The Speaker and Language Recognition Workshop (Odyssey 2022)
28 June - 1 July 2022, Beijing, China

Progressive contrastive learning for self-supervised text-independent speaker verification

Junyi Peng[1], *Chunlei Zhang*[2], *Jan "Honza" Černocký*[1], *Dong Yu*[2]

[1]Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia
[2]Tencent AI Lab, Bellevue, WA 98004, USA

Abstract

Self-supervised speaker representation learning has drawn attention extensively in recent years. Most of the work is based on the iterative "clustering-classification" learning framework, and the performance is sensitive to the pre-defined number of clusters. However, the cluster number is hard to estimate when dealing with large-scale unlabeled data. In this paper, we propose a progressive contrastive learning (PCL) algorithm to dynamically estimate the cluster number at each step based on the statistical characteristics of the data itself, and the estimated number will progressively approach the ground-truth speaker number with the increasing of step. Specifically, we first update the data queue by current augmented samples. Then, eigendecomposition is introduced to estimate the number of speakers in the updated data queue. Finally, we assign the queued data into the estimated cluster centroid and construct a contrastive loss, which encourages the speaker representation to be closer to its cluster centroid and away from others. Experimental results on VoxCeleb1 demonstrate the effectiveness of our proposed PCL compared with existing self-supervised approaches.

1. Introduction

Speaker verification (SV) aims to verify whether an unknown speech utterance belongs to a specific speaker. According to the restriction of content, speaker verification can be categorized into text-dependent speaker verification (TD-SV) and text-independent speaker verification (TI-SV) [1].

In recent years, deep neural networks have demonstrated the impressive successes in extracting discriminative speaker representation for SV task [2, 3, 4]. In order to enhance the discrimination of the speaker representation, most existing methods focus on designing effective neural network structures [5, 6] and loss functions (e.g. triplet loss [7], angular softmax loss [8], affinity loss [9]). These methods rely on speaker identity label of all training utterances. However, it is expensive to annotate on large-scale speech data. The insufficient labeled training data may limit the performance of SV systems, especially when the utterances are recorded with different devices.

To solve this problem, researchers started following a self-supervised learning (SSL) framework which leverages large amount of unlabeled data to learn the speaker representation from speech. The training objective of these methods is to discriminate the augmented sample against all other samples in the training dataset. In [10, 11], the momentum contrast (MoCo) was utilized to learn speaker embedding from speech segments with contrastive loss, where a queue and a moving averaged encoder were employed to maintain a dynamic memory queue. It provided a good performance both on self-build phone-recorded and VoxCeleb corpus [12, 13]. In [14], a mutual information based objective was proposed to identify the distances among data samples by reducing the gap between two different segments from the same utterance to the same class (positive) while separating two segments belonging to different classes (negative). Essentially, their methods rely on the instance discrimination based on binary classification: each training speech segment is regarded as an independent class, and the training criterion is set to discriminate its own transformed segment from a large number of other speech segments and their augmentations. Obviously, in a large-scale randomly composed set of negative pairs, it is likely that it contains some pairs with the same speaker identity. However, when calculating the contrastive loss, the importance of segments sharing similar speaker identity characteristics is neglected. These positive samples may have the potential to improve the quality of learned speaker representation.

In order to make full use of the unlabeled data and encode the similar speaker identity characteristics, some researchers turned to leverage a clustering-based training objective. In [15, 16], an iterative "clustering-classification" framework is adopted, where the pseudo labels are firstly produced by a clustering algorithm and then classification is performed based on these pseudo labels. This approach and its variants achieved comparable performances to fully supervised methods on the VoxCeleb dataset. In [17], a prototypical memory queue

[†] Work done during an internship at Tencent

10.21437/Odyssey.2022-3

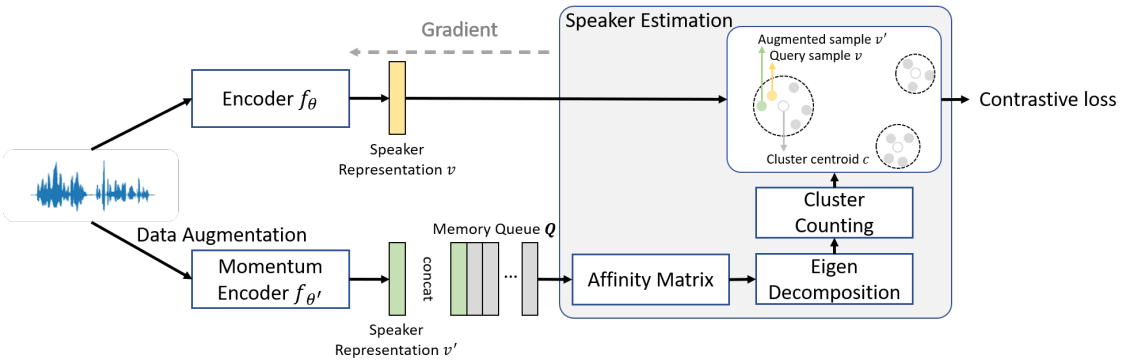


Fig. 1. The overview of the proposed self-supervised progressive contrastive learning framework.

(bank) was introduced to enforce the speaker embedding to be closer to their assigned prototype (centroid) with an intermediate clustering step. The prototype-based system outperforms MoCo-based system on the Voxceleb corpus. A potential shortcoming of these clustering-based methods is that the system performance is unstable and sensitive to the pre-defined cluster numbers. Moreover, the dogmatic cluster updating strategy leads to a poor generalization performance on unseen unlabeled datasets. In addition, clustering the whole large-scale dataset is time-consuming, and storing these pseudo labels and cluster centroids requires large amount of memory.

In this paper, based on our previous work [17], we propose a novel self-supervised framework, named progressive contrastive learning (PCL). PCL dynamically and efficiently estimates the number of clusters at each step based on the statistical characteristics of queued data itself, instead of a pre-defined number or generating from the whole dataset. Specifically, we first update the memory queue using current augmented mini-batch data. Secondly, we decompose the affinity matrix of the updated memory queue into eigenvalues to find potential factors for distinguishing speaker identity. Then, the number of speaker is estimated according to the maximum eigengap in each step. Finally, based on the estimated number, we cluster the queue data by K-means to form the contrastive loss which compresses the diversity of samples to their corresponding factors and disperses the difference of a sample to other factors, simultaneously. Extensive experiments are conducted on the VoxCeleb1 SV task. Results show that our proposed PCL outperforms typical self-supervised system based on MoCo [18], MOBY [19] and ProtoNCE [20].

The rest of paper is organized as follows: Section 2 gives a brief introduction to the instance-based learning (i.e. InfoNCE [21]). Section 3 describes the proposed progressive constrastive learning in detail. Experimental setup including database description, training paradigm, and result analysis are described in Section 4 and 5. Section 6 concludes the paper.

2. Related work

InfoNCE. The objective of self-supervised speaker representation learning is to learn a mapping $f(\cdot)$ that can effectively encode the intrinsic speaker information of an utterance for downsteam tasks, such as speaker verification, speaker diarization, etc. Instance-wise contrastive learning can accomplish this objective through optimizing a contrastive loss function (i.e. InfoNCE [21]), defined as:

$$L = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp(\mathbf{v}_i \cdot \mathbf{v}_i^{'}/\tau)}{\exp(\mathbf{v}_i \cdot \mathbf{v}_i^{'}/\tau) + \sum_{j=1}^{M} \exp(\mathbf{v}_i \cdot \mathbf{v}_j^{'}/\tau)}, \tag{1}$$

where $\mathbf{v}_i$ and $\mathbf{v}_i^{'}$ are query sample and corresponding transformed sample of instance i respectively, and $\mathbf{v}_j^{'}$ contains M negative samples for other samples. M is the size of a memory queue, and τ is a hyper-parameter temperature, N is the mini-batch size. In MoCo based SSL framework, $\mathbf{v}_j^{'}$ is extracted by a momentum encoder parameterized by $\theta^{'}$, which can be updated by a query encoder θ as $\theta^{'} = m\theta^{'} + (1-m)\theta$, where m is the momentum coefficient.

3. Progressive contrastive learning

The framework of our proposed progressive contrastive learning based SV system is illustrated in Fig.1. At first, the original speech segment and its augmented segment are fed into the encoder f_θ and momentum encoder $f_\theta^{'}$ to generate D-dimensional unit-norm speaker embedding vectors $\mathbf{v} \in \mathbb{R}^{1\times D}$ and $\mathbf{v}^{'} \in \mathbb{R}^{1\times D}$, respectively. Secondly, spectral clustering algorithm is used to estimate the number of speakers in the updated memory queue. Finally, contrastive loss is constructed based on the generated pseudo labels in the queue, aiming to encourage the query sample to be closer to its corresponding cluster centroid, compared to centroids of other clusters.

3.1. Speaker estimation

Since negative samples in the queue are randomly collected, the distribution of these samples may not be completely independent of the query sample, causing many negative pairs that may share the similar speaker identity information to be forced pulled away. Thus, the speaker representation is limited to encode the inherent speaker information of speech. To alleviate this issue, instead of obtaining a prior knowledge of the training data (i.e. an approximate estimation of total speaker number), in this paper, we first enqueue the current augmented mini-batch containing speaker vectors extracted from momentum encoder to the memory queue, and remove the oldest mini-batch in the queue. Then, we investigate the statistical characteristics of the updated queue data $\mathbf{Q} \in \mathbb{R}^{M\times D}$ by computing its affinity matrix $\mathbf{A} \in \mathbb{R}^{M\times M}$:

$$\mathbf{A} = \mathbf{Q}\mathbf{Q}^T, \tag{2}$$

where we set the diagonal elements of $\mathbf{A}$ to 0. To extract the main potential speaker identity characteristics, the normalized graph Laplacian matrix of the affinity matrix $\mathbf{A}$ is decomposed into eigenvalues and corresponding eigenvector matrix, where the sorted eigenvalues and eigenvector matrix are denoted as $s_1, ...s_M$ and $\mathbf{S} \in \mathbb{R}^{M\times M}$, respectively. Instead of a pre-defined cluster number, in this paper, we determine the cluster number k with the maximum eigenvalue gap:

$$k = \underset{1\le k\le M-1}{\operatorname{argmax}} \left(s_k - s_{k+1}\right), \tag{3}$$

In this way, the speaker number is dynamically generated from the queued data itself, instead of a well-designed hyper-parameter based on prior knowledge. Finally, we employ a clustering algorithm, such as K-means, to assign the queued samples into k clusters and estimate their corresponding centroids. During training, the model can learn more robust speaker representations progressively with a gradually precised estimation. In this way, the system may have a great generalization ability on additional or unseen large-scale data.

3.2. Contrastive learning

In order to simultaneously maximize the separability of speaker representations from different clusters and the compactness of those within the same cluster, based on our previous work [17], we redesign the ProtoNCE loss. The prototypes (centroids of clusters) are no longer generated from the whole dataset, which is time-consuming and crude. In this paper, we replace those with the queue-level centroids, which are more portable and effective, to investigate the speaker characteristics under the self-supervised learning framework. The modified ProtoNCE is formulated as:

$$L_{\mathrm{PCL}} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp(\mathbf{v}_i \cdot \mathbf{c}_i/\phi)}{\sum_{j=1}^{k}\exp(\mathbf{v}_i \cdot \mathbf{c}_j/\phi)}, \tag{4}$$

where $\mathbf{c}_i$ denotes the centroid of cluster i to which augmented sample $\mathbf{v}_i$ belongs, named positive centroid; otherwise negative centroid, and ϕ is the dynamically estimated temperature coefficient, suggesting the concentration of speaker representation around their centroids. ϕ is set to $\frac{\sum_{z=1}^{Z}\|\mathbf{v}_z - c\|}{Z\log(Z+\beta)}$, where Z is the total number of speaker embedding in this cluster, c is the cluster centroid, β indicates the smooth constant and is set to 10. Furthermore, following [20], we combine the InfoNCE and Eq.4 to retain the property of local smoothness and assist bootstrap clustering by a weight α. In our experiments, the α is fixed to 0.2, which results in the best performance in previous work [17]. The total objective can be written as:

$$L = \alpha L_{\mathrm{PCL}} + (1-\alpha)L_{\mathrm{InfoNCE}}, \tag{5}$$

3.3. Properties

In summary, the proposed self-supervised progressive contrastive learning has some interesting properties:
Dynamic cluster estimation. Different from the iterative learning framework, whose performance is sensitive to the pre-defined cluster number, our approach dynamically estimates the number of clusters in each step based on the statistical characteristics of training data itself. This progressive process makes the training more stable and has better generalization performance on the large-scale dataset.
More efficient. Our approach inherits the advantages of our previous work [17], which assembles the representations of utterance from a speaker to its estimated centroid. In addition, the proposed PCL only focuses on clustering the data in the memory queue rather than the whole dataset. With the increasing of iterations, the estimated cluster centroids will progressively be more robust, which makes the training process efficient and time-saving.

4. Experiments and Discussion

In order to fairly compare the experimental results, The experiments settings are consistent with the baseline from [17], only except for the objective function. Thus, the same network structure, data augmentation procedure, acoustic feature, training and testing strategies are utilized in our experiments.

4.1. Datasets

The SV performance is evaluated on the VoxCeleb corpus [12, 13] and CNCeleb [22] datasets, which both are widely used large-scale text-independent speaker verification dataset. The entire VoxCeleb dataset involves two

Table 1. Results for SV systems on the CNCeleb1 eval/test dataset. * It is noted that the ground-truth number of speakers is 800.

Model	Self-supervised Loss	Predefined Cluster	EER(%)	minDCF
ECAPA-TDNN	AAM-Softmax [supervised]		13.27	0.54
ECAPA-TDNN	without fine-tuning		16.56	0.77
	MoCo	-	15.58	**0.63**
	ProtoNCE	500	15.02	0.67
		800	15.18	0.68
		1500	17.32	0.77
		2000	41.98	0.99
	PCL	-	**14.95**	0.64

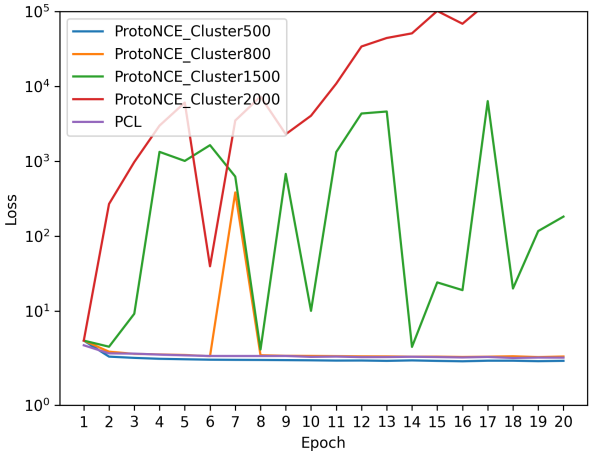


Fig. 2. The training loss on CNCeleb1-dev with different predefined number of clusters.

parts: VoxCeleb1 and VoxCeleb2. The utterances are collected from YouTube videos, where the celebrities belong to different nationalities and have a wide range of accents. The training set is derived from the development set of VoxCeleb2. In terms of CNCeleb dataset, the dataset contains more than 130,000 utterances from 1,000 Chinese celebrities. The total duration of utterance is around 274 hours. The training subset involves 800 speakers.

4.2. Implementation details

Network structure: For speaker verification, we use ECAPA-TDNN [4] as the trunk architecture, which is a modified version of the original Time Delay Neural Network (TDNN) [2]. To be specific, the frame-level feature extractor is restructured by 1-dimensional Res2Net modules and the utterance-level feature is aggregated from different hierarchical levels. The L2-norm output of the last hidden layer is extracted as the speaker embedding. No Linear Discriminant Analysis (LDA) nor Probabilistic Linear Discriminant Analysis (PLDA) is used.

Augmentation: Under the self-supervised framework, the learned speaker representation is encouraged to encode the intrinsic speaker information, which is robust to the noise. In this paper, to increase the diversity, we augment the original Voxceleb2 dataset using RIRs [23] and MUSAN datasets. Specifically, after random reverberation augmentation, we utilize three kinds of additive noise (noise, music, and babble) with different SNRs, to enrich the training data.

Feature: The acoustic features are 30-dimensional MFCCs with a frame length of 25ms. Mean-normalization is used at each feature dimension of the MFCCs. Also, an energy-based voice active detection (VAD) is used to detect speech frames.

Training: For VoxCeleb, our systems are optimized by SGD with learning rate of 0.01, which finally will be reduced to 0.0001 with a cosine learning rate scheduler (CosLR). The CosLR uses the cosine function to decrease the learning rate. The mini-batch size N is 1000, and the size of memory queue M is 10000. The momentum coefficient m is 0.999. The models are trained on 8 NVIDIA Tesla V100 GPUs for 150 epochs. In terms of unsupervised fine-tuning experiments on CNCeleb, we fine-tune all the models from a pretrained model, which is trained on VoxCeleb2-dev using self-supervised loss (ProtoNCE) with 20 epochs. We choose SGD as the optimizer with an initial learning rate of 0.005.

Metric: Equal error rate (EER) and minimum detection cost function (minDCF) are used to measure the performance of SV system. We use the same parameters as [12], where the target probability P_{tar} is 0.01, C_{fa} and C_{fr} have the same weight of 1.0. We adopt the cosine similarity as backend for all comparison systems.

5. Results and Discussions

5.1. Comparison among different loss functions using pre-trained model

Table 1 presents results of the our systems with different SSL objectives. Our pre-trained model is trained on VoxCeleb2-dev dataset with the ProtoNCE, which achieves 7.21 % EER on VoxCeleb1-O trial. Interestingly, the performance of the ProtoNCE based system fluctuates with the different number of predefined clus-

Table 2. Results for SV systems on the Voxceleb1-O test dataset. CLS: contrastive self-supervised learning. *When using ProtoNCE, the number of clusters is set to 5000.

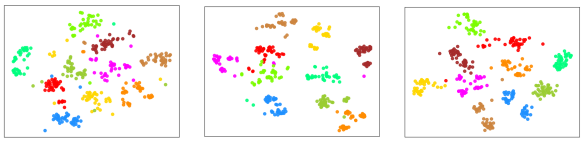
Model	Objective	EER	minDCF
ResNet34 [11]	Moco	13.48	-
ResNet [15]	CSL	8.86	0.51
TDNN[17]	MoCo	8.63	0.64
TDNN[17]	ProtoNCE	8.23	0.59
ECAPA-TDNN[24]	MoCo	7.3	-
ECAPA-TDNN	MoCo	8.31	0.67
ECAPA-TDNN	MoBY	8.20	0.66
ECAPA-TDNN	ProtoNCE*	7.21	0.61
ECAPA-TDNN	PCL	7.11	0.61

ters. When the number of clusters is significantly larger than the ground-truth number (800), there is a dramatic drop in performance. It is also shown in Fig 2, the training losses of ProtoNCE with predefined cluster 1500 and 2000 show a significant upward and unstable trend, respectively. This denotes that ProtoNCE is highly sensitive to the predefined number of clusters. When dealing with real unlabeled utterances, this prior information about the dataset is usually missing which may limit generalization. Compared with ProtoNCE and MoCo, the proposed PCL achieves the best performance in terms of EER and is slightly worse than MoCo in minDCF. In addition, the training loss is more stable than ProtoNCE based systems. This means the dynamic estimation of the potential number of speakers has the potential to extract more discriminative speaker embedding.

5.2. Comparison with state-of-the-art systems

To demonstrate the effectiveness of the proposed PCL, we compare it with other state-of-the-art systems with the self-supervised framework. Table 2 reports the our implementation results of ECAPA-TDNN with proposed PCL, as well as other SSL SV systems using different types of objective, including MoCo, MOBY (MoCo+BYOL) and ProtoNCE. We also list the state-of-the-art self-supervised SV systems for comparison. In order to achieve a good performance, in terms of the ProtoNCE based system, we finally set the total number of clusters to 5,000 after several attempts. Noted the ground-truth number of speakers in VoxCeleb2-dev is 5,994.

We notice that using the same objective function (MoCo), the ECAPA-TDNN based system outperforms the TDNN based system by 0.3% EER. This confirms the effectiveness of the ECAPA-TDNN. Moreover, using the same feature extractor, the MoBY based system achieves a slight relative improvement of 1.3% in EER. This is because the MoBY based approach has to maintain two independent memory queues, limited by the GPU memory size, we have to choose a smaller mini-batch size. As [18] pointed out, this may limit the performance of learned representation. Replacing the MoBY with the ProtoNCE, a further improvement is achieved. This indicates that encoding the samples with similar speaker identity information in the queue can significantly improve the discrimination of representation. Compared to the ProtoNCE, PCL achieve a relative improvement (i.e. 7.11% v.s. 7.21%) in EER. This means dynamically estimating the cluster number by data itself can improve the generalization of learned speaker representation. Finally, the proposed PCL optimized ECAPA-TDNN system outperforms all other self-supervised systems and obtains state-of-the-art performance on VoxCeleb1 dataset.



(a) MoCo (b) ProtoNCE (c) PCL

Fig. 3. T-SNE visualizations of speaker representations learned with different SSL frameworks. Different colors indicate different speakers.

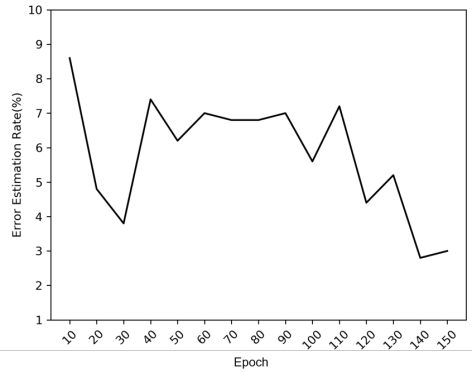
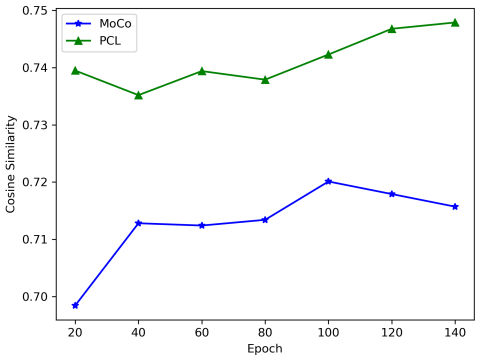


Fig. 4. The speaker estimation error rate on VoxCeleb2 versus training epoch.

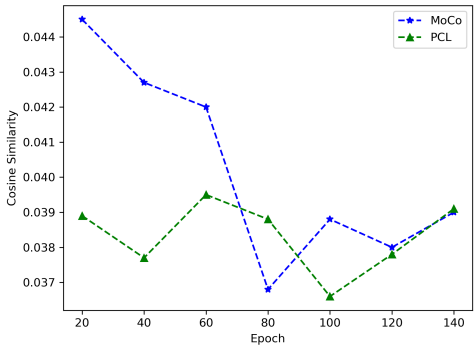
5.3. Visualization of learned speaker representation

In Fig.3, we use t-SNE to visualize the learned speaker representation of 10 random speakers, each with 40 utterances in the Voxceleb1 test dataset.

Compared to Moco and ProtoNCE based speaker representation, the proposed PCL optimized SV system learns to make tighter speaker clusters and increase the separateness between speakers. This suggests that the progressive contrastive learning based on the data characteristics can generate a more discriminative speaker representation, and is robust to the channel and noise variability.



(a) Positive samples



(b) Negative samples

Fig. 5. Cosine similarity between speaker representation and its corresponding positive/negative samples (or cluster centroids).

5.4. Effectiveness of speaker estimation

Before calculating the modified ProtoNCE, we dynamically estimate the speaker number K, which plays a key role in our framework. It determines the distribution of positive/negative centroids. And the SV model is expected to capture high-level speaker identity information among these centroids. The speaker number estimation error rate, which is computed as $1-K/H$ with the ground-truth speaker number H, can be found in Fig.4.

As shown in Fig.4, generally, the error estimation rate shows a descending trend, and the estimated speaker number progressively approaches the ground-truth number as the increasing of epoch. It suggests the proposed PCL can effectively cluster the speaker representations with the same speaker identity information.

5.5. Effectiveness of learned cluster centroid

In order to further analyze how and why our proposed PCL is effective, we plot the average cosine similarity between samples and its corresponding cluster centroid, and the similarity of positive pairs during training process in Fig.5(a) (higher is better). We also calculate the average similarity between speaker representation and its negative pairs (or other cluster centroids) versus epoch in Fig.5(b) (lower is better).

As shown in Fig.5 (a), compared to pair-wise objective (MoCo), speaker representation optimized with proposed PCL has a higher similarity with positive samples owing to the effect of clustering. This denotes the cluster centroid has the potential to be more stable and discriminative than randomly composed pairs. In addition, as the iteration goes, the MoCo-based system shows a small fluctuation, while the PCL-based system can steadily increase in similarity. This suggests a better speaker representation can be learned from a cluster with higher centralized speaker information. As we can see from Fig.5(b), the proposed PCL optimized system exhibits lower similarity of negative samples at most epochs. The MoCo optimized system appears a rapid decrease at the beginning and a fluctuation in the last epochs. The PCL optimized system shows fluctuations due to the progressively accurate speaker estimation.

6. Conclusion

In this work, we proposed a progressive contrastive learning framework, which dynamically estimated the queue-level speaker number for text-independent speaker verification. Several centroids were assigned to the updated queue data instead of the whole dataset, which was more effective and time-saving. Moreover, we derived a new objective based on the queue-level centroids to encourage samples to be close to their corresponding centroid in each step to obtain further performance gain. The promising performance improvement confirmed the effectiveness of the proposed PCL.

7. Acknowledgment

The work was partly supported by Czech National Science Foundation (GACR) project NEUREM3 No. 19-26934X, and Czech Ministry of Education, Youth and Sports from project No. LTAIN19087 "Multi-linguality in speech technologies". Computing on IT4I supercomputer was supported by the Czech Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project "e-Infrastructure CZ-LM2018140".

8. References

[1] Joseph P Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acous-*

tics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5329–5333.

[3] Junyi Peng, Rongzhi Gu, Yuexian Zou, and Wenwu Wang, "Speaker-discriminative embedding learning via affinity matrix for short utterance speaker verification," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019.

[4] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech2020*, 2020, pp. 1–5.

[5] Junyi Peng, Xiaoyang Qu, Jianzong Wang, Rongzhi Gu, Jing Xiao, Lukáš Burget, and Jan Černockỳ, "Icspk: Interpretable complex speaker embedding extractor from raw waveform," *Proc. Interspeech 2021*, pp. 511–515, 2021.

[6] Junyi Peng, Yuexian Zou, Na Li, Deyi Tuo, Dan Su, Meng Yu, Chunlei Zhang, and Dong Yu, "Syllable-dependent discriminative learning for small footprint text-dependent speaker verification," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 350–357.

[7] Chunlei Zhang and Kazuhito Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances.," in *Interspeech*, 2017, pp. 1487–1491.

[8] Weicheng Cai, Jinkun Chen, and Ming Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 74–81.

[9] Junyi Peng, Rongzhi Gu, and Yuexian Zou, "Deep speaker embedding with long short term centroid learning for text-independent speaker verification," *Proc. Interspeech 2020*, pp. 3246–3250, 2020.

[10] Ke Ding, Xuanji He, and Guanglu Wan, "Learning speaker embedding with momentum contrast," *arXiv preprint arXiv:2001.01986*, 2020.

[11] Jangho Lee, Jaihyun Koh, and Sungroh Yoon, "Momentum contrast speaker representation learning," *arXiv preprint arXiv:2010.11457*, 2020.

[12] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech 2017*, pp. 2616–2620, 2017.

[13] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *Proc. Interspeech 2018*, pp. 1086–1090, 2018.

[14] Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio, "Multi-task self-supervised learning for robust speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6989–6993.

[15] Danwei Cai, Weiqing Wang, and Ming Li, "An iterative framework for self-supervised deep speaker representation learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6728–6732.

[16] Yuki Takashima, Yusuke Fujita, Shota Horiguchi, Shinji Watanabe, Paola García, and Kenji Nagamatsu, "Semi-supervised training with pseudo-labeling for end-to-end neural diarization," *arXiv preprint arXiv:2106.04764*, 2021.

[17] Wei Xia, Chunlei Zhang, Chao Weng, Meng Yu, and Dong Yu, "Self-supervised text-independent speaker verification using prototypical momentum contrastive learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6723–6727.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[19] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu, "Self-supervised learning with swin transformers," *arXiv preprint arXiv:2105.04553*, 2021.

[20] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi, "Prototypical contrastive learning of unsupervised representations," *arXiv preprint arXiv:2005.04966*, 2020.

[21] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[22] Yue Fan, JW Kang, LT Li, KC Li, HL Chen, ST Cheng, PY Zhang, ZY Zhou, YQ Cai, and Dong Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and*

Signal Processing (ICASSP). IEEE, 2020, pp. 7604–7608.

[23] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[24] Jejin Cho, Jesus Villalba, and Najim Dehak, "The jhu submission to voxsrc-21: Track 3," *arXiv preprint arXiv:2109.13425*, 2021.