HYU Submission for the VoxCeleb Speaker Recognition Challenge 2022

Jeong-Hwan Choi[*]*, Ye-Rin Jeoung*[*]*, Jehyun Kyung, Ju-Seok Seong, and Joon-Hyuk Chang*

Department of Electronic Engineering, Hanyang University, Seoul, Republic of Korea

{brent1104, jyr0328, jehyunkyung, as2835510, jchang}@hanyang.ac.kr

Abstract

This report describes HYU submission to track 3 and 4 of the Voxceleb Speaker Recognition Challenge 2022 (VoxSRC-22). Track 3 focuses on semi-supervised domain adaptation for speaker verification. We fine-tune the pre-trained models that fit with English with Chines labeled data and generate the pseudo labels of Chines unlabeled data with iterative clustering. We fine-tune the pre-trained models again for domain adaptation by using the real and pseudo labels. In track 4, a speaker diarization task, we apply energy-based voice activity detection to overlapped speech and extract speaker embeddings by sliding the time frames using the pre-trained model of speaker embedding extractor. We employ spectral clustering with an attention-based embedding aggregation method to log the speech timestamps and tagging with speaker-specific labels. Our best-submitted score to the challenge achieved 11.23% and 9.44% in equal error rate and diarization error rate, respectively, on the VoxSRC-22 track 3 and 4 test set.

Index Terms: VoxSRC, speaker verification, semi-supervised learning, speaker diarization

1. Introduction

The VoxCeleb Speaker Recognition Challenge 2022 (VoxSRC-22) has been held with various speaker verification tasks, including a speaker diarization task [1, 2, 3]. It contributes to the development of advanced technology that highly represents the characteristics of each speaker in a segment and frame level of speech. This year four tracks are opened, track 1 and track 2 are fully supervised speaker verification tasks, and the tracks are split into whether there are restrictions on training data. Track 3 is a semi-supervised speaker verification task using the Chinese dataset, and domain adaptation from English to Chinese is designated a sub-goal. Track 4 order to determine "who talked when," speaker diarization divides multi-speaker audio into homogeneous segments representing a single speaker. Track 4 is an open track that allows all data for training, and the difference from last year is that some errors in the valid set are corrected.

This report summarizes the Hanyang University solution for the VoxSRC-22 tracks 3 and 4. Our training strategies for track 3 is divided into three steps. 1) prepare the pre-trained model using source domain data and adapting it with small source domain data that are labeled. 2) generate a pseudo-label to utilize unlabeled and large source domain dataset. 3) fine-tune the various pre-trained models with real and pseudo labels for all source domain data as targets, then perform score fusion. Moreover, in track 4, we develop our diarization system with four steps. 1) prepossess the audio mixture using the energy-based voice activity detection (VAD) module. 2) extract the speaker embedding from the pre-trained speaker embedding extractor. 3) cluster the speaker embeddings based on spectral clustering with applying attention-based embedding aggregation (AA) method [4] and scaling the affinity matrix. 4) estimate the final cluster labels and perform the scoring with dscore tool [5]. The following sections will describe our works in detail.

[*]Equal contributions.

2. Track 3 : Semi-supervised domain adaptation

2.1. Speaker embedding extractor

This work considers ECAPA-TDNN-L for all training steps and three different speaker embedding extractor architectures for the final training step. The ECAPA-TDNN-L was suggested in [6], which is constructed with TDNN architectures and emphasizes channel attention, propagation and aggregation. It also incorporated the squeeze-excitation blocks, multi-scale Res2Net features, and a different multi-stage aggregation method with channel-dependent attentive statistics pooling. We download the pre-trained model, which used 80-dimensional (80D) mel-filterbank energies (MFBEs), and the base channel and embedding size are set to 1,024 and 192, respectively, [7]. Extra speaker embeddings are ResNet-34 [8], Res2Net-34 [9], and BC-CMT-Base [10]. The ResNet-34 was widely used in image and sound classification and adopted good performance in the speaker verification field [11]. We download the pre-trained model from the [12], which is trained using 64D MFBEs. It has 512D speaker embedding and incorporates the squeeze-excitation blocks and original ResNet blocks. The Res2Net model advanced residual learning by dividing the channel dimension of ResNet into multiple scales. The model used in this work used hyperparameters to have four scales and widths of 16. This Res2Net-34 is identical to that described in [13]. we consider the BC-CMT as the last model which is a CNN-Transformer hybrid algorithm that is Incorporated with broadcasted residual learning [14] and computer-meets-vision-Transformer [15]. BC-CMT also proposed frequency-statistics-dependent attentive statistics pooling to effectively capture the speaker information in the frequency dimension. We use the BC-CMT-Base. Further details can be found in [10]. Both Res2Net-34 and BC-CMT-Base require 80D MFBEs and 256D speaker embeddings that are extracted from the penultimate layer with batch normalization. Note that VAD is not applied when extracting the MFBEs that are used in all models.

2.2. Dataset

VoxSRC-22 track 3 allows using the source and target domain datasets. The source domain dataset was the VoxCeleb2, which mainly comprises English. VoxCeleb2 [16] dev set has 5,994 speakers and 1,092,009 utterances that were allowed to use for training speaker embedding extractor. The target domain dataset is a portions of CN-Celeb2 [17] that are divided into an unlabeled and labeled data according to the challenge protocol. The unlabeled dataset contains over 450,000 utterances, and the

Table 1: *Results on HYU submission on VoxSRC-22 track 3 valid and test set*

Speaker embedding extractor	Model size	Training Step			VoxSRC-22 track 3 valid set				VoxSRC-22 track 3 test set			
					CSS		PLDA		CSS		PLDA	
		1	2	3	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
ECAPA-TDNN-L	15.7 M	✓	-	-	13.11	0.503	-	-	**12.98**	**0.619**	-	-
		-	1	-	12.98	0.511	-	-	-	-	-	-
		-	2	-	12.78	0.510	-	-	-	-	-	-
		-	3	-	12.64	0.520	-	-	-	-	-	-
		-	4	-	12.34	0.524	-	-	13.21	0.650	-	-
		-	-	✓	11.62	0.534	11.06	0.540	-	-	-	-
ResNet-34	8.0 M	-	-	✓	12.74	0.565	12.22	0.552	-	-	-	-
Res2Net-34	12.2 M	-	-	✓	12.22	**0.485**	10.24	**0.469**	-	-	-	-
BC-CMT-Base	6.3 M	-	-	✓	12.67	0.545	11.56	0.545	-	-	-	-
Fusion	-	-	-	-	**11.36**	0.500	**9.68**	0.475	13.48	0.647	**11.23**	**0.578**

labeled dataset contains 1,000 utterances of 50 speakers. Note that we does not consider the short utterances of less than 1 second in training. The number of utterances of the valid and test set provided by the organizer is 2,500 and 18,000, respectively, and the number of trials is 40,000 and 30,000, respectively. Data augmentation was performed by using MUSAN noise [18], babble, and music samples and simulation room impulse responses [19] in an on-the-fly manner.

2.3. Training strategies

*2.3.1. Step*1

VoxCeleb2 conducts mainly in English interview; thus, we adopted the ECAPA-TDNN-L to Chinese dataset with multiple genre to achieve the goal. We fine-tuned the model using labeled dataset of CN-Celeb2 to adjust with Chinese domain. We removed the last linear layer that had a size of number of Voxceleb2 speakers and connected a randomly initialized 50D linear layer. First, we freezed the rest of the ECAPA-TDNN-L, and then fine-tuned the last layer. This training schemes similar to those described in [20, 21]. For the fine-tuning, angular additive margin softmax (AAMsoftmax) loss function with scale of 30 was used for objective function. After last layer training 50 epochs with zero margin, we unfreezed the entire network and further proceeded the fine-tuning with the margin of 0.2. The total 120 epochs were set for fine-tuning with warm-up cosine scheduling, the learning rate linearly increased from 0 to 0.0001 in the first ten epochs and decreased to 0.

*2.3.2. Step*2

Although there were many studies on effectively using a large unlabeled dataset, there was a limit to improving performance without speaker information. To overcome this problem, we used the iterative clustering method that generates pseudo speaker labels, sets them as targets, performs training similar to supervised learning, and repeats the generation/learning process several times until valid performance converges. We used the ECAPA-TDNN-L trained in Stage 1 to extract the speaker embeddings for each utterance of the labeled and unlabeled dataset of CN-Celeb2. One embedding representing each speaker is obtained by averaging from labeled data, and the k-means clustering [22] is performed with embeddings of each utterance extracted from unlabeled data to get a pseudo label. We set the trained model as an initial model. The pseudo and real labels that form the unlabeled and labeled datasets were given as targets. Identically equal to step 1, we replaced the last layer with a linear layer of 2000 size in which the cluster number was set to 2000. Subsequently, all layers except for the last layer were frozen until the 50 epochs; after that, 70 epochs were learned by unfreezing all layers. The margin of AAMsoftmax was increased from zero to 0.2 at unfreeze for training when the layer was unfreezing. Learning rate warm-up was applied for the ten epochs, then decreased to zero with cosine annealing. We created a label and repeat the same process of learning using the label four times. After the last learning, we determined the final label for all the unlabeled dataset.

*2.3.3. Step*3

Returning to step 1, domain adaptation for CN-Celeb2 was performed for several speaker recognition models. The difference was that the training data size was very large compared to step1 because the final label was generated in step2. We used the ECAPA-TDNN-L, ResNet-34, Res2Net-16s4w, BC-CMT-Base for the speaker extractor model. The training strategy was identically same as in step 1, just like step 2.

2.4. Results

We extracted the speaker embeddings to test set and all training set that used in step 3. The speaker embeddings were divided by the overall average, then length normalization was applied. We evaluated the valid and test trials with cosine similarity (CSS), and probabilistic linear discriminant analysis (PLDA)[23, 24] for scoring methods. The PLDA was trained with the limitation that the within-speaker covariance was set to be a diagonal matrix in each iteration of the expectation-maximization step, as suggested in [25]. The results were evaluated in terms of the equal error rate (EER) and minimum of the detection cost function (minDCF) with target probabilities of 0.05. The experimental results were shown in Table 1. Comparing the experiment results with training step 3, ECAPA-TDNN-L was the best in the valid set, followed by Res2Net-34, BC-CMT-Base, and ResNet-34. In the test set, Res2Net-34 significantly outperformed the others, and ECAPA-TDNN-L, BC-CMT-Base, and ResNet-34 followed the Res2Net-34. We calculated the fusion score by adding the scores of all models to one, and the PLDA of the fusion model achieved the best performance by achieving 11.23% in EER.

Table 2: *Results on HYU submission on VoxSRC-22 track 4 valid set*

Case	Range	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	DER(%)	JER(%)
1		-	-	-	-	+0.4	+0.3	-	-	13.08	34.79
2		-	-	-	-	+0.4	+0.3	+0.2	-	12.25	34.63
3		-	-	-	-	+0.4	+0.3	+0.2	+0.1	12.19	34.12
4	scaling	-	-	-	+0.5	+0.4	+0.3	+0.2	+0.1	12.52	34.27
5	value	-0.1	-	-	-	+0.4	+0.3	+0.2	+0.1	12.19	34.54
6		-0.1	-0.2	-	-	+0.4	+0.3	+0.2	+0.1	**12.18**	**33.59**
7		-0.1	-0.2	-0.3	-	+0.4	+0.3	+0.2	+0.1	12.71	34.40
8		-0.1	-0.2	-0.3	-0.4	+0.4	+0.3	+0.2	+0.1	12.63	34.65

3. Track 4 : Speaker diarization

3.1. Speaker diarization based on spectral clustering

For clustering-based speaker diarization, our proposed model is composed of VAD, speaker embedding extractor, and clustering modules. Before extracting the speaker embeddings, the input audio frames are pre-processed using the energy-based VAD module. After then, we used ResNet-34 for the speaker embedding extractor proposed in [26]. This model uses 64D MFBEs as input and applies channel-dependent attentive statistics pooling to the multiple hierarchies of feature maps for extracting speaker embeddings. We extracted them from the penultimate layer after the pooling of this model.

Finally, we applied the clustering technique called spectral clustering (SPC) [27]. SPC is the algorithm for grouping the features using the manifold of embedding spaces. This algorithm calculates the affinity matrix with cosine similarity between two inputs. Then, eigen-decomposition to affinity matrix is carried out. Finally, k-means clustering is performed on the spectral embeddings to estimate the final cluster labels.

3.2. Technique for affinity matrix : AA and scaling

The SPC algorithms for diarization have limitations depending on the input features. For example, the SPC is sensitive to noises in the affinity matrix. Therefore we referred to [4], and applied the attention-based embedding aggregation (AA) to remove noises and outliers. To use this technique, We calculate the affinity matrix with cosine similarity for each embedding using the softmax function with temperature τ. Then, the embeddings are aggregated based on this affinity matrix. We expect this technology to be able to form appropriate clusters.

Furthermore, we added a process of scaling the affinity matrix, which completed the AA process to help it be divided more efficiently in the clustering process. The values of each element constituting the affinity matrix were scaled using a method of adding or subtracting a specific value from the value of a certain interval. Through this, the ambiguous values changed to clearly distinguished ones, and the embedding existence interval vaguely clustered in the clustering process was reduced.

3.3. Dataset

The speaker embedding extractor were trained using the dev partition of VoxCeleb2 dataset with data augmentation as mentioned in chapter 2.

We used dev set of VoxConverse [28] for valid set which provided from challenge which consists of 216 recordings and 20.3 hours in total. The number of speakers in one recording varies from 1 to 20. Based on the score result from the dev set, the result was evaluated with a test set and submitted it.

3.4. Experiments

The training epoch of the speaker embedding extractor was defined as the iterations over 30,000 mini-batches. The model weights were subject to ℓ_2-regularization with a scale of 0.01. This model was trained using the stochastic gradient descent optimizer with a learning rate of 0.01.

Based on various experiments, the hyperparameters for the AA technique and the range and degree of scaling were determined. as the iterations over 30,000 mini-batches. We used temperature τ as 20 to calculate the softmax function of the AA technique. Furthermore, unlike the paper [4] proposed by AA, we applied this process only once without iteration to prevent the affinity matrix from being severely deformed. Then we added and subtracted the value of the affinity matrix, which processed this AA technique. As shown in Table 2, the range of elements constituting the matrix was divided into 0.1 units, and scaled so that ambiguous values were converted into definite values. Through this scaling process, we tried to help the affinity matrix cluster the embedding of each speaker.

3.5. Results

We introduced a diarization system combining each module to produce a suitable result for this task. Diarization Error Rate (DER) which consists of miss speech, false alarm, and speaker confusion error and Jaccard error rate (JER) which is introduced for DIHARD II [29] that is based on the Jaccard index used as the evaluation measure. The dscore tool provided by the challenge was used to evaluate and score the model with DER and JER. We selected a system that achieved a DER of 12.18% based on its performance on the development set of VoxConverse used for verification. Finally, our system has achieved DER of 9.44% and JER of 42.23% on the test set, which shows a performance improvement of about 52% over baseline from the DER perspective.

4. Discussion and conclusions

Our work achieved EER and DER of 11.23%, and 9.44%, respectively, in tracks 3 and 4, which were the evaluation metrics of the challenge. We analyzed the experimental results and looked for points for improvement. As for ECAPA-TDNN-L results in Track 3, the test EER of step 1 was superior to step 2, and the valid EER of step 2 was better than that of step 3. We considered that selection of the pseudo label was not accurate. In step 3, when selecting a model, Res2Net was better than ECAPA-TDNN-L. In the recently conducted first-

place presentation of the CN-SRC speaker verification track, ECAPA-TDNN-L showed lower performance in CN-Celeb than the ResNet. We believe that experimenting with Res2Net will show better performance than ECAPA-TDNN-L. In track 4, the role of the affinity matrix is important in performing spectral clustering for diarization. Therefore, we made sure that the affinity matrix had clearer values by applying AA and scaling so that clustering was well performed for speaker embeddings located in ambiguous boundaries. Through these techniques, our proposed diarization system showed improved performance, showing that the affinity matrix with definite values is more helpful for the subsequent clustering process.

5. References

[1] J. S. Chung, A. Nagrani, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, and A. Zisserman, "Voxsrc 2019: The first voxceleb speaker recognition challenge," *arXiv preprint arXiv:1912.02522*, 2019.

[2] A. Nagrani, J. S. Chung, J. Huh, A. Brown, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, and A. Zisserman, "Voxsrc 2020: The second voxceleb speaker recognition challenge," *arXiv preprint arXiv:2012.06867*, 2020.

[3] A. Brown, J. Huh, J. S. Chung, A. Nagrani, and A. Zisserman, "Voxsrc 2021: The third voxceleb speaker recognition challenge," *arXiv preprint arXiv:2201.04583*, 2021.

[4] Y. Kwon, J.-w. Jung, H.-S. Heo, Y. J. Kim, B.-J. Lee, and J. S. Chung, "Adapting speaker embeddings for speaker diarisation," *arXiv preprint arXiv:2104.02879*, 2021.

[5] https://https://github.com/nryant/dscore/.

[6] K. D. Brecht Desplanques, Jenthe Thienpondt, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech*, 2020.

[7] https://github.com/lawlict/ECAPA-TDNN/.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 2016, pp. 770–778.

[9] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, p. 652–662, Feb. 2021.

[10] J.-H. Choi, J.-Y. Yang, Y.-R. Jeoung, and J.-H. Chang, "Improved CNN-Transformer using broadcasted residual learning for text-independent speaker verification," in *Proc. Interspeech*, 2022.

[11] Y. Kwon, H. S. Heo, B.-J. Lee, and J. S. Chung, "The ins and outs of speaker recognition: lessons from VoxSRC 2020," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2021.

[12] https://github.com/clovaai/voxceleb-trainer/.

[13] J.-H. Choi, J.-Y. Yang, Y.-R. Jeoung, and J.-H. Chang, "HYU submission for the SASV challenge 2022: Reforming speaker embeddings with spoofing-aware conditioning."

[14] B. Kim, S. Yang, J. Lee, and D. Sung, "Broadcasted residual learning for efficient keyword spotting," in *Proc. Interspeech*, 2021, p. 4538–4542.

[15] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, "Cmt: Convolutional neural networks meet vision transformers," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), June 2022, pp. 12 175–12 185.

[16] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018.

[17] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang, "Cn-celeb: multi-genre speaker recognition," *Speech Communication*, 2022.

[18] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[19] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2017, pp. 5220–5224.

[20] J. Thienpondt, B. Desplanques, and K. Demuynck, "The idlab voxsrc-20 submission: Large margin fine-tuning and quality-aware score calibration in dnn based speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2021.

[21] J.-Y. Yang, J.-H. Choi, and J.-H. Chang, "The HYU speaker recognition system for the SdSV challenge 2020," 2020.

[22] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, p. 129–137, 1982.

[23] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE/CVF International Conference on Computer Vision* (ICCV), 2007, pp. 1–8.

[24] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proc. European Conference on Computer Vision* (ECCV), 2006, p. 531–542.

[25] Q. Wang, K. A. Lee, and T. Liu, "Scoring of large-margin embeddings for speaker verification: Cosine or plda?" in *Proc. Interspeech*, 2022.

[26] J.-Y. Yang and J.-H. Chang, "Task-specific optimization of virtual channel linear prediction-based speech dereverberation front-end for far-field speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

[27] H. Ning, M. Liu, H. Tang, and T. S. Huang, "A spectral clustering approach to speaker diarization," in *Ninth International Conference on Spoken Language Processing*, 2006.

[28] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the conversation: speaker diarisation in the wild," in *Proc. Interspeech*, 2020.

[29] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The second dihard diarization challenge: Dataset, task, and baselines," in *Proc. Interspeech*, 2019.