Extended Abstracts of RoyalFlush Submission in Third DIHARD Challenge

Chenyi Yu, Guishan Wang, Yuhang Chen, Xinhui Hu, Xinkang Xu

Hithink RoyalFlush AI Research Institute, Zhejiang, China

{yuchenyi,wangguishan,chenyuhang,huxinhui,xuxinkang}@myhexin.com

Abstract

This paper describes the speaker diarization systems submitted for the Third DIHARD Challenge from the RoyalFlush team. The speaker diarization systems include multiple modules. For each module, we explored different techniques to enhance performance. Our final submission employed the TDNN-F based SAD, the Deep x-vector based speaker embedding, the probabilistic linear discriminant analysis (PLDA) estimation based similarity measure and agglomerative hierarchical clustering (AHC). Variational-Bayes hidden Markov model (VB-HMM) is applied in the resegmentation stage. Furthermore, overlap detection also brings slight improvement. In Track 1, our system achieved the DER of 17.59% in the CORE set and 15.59% in the FULL set of the evaluation. In Track 2, our system achieved the DER of 19.56% in the CORE and 17.68% in the FULL sets of the evaluation, ranking the first and second on the leaderboard among all participants.

Index Terms: DIHARD, SAD, speaker embedding, similarity measure, clustering, resegmentation, overlap detection

1. Introduction

Speaker diarization is the task of determining "who spoke when" in an audio file that usually contains an unknown number of speakers with variable speech duration. In Third DIHARD Challenge, tracks 1 and 2 consisted of performing diarization on single-channel recordings from different domains with and without oracle speech activity detection (SAD). We took part in these tracks, all components of our submitted systems will be correspondingly introduced in the following parts.

2. System for Track 1

System of Track 1 was provided with a reference speech segmentation from the challenge committee. The following components are used in our system.

2.1. Data preparation

The clean training data is composed of VoxCeleb1 and 2 containing over 1.2 million utterances from 7323 speakers. Kaldi data augmentation was performed on the clean data with the MUSAN and RIRS-NOISES corpora. We randomly selected utterances of the same scale as the clean sets from the above augmented data set and then combined them with the clean sets. Then, utterances that are shorter than 4 seconds and spoken by speakers with less than 8 utterances were all discarded. In all, the training set for x-vector extractor contains over 2.5 million utterances and 5,500 hours.

2.2. X-vector extractor

The x-vector extractor system was constructed using the Kaldi toolkit. The acoustic feature for it is the FBANK with 40 dimensions and 16kHz sampling frequency. At the training stage, an energy-based VAD was used since allowing a certain amount of noise during training helps improve the robustness of neural networks. We used 200 frames in all training segments and generated 120 Kaldi archives. The Deep Neural Network-based x-vector extractor was trained with a minibatch size of 128, an initial learning rate of 0.001 and a final learning rate of 0.0001 for 6 epochs. At the testing stage, from the input test recordings, x-vectors were extracted every 0.25s from 1.2s overlapping subsegments for tracks 1 and 2 since we found that it improved the performances significantly. The architecture of the network for x-vector extraction basically followed the BUT System Description for DIHARD Speech Diarization Challenge 2019.

Table 1: *Results in Track 1*

		Dev		Eval	
		DER	JER	DER	JER
Baseline	CORE	20.25	46.02	20.65	47.74
	FULL	19.41	41.66	19.25	42.45
RoyalFlush	CORE	15.16	41.87	17.59	42.50
	FULL	13.78	33.03	15.59	37.40

2.3. PLDA and AHC

The PLDA model was used to calculate the similarity scores for AHC. AHC was performed as the binary-tree building process. The PLDA model was trained on x-vectors extracted from the 3s speech segments of 256,000 utterances which are a subset of the clean training data. Before the PLDA training, the x-vectors were centered, whitened, and length-normalized. The centering and whitening transformation were estimated on the joint set of DIHARD development and evaluation data. We used PLDA model to calculate log-likelihood ratio scores as a similarity metric for each pair of x-vector from recording, and then put the similarity metric into AHC to merge similar items according to the scores. However, in order to process these steps properly, we need to set a threshold to end the AHC process. By our experiments, we found that the best threshold is -0.1 for this system.

2.4. Variational Bayes HMM (BHMM)

Variational BHMM at x-vector level was used to cluster x-vectors. We take the AHC results as initial labels for each segments in BHMM. Before that, the x-vector was projected using Latent Discriminant Analysis (LDA) into a 260-dimensional space. Iterative VB inference was run until convergence to update the assignment of x-vectors to speaker clusters. Through the process of BHMM, we resegment the assignment for each speaker clusters. In our system,we got a group of parameters of the BHMM model that get the best value in the development set: the speaker regularization coefficient F_b was 12, which controls

Table 2: *The architecture of SAD model*

Layer	Layer Type	Context factor	Size
1	TDNN-ReLU	t-2:t+2	256
2	TDNNF-ReLU	t-3,t,t+3	256
3	TDNNF-ReLU	t-3,t,t+3	256
4	TDNNF-ReLU	t-3,t,t+3	256
5	TDNNF-ReLU	t-3,t,t+3	256
6	TDNNF-ReLU	t-3,t,t+3	256
7	Stats Pooling	full-seq	2*256
8	TDNN-ReLU	Layer6(t-6,0, t+6,t+12),Layer7	256
9	TDNNF-ReLU	t-3,t,t+3	256
10	TDNNF-ReLU	t-3,t,t+3	256
11	TDNNF-ReLU	t-3,t,t+3	256
12	TDNNF-ReLU	t-3,t,t+3	256
13	TDNNF-ReLU	t-3,t,t+3	256
14	Stats Pooling	full-seq	2*256
15	TDNN-ReLU	Layer13(t-12,0, t+12,t+24),Layer14	256
16	Dense-Softmax		3

the BHMM to find a right number of speaker clusters; acoustic scaling factor F_a was 0.4, which was compensated for the incorrect assumption of statistical independence between observations; the probability of not changing speakers when moving to the next observation P_{loop} is 0.7.

2.5. Overlap detection

Our overlap detection system included two parts: a deep neural network for obtaining frame-level scores and a back-end processing for scoring results. The structure of the neural network in the overlap system was the same as the SAD module which will be introduced in Section 3. The overlap detection model uses the overlapped information related to the development set as the training label when fine-tuning it. The rest of the training process and parameters were the same as the SAD module. At the back end, we used empirical knowledge and considered the segment of the overlap as the speech overlap interval after the softmax output threshold is greater than 0.8 and the number of consecutive frames exceeds 30 frames.

2.6. Results and analysis

The performances of our submitted system on the development and evaluation sets are shown in Table 1. We achieved the DER of 15.16% in the CORE and 13.78% in the FULL sets of the development set, outperforming the baseline with 22.91% and 29.01% relative reductions, respectively. Our proposed system achieved the DER of 17.59% in the CORE and 15.59% in the FULL sets of the evaluation set, outperforming the baseline with 14.82% and 19.01% relative reductions, respectively.

3. System for Track 2

System of Track 2 was provided with just the raw audio input for each recording session. Different from the Track 1 in which the reference speech segmentations were provided by the challenge, in Track 2, the speech segmentation was performed by the challenge participants themselves. Our submitted system employed an NN-based SAD model to generate speech or nonspeech labels for the development and evaluation sets. The other components of the submitted system in Track 2 were the same as the Track 1.

Table 3: *Results in Track 2*

		Dev		Eval	
		DER	JER	DER	JER
Baseline	CORE	22.28	47.75	27.34	51.91
	FULL	21.71	43.66	25.36	46.95
RoyalFlush	CORE	17.22	43.61	19.56	44.02
	FULL	16.01	39.22	17.68	39.37

3.1. SAD module

The SAD module includes following two parts: a neural network for obtaining frame-level scores and a back-end processing for scoring results. The neural network mainly used Fisher corpora upsampled to 16kHz for pre-training, and then it used development and evaluation sets for fine-tuning. The architecture of the network for SAD model is shown in Table 2, and the bottleneck neuron numbers of TDNN-F layers were all set to 96. In the pre-training stage, the SAD deep neural networks were trained with a minibatch size of 128, an initial learning rate of 0.0003 and a final learning rate of 0.00003 for 40 epochs. In the fine-tuning stage, we used the same training parameters as the pre-training stage except for the fixed learning rate of 0.00003. At the back end, we basically followed the methods used in the LEAP submission system. However, in this system, we used the relevant decoding parameters with the acoustic likelihood threshold greater than 0.5 and the minimum speech segment 0.1 in the segmentation of valid speech.

3.2. Results and analysis

The performances of our submitted system on the development and evaluation sets are shown in Table 3. We achieved the DER of 17.22% in CORE and 16.01% in FULL sets of the development set, outperforming the baseline with 22.71% and 26.26% relative reduction, and achieved the DER of 19.56% in CORE and 17.68% in FULL sets of the evaluation set, outperforming the baseline with the 28.46% and 30.28% relative reduction.

4. Conclusions

We provide an overview of the submitted systems to the Third DIHARD Challenge. We have built a relatively complete system of speaker diarization and optimized each sub-module. The details about our systems are described. On the development set, we achieved the DER of 15.16% in CORE and 13.78% in FULL sets for the Track 1, and we achieved the DER of 17.22% in the CORE and 16.01% in the FULL sets for the Track 2, respectively. On the evaluation set, we achieved the DER of 17.59% in CORE and 15.59% in FULL sets for the Track 1, and we achieved the DER of 19.56% in the CORE and 17.68% in the FULL sets for the Track 2, ranking the first and second on the leaderboard among all participants. Although we have achieved promising results in the Track 2, there are still a lot of challenges in many complex scenarios such as multiple speakers and noisy environment, which will be the direction of our future work.