LEAP Diarization System for the Second DIHARD Challenge

Prachi Singh[1], *Harsha Vardhan M A*[1], *Sriram Ganapathy*[1], *Ahilan Kanagasundaram*[2]

[1]Learning and Extraction of Acoustic Patterns (LEAP) Lab, Indian Institute of Science.
[2]University of Jaffna, Sri Lanka.

{prachisingh,sriramg}@iisc.ac.in, harshavardhan.ma@gmail.com,ahilan@eng.jfn.ac.lk

Abstract

This paper presents the LEAP System, developed for the Second DIHARD diarization Challenge. The evaluation data in the challenge is composed of multi-talker speech in restaurants, doctor-patient conversations, child language acquisition recordings in home environments and audio extracted YouTube videos. The LEAP system is developed using two types of embeddings, one based on i-vector representations and the other one based on x-vector representations. The initial diarization output obtained using agglomerative hierarchical clustering (AHC) done on the probabilistic linear discriminant analysis (PLDA) scores is refined using the Variational-Bayes hidden Markov model (VB-HMM) model. We propose a modified VB-HMM model with posterior scaling which provides significant improvements in the final diarization error rate (DER). We also use a domain compensation on the i-vector features to reduce the mis-match between training and evaluation conditions. Using the proposed approaches, we obtain relative improvements in DER of about 7.1% relative for the best individual system over the DIHARD baseline system and about 13.7% relative for the final system combination on evaluation set. An analysis performed using the proposed posterior scaling method shows that scaling results in improved discrimination among the HMM states in the VB-HMM.

Index Terms: Speaker Diarization, i-vector, x-vector, HMM-VB, PLDA.

1. Introduction

Speaker diarization, the task of identifying who spoke when in a multi-talker speech recording, is receiving increased attention in the recent years. It has several potential applications such as surveillance, forensics, information retrieval, rich transcription for automatic speech recognition (ASR) systems and in call center applications. The task of speaker diarization is challenging in noisy and channel degraded environments where speech is corrupted by background noise. The variations in domain/speaking style of the speech also affect the diarization performance [1].

In the previous decade, NIST had performed a series of evaluations in the topic of speaker diarization on meeting room conditions [2]. The diarization tasks on other domains like broadcast news were also pursued [3]. However, most of these diarization system development efforts focused on systems that were aimed to operate on isolated domains. The first among the series of recent initiatives to benchmark diarization systems, the first DIHARD challenge, proposed a challenge where the performance of a diarization system was evaluated on a range of complex realistic operational scenarios. The foundational work for this evaluation came up from an analysis of the challenging scenarios for state-of-art diarization systems [4] from doctor-patient conversations, child language acquisition data [5], meeting speech, dinner party conversations in restaurants etc. In addition, the more comfortable evaluation criterion which simply ignored overlapping speech regions and had a liberal collar were updated for the DIHARD challenge with an error evaluation on overlapping regions without any collar. This initiative has been revamped with more challenging recordings in the second DIHARD challenge [6]. This paper describes the speaker diarization system development efforts by the LEAP team.

For the first DIHARD challenge, many of the successful systems used the neural network based speech embedding (x-vector) representation [7]. The x-vector representations replaced the previously employed GMM based i-vector features and were extracted for fixed length chunks of duration 1.5sec. The pairwise scores on segment x-vectors are computed using a probabilistic linear discriminant analysis (PLDA) scoring [8]. The PLDA score matrix is clustered with a agglomerative hierarchical clustering (AHC) [9]. The AHC clusters are further refined using a Variational-Bayes hidden Markov model (VB-HMM) to improve the diarization performance [10].

In the DIHARD-I challenge, a previous effort had explored the use of i-vectors for estimating the number of speakers [11]. Another approach used a neural network based domain classifier and optimized the diarization system on each domain separately [12]. The speaker diarization approach using binary shift keying was explored by [13]. The x-vector approach with HMM-VB refinement and the fusion with i-vector representations was attempted in [14].

In this paper, our major contributions are,

- Posterior scaling for VB-HMM - In this approach, we boost the zeroth order statistics before the VB-HMM likelihood computation. This posterior scaling improves the speaker separation in the VB-HMM model which results in significant reduction in the overall DER.
- Domain compensation - The i-vector embeddings have a mismatch between training conditions and the DIHARD development dataset which can be compensated using variance normalization.

The rest of the paper is organized as follows. Section 2 introduces the dataset used in training and testing the models. Section 3 describes the x-vector baseline system. In Section 4, we describe the proposed approaches which improves over baseline including posterior scaled HMM-VB model (section 4.1) and i-vector domain mismatch compensation (section 4.2) for speaker diarization. Section 5 describes the systems implemented by LEAP team using proposed approaches. In Section 6, we report the experiments and results on the DIHARD challenge. This is followed by a summary of the work in Section 7.

2. Dataset

DIHARD II track 1 single channel development dataset, which is a superset of DIHARD I development dataset is garnered