LEAP Submission for DIHARD 2018

Shobhana Ganesh, Bharat P, Neeraj Sharma, Prachi Singh, Sriram Ganapathy

Learning and Extraction of Acoustic Patterns (LEAP) Lab, Department of Electrical Engineering
Indian Institute of Science, Bangalore, India

Abstract

The problem of associating segments in an audio signal with a particular speaker to answer the question of 'who spoke when', also referred to as speaker diarization, has gained considerable interest owing to its significance as a pre-processing step in automatic speech recognition applications. While diarizing systems perform well on clean datasets such as telephone conversations and interviews, the performance on datasets associated with meetings, child speech, multiple number of speakers with short turns in conversations, etc., remains a byzantine task. In this report, we describe our system designed for diarization of data drawn from the later kinds of datasets. This system was submitted to the First DIHARD Speech Diarization Challenge, 2018. The system makes use of MFCCs as front-end features, followed by an i-vector modeling and subsequently, PLDA scoring followed by agglomerative clustering. This set up uses i-vectors, obtained from a GMM-UBM to detect change in speakers and an unsupervised calibration method to estimate the number of speakers. We obtain a DER of 28.52 (and MI of 8.32) and 53.4 (and MI of 7.6) on the evaluation data in track 1 and track 2, respectively.

Index Terms: DIHARD, diarization, PLDA.

1. Data Resources

- AMI Corpus - The AMI corpus is a series of recordings of meetings involving far field microphones with an average of four people per meeting. Each meeting is recorded using a set of different devices, namely, microphone array consisting of eight single distant microphones. We used the first microphone channel from the microphone array for our training set.
  Link to dataset: http://groups.inf.ed.ac.uk/ami/download/
- LibriSpeech Corpus [1] - This is an audio book corpus containing contributions from huge collection of speakers, more than 1000 speakers and 1000 hrs of speech. We used a subset of this dataset, named train-clean-set, containing 193 English audio books, each close to 10 mins and read by a different speaker.
  Link to dataset: http://www.openslr.org/12/
- Paidologos dataset - This dataset consists of laboratory recordings of words in isolation spoken by children in English, Japanese, Greek, and Cantonese.
  https://phonbank.talkbank.org/browser/index.php
- Switchboard cellular - These are audio recordings[2] of telephone conversations between two individuals. To balance a predominantly male dataset, we utilized all recordings with a female speaker. We also took the conversations occurring either in an outdoor or indoor setting.
  Catalog ID - LDC2001S13
- Distress Analysis Interview Corpus (DAIC) - These are a set of audio recordings of clinical interviews involving a robot Ellie and a patient with psychological disorder.
  Link to dataset: http://dcapswoz.ict.usc.edu/

2. Algorithm

The system used for the challenge is the i-vector based model with PLDA scoring which is then followed by agglomerative hierarchical clustering [3]. In this system, the first step is extracting MFCC features from the audio files. These are used as input for training a GMM-UBM, and the parameters of this are used in building an i-vector extractor. For an input speech file, we obtain i-vectors over 150 ms segments, with a 75 ms shift. A PLDA scoring is then performed on these i-vectors to determine the similarity between i-vectors in the file. Once these scores are obtained, the agglomerative hierarchical clustering is used to obtain speaker specific segments.

2.1. Signal Processing

Out of the datasets used for training, only the AMI corpus and switchboard cellular have a natural reverberation and background noise. Our analysis on dev set showed that the DER was low for files which were clean and high for noisy audio files. To circumvent this issue we made use of data augmentation. To generate noisy data at different SNRs we used the MUSAN[4] corpus for adding noise (such as babble, music, etc) at different SNRs and reverberation corresponding to a range of RT60s.

2.2. Acoustic Features

The use of pitch as an additional feature was also looked at with child data. However, this gave a marginal improvement only in those files where a male speaker was present. Overall, the experiments fared well when MFCC features were extracted from the audio files. Hence, we used these as front-end features. The MFCC features were extracted at a window size of 25 ms and a 10 ms shift. The sampling frequency was set to 16 khz and the first 20 cepstral coefficients were extracted.

2.3. Speech Activity Detection

For the system submitted in Track 2 we have used speech activity detection (SAD) algorithm[5] to generate the segment boundaries. This is available in the Voicebox Matlab toolkit.

2.4. Segment Representation

The MFCC features extracted were input to GMM background model of 2048 mixtures. This model was used to obtain the super vectors needed for the T-matrix training from which the i-vectors were extracted. These i-vectors were obtained at every 150 ms with a 75 ms hop for development and evaluation data, and for the data used in training, we extract 300 ms i-vectors for every 10 s. The i-vectors are of dimension 128. These i-vectors contain speaker dependent characteristics and can separate out distinct characteristics of an individual from an audio source. The i-vectors were all obtained from the segments defined as speech in the development and evaluation data for track 1. The segments for track 2 were defined with a speech activity detector as explained above (sec. 2.3).

2.5. Speaker Estimation and Clustering

Once the i-vectors were obtained a technique to estimate different speakers and cluster them was required. We used the plda scoring method to determine how similar each i-vector was to another. This created a matrix containing the similarity score of each i-vector with every other i-vector. During clustering the i-vectors having a high similarity were merged into a single cluster which indicated a speaker identity. To stop the merging a threshold value was set. We used a calibration method having a 2 mixture GMM to compute a threshold value. We then tweaked this threshold to improve the performance. A higher threshold would allow the algorithm to merge more i-vectors for a given speaker. On the whole we observed that the system underestimated the number of speakers when it was a high number (e.g $10-12$). The clustering technique used to identify speaker clusters from the plda scores was the agglomerative hierarchical clustering algorithm.

In addition, we also experimented with a combination of systems. Taking the PLDA scores from the two systems which give the lowest DER on the development data, we compared the two scores to obtain a combined matrix derived from the two individual matrices for each audio segment. We took the average scores as well as the minimum score of the two matrices. Our results indicated that the minimum scores gave the best result.

3. Hardware Requirements

The hardware requirements reported were common for both, the training and testing phase.

3.1. CPU information

Model - Acer F380 series
Number of cores - 64
Memory - 256 GB

3.2. GPU information

Model - NVIDIA Quadro P5000
CUDA cores - 2560
GPU memory - 16 GB

3.3. Toolkit used

The experiments were all conducted using the Kaldi[6] framework. The data for training and testing were prepared according to the data formats required by kaldi. Other toolkits and software used are the HTK toolkit, Miniconda for Python 2.7, and MATLAB.

3.4. Execution times

The wall clock time of the total execution time is reported here. For training - 1 hour 30 minutes Testing (one 10 minute file) - 5 minutes

4. References

[1] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[2] J. J. Godfrey, E. C. Holliman, and J. McDaniel, “Switchboard: Telephone speech corpus for research and development,” in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[3] G. Sell and D. Garcia-Romero, “Speaker diarization with plda i-vector scoring and unsupervised calibration,” in *2014 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2014, pp. 413–417.

[4] D. Snyder, G. Chen, and D. Povey, “MUSAN: A music, speech, and noise corpus,” *CoRR*, vol. abs/1510.08484, 2015. [Online]. Available: http://arxiv.org/abs/1510.08484

[5] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, 1999.

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The kaldi speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.