Literature Review 3

Primary paper:

Real-time Speaker Recognition System using Multi-stream i-vectors for AI Assistant

```
@INPROCEEDINGS{8326161,
author={K. Cho and J. Roh and Y. Han and N. Kim and J. Lee},
booktitle={2018 IEEE International Conference on Consumer Electronics (ICCE)},
title={Real-time speaker recognition system using multi-stream i-vectors for AI assistant},
year={2018},
volume={},
number={},
pages={1-4},
keywords={Acoustics;Feature extraction;Neural networks;Real-time systems;Speaker recognition;Speech;Speech recognition},
doi={10.1109/ICCE.2018.8326161},
ISSN={},
month={Jan},}
```

Secondary paper:

TIME DELAY DEEP NEURAL NETWORK-BASED UNIVERSAL BACKGROUND MODELS FOR SPEAKER RECOGNITION

```
@INPROCEEDINGS{7404779,
author={D. Snyder and D. Garcia-Romero and D. Povey},
booktitle={2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)},
title={Time delay deep neural network-based universal background models for speaker recognition},
year={2015},
volume={},
number={},
pages={92-97},
```

keywords={neural nets;speaker recognition;DNN;EER;GMM baseline;Gaussian mixture model;LVCSR task;equal error rate;i-vector-based speaker recognition systems;speaker recognition;time delay deep neural network;universal background models;Acoustics;Computational modeling;Delay effects;Feature extraction;Neural networks;Speaker recognition;Training;deep neural networks;i-vector;speaker recognition;time delay neural networks},
doi={10.1109/ASRU.2015.7404779},
ISSN={},
month={Dec},}

Summary:

The primary paper used a real-time speaker recognition that works in sync with ASR engine. In order to improve recognition accuracy, they employed multi-stream i-vectors. The proposed multi-stream i-vectors compensate the limited amount of information contained in a single i-vector. Their approach outperforms than the convensional i-vector method in terms of EER. And this is based on an existing i-vector approach, which is mentioned on secondary paper. The deep neural networks (DNN) have been incorporated into i-vector-based speaker recognition systems, where they have significantly improved state-of-the-art performance. In these systems, a DNN is used to collect sufficient statistics for i-vector extraction. In this study, the DNN is a recently developed time delay deep neural network (TDNN) that has achieved promising results in LVCSR tasks. They believe that the TDNN-based system achieves the best reported results on SRE10 and it obtains a 50% relative improvement over our GMM baseline in terms of equal error rate (EER). For some applications, the computational cost of a DNN is high. Therefore, they also investigated a lightweight alternative in which a supervised GMM is derived from the TDNN posteriors. This method maintains the speed of the traditional unsupervised-GMM, but achieves a 20% relative improvement in EER. According to this method, in first paper, they designed a new multi-stream i-vectors approach which used multiple extractors by training multiple T-matrixes of different dimensions for improving performance by combining the information contained in each i-vector. In addition, the acoustic score based VAD and splice normalization were applied for real-time processing. Experimental results demonstrate the superior performance and real-time implementation of our speaker recognition system. In addition, they had improved the usability of various applications by enabling the ASR system and our speaker recognition system to be linked in real time.

Relationship:

1. In primary paper, on the Introduction part, they mentioned a method that recurrent neural network - long short-term memory (RNN-LSTM) and time delay deep neural network (TDNN) are superior to DNN in the field of speech recognition, and they are being used more and more in the field of speaker recognition, which was proposal in second paper the DNN-based speaker recognition methods achieve excellent results, but the performance comes at the cost of increased computational complexity.
2. In primary paper, they used an approach which is the existing i-vector approach, generation of UBM and T-matrix through Baum-Welch statistics based on GMM or neural network such as DNN are performed in Multi-stream I-vectors approach part. The method was described on second paper. The goal of the supervised-GMM (shortened to sup-GMM) is to model phonetic content in a lightweight model. This is achieved by creating a GMM based on DNN posteriors and speaker recognition features. In contrast to the similar model, their sup-GMM is full-covariance. The supervised and unsupervised GMMs differ only in the UBM training procedure.

3. In primary paper, on the Experiment and Results part, they indicated the i-vector approach based on neural network showed superior speaker recognition performance than GMM based i-vector approach. The good strategy was talked in GMM baseline part on second paper. They used The UBM in our baseline system is a full-covariance GMM with several thousand mixture components. They compared systems with 2048, 4096, and 5297 components. The front-end consists of 20 MFCCs with a 25ms frame-length. The features are mean-normalized over a 3 second window. Delta and acceleration are appended to create 60 dimensional frame-level feature vectors. The non-speech frames are then eliminated using energy-based voice activity detection (VAD).
4. In primary paper, on the Experiment and Results part, they confirmed that TDNN or RNN-based i-vector approach is superior to DNN-based speaker recognition. The approach was showed in TDNN-UBM part on second paper. the TDNN posteriors create the sufficient statistics for i-vector extraction. As in the other systems, the speaker recognition features are filtered using a frame-level VAD. However, in order to maintain the correct temporal context, they could not remove frames from the TDNN input features. Instead, the VAD results are reused to filter out posteriors corresponding to non-speech frames.

The four portions are relationship between these two papers.