XMUSPEECH System for the Third DIHARD Challenge

Jie Wang[1], Hao Lu[2], Lin Li[1], Qingyang Hong[2]
[1]*School of Electronic Science and Engineering, Xiamen University, China*
[2]*School of Informatics, Xiamen University, China*
lilin@xmu.edu.cn,qyhong@xmu.edu.cn

***Abstract*—The main task of speaker diarization is "who spoke when". As the front-end of many speech systems, speaker diarization system plays an important role in the field of Automatic Speech Recognition. Speakers diarization system used to take i-vector,d-vector,x-vector as embeddings, of which x-vectors were extracted by time-delay neural network (TDNN) extractor. Then, the resulting PLDA model is used to calculate log-likelihood ratio verification scores as a similarity metric for each pair of x-vectors from the test records. According to the score metric, those x-vectors were clustered into many classes. In this paper, we used Residual Neural Network (ResNet) instead of time-delay neural network to extract ResNet vectors of segments. Testing on DIHARD3 corpus shows that the performance of the system is better than that of the system using time-delay neural network model as embeddings extractor.**

***Index Terms*—speaker diarization, Residula Neural Network, dihard3, ResNet vectors**

I. NOTABLE HIGHLIGHTS

The speaker diarization system use ResNet vectors as embeddings. ResNet vectors are similar to x-vectors, but replace Time-delayed Neural Networks (TDNN) [1] with ResNet as front-end. ResNet extractor has three components including Resnet34 [2] front-end, statistical pooling layer and AM-softmax layers. Embedding layers, between statistical pooling layer and AM-softmax layer, include near embeddind and far embedding. After experiment, near embedding was chosen as ResNet vectors. ResNet solves the problem of gradient disappearance caused by increasing the depth of the neural network. Because ResNet combine the output of the network with the original input as the current output of neural networks.

II. DATA RESOURCES

A. training dataset

We take VoxCeleb2 [3] dataset to train ResNet extractor.VoxCeleb2 contains over 1 million utterances for 6,112 celebrities, extracted from videos uploaded to YouTube. The development set of VoxCeleb2 has no overlap with the identities in the VoxCeleb1 or SITW datasets. The dataset is fairly gender balanced, with 61% of the speakersmale. The speakers span a wide range of different ethnicities, accents, professions and ages. Videos included in the dataset are shot in a large number of challenging visual and auditory environments. In order to judge if two audio segments belong to the same speaker or not, we need to train a probabilistic linear discriminant analysis (PLDA) [4] model. PLDA model also trained with VoxCeleb2 development dataset.

B. test dataset

We used DIHARD3 [5] as the test dataset which is the third in a series of diarization challenges focusing on "hard" diarization. The dataset including devlopment dataset consist of selections of 5-10 minute duration samples drawn from 11 domains. A development set contains 34.15 hours of utterances and a evaluation data includes 33.01 hours of utterance. Significantly, DIHARD3 dataset was defined as two parts, one named the core data is a "balanced" dataset in which the total duration of each domain is approximately equal, the others named full data is a larger evaluation set that uses all available selections for each domain and which is, thus, unbalanced with some domains having more audio than others.

III. DETAILED DESCRIPTION OF ALGORITHM

A. signal processing

The VoxCeleb2 dataset were augmented in different ways including noise, reverberation and perturbation. 16khz dataset was augmented with noise, babble and music from the MUSAN corpus. In addition, we use the MUSAN dataset, which consists of over 900 noises, 42 hours of music from various genres and 60 hours of speech from twelve languages. What's more, we use the simulated RIRs, and the reverberation itself is performed with the multi-condition training tools in the Kaldi ASpIRE recipe. Meanwhile, we perturbed the speed of original data with 0.9 times and 1.1 times.

B. acoustic features

We extract 81-dimensional FilterBank(fbanks) from 16khz augmented VoxCeleb2 dataset, with 25 length and 10 ms step. Before training ResNet34 model, cepstral mean and variance normalization (CMVN) methods had been applied to the feature.

C. segments representation

- Segments: before extracting ResNet vector, utterance are divided into sub-segments(as given by the reference Voice Activity Detection (VAD)) every 0.75 seconds from windows of 1.5 seconds. That means those segments were overlapped by 0.75 seconds.

- ResNet-vector extractor: deep ResNet vector is employed as the speaker embedding. We take VoxCeleb2 corpora and their augmented data to train ResNet34 2D, witch include 5994 speakers in total. Statistics pooling layer connects after ResNet34 front-end. We chose near layer as embeddings layer of which dimension is 256. Specially, we used additive margin softmax (AM-softmax) [6] function as the measurement loss. The networks train 6 epoch.
- Back-ends system: the back-ends consisted of zero-means, whitening, length normalization and PLDA. The PLDA model is used to calculate the similarity scores for agglomerative hierarchical clustering(AHC) [7]. PLDA model parameters were also learned from the VoxCeleb2 dataset. Then we used the similarity measurement algorithms to compute scores between any two speaker embeddings in the sequence.

D. clustering method

The output of PLDA model is used as the input of AHC algorithm.The clustering is essentially the same as implemented in the offificial Kaldi diarization recipe. We applied AHC on the score matrix. In order to stop the AHC algorithm, we set a threshold by assuming that there are no more than 10 speakers.Then we can make sure all the segments clustered into less than 10 class.

E. Results on the development set

a) Results using our system in the third DIHARD evaluation: We used to test the performance of time-delay neural network (TDNN) extractor system of which EER is 26.74 on evaluation full Dataset.

TABLE I
RESULTS USING OUR SYSTEM IN THE THIRD DIHARD EVALUATION

Task	DER	JER
Evaluation Full	19.870%	42.320%
Evaluation Core	22.070%	48.390%

b) Other systems perform on DIHARD3 development: We also used to test the performance of resegmentation on the third DIHARD development dataset. The algorithms of resegmentation are Variational Bayes (VB) and Bayesian HMM clustering of x-vector sequences (VBx). VB and VBx use clustered result to initial resegmentation model in common, but VBx replaced the i-vectors which was applied in VB model with x-vectors. The details of VB and VBx were refered to [8], [9]. The result of x-vector tdnn with VBx was the best system on DIHARD3 development dataset. We can infer that VBx resegmentation can improve the system performance well.

TABLE II
OTHER SYSTEMS PERFORM ON THE THIRD DIHARD DEVELOPMENT

System	DER
x-vector tdnn	26.74%
ResNet-vector + vb	28.24%
x-vector tdnn + vbx	20.31%

F. Hardware requirements

- The infrastructure used to run the experiments was a CPU,Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz, with a total memory of 35GB.
- ResNet extractor model was trained by GPUs(Tesla P100 PCIe 16GB) for 4 days. We train the thin Resnet34 model with AM-softmax loss and 4 GPUs were used to accelerate training. In addition, we used PyTorch 1.6.0 as machine learning frameworks.

REFERENCES

[1] Snyder D , Garcia-Romero D , Sell G , et al. X-Vectors: Robust DNN Embeddings for Speaker Recognition[C]// ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[2] He K , Zhang X , Ren S , et al. Deep Residual Learning for Image Recognition[J]. 2016.

[3] J. S. Chung*, A. Nagrani*, A. Zisserman ¡VoxCeleb2: Deep Speaker Recognition¿

[4] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in 2014 IEEE Spoken Language Technology Workshop (SLT), Dec 2014, pp. 413–417.

[5] Ryant N , Church K , Cieri C , et al. Third DIHARD Challenge Evaluation Plan[J]. arXiv e-prints, 2020.

[6] Feng, Wang, Jian, et al. Additive Margin Softmax for Face Verification[J]. IEEE Signal Processing Letters, 2018.

[7] Patrick Kenny, Douglas Reynolds, and Fabio Castaldo, "Diarization of Telephone Conversations using Factor Analysis,"IEEE Journal of Special Topics in Signal Processing, vol. 4,no. 6, pp. 1059–70, December 2010.

[8] Diez, Mireia, et al. "Analysis of speaker diarization based on bayesian hmm with eigenvoice priors." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2019): 355-368

[9] Landini, Federico, et al. "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks." arXiv preprint arXiv:2012.14952 (2020).