DIHARD 2018 I3A (ARAGON INSTITUTE FOR ENGINEERING RESEARCH - UNIVERSITY OF ZARAGOZA) SYSTEM DESCRIPTION

Ignacio Viñals, Pablo Gimeno, Alfonso Ortega, Antonio Miguel and Eduardo Lleida

March 31, 2018

Abstract

The I3A submission consisted a modification of the standard i-vector PLDA approach in speaker verification adapted to diarization. Following a bottom-up approach, i-vectors are clustered by means of a Fully Bayesian PLDA solved by Variational Bayes. The whole submission is integrated by five systems, all of them following the same principles. The best obtained scores with this strategy are DER=26.02% and MI=8.52 in track1, whereas in track2 DER=38.00% and MI=8.12 are obtained.

1 Data Resources

The considered data pool for the DIHARD challenge tried to provide the widest possible variety of data but also adapting it to the evaluation conditions. Our main dataset is the Multi-Genre Broadcast Challenge 2015 (MGB) dataset [1]. 1600 hours of broadcast data under different conditions were considered for the evaluation scenario variability. These data were complemented with different meetings corpora (AMI corpus[2], ICSI meeting corpus [3] and the Rich Transcription 2009 dataset (RT09)) were also included to add more variability, and include some knowledge about the meetings scenario.

2 Detailed description of the algorithm

2.1 Feature Extraction

The front end extracts acoustic feature vectors of 20 MFCC including C0 (C0-C19) over a 25 ms hamming window every 10 ms (15 ms overlap). No derivatives are considered. The obtained features are normalized according to a Short-Time Cepstral Mean and Variance Normalization (STCMVN) with a 1.5-second analysis window.

2.2 Voice Activity Detection

Voice Activity Detection (VAD) is performed by means of a 1-layer BLSTM network of 128 neurons, trained with DIHARD development set. This network studies the data in 3-second duration sequences, providing one label each 10 ms of audio.

2.3 Segment Representation

The segmentation step in the submission is based on a BIC analysis [4], considering a 3-second sliding window. Each acoustic segment is represented by an i-vector [5], with models described as follows. Two i-vector extractors were considered. Baseline system and System 1 work according to a 256-Gaussian 200-dimension i-vector extractor exclusively trained with Multi-Genre Broadcast dataset. Systems 2,3 and 4 work in terms of a 512-Gaussian 200-dimension i-vector model, trained considering the whole pool of datasets (MGB, AMI ISCI meetings and RT09). With both models, Centering, whitening [6] and length normalization [7] are applied. While Baseline and System 1 works by means of local centering (each evaluation episode is centered according to itself), systems 2,3 and 4 rely on some universal centering, trained with the described pool of datasets.

2.4 Clustering Method

The i-vector clustering is performed by a Fully Bayesian PLDA solved by Variational Bayes [8][9]. While our Baseline and system 1 work with a 50 dimension PLDA only trained with MGB, systems 2,3 and 4 consider a 200 dimension PLDA trained with all the available data. This clustering is initialized in two different ways: Baseline and systems 1,2,3 are initialized by means of Agglomerative Hierarchical Clustering. System 4 considers the log-likelihood PLDA ratio matrix as an image, applying different thresholds to determine initial clusters. For systems baseline and 1 unsupervised in-domain adaptation [10] is performed.

2.5 Speaker estimation

The Variational Bayes solution has the ability to recombine and eliminate speakers at will, only depending on the initial clustering. Therefore, multiple initializations are considered, and the final diarization solution is chosen taking into account the Evidence Lower Bound (ELBO) penalized by the VB model complexity, i.e. the number of speakers and the number of free parameters.

3 Computational Resources

The system was developed using Intel® Xeon® E5520 2.27 GHz and Intel® Xeon® Processor E3-1231 v3. The approximated resources by a 10-minute audio are given in the next Table

Table 1: *Processing Times and Memory*

	Time by Audio (secs)	Memory (MB)
VAD	60	1000
MFCC	30	200
i-vectors	60	2000
Whitening + Lnorm	0.02	2000
Clustering		
Baseline & System 1	10	4000
Systems 2 , 3 and 4	30	4000

References

[1] P. Bell, M. J. F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, and P. C. Woodland, "The MGB Challenge: Evaluating Multi-Genre Broadcast Media Recognition," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015Scottsdale, Arizona, USA, Dec. 2015, IEEE.*, vol. 1, 2015.

[2] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, L. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI Meeting Corpus: A pre-announcement," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3869 LNCS, pp. 28–39, 2006.

[3] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," *Proceedings of the first international conference on Human language technology research - HLT '01*, pp. 1–7, 2001.

[4] S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering Via the Bayesian Information Criterion," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, vol. 6, pp. 127–132, 1998.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[6] J. Villalba and E. Lleida, "Handling i-vectors from different recording conditions using multichannel simplified plda in speaker recognition," 2013.

[7] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 249–252, 2011.

[8] J. Villalba and E. Lleida, "Unsupervised Adaptation of PLDA By Using Variational Bayes Methods," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 744–748, 2014.

[9] J. Villalba, A. Ortega, A. Miguel, and E. Lleida, "Variational Bayesian PLDA for Speaker Diarization in the MGB Challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 667–674, 2015.

[10] I. Viñals, A. Ortega, J. Villalba, A. Miguel, and E. Lleida, "Domain Adaptation of PLDA models in Broadcast Diarization by means of Unsupervised Speaker Clustering," *Interspeech*, pp. 2829–2833, 2017.