UWB-NTIS Speaker Diarization Systems for the DIHARDIII 2020 Challenge

Zbyněk Zajíc
University of West Bohemia, Faculty of Applied Sciences
NTIS - New Technologies for the Information Society,
Univerzitní 8, 306 14 Plzeň, Czech Republic
zzajic@ntis.zcu.cz

***Abstract*—This document describes the diarization system developed by the team from the New Technologies for the Information Society (NTIS) research center of the University of West Bohemia (team "UWB-NTIS"), for the Third DIHARD Speech Diarization Challenge.**

I. Notable highlights

This system is classical approach for diarisation based on x-vectors clustering.

II. Data Resources

- VoxCeleb1 and 2 (http://www.robots.ox.ac.uk/~vgg/data/voxceleb/),
- Czech Speecon database (ELRA-S0298),

Additional data augmentation (additive noise, music, babble and reverberation).

III. Detailed Description of the Algorithm

Our systems are based on the standard approach of segmentation, x-vector [1]–[3] extraction and clustering. We used segmentation and clustering proces based on a Kaldi diarization recipe[1].

1) Feature extraction: We used Linear Frequency Cepstral Coefficients (LFCCs), Hamming window of length 25 ms with 10 ms shift. There are 40 triangular filter banks linearly spread across the frequency spectrum, and 25 LFCCs are extracted. The resultant 50-dimensional feature vector ($D_f = 50$) also includes delta coefficients.

2) Speech activity detection: We only participated in Track 1, so information about speaker activity was received as gold speech segmentation.

3) Segment Representation: Kaldi recipe for diarization [4] was used. The segmentation provides chunks of speech between important non-speech events (Kaldi SAD segmentation) and subsequently divides these segments into sub-segments with constant length 1.5 s and overlap 0.75 s (the minimum length of a segment is 0.5 s). A Time Delay Neural Network is used as an x-vector extractor, and x-vectors are extracted from the affine component of the second-to-last layer with dimension 500. We also whitened the vectors.

[0]https://github.com/kaldi-asr/kaldi/tree/master/egs/callhome_diarization/v1 and /v2

4) Clustering: The x-vectors are clustered using AHC, with the stopping threshold set on development data. This threshold was found for the entire development set. We use a Probabilistic Linear Discriminant Analysis (PLDA) model [5] to evaluate the distance between the x-vectors.

No resegmentation is used.

IV. Results on the development set

The results for development set is 22.95% DER

V. Hardware Requirements

The code was not optimized with regards to execution time (e.g., intermediate results were saved to the disk).

A. Training x-vector extraction and FA models:

Hardware:

- CPU 8-core Intel Xeon E5-2650v2 2.60 GHz
- GPU 2x nVidia Tesla K20, 5GB, 1000 GFLOPS
- 4 GB RAM
- 10 GB required storage

Total training time: approx. 96 hours

B. Execution times to process the entire development set:

Hardware (main system):

- CPU Intel(R) Core(TM) i7 cpu - 4 cores used, 3.07GHz
- GPU NVIDIA GeForce 1080 Ti, 11 GB VRAM, 11,340 GFLOPS
- 32 GB RAM
- 80 MB required storage

Total training time: approx. 3 hours

References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *ICASSP*, 2018, pp. 5329–5333.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[3] L. Machlica and Z. Zajíc, "Factor Analysis and Nuisance Attribute Projection Revisited," in *Interspeech*, vol. 2, Portland, 2012, pp. 1570–1573.

[4] G. Sell, D. Snyder, A. Mccree, D. Garcia-romero, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur, "Diarization is Hard : Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge," in *Interspeech*, Hyderabad, 2018, pp. 2808–2812.

[5] S. Ioffe, "Probabilistic Linear Discriminant Analysis," *Lecture Notes in Computer Science*, vol. 3954 LNCS, pp. 531–542, 2006.