The DKU-Duke-Lenovo System Description for The Third DIHARD Speech Diarization Challenge

Weiqing Wang, Qinjian Lin, Danwei Cai, Ming Li

***Abstract*—In this paper, we present the submitted system for the third DIHARD Speech Diarization Challenge from the DKU-Duke-Lenovo team. Our system consists of several modules: voice activity detection (VAD), segmentation, speaker embedding extraction, agglomerative hierarchical clustering. In addtion, the target speaker VAD (TSVAD) is used for the phone call data to further improve the performance. Our final submitted system achieves a DER of 15.43% for core evaluation set and 13.39% for full evaluation set on task 1, and we also get a DER of 21.63% for core evaluation set and 18.90% for full evaluation set on task 2.**

For task 1, we use Voxceleb 1 & 2 to training a model to extract the embedding for 8k and 16k wav. Ami meeting corpus and icsi meeting corpus are used for attention-based similarity measurement, and the DIHARD III development set is used for finetuning. For TSVAD, the model is trained on a collection of SRE-databases including SRE 2004, 2005, 2006, 2008 and Switchboard. Then we finetune this model on the first 41 telephone speech data and validate on the remaining 20 data in the DIHARD III development set.

For task 2, an additional VAD model is directly trained on the DIHARD III development set, where 90% of the dev set is for training and the remaining is for validation.

***Index Terms*—Speaker Diarization, Speaker Recognition, Deep learning**

I. Introduction

References