The QUT speaker diarization system for the First DIHARD challenge

I. Himawan, A. Kanagasundaram, H. Ghaemmaghami S. Sridharan, C. Fookes

Queensland University of Technology, Brisbane, Australia

{i.himawan, s.sridharan, c.fookes}@qut.edu.au, ahilan.eng@gmail.com

Abstract

We present details of the QUT submission to the First DIHARD challenge, which is focussed on speaker diarization on a diverse set of challenging domains. Our i-vector/GMM system achieves a diarization error rate (DER) of 33.15% on the track one evaluation data that is diarization using gold speech segmentation.

Index Terms: speaker diarization, i-vector

1. Introduction

Speaker diarization aims to provide annotation labels to each speaker-homogeneous segments in a recording that correspond to the unique speaker identity. The ultimate goal for this task is to answer '*who spoke when*' in a multi-speaker environment. There are many potential applications which can benefit from speaker diarization system such as automatic speech recognition (ASR), forensics, and information retrieval. For example, since the ASR typically assume one speaker is speaking in an utterance, having information regarding speakers and their speaking boundaries could potentially improve the ASR accuracy [1].

Recent advances in speaker diarization is influenced by the techniques proposed for speaker recognition that is to determine if the segment containing speech in an utterance was spoken by the same person as in the previous segments or not. This process can be repeated for all spoken segments in the conversation. Using the i-vector to represent speakers have proven to yield state-of-the-art speaker verification performance, and now the same technology has been applied for speaker diarization as well. Comparison can be made between two i-vectors to determine if they are spoken by the same speaker by producing verification scores [1].

Researchers have discussed the current speaker diarization problems in Jelinek Summer Workshop on Speech and Language Technology (JSALT) 2017. It was found that the existing state-of-the-art diarization systems perform poorly, when the recordings with a wide range of acoustic conditions and a large amounts of noise, and high levels of speaker overlap. The First DIHARD challenge has been proposed to develop state-of-the-art diarization system to address this issue.

2. Data resources

We used the NIST SRE 2004, 2005, 2006, and 2008 datasets (Catalog IDs: LDC2006S44, LDC2011S01, LDC2011S09, LDC2011S05, LDC2011S07) and Switchboard 2 phase II and III, Switchboard Cellular Part 2 corpora (Catalog IDs: LDC99S79, LDC2002S06, LDC2004S07) for training the i-vector extractor system. Only the NIST SRE datasets are used for PLDA training.

3. Detailed description of algorithm

In i-vector framework, the speaker can be represented as fixed-dimensional vectors extracted from a recording of speech. The i-vector extraction can be viewed as a probabilistic compression process that reduce GMM supervectors to a low dimensional subspace, named total variability space [2],

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{1}$$

where $\mathbf{m}$ is the speaker and session independent UBM mean supervector, $\mathbf{T}$ is called Total Variability matrix which has a low rank, and $\mathbf{w}$ is the resulting i-vectors.

In general, i-vectors based speaker diarization involves four main tasks: (1). Speech activity detection to segment the input utterance into speech and non-speech segments. (2) Extract fixed-dimensional representation (i.e., i-vectors) from the short-term speech. (3). Cluster the representations and assign speaker labels to the audio segments those representation were extracted from. (4). Refine the speaker segmentation at a more fine-grained level (e.g., frame-level).

The GMM/i-vector system used a 2048 component GMM trained on 40 cepstral coefficients (20 MFCCs including their first derivatives) to produce 128-dimensional i-vectors. For speaker recognition, the i-vectors are usually estimated on long utterances. However, in diarization system, the analysis is performed on short segments and usually i-vectors are extracted from 1-2 seconds of speech. PLDA models are applied on the i-vectors for i-vectors scoring [3, 4]. Speaker clustering is then performed using an unsupervised method on the extracted i-vectors using agglomerative hierarchial clustering (AHC) [5]. The AHC stopping criterion is determined based on the threshold discovered from unsupervised calibration of PLDA scores using development data.

3.1. Performance Metric

The performance of our systems is measured in terms of Diarization Error Rates (DER) [1]. This metric takes into account both the incorrect speaker assignment and segmentation errors. The Diarization Error Rate is computed as,

$$DER = E_{FA} + E_{miss} + E_{spk} \tag{2}$$

where E_{FA} refers to false alarm speech which is the amount of time incorrectly detected as speech divided by the total reference speaker time, E_{miss} refers to miss speech, defined as the amount of speech time that has not been detected as speech divided by the total reference speaker time, and E_{spk} is speaker error, defined as the time assigned to incorrect speakers divided by the total reference speaker time.

3.2. Results

The results on the track one evaluation data are shown in Table 1.

Table 1: *DER (%) We reported results of two GMM/i-vector systems with different length of speech segments when extracting i-vectors: (1) 1.5 seconds segment length with 0.75 seconds overlap, and (2) 3 seconds segment with 1.5 seconds overlap.*

	Data	
Methods	Dev	Eval
GMM/i-vector (1.5sec,0.75sec)	30.28	33.15
GMM/i-vector (3.0sec,1.5sec)	27.23	-

4. Hardware requirements

Table 2: *Hardware requirements for system development (CPU: 64-bit Intel Xeon E5-2680 2.5GHz).*

Development stages	No. CPU	RAM (GB)	Time (h)
Diagonal UBM training	40	25	0.7
Full UBM training	40	25	1.3
Training i-vectors extractor	40	90	11
Extracting i-vectors for PLDA	40	10	0.5

In this section, we report the hardware requirements for system development and evaluation. Table 2 presents the total number of CPU cores with RAM and the execution time for each step in the training stage. The i-vectors are extracted for every 1.5 seconds with 0.75 seconds overlap. The time taken to estimate i-vectors from ten minutes recordings (about 800 i-vectors) is 55 seconds.

5. Conclusions

The paper presents description and results for the QUT speaker diarisation systems submitted to the First DIHARD speaker diarisation challenge. Future works will investigate speaker diarization from scratch whereby our preliminary system achieved 57.14% DER.

6. References

[1] X. A. Miró et al., "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 2-, no. 2, pp. 356–370, 2012.

[2] N. Dehak et al., "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language*, vol. 19, pp. 788–798, 2011.

[3] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *in Proceedings of IEEE Spoken Language Technology Workshop*, 2014, pp. 413–417.

[4] I. Salmun et al., "On the use of PLDA i-vector scoring for clustering short segments," in *in Proceedings of Odyssey*, 2016, pp. 407–414.

[5] G. Sell et al., "Priors for speaker counting and diarization with AHC," in *in Proceedings of Interspeech*, 2016, pp. 2194–2198.