University of Sheffield

Speaker Diarization System for the DIHARD Challenge

Mangesh Hambarde

Supervisor: Dr. Thomas Hain

A report submitted in fulfilment of the requirements
for the degree of MSc Computer Science with Speech and Language Processing

in the

Department of Computer Science

August 21, 2019

Declaration

All sentences or passages quoted in this report from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this project and the degree examination as a whole.

Name: _____

Signature: _____

Date: _____

Abstract

Speaker Diarization is the task of finding out "who spoken when?" given an audio recording. It is an important field because it is a crucial preprocessing step for many areas in speech processing like speech recognition. Over recent years, the availability of fast computing power and emergence of massive amounts of multimedia data has boosted the field. The demand for various kinds of speech technology, and hence diarization, has become greater than ever. The performance of diarization systems has also improved significantly in the past decade or so thanks to the continued efforts of the research community. But even so, the absolute numbers tell a different story and it seems that there is still a long way to go. There is still a need for big improvements so that speaker diarization technology can be deployed in real world applications. This makes it an exciting research area which has a lot of potential to improve in the next few years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speaker diarization is the task of finding out "who spoke when?" in an audio recording with an unknown amount of speakers. It aims to find all segments of speech within the recording, possibly overlapping, along with their intra-recording speaker identities. It acts as an important upstream preprocessing step for most tasks in speech processing, like speech recognition, speech enhancement, speech coding etc.

With increase in computing power, speech processing technologies have achieved incredible advances in the past decade that were not possible earlier. This has increased interest in Rich Transcription (RT) technologies that can be used to automatically index the enormous amount of audio and video information that is generated in the modern world. Since speaker diarization is an important part in any RT system, there is a great deal of research interest in the area.

Diarization is not an easy problem since the output is affected by several factors like the application domain (broadcast news, meetings, telephone audio, internet audio, restaurant speech, clinical recordings etc), types and quality of microphones used (boom, lapel, far-field), inter-channel synchronization problems, overlapping speech, etc. These days, most of the research focuses on the meeting speech domain, since most problems that exist in speech recognition are encountered in this domain. The meeting scenario is thus often termed as "speech recognition complete".

The DIHARD challenge was created to establish standard datasets for diarization and create performance baselines for comparison, thus encouraging further research. The challenge focuses on "hard" diarization, combining several domains of speech like broadcast speech, meeting speech, telephone speech, and many more. Creating a system for the challenge can be a rewarding experience since it gives a chance to learn about state-of-the-art speaker diarization techniques.

1.1 Speaker Diarization

1.2 Motivation and Objectives

1.3 Report Outline

Chapter 2

Literature Review on Speaker Diarization

2.1 Introduction

2.2 Feature Extraction

2.3 Speech Activity Detection

2.4 Segmentation

2.5 Clustering

2.5.1 Speaker Representation

GMM

i-vectors

x-vectors

2.5.2 Agglomerative Clustering

2.5.3 Distance metrics

BIC

PLDA

2.6 Evaluation

Diarization Error Rate

Jaccard Error Rate

Chapter 3

DIHARD challenge setup

3.1 Task definition

3.2 Evaluation Tracks

3.3 Scoring

3.4 Datasets

Chapter 4

Baseline setup

4.1 Overview

4.2 Kaldi toolkit

4.3 Speech Activity Detection

4.4 Speaker Representation

4.5 Segmentation and Clustering

Chapter 5

Experiments and Results

5.1 Baseline results

5.2 Using Existing Pre-trained Models

5.2.1 Kaldi VoxCeleb x-vector model

5.2.2 Kaldi VoxCeleb i-vector model

5.3 Training Models on In-Domain Data

5.3.1 No Augmentation

5.3.2 Augmentation with Noise and Reverberation

5.4 Manual Speech Activity Detection

5.4.1 WebRTC

5.4.2 SHoUT toolkit

5.4.3 Energy-based

5.4.4 DNN-based

5.5 Speech Enhancement

5.5.1 Baseline

5.6 Alternative Distance Metrics for Clustering

5.6.1 BIC

5.6.2 Cosine

Chapter 6

Conclusions