discovery in Broadcast TV at MediaEval 2015,” *MediaEval Workshop* 2015.

[2] F. Nishi *et al.* “Speaker Diarization Using Multi-Modal i-vectors,” *ITC-CSCC*, pp. 27-30, 2015.

[3] G. Friedland *et al.* "Multi-modal speaker diarization of real-world meetings using compressed-domain video features," *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, pp. 4069-4072, 2009

[4] F. Vallet *et al.* "A multimodal approach to speaker diarization on TV talk-shows," *Multimedia, IEEE Transactions on* vol. 15, num. 3, pp. 509-520, 2013

[5] B. Xavier, and G. Linares. "Constrained speaker diarization of TV series based on visual patterns," *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014.

[6] N. Dehak *et al*. “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol.19, no.4, pp.788-798, 2011.

[7] S. Shum *et al*. “Exploiting Intra-Conversation Variability for Speaker Diarization” Proc. Interspeech, pp.945-948, 2011.

[8] O. Deniz *et al*. “Face recognition using histograms of oriented gradients,” *Pattern Recognition Letters*, vol.32, num.12, pp.1598-1603, 2011.

[9] J. Poignant *et al*. “Unsupervised speaker identification using overlaid texts in TV broadcast,” *Proc. Interspeech*, 2012.

[10] J. Poignant *et al*. “From text detection in videos to person identification,” *Proc. Multimedia and Expo*, pp.854-859, 2012.

[11] P. Kenny *et al*. “Eigenvoice modeling with sparse training data”, *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 3, pp. 345-354, 2005.