Self-Supervised Learning With Cluster-Aware-DINO for High-Performance Robust Speaker Verification

Bing Han, *Member, IEEE*, Zhengyang Chen, *Student Member, IEEE*, and Yanmin Qian, *Senior Member, IEEE*

***Abstract*—The automatic speaker verification task has achieved great success using deep learning approaches with a large-scale, manually annotated dataset. However, collecting a significant amount of well-labeled data for system building is very difficult and expensive. Recently, self-supervised speaker verification has attracted a lot of interest due to its no dependency on labeled data. In this article, we propose a novel and advanced self-supervised learning framework based on our prior work, which can construct a powerful speaker verification system with high performance without using any labeled data. To avoid the impact of false negative pairs, we adopt the self-distillation with no labels (DINO) framework as the initial model, which can be trained without exploiting negative pairs. Then, we further introduce a cluster-aware training strategy for DINO to improve the diversity of data. In the iterative learning stage, due to a mass of unreliable labels from unsupervised clustering, the quality of pseudo labels is important for the system performance. This motivates us to propose dynamic loss-gate and label correction (DLG-LC) methods to alleviate the performance degradation caused by unreliable labels. Furthermore, we extend the DLG-LC from single-modality to multi-modality on the audio-visual dataset to further improve the performance. The experiments were conducted using the widely-used Voxceleb dataset. Compared to the best-known self-supervised speaker verification system, our proposed method achieve relative EER improvement of 22.17%, 27.94% and 25.56% on Vox-O, Vox-E and Vox-H test sets, even with fewer iterations, smaller models, and simpler clustering methods. Importantly, the newly proposed self-supervised learning system even achieves comparable results with the fully supervised system, but without using any human-labeled data.**

***Index Terms*—Self-supervised speaker verification, cluster-aware dino, dynamic loss-gate, label correction, multi-modality.**

I. Introduction

RECENTLY, deep learning methods have been widely applied for speaker verification tasks and many efforts have been made such as various model architecture [2], [3], [4], [5], [6], training objection [7], [8], [9], pooling methods [10], [11] and so on, to achieve performance improvement compared with traditional methods such as Gaussian Mixture Model-Universal Background Model (GMM-UBM) [12], i-vector [13]. However, all of these methods are based on fully-supervised training and usually require large amounts of training data with accurate human annotations. As we know, the collection of large-scale, well-labeled data is actually very difficult and expensive.

Self-supervised learning is gaining traction as a means of reducing dependency on labeled data. Currently, some researchers are investigating its applicability to speaker verification tasks. Inspired by the great success of speech pre-trained models, e.g. wav2vec 2.0 [14] and HuBERT [15] in automatic speech recognition (ASR) tasks, some researchers [16] utilized them to extract the universal speech representation and apply to SV task. However, since these pre-trained models lack explicit speaker information, simply fine-tuning them on speaker verification task does not yield optimal results. In the work [17], the speech representation learned from large-scale unlabeled data was explored to replace the acoustic features, and then a deep neural network was trained in a supervised way. Although a promising performance was obtained, it still requires labeled data for training, and the parameter size is too large for real applications due to the large pre-trained model.

To fully leverage large-scale unlabeled data, inspired by text-to-speech (TTS) task, a generative method has been investigated in [18] to separate speaker representation with the help of phone information. Subsequently, some researchers came up with a hypothesis that speech segments truncated from the same utterance belong to the same speaker, while those from different utterances belong to different speakers. This hypothesis is approximately true for speaker verification datasets. Based on this hypothesis, several efforts [19], [20], [21], [22], [23] have been made to obtain discriminative speaker representations by maximizing information between different segments from the same utterance via contrastive-learning. Then, inspired by [24], an iterative learning framework [25] was developed to further improve the performance of self-supervised SV systems. This state-of-the-art system typically consists of two stages. In the first stage, a contrastive learning-based objective function is applied to train a speaker encoder. In the second stage, the pre-trained model from stage I is used to estimate pseudo-labels through clustering, which are then used as the supervised signal to train a new encoder. This process is iteratively performed to continuously improve the performance.

While this two-stage framework has shown performance improvement [26], [27], [28], [29], [30], it has several shortcomings which impede the further improvement of the system performance. For contrastive learning methods in stage I, speech

Manuscript received 17 May 2023; revised 5 October 2023; accepted 6 November 2023. Date of publication 10 November 2023; date of current version 1 December 2023. This work was supported in part by China NSFC Projects under Grants 62122050 and 62071288 and in part by Shanghai Municipal Science and Technology Commission Project under Grant 2021SHZDZX0102. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Omid Sadjadi. *(Corresponding author: Yanmin Qian.)*

The authors are with the Department of Computer Science and Engineering & MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: hanbing97@sjtu.edu.cn; zhengyang.chen@sjtu.edu.cn; yanminqian@sjtu.edu.cn).

Digital Object Identifier 10.1109/TASLP.2023.3331949

2329-9290 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 06,2024 at 09:07:02 UTC from IEEE Xplore. Restrictions apply.

segments cropped from different utterances are regarded as negative pairs to be pushed away from each other in speaker space. However, different utterances may belong to the same speaker in the real situation, which means that this inaccurate assumption might lead to errors. To tackle this problem, we introduced a negative-pairs free framework named DINO [31] to avoid the impact of the false negative pairs in our previous work [1]. During the second iterative stage, [24], [25] have proved that many pseudo labels generated by the clustering algorithm lack reliability, which would confuse and degrade the model. Researchers have conducted numerous studies on speaker recognition with noisy labels [29], [32], [33], [34], and one key approach is finding a way to select high-quality pseudo labels for enhancing model performance. In [29], they observed that the data with lower loss is more reliable than those with unreliable labels, and then proposed a loss-gate learning strategy to distinguish between reliable labels and unreliable labels by setting a loss threshold. The network is only updated using data with loss values below this threshold, ensuring the usage of reliable data. Although this approach led to further improvements, the use of manually set thresholds in each iteration limits flexibility and fails to make use of data with unreliable labels.

To solve these problems, this paper further extends our prior study [1] by focusing on novel algorithmic enhancements and additional analyses. The main contributions are summarized as follows:

1) In our previous work [1], we introduced DINO [31] as the self-supervised learning framework to obtain the initial pre-trained model, which is negative-pairs free to avoid the impact of the false negative pairs. Here, we propose a cluster-aware (CA) training strategy for algorithmic enhancement of DINO, which can improve the diversity of data and then obtain better performance.
2) In addition, we provide several additional analyses about dynamic loss-gate and label correction(DLG-LC).
3) Then, the DLG-LC method is further extended from audio single-modality to audio-visual multi-modality. Multi-modal data utilize multi-modal knowledge and make reliable label selection more efficient.
4) With these strategies, we achieve a great performance leap compared with the state-of-the-art (SOTA) system with self-supervised learning nowadays, even with fewer iterations, smaller models, and simpler clustering methods. More promisingly, this newly proposed self-supervised learning framework can approach the fully supervised system in performance which is trained in the same setup.

Recently, several DINO-based contemporaneous works have been released. It's noted that our work is done independently and concurrently with [35], [36], [37], [38], [39] related methods. No other papers were published when we submitted our work. We also compare our model to these concurrent works in the experiments and showcase the superiority of our methods.

II. Self-Supervised Learning for Speaker Verification

In this section, the commonly utilized two-stage self-supervised speaker verification framework is reviewed, including the first contrastive-learning stage for pre-trained model and the second iterative learning stage.

A. *Contrastive Based Self-Supervised Speaker Verification*

Self-supervised learning (SSL) is a type of unsupervised training that utilizes pretext or proxy tasks to learn the representations from the data itself. Common SSL methods can be broadly classified into two categories: generative [18] and contrastive [20], [21], [22], [23] methods. In speech application, based on the hypothesis that segments sampled from the same utterance belong to the same speaker while those from different utterances come from different speakers, most studies of SV tasks focus on contrastive learning approaches. Among them, SimCLR [40] is one of the most popular contrastive learning frameworks. Its basic idea is to minimize the distance between the representations of augmented segments cropped from the same utterance as well as maximize the distance between negative pairs from different utterances. Besides, the MoCo [41] framework further improves performance by incorporating a dynamic dictionary with a queue and a moving-averaged encoder. Based on these frameworks, many works such as equilibrium learning [20], augmentation adversarial training [21], channel-invariant training [22], prototype momentum [23] are proposed to learn more discriminative speaker representation.

B. *Iterative Framework for Self-Supervised Speaker Verification*

Considering that contrastive learning can naturally introduce label error and might degrade the model performance when the cluster indices of its learned embeddings are directly used as pseudo labels, in [25], they proposed an iterative, self-evolving framework to further improve the performance of self-supervised speaker verification systems. This framework is mainly divided into two stages, and they are illustrated as follows:

- Stage I: Pre-training
  1) Use contrastive learning or other self-supervised learning methods to pre-train a speaker encoder as the initial model.
  2) With the pre-trained model, extract the speaker embeddings for the training set and then apply a clustering algorithm to assign pseudo labels.
- Stage II: Iterative training and pseudo labeling.
  1) Train a new encoder with the pseudo labels generated by the previous step.
  2) Perform a clustering algorithm to update pseudo labels with the new encoder.
  3) Repeat stage II several times until the model converges.

Although this framework requires high computing resources due to the several iterations, it is widely used in [26], [27], [28], [29], [42] for its advanced performance. In addition, this framework is extended to the audio-visual dataset in [30] and achieves better performance with the help of multi-modal information in the clustering algorithm.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 06,2024 at 09:07:02 UTC from IEEE Xplore. Restrictions apply.

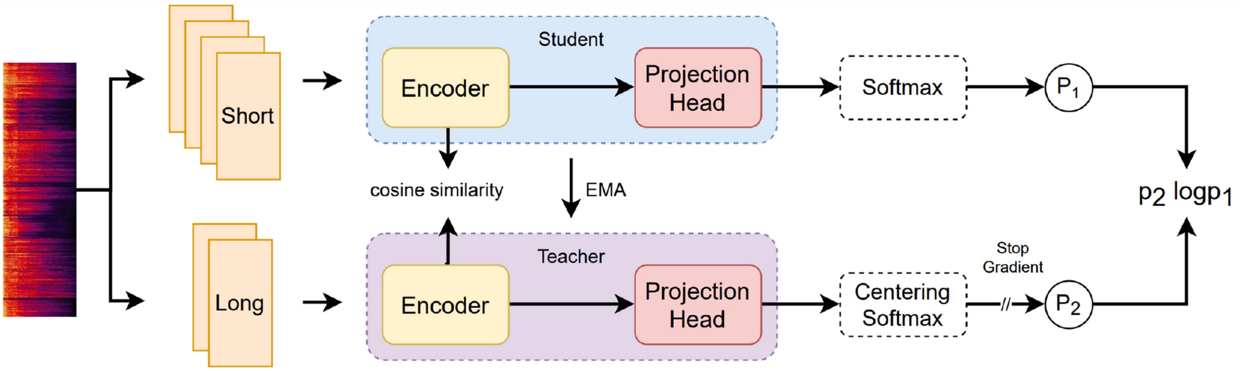

Fig. 1. Framework of distillation with no label (DINO) for self-supervised speaker representation learning.

TABLE I
PROBABILITY OF REPEAT SPEAKER IN A BATCH

Batch Size (B)	16	32	64	128	256
Probability	0.020	0.080	0.286	0.745	0.996

III. CLUSTER-AWARE-DINO FOR SPEAKER VERIFICATION

For contrastive learning-based methods in previous works, they shared the same assumption that segments cropped from the same utterances form positive pairs and those from different utterances in a batch belong to different speakers. But this assumption does not hold all the time because repeat speakers might appear in the same batch. Taking the statistics on Voxceleb 2 as an example, we can compute the probability of repeat speakers on Voxceleb 2 by (1) and the results are listed in Table I.

$$p_{repeat}(N,B) = 1 - \frac{P(N,B)}{N^B} = 1 - \frac{N!}{N^B(N-B)!} \quad (1)$$

where N is the speaker number in training set (N = 5994 for Voxceleb 2), B is batch size and $P(N,B) = \frac{N!}{(N-B)!}$ is B-permutation of N.

According to the Table I, a larger batch size leads to a higher probability of repeating which will cause a negative impact on the model. We can use a small batch size to alleviate this problem, but it will degrade the performance [41].

A. DINO Based Self-Supervised Learning

To tackle this problem, negative-pairs free DINO [31] is introduced to the self-supervised speaker verification task, and the whole framework is shown in Fig. 1.

Firstly, 4 short $\{x_1^s, x_2^s, x_3^s, x_4^s\}$ and 2 long segments $\{x_1^l, x_2^l\}$ are randomly sampled from an utterance using a multi-crop strategy [43]. The long segments allow for the extraction of more stable speaker embeddings. It is notable that when sampling, these segments should overlap as little as possible. Same as the previous works [20], [21], [22], [23], we still obey the assumption that the segments cropped from the same utterance belong to the same speaker and then apply different kinds of data augmentation on them by adding noise or room impulse response for robust performance. Unlike SimCLR [40], which only uses one encoder to do contrastive learning, our model consists of not only a *student* encoder but also a momentum *teacher* encoder whose architecture is similar to knowledge distillation [44]. After augmentation, all segments pass through the *student* encoder while only the long segments pass through the *teacher* encoder, thus encouraging the *short-to-long* correspondences by minimizing the cross-entropy $H(\cdot)$ between two distributions, as shown in the following (2):

$$L_{ce} = \sum_{x \in \{x_1^l, x_2^l\}} \sum_{x' \in \{x_1^l, x_2^l, x_1^s, \ldots, x_4^s\}} H(P_t(x) \mid P_s(x')) \quad (2)$$

where output distributions of momentum *teacher* network f_{θ_t} and *student* network f_{θ_s} are denoted by P_t and P_s respectively. And P can be computed by using a softmax function to normalize the output:

$$P_s(x) = Softmax(\frac{f_{\theta_s}(x)}{\epsilon_s}) \quad (3)$$

where $\epsilon_s > 0$ is the temperature parameter that can control the sharpness of the output distribution. Similarly, there is a formula that holds for P_t with temperature $\epsilon_t > 0$, too. Moreover, a mean statistic c computed over batches is used for centering *teacher* model's output distribution by $f_{\theta_t}(x) = f_{\theta_t}(x) - c$, and the statistic c is updated during the training process with a moving average strategy. During the training, both sharpening and centering are applied to avoid trivial solution [31].

The *teacher* and *student* models own the same architecture but different parameters due to the different update methods. The *student* model is updated by gradient descent while the *teacher* model is updated by the exponential moving average (EMA) of the *student*'s parameters. The EMA update rule is defined as follows:

$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s \quad (4)$$

where λ is adjusted by a cosine schedule [45] from 0.996 to 1 during training. Speaker embeddings are extracted by Encoders and then passed through the Projection Head, which is composed of a 3-layers perceptron with a hidden dimension of 2048, followed by ℓ_2 normalization and a weight normalized fully connected layer with K dimensions. The whole architecture is similar to [31].

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 06,2024 at 09:07:02 UTC from IEEE Xplore. Restrictions apply.

In addition, a cosine-based consistency loss is added to ensure that the speaker embedding is encoded into cosine space which is more suitable for the scoring and clustering in the following. It works by maximizing the cosine similarity among the embeddings extracted from the same speaker. Finally, the total loss is summarized with coefficient α:

$$L_{dino} = L_{ce} + \alpha \sum_{e\in\{e_1^l,e_2^l\}} \sum_{e'\in\{e_1^l,e_2^l,e_1^s,\ldots,e_4^s\}} \left(1 - \frac{e\cdot e'}{\|e\|\,\|e'\|}\right) \tag{5}$$

where e represents the extracted speaker embedding from the encoder.

B. Cluster-Aware Training on DINO

In traditional DINO, all segments are sampled from the same utterance to form positive pairs. However, due to the limited duration of the utterances, these segments usually have a great degree of overlaps. As mentioned above, the optimization of DINO is encouraging *short-to-long* correspondences by minimizing the cross-entropy between two distributions of positive pairs. Because there are a lot of overlapped parts in the segments, the model might pay more attention to the linguistic content, channel and other irrelevant information of the overlapped parts, and ignore the speaker information in the audio. Although we can add different types of data augmentation to segments, the data still lacks diversity, which could lead the model optimization in the wrong direction.

In order to reduce the overlaps of segments and increase the diversity in the segments for long-short correspondence, we propose to crop segments with being aware of cluster information, which is named CA-DINO in the following. More specifically, we divide model training into two stages. In the early stage of training, we optimize the model according to the traditional DINO strategy. Once the model has converged and can extract discriminative speaker representations, the training process will enter the next stage. Here, we perform a clustering algorithm using the extracted speaker embeddings, assuming that utterances in the same cluster belong to the same person. As shown in Fig. 2, the positive pairs are sampled from several utterances belonging to the same cluster rather than a single utterance. These pairs may come from the same speaker but with different speaking contents and channels, which leads to a high data diversity and makes the model pay more attention to the speaker's information instead of irrelevant factors. Considering the resource consumption of extracting the speaker embeddings, the clustering operation will be done every few rounds.

IV. Iterative Learning With Dynamic Loss-Gate and Label Correction

Based on the proposed CA-DINO self-supervised learning, we then apply the iterative learning framework [25] to further improve the performance of self-supervised SV. During the iterative process, a serious problem is that the generated pseudo labels contain a lot of noises which will confuse and degrade the network. Considering this limitation, several works have been done to select high-quality pseudo labels. In [25], an aggressive training method is applied to purify the labels using clustering confidence but achieves minor profit. In [29], they conducted a toy experiment and observed that data samples with lower loss is more reliable. Then, they propose a loss-gate (LG) strategy to select the data with lower loss by setting a fixed threshold and only use these data to update the model. With the LG strategy, the system achieved obvious improvement, but the threshold setting in this method is heavily dependent on human experience, and unreliable data are not fully utilized.

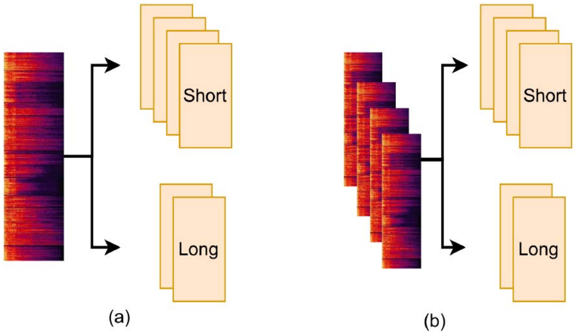

Fig. 2. Difference between traditional DINO and cluster-aware training DINO. (a) Traditional DINO:long and short segments are sampled from the same utterance to compose the positive pairs. (b) Cluster-aware training DINO: through a simple clustering algorithm, we consider that the same speaker in the same cluster shares the same identity and segments are cropped from the corresponding cluster.

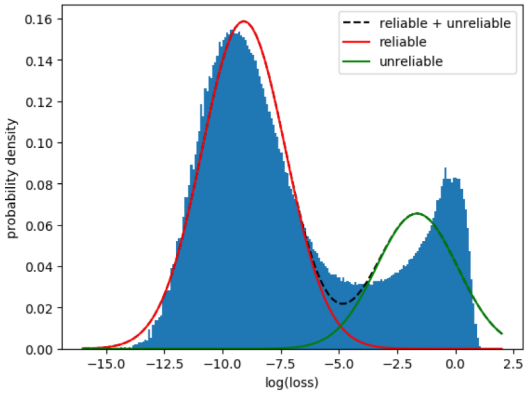

Fig. 3. Loss distribution of Loss-gate (LG) learning [29] on Voxceleb 2 [46]. Loss value is scaled by log function, and the lines are estimated by GMM with two components.

In this section, we will introduce our proposed DLG-LC to adjust the loss-gate threshold dynamically and correct the unreliable pseudo label to fully utilize the data, and then this DLG-LC approach is extended to utilize the multi-modality for further improvements.

A. Dynamic Loss-Gate

In order to determine an appropriate loss-gate threshold, we implemented the LG learning and visualized it to analyze the distribution of loss values on Voxceleb 2 [46] dataset. The histogram of loss values is provided in Fig. 3. According to the figure, there exist two sharp peaks in the distribution obviously. And similar experiments conducted in [47] have shown that data

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 06,2024 at 09:07:02 UTC from IEEE Xplore. Restrictions apply.

with reliable and unreliable labels can be represented by two peaks respectively. If we can find a way to model the distribution, then the loss-gate threshold can be determined dynamically as the loss distribution varies, which can avoid laborious manual tuning.

Gaussian distribution is an important continuous probability distribution of real-valued random variables, whose general form of the probability density function is defined in (6).

$$\mathcal{N}\left(\mu,\sigma^2\right)=\frac{1}{\sigma\sqrt{2\pi}}exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{6}$$

where location parameter and scale parameter are denoted by μ and σ respectively. The Gaussian distribution is known for its bell-shaped curve, with low values on both sides and high values in the middle, which is very similar to the "peaks" of loss observed in Fig. 3. In this case, Gaussian Mixture Model (GMM) with two components can be applied to model the loss distribution of reliable and unreliable samples respectively:

$$p(x)=\lambda_1\mathcal{N}\left(\mu_1,\sigma_1^2\right)+\lambda_2\mathcal{N}\left(\mu_2,\sigma_2^2\right) \tag{7}$$

where λ_1 and λ_2 represent the weights for two Gaussian components. After fitting, the fitted curves are plotted in Fig. 3, it's obvious to find that the two weighted Gaussian components can be used to approach these two "peaks". Then, by computing the loss values whose probabilities belonging to the two components are equal, the loss-gate threshold τ_1 can be obtained easily to distinguish between the reliable and unreliable data:

$$\tau_1 : p_1(\tau_1)=p_2(\tau_1) \tag{8}$$

where $p_1(x)=\lambda_1\mathcal{N}(\mu_1,\sigma_1^2)$ and $p_2(x)=\lambda_2\mathcal{N}(\mu_2,\sigma_2^2)$. For each epoch, all loss values are recorded for re-estimating the parameters of GMM, so τ_1 can be tuned dynamically according to the current training condition.

Our DLG introduces this dynamical loss-gate threshold τ_1 into the speaker classification loss function ArcMargin Softmax (AAM) [48] to select the data and only these retained data with losses under the threshold are used to update the parameters of the network.

$$L_{DLG}=-\frac{1}{B}\sum_{i=1}^{B}\mathbb{1}_{l_{i,clean}<\tau_1}\log\frac{e^{s(\cos(\theta_{y_i,i}+m))}}{Z} \tag{9}$$

where $Z=e^{s(\cos(\theta_{y_i,i}+m))}+\sum_{j=1,j\neq i}^{c}e^{s(\cos(\theta_{y_i,i}))}$, $\theta_{j,i}$ is the angle between the column vector W_j and speaker embedding $e_{i,aug}$ of the augmented segment. $\mathbb{1}$ here denote the Indicator function, B is batch size, c is number of speaker, s is the scaling factor and m is hyper-parameter to control the margin. AAM can enforce larger gaps between the nearest speakers and is widely adopted in speaker recognition tasks.

B. Label Correction

For samples with larger loss values than loss gate threshold τ_1, it's hard to assign reliable pseudo labels by the unsupervised clustering algorithm, and we call them hard samples. Instead of dropping them away directly [29], we propose the label correction (LC) strategy to correct unreliable pseudo labels automatically during the training process so that we can utilize these hard samples effectively. Researchers in [33], [34] have indicated that the network is capable of clustering noisy samples into their correct classes. To leverage this ability, we hypothesize that the output prediction of the model is more reliable than pseudo labels generated by clustering. Thus the predicted posterior probability is regarded as the target labels and incorporated into the objective loss function to prevent the model from fitting into inaccurate labels. However, not all prediction labels are suitable for training. Inspired by [49], [50], we assume that the prediction label owns high confidence if the model assigns a high probability to one of the possible classes. Then, another fixed threshold τ_2 is introduced to retain the prediction whose probability of largest class is above τ_2, and the label correction loss is defined as the following (10):

Algorithm 1: The Proposed Dynamic Loss-Gate and Label Correction.

```
Input: mini-batch $D_m=\{(x_1,x_2,y)\}_{i=1}^n$ where $x_1$ and $x_2$ are
       different crops from the same utterance; two threshold $\tau_1$ and
       $\tau_2$; Network $g(\cdot)$ including a speaker encoder and a classifier
; sharpness factor $\epsilon_c$ Output: the loss of the mini-batch
for $(x_1,x_2,y)\in D_m$ do
    $x_{clean}$, $x_{aug}$ = $x_1$, augment($x_2$)        # augment one segment
    $p_{clean}$, $p_{aug}$ = $g(x_{clean})$, $g(x_{aug})$    # output distribution
    Compute the AAM-softmax loss $l_{clean}$ and $l_{aug}$ according the
      pseudo label $y$
    Record the $l_{clean}$ value
    if $l_{clean}<\tau_1$ then
        return $l_{aug}$                          # pseudo label $y$ is reliable
    else
        if max($p_{clean}$) > $\tau_2$ then
            $p_{\hat{clean}}$ = sharp($p_{clean}$, $\epsilon_c$)   # sharpen the distribution
            compute the cross-entropy $l$ between $p_{\hat{clean}}$ and $p_{aug}$
            return $l$
        else
            return 0                          # prediction isn't reliable
        end if
    end if
end for
After one epoch, re-estimate the GMM on the recorded loss values
  and then update the $\tau_1$
```

$$L_{LC}=\frac{1}{B}\sum_{i=1}^{B}\mathbb{1}_{l_{i,clean}>\tau_1,\max(p_{i,clean})>\tau_2}H(p_{i,\hat{clean}}\mid p_{i,aug}) \tag{10}$$

where $p_{i,aug}$ represents the output probability of augmented segments and $p_{i,clean}$ represents their corresponding clean version (without any data augmentation strategies). $H(\cdot)$ here denotes the cross-entropy loss function between two probability distributions. In addition, to encourage a peaky distribution, $p_{i,\hat{clean}}$ is obtained by applying a sharpening operation with sharpness factor ϵ_c which is described in (3).

Then, the DLG loss and LC loss are combined to optimize the speaker model as (11).

$$L=L_{DLG}+L_{LC} \tag{11}$$

More specifically, the pseudo-code for describing the flow of the DLG-LC algorithm is provided in detail and shown in Algorithm. 1.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 06,2024 at 09:07:02 UTC from IEEE Xplore. Restrictions apply.

C. Incorporate With Multi-Modality

The researchers in [30] have introduced the multi-modality information into the data clustering step to generate more accurate pseudo labels in self-supervised speaker verification. In our work, considering that the audio and visual features from the same video share the same speaker identity, we also try to enhance our DLG-LC method by integrating visual modality to utilize data more effectively and achieve better performance. Our fusion of visual information is mainly divided into two aspects: firstly, to aid DLG-LC in selecting more dependable data, and secondly, to improve the clustering results during data clustering.

1) Multi-Modal Based DLG-LC: Different from the single-modal DLG-LC, our strategy of selecting reliable data has been slightly adjusted. For multi-modal data, we will use two independent encoders to encode audio and visual data separately. Then, by recording the loss values, we can obtain two loss-gate thresholds for audio and visual respectively. For an audio-visual instance, it can be regarded having a reliable label only if its loss values are both under these two loss-gate thresholds. We then optimize these instances using AAM softmax, as defined in (9).

For unreliable data, the multi-modal label correction (LC) will be performed on it. First, we compare whether the predicted labels of the two modal networks are consistent. If the predictions of the two models belong to the same class, it indicates that the accuracy of the prediction is relatively high. Unlike single-modal LC, which uses soft labels for training, our output is verified by multi-modal, which has higher reliability. As a result, we use the "hard" labels (i.e. the $\arg\max$ of the model's output distribution p_{clean}) to optimize the p_{aug} using AAM softmax. If the network disagrees with the predicted labels, then we use the soft labels to optimize models separately based on (10).

2) Multi-Modal Based Data Clustering: During the previous training step, the multi-modal information was only used to select reliable data, and the models of the two modalities were not structurally related. As a result, we can obtain audio $g_a(\cdot)$ and visual encoders $g_v(\cdot)$ independently. Given a dataset with audio x_a and visual modality x_v, we can use a trained encoder to extract audio embedding e_a and visual embedding e_v respectively. To leverage the complementary information present in both modalities, we apply an additional clustering on the joint representation $e_{av} = (e_a, e_v)$, which is formed by concatenating the audio and visual embeddings. With the joint operation, the representation will be more discriminative and the cluster will be more robust. Then, pseudo labels for the next iteration will be generated by k-means on these audio-visual joint embeddings.

V. Experiments Setup

A. Dataset

The experiments are conducted on Voxceleb [46], [51] which is a large-scale audio-visual dataset for the speaker recognition task. For the model training in stages I and II of self-supervised learning, we adopt the development set of Voxceleb 2 [46] for training the networks, and no speaker identity information is used during this process. Because we introduced visual features into the iterative learning stage, we excluded some utterances with the video missing in the data set. Then, the final audio-visual training set includes 1,091,251 utterances among 5,994 speakers, extracted from YouTube.

For the evaluation, we report the experimental results on 3 trials as defined in [46]: the Original, Extended, and Hard Voxceleb test sets. **Vox-O** is the original test set of Voxceleb 1 contains 37,720 trials from 40 speakers. **Vox-E** is a trial list which (using the entire dataset) contains 581,480 trials from 1251 speakers. **Vox-H** is a hard evaluation list consisting of 552,536 pairs sampled from 1190 speakers in Voxceleb 1, all of which are from the same nationality and gender.

B. Metrics

The main metrics adopted in this paper are (i) Equal Error Rate (EER) which is the error rate when both acceptance and rejection rates are equal, and (ii) the normalized minimum Detection Cost Function (minDCF) which is defined by (12) :

$$C_{det} = C_{miss} \times P_{miss} \times P_{tar} + C_{fa} \times P_{fa} \times (1 - P_{tar}) \tag{12}$$

where we set the prior target probability P_{tar} as 0.01 and equal weights between misses C_{miss} and false alarms C_{fa}. Both EER and minDCF are commonly used as evaluation metrics for speaker verification systems.

C. Data Augmentation

1) Audio: To generate extra training samples and increase the diversity of data, we perform online data augmentation strategy [52] by adding background noise or convolutional reverberation noise from MUSAN [53] and RIR dataset [54] respectively. The noise types in MUSAN include ambient noise, music, television, and babble noise for the background additive noise. We can obtain augmented data by mixing the noise with the original speech in time-domain waveform directly and the signal-to-noise ratios (SNR) are randomly applied between 5 to 20 dB. For the reverberation, the convolution operation is performed with 40,000 simulated room impulse responses (RIR) [54]. After applying the augmentation, we normalize the waveform value for stable training. We used 80-dimensional log Mel filter-bank energies with 25 ms length Hamming windows and 10 ms window shift as the acoustic features, while no voice activity detection (VAD) is involved in our experiments.

2) Visual: For each video segment in VoxCeleb 1 & 2 datasets, images are extracted at one frame per second. Then, we align the faces in extracted frames using the landmarks predicted by MTCNN [55] and after that, the similarity transformation is used to map the face region to the same shape ($3 \times 112 \times 96$). To enhance the quality of the visual features, we resize each image to the most common size of the model ($3 \times 224 \times 224$). In the following, several data augmentation strategies including random color distortion, random horizontal flipping, random grey scaling, and random Gaussian blur are applied to the original images with 0.8 probability. Finally, we normalize the pixel value of each image to the range of [−0.5, 0.5] before feeding it into the model.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 06,2024 at 09:07:02 UTC from IEEE Xplore. Restrictions apply.

TABLE II
MODEL ARCHITECTURE OF VISUAL ENCODER RESNET34 [60]

Layer	Structure	Output Size
Input Conv2D	- $\mathbf{C}(3\times3,32)$	$3\times L\times L$ $32\times L\times L$
Residual Block 1	$\begin{bmatrix}\mathbf{C}(3\times3,32)\\ \mathbf{C}(3\times3,32)\end{bmatrix}\times 3$, stride 2	$32\times\frac{L}{2}\times\frac{L}{2}$
Residual Block 2	$\begin{bmatrix}\mathbf{C}(3\times3,64)\\ \mathbf{C}(3\times3,64)\end{bmatrix}\times 4$, stride 2	$64\times\frac{L}{4}\times\frac{L}{4}$
Residual Block 3	$\begin{bmatrix}\mathbf{C}(3\times3,128)\\ \mathbf{C}(3\times3,128)\end{bmatrix}\times 6$, stride 2	$128\times\frac{L}{8}\times\frac{L}{8}$
Residual Block 4	$\begin{bmatrix}\mathbf{C}(3\times3,256)\\ \mathbf{C}(3\times3,256)\end{bmatrix}\times 3$, stride 2	$256\times\frac{L}{16}\times\frac{L}{16}$
Embedding	-	192

C (kernal size, channel) denotes the convolutional 2D layer. [·] represents the residual block and L is the image size of input.

TABLE III
MODEL ARCHITECTURE OF AUDIO ENCODER ECAPA-TDNN [56]

Layer	Structure	Output Size
Input Conv1D	- $\mathbf{C}(5,512)$	$F\times T$ $512\times T$
SE-Res2Block 1	$\mathbf{C}(1,512)$ $\mathbf{C}(3,64)\times 8$, dilation 2 $\mathbf{C}(1,512)$	$512\times T$
SE-Res2Block 2	$\mathbf{C}(1,512)$ $\mathbf{C}(3,64)\times 8$, dilation 3 $\mathbf{C}(1,512)$	$512\times T$
SE-Res2Block 3	$\mathbf{C}(1,512)$ $\mathbf{C}(3,64)\times 8$, dilation 4 $\mathbf{C}(1,512)$	$512\times T$
Conv1D Pooling Layer Embedding	$\mathbf{C}(1,1536)$ Attentive Stat Pooling -	$1536\times T$ 3072×1 192

C (kernal size, channels) denotes the convolutional 1D layer. F is the dimension of the input acoutic features which is determined by the number of frequency bins of the Mel spectrogram. T relates to the frames of the speech segments.

D. CA-DINO Setup

1) DINO: For DINO, considering the training time and memory limitation, we adopt ECAPA-TDNN [56] as an audio encoder to learn discriminative speaker representation, which is a time-delay neural network (TDNN) [3] based backbone with emphasized channel attention, propagation, and aggregation. It employs a channel- and context-dependent attention mechanism [57], Multi-layer Feature Aggregation (MFA), as well as Squeeze-Excitation (SE) [58] and residual blocks. The model architecture of ECAPA-TDNN is shown in Table III . For each utterance, two long (3 seconds) and four short (2 seconds) segments are randomly cropped and regarded as positive pairs. It is worth noting that all the segments will be applied with data augmentation, and after that, they are encoded into 192-dimensional speaker embeddings by the encoder. Similar to the configuration in [31], the K in the DINO projection head is set as 65,536. Temperatures for the teacher ϵ_t and the student ϵ_s are 0.04 and 0.1 respectively. In addition, we set cosine loss weight α as 1.0 to balance two losses. The whole training process will last 150 epochs. Model parameters are updated using the stochastic gradient descent (SGD) algorithm with weight decay $5e$-5. The learning rate is linearly ramped up from 0 to 0.2 in the first 20 epochs, and then it decays to $1e$-5 with the cosine schedule [45]. Moreover, the momentum also follows the cosine schedule from 0.996 to 1.0.

2) Cluster-Aware Training: For cluster-aware training strategy, we train the model normally in the first 90 epochs. After that, a clustering algorithm is applied on the whole training set every 5 epochs, which is supported by faiss library [59]. Considering the time complexity and the amount of training data, we only utilize k-means here which requires a few extra computations. The results of clustering are used for the generation of training data. Positive pairs are sampled from utterances belonging to the same cluster rather than the single one.

E. DLG-LC Setup

1) Single Modality: In this stage, for a fair comparison with [29], we also adopt ECAPA-TDNN [56] as our audio encoder to extract speaker embedding. For clustering, we choose k-means algorithm to assign the pseudo label to the training set. Unlike some works [29], [30], [42] that directly regard the number of real speakers as the number of clusters, we choose 7500 as the cluster number to verify the robustness of our method. For LC, sharpening parameters ϵ_c and threshold τ_2 are set as 0.1 and 0.5 respectively. The learning rate decays from 0.1 to $5e$-5 exponentially and we set the momentum and weight decay as 0.9 and $1e$-4. Finally, the training process will last 100 epochs.

2) Multi Modality: For audio-visual based DLG-LC, except for the addition of an image encoder, other configurations are consistent with the single-modal. We employ the ResNet34 [60] as the backbone network for the visual encoder, which is similar to the recent works [61], [62]. More detail is shown in Table II.

VI. EXPERIMENTAL RESULTS

The experiments are performed in six parts. In Section VI-A, a performance comparison of the proposed Cluster-aware DINO with previous works in stage I are reported, and we discuss how the number of clusters affects the cluster-aware training strategy. In Section VI-B, we report the speaker verification performance of CA-DINO finetuned on the small-scale labeled data. In Section VI-C, an ablation study of our proposed DLG-LC is given to demonstrate its effectiveness. Then, Section VI-D and Section VI-E showcase how our proposed DLG-LC can improve the performance in both single-modal and multi-modal scenarios. Finally, in Section VI-F, we provide a comprehensive comparison between our newly proposed self-supervised learning method and previous works to demonstrate the superiority and robustness of our system.

A. Evaluation of CA-DINO Based Speaker Verification

Table IV summarizes the speaker verification performance of our proposed methods and compares them with other previous self-supervised speaker models. All the methods are trained on Voxceleb 2 without any speaker label and evaluated on the

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 06,2024 at 09:07:02 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
PERFORMANCE COMPARISON OF THE PROPOSED CA-DINO WITH OTHER SELF-SUPERVISED SPEAKER VERIFICATION METHODS

SSL Methods	EER (%)	minDCF
Disent [63]	22.090	-
CDDL [64]	17.520	-
GCL [19]	15.260	-
i-vector [21]	15.280	0.63 (p=0.05)
AP + AAT [21]	8.650	0.45 (p=0.05)
SimCLR + uniform [22]	8.280	0.610
MoCo + WavAug [23]	8.230	0.590
Unif+CEL [20]	8.010	-
*DINO [35], [37]	4.830	-
*C3-DINO [36]	3.300	-
*DINO (Raw waveform) [38]	5.400	0.340 (p=0.05)
*DINO + Curriculum Learning [39]	4.470	0.306 (p=0.05)
DINO	31.233	0.990
+ EMA	4.404	0.434
+ + Cluster Aware (CA)	3.585	0.353

* Contemporaneous works when this paper is under review.
SSL means Self-supervised learning. EER (%) and mindcf (p=0.01) are evaluated on Vox-O test set.

Vox-O test set. According to the results, we can find that the methods based on contrastive learning [20], [21], [22], [23] have greatly improved the performance compared with the traditional works [19], [63], [64]. And negative-pairs-free DINO-based methods also achieve a great performance leap again compared to contrastive learning based methods which rely on and positive and negative pairs. It shows that negative pairs are indeed a bottleneck for performance improvement. It's noted that our baseline achieves the comparable result of contemporaneous works [35], [37], [38], [39] except C3-DINO [36] because it's trained with larger batch size and longer segments which is high demand for large computation resources.

In addition, we also provide the ablation study at the bottom of Table IV. When we train the DINO without the exponential moving average (EMA), it's difficult to converge and only obtains a very bad result which demonstrates that EMA is the key to preventing the model from collapsing. Then we apply the cluster-aware (CA) strategy when training the DINO, the performance has been further improved. The proposed CA-DINO achieves an EER of **3.585%**, with **55.24%** relative EER improvement compared with the best previously published performance of contrastive learning based self-supervised SV system [20].

During the cluster-aware training, there exists a k-means clustering operation. We also conducted an experiment to explore the influence of the number of clusters on the performance and the results are reported in Table V . The motivation of cluster-aware training is to increase the diversity of positive pairs. Due to the lack of ground truth, it is inevitable to introduce false positive pairs when increasing diversity. Normalized Mutual Information (NMI) is a good measure for determining the quality of clustering, which can be used to evaluate the trade-off between the diversity and false positive pairs. Assume that the true speaker labels is U and the predicted pseudo labels is V, NMI measures the clustering quality by computing the information shared between U and V:

$$\text{NMI}(U,V) = \frac{2 \times I(U;V)}{H(U) + H(V)} \tag{13}$$

TABLE V
PERFORMANCE COMPARISON OF CLUSTER-AWARE TRAINING WITH DIFFERENT CLUSTER NUMBERS

# Cluster	1080k	30k	20k	10k	5k
NMI	0.753	0.891	0.902	0.912	0.898
P$_{fp}$	0.000	0.054	0.074	0.127	0.233
N$_{avg}$	1.00	36.37	54.56	109.13	218.25
EER(%)	4.404	3.909	3.946	**3.585**	3.978

EER (%) is evaluated on VOX-O test set. 1080 k here means that one utterance is one class, which is equivalent to training without the cluster-aware strategy. NMI denotes normalized mutual information. P$_{fp}$ denotes the probability of false positive pairs from the same cluster. N$_{avg}$ denotes the average number of utterances belonging to same cluster, which can reflect diversity.

TABLE VI
EER(%) COMPARISON OF FINETUNING THE PRE-TRAINED SELF-SUPERVISED MODEL WITH DIFFERENT AMOUNT OF LABELED DATA FROM VOXCELEB 1

Initial Model	None	10%	20%	50%	100%
Random	32.78	6.893	5.276	3.691	2.755
SimCLR	8.547	4.388	3.797	3.266	2.936
CA-DINO	**3.585**	**2.393**	**2.356**	**2.016**	**1.835**

Results are evaluated on Vox-O which is the test set of voxceleb 1.

where $I(U;V)$ is the mutual information between U and V, and $H()$ denotes entropy. From the results, it is observed that NMI is positively correlated with the results, and our proposed cluster-aware training strategy can improve the NMI effectively and bring stable improvements for all the given number of clusters compared with the baseline system (1080 k). Meanwhile, CA-DINO with 10 k cluster number outperforms other systems which shows that the reasonable setting for the number of clusters can maximize the performance improvement.

B. Evaluation of CA-DINO With Pretrain-Finetune Framework With Labeled Data

To better illustrate the superior performance of our proposed CA-DINO, we conduct an exploration of self-supervised learning using the pretrain-finetune framework, i.e. fine-tuning the self-supervised model with a small amount of labeled data in the downstream speaker verification task. We randomly sample 10%/20%/50%/100% labeled utterances from Voxceleb1 [51] as the supervision and finetune the self-supervised models with these data.

As shown in Table VI , it is observed that self-supervised models, both SimCLR and proposed CA-DINO, made great improvements compared with model training from scratch, which shows that a pretraining model with better initialization is very important in low-resource conditions. Moreover, comparing the proposed CA-DINO with SimCLR, the proposed non-contrastive CA-DINO outperforms SimCLR obviously and can obtain a good performance position only with few labeled data in downstream speaker verification tasks. Moreover, with only 10% part of the labeled data, CA-DINO even achieves a better performance than the fully supervised system, i.e. 2.393% vs. 2.755%, which is meaningful to economize lots of manual annotation.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 06,2024 at 09:07:02 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
EER (%) Comparison on Vox-O, E, H of the Proposed DLG-LC in Iteration 1

Method	Threshold	Vox-O	Vox-E	Vox-H
SimCLR	-	6.281	7.428	11.54
DINO	-	3.287	3.613	6.039
CA-DINO	-	**2.909**	**3.315**	**5.692**
CA-DINO				
+ LG [29]	0.5	2.684	3.129	5.277
+ LG [29]	1	**2.441**	**2.930**	**4.892**
+ LG [29]	3	2.516	3.037	5.094
+ LG [29]	5	2.553	3.052	5.173
CA-DINO				
+ DLG	Dynamic	2.186	2.473	4.306
++ LC	Dynamic	**2.021**	**2.331**	**4.012**

In this experiment, pseudo labels are estimated from our pre-trained CA-DINO system. SimCLR and CA-DINO here mean we used all the data with the estimated pseudo labels as the supervisory signal without any data selection strategy during the system training.

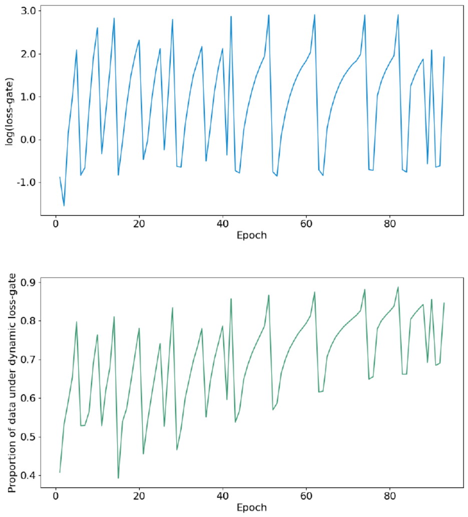

Fig. 4. Dynamic loss-gate threshold versus epoch (Up) and selected data proportion under loss-gate versus epoch (Down).

C. Evaluation of Proposed DLG-LC

Based on pseudo labels generated by pre-trained models in stage I, we conducted some experiments to illustrate the effectiveness of our proposed methods. The corresponding results are presented in Table VII . Firstly, following the iterative learning framework proposed by [25], we estimate the pseudo labels based on the speaker embedding extracted by CA-DINO and train a new encoder using these labels. To showcase the superiority of our method, we also trained a model based on SimCLR which is the most popular self-supervised speaker verification method [22]. From the results in the Table, we can see that the model based on CA-DINO outperforms SimCLR and DINO on all test sets with a significant improvement. Then based on pre-trained CA-DINO, we also conduct an exploration of DLG-LC in Iteration 1. According to the results, it can be observed that the loss-gate (LG) learning with fixed thresholds to select data can bring significant improvement compared with the system trained without any data selection. This indicates that loss-gate can effectively select reliable labels, which are of benefit to the model. However, we also try to set different thresholds (0.5, 1, 3, 5), and find that the choice of threshold also has a non-negligible impact on model performance [29]. Based on the estimated GMM, our proposed dynamic loss-gate (DLG) can adjust the threshold dynamically to consider the current training situation and achieve better performance than LG which only adopts a fixed threshold during the whole training process. In addition, we apply the label correction (LC) strategy to make full use of data with unreliable labels, and the results are further improved. Compared with the baseline system (SimCLR without data selection), the proposed CA-DINO with DLG-LC outperforms it with a relative **70.05%**, **68.61%**, **65.23%** EER reduction on Vox-O, Vox-E, and Vox-H sets, respectively.

To understand how the dynamic loss-gate (DLG) changes and filters the reliable data, we visualize it and present it in Fig. 4. Different from LG [29], which pre-trains the model on whole data and then fine-tunes with a fixed threshold, our DLG can repeat this process automatically by increasing and decreasing the threshold. In this repetitive process, more and more reliable data are filtered by dynamic threshold and utilized to optimize the model, which leads to performance improvement over multiple cycles of iterative training pipeline.

D. Iterative Learning With DLG-LC

In order to further demonstrate the superiority of our proposed method, we carried out several rounds of iterative training following [25]. In this training process, we trained the speaker network with the pseudo labels iteratively and updated the pseudo labels using the new converged network. Table VIII summarizes the performance of EER and minDCF of each iteration with and without the proposed DLG-LC strategy on Vox-O, Vox-E, and Vox-H test sets. Firstly, we compare the iterative results of SimCLR and CA-DINO respectively, both of which were trained without any loss-gate strategies. According to the results, it is observed that the iterative learning method can continuously improve the performance of the system with the increase of iteration number. However, the convergence speed based on SimCLR is significantly slower than that based on CA-DINO. SimCLR does not converge even in the 5th round, while CA-DINO has achieved the best performance position in the 3rd round. In addition, the final performance of SimCLR with iterative learning is still worse than the initial performance of CA-DINO. The proposed CA-DINO owns consistently large advantages over SimCLR in each iteration which further demonstrates the superiority of the proposed CA-DINO in self-supervised speaker verification.

Based on the pseudo-labels generated by CA-DINO, we applied the proposed strategy of DLG-LC, and the performance significantly improved further. It only took one round of iteration to obtain better results than three rounds of iterations without DLC-LC, showing the importance of dynamic threshold filtering and label correction on data usage. After convergence with more iterations, its performance is much better than the system without DLG-LC. It shows that the proposed DLG-LC can not only

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 06,2024 at 09:07:02 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
EER (%) AND MINDCF (P=0.01) COMPARISON ON VOX-O, VOX-E, AND VOX-H TEST SETS FOR DIFFERENT ITERATIONS OF THE PROPOSED DLG-LC WITH OTHER STRATEGIES

Initial Model	DLG-LC	Iteration	Vox-O		Vox-E		Vox-H	
			EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
SimCLR	×	Initial	8.547	0.6453	9.228	0.6912	14.21	0.7757
		1	6.281	0.5811	7.428	0.6221	11.54	0.7213
		2	5.914	0.5299	6.745	0.5880	10.54	0.6971
		3	5.547	0.5259	6.407	0.5580	10.14	0.6698
		4	4.872	0.4651	5.593	0.5144	8.923	0.6408
		5	**4.484**	**0.4545**	**5.225**	**0.5055**	**8.501**	**0.6321**
CA-DINO	×	Initial	3.585	0.3529	3.852	0.4182	6.918	0.5743
		1	2.909	0.3000	3.315	0.3372	5.692	0.4654
		2	2.606	0.2887	3.181	0.3211	5.403	0.4489
		3	**2.558**	0.3054	**3.064**	**0.3176**	5.342	**0.4482**
		4	2.643	**0.2825**	3.065	0.3200	**5.291**	0.4483
CA-DINO	✓	Initial	3.585	0.3529	3.852	0.4182	6.918	0.5743
		1	2.021	0.2171	2.331	0.2419	4.012	0.3484
		2	1.596	0.1665	2.004	0.2089	3.484	0.3083
		3	**1.585**	0.1671	**1.879**	**0.1963**	3.293	**0.2941**
		4	1.606	**0.1636**	1.906	0.2028	**3.274**	0.2955

SimCLR and CA-DINO without DLC-LC mean that we used all the estimated pseudo labels of the data without data selection in the training process.

TABLE IX
EER (%) AND MINDCF (P=0.01) COMPARISON ON VOX-O, VOX-E, AND VOX-H TEST SETS FOR DIFFERENT ITERATIONS OF THE PROPOSED DLG-LC WITH SINGLE- OR MULTI-MODALITY

Training Modality	Iteration	Vox-O		Vox-E		Vox-H	
		EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
Audio	Initial	3.585	0.3529	3.852	0.4182	6.918	0.5743
Audio	1	2.021	0.2171	2.331	0.2419	4.012	0.3484
	2	1.596	0.1665	2.004	0.2089	3.484	0.3083
	3	**1.585**	0.1671	**1.879**	**0.1963**	3.293	**0.2941**
	4	1.606	**0.1636**	1.906	0.2028	**3.274**	0.2955
Audio-Visual	1	1.537	**0.1326**	1.789	0.1910	3.235	0.3007
	2	**1.292**	0.1565	**1.571**	**0.1688**	**2.799**	**0.2676**
	3	1.356	0.1553	1.602	0.1711	2.839	0.2712

It's noted that they are both initialed with CA-DINO in the first self-supervised pretraining stage. Both our audio and visual encoders are trained independently, and the fusion of multi-modal information only performs when clustering data and selecting data in iterative learning. We do the testing still with the single audio modality.

speed up the model convergence and reduce the training time but also significantly boost the performance upper limit of the self-supervised learning model.

E. Incorporate With Multi-Modality

Then we introduce visual information in the iterative learning process. The difference from the work in [30] is that we not only use multi-modality when doing the data clustering but also utilize multi-modality information when applying data selection through DLG-LC. Table IX illustrates the EER and minDCF performance comparison of DLG-LC with single- and multi-modality.

It is observed that incorporating both audio-visual modality knowledge in iterative learning can obtain additional performance improvement, which demonstrates that extra visual information can make the data usage better. Take the EER of Vox-H as an example, with only single modality audio data, the relative EER reduction of the current and previous iterations are **42.01%**, **13.16%**, and **5.48%** on Vox-H trials for the first three iterations. If iterative learning with audio-visual data, the relative EER reduction percentages are **53.24%**, **13.48%** for the first two iterations.

F. Comparison With Other Systems

In this section, a performance comparison among our proposed CA-DINO with DLG-LC and other self-supervised speaker verification systems is given in Table X, and most of them are from the latest Voxceleb Speaker Recognition Challenge (VoxSRC) [66], [67] which represent the most advanced systems nowadays. Besides, the fully supervised system is also illustrated as the first line of Table X for comparison.

Compared with the previous works using large-size models, the model we adopt is ECAPA-S (Small, C=512) which has fewer parameters and requires fewer computation resources. Compared to systems with AHC (Agglomerative Hierarchical

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 06,2024 at 09:07:02 UTC from IEEE Xplore. Restrictions apply.

TABLE X
EER (%) Comparison on Vox-O, Vox-E, Vox-H Among the Proposed CA-DINO With DLG-LC and Other Most Advanced Self-Supervised Systems

Methods	Model	**# Iteration**	**# Clusters**	**Cluster**	**Vox-O (EER)**	**Vox-E (EER)**	**Vox-H (EER)**
Fully Supervised [56]	ECAPA-S	-	-	-	1.010	1.240	2.320
IDLab [26]	ECAPA-L	7	7500	AHC	2.100	-	-
JHU [27]	Res2Net50	5	7500	AHC	1.890	-	-
SNU [28]	ECAPA-L	5	7500	AHC	1.660	-	-
LG [29]	ECAPA-L	5	6000	K-M	1.660	2.180	3.760
DKU + single-modal [30]	ResNet34	5	6000	K-M	2.740	3.080	5.480
DKU + multi-modal [30]	ResNet34	5	6000	K-M	1.920	2.030	3.720
C3-DINO [36]	ECAPA-S	-	-	-	2.200	-	-
CA-DINO	ECAPA-S	3	7500	K-M	2.558	2.129	5.148
CA-DINO + DLG-LC + single-modal	ECAPA-S	3	7500	K-M	1.585	1.879	3.293
CA-DINO + DLG-LC + multi-modal	ECAPA-S	2	7500	K-M	**1.292**	**1.571**	**2.799**
CA-DINO + DLG-LC + multi-modal*	ECAPA-S	2	7500	K-M	**1.191**	**1.474**	**2.543**

* The results are given with adaptive s-norm [65] which requires label information for a fair comparison with fully supervised system [56].
The model architecture, clustering number, method and iteration rounds of each system are listed in detail. Note that AHC and K-M here mean agglomerative hierarchical clustering and k-means. ECAPA-S (small) and ECAPA-L (large) here denote the ECAPA-TDNN with 512 channels and 1024 channels respectively.

Clustering), to make it easier to implement, we adopt a simpler and more convenient clustering method K-M (k-means) to generate pseudo labels. Moreover, when clustering data, we set the number of clusters to 7500 instead of 6000, because 6000 is closer to the real number of speakers (5994) in the training set which is more opportunistic. From the results, it is observed that our proposed new self-supervised speaker verification framework is far superior to all the existing methods in both single- and multi-modality, even with fewer iterations, smaller model, and simpler clustering method. For the single modality condition, the proposed CA-DINO with DLG-LC outperforms the best system (LG) [29] by relative **4.52%**, **13.81%** and **12.42%** on Vox-O, Vox-E, and Vox-H sets respectively with only 3 iterations. If we use audio-visual data in the iterative learning stage, the corresponding improvement is enlarged to relative **22.17%**, **27.94%** and **25.56%**, which is a great performance leap. As for C3-DINO [36], it trains the model by DINO loss with ProtoNCE [23] initialed weights, which is a new approach different from typical iterative training methods. However, our method still has superior performance compared to C3-DINO.

In summary, our proposed system achieves the new **state-of-the-art** performance for self-supervised speaker verification with a large performance improvement, despite we train the systems with fewer iterations, smaller model, and simpler clustering method. More promisingly, compared to the conventional fully supervised system with the same training configuration, our newly proposed self-supervised learning system even obtains a comparable performance with the supervised system, but without using any ground-truth labels.

VII. Conclusion

In this work, we proposed an advanced self-supervised speaker verification system called Cluster-Aware DINO (CA-DINO) with Dynamic Loss-Gate and Label Correction (DLG-LC). Based on the DINO framework we introduced before, the cluster-aware training strategy is incorporated. More specifically, positive samples are collected from the same category rather than single sentences, so that the model can utilize more diverse data and obtain stable improvement. In the iterative learning stage, DLG-LC is adopted here with additional analyses and then extended to multi-modality for further improvements. The experiments on Voxceleb showed that our newly proposed CA-DINO with DLG-LC is superior and achieves the new **state-of-the-art** performance for self-supervised speaker verification. More promisingly, the gap between unsupervised and supervised representation learning is dramatically reduced for speaker verification, achieving close performance to the fully supervised system with our self-supervised earning method.

Acknowledgment

This paper is an extension of our prior conference paper [1].

References

[1] B. Han, Z. Chen, and Y. Qian, "Self-supervised speaker verification using dynamic loss-gate and label correction," in *Proc. ISCA Interspeech*, 2022, pp. 4780–4784.

[2] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 4052–4056.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5329–5333.

[4] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," 2019, *arXiv:1910.12592*.

[5] B. Han, Z. Chen, and Y. Qian, "Local information modeling with self-attention for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 6727–6731.

[6] B. Han, Z. Chen, B. Liu, and Y. Qian, "Mlp-svnet: A multi-layer perceptrons based network for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7522–7526.

[7] Y. Liu, L. He, and J. Liu, "Large margin softmax loss for speaker verification," in *Proc. ISCA Interspeech*, 2019, pp. 2873–2877.

[8] J. S. Chung et al., "In defence of metric learning for speaker recognition," in *Proc. ISCA Interspeech*, 2020, pp. 2977–2981.

[9] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5115–5119.

[10] S. Wang, Y. Yang, Y. Qian, and K. Yu, "Revisiting the statistics pooling layer in deep speaker embedding learning," in *Proc. IEEE 12th Int. Symp. Chin. Spoken Lang. Process.*, 2021, pp. 1–5.

[11] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification," in *Proc. ISCA Interspeech*, 2018, pp. 3573–3577.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 06,2024 at 09:07:02 UTC from IEEE Xplore. Restrictions apply.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digit. Signal Process.*, vol. 10, no. 1/3, pp. 19–41, 2000.

[13] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[14] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 12449–12460 .

[15] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Trans. Audio Speech Lang. Process..*, vol. 29, pp. 3451–3460, 2021.

[16] Z. Fan, M. Li, S. Zhou, and B. Xu, “Exploring wav2vec 2.0 on speaker verification and language identification,” in *Proc. ISCA Interspeech*, 2021, pp. 1509–1513.

[17] Z. Chen et al., “Large-scale self-supervised speech representation learning for automatic speaker verification,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 6147–6151.

[18] T. Stafylakis, J. Rohdin, O. Plchot, P. Mizera, and L. Burget, “Self-supervised speaker embeddings,” in *Proc. ISCA Interspeech*, 2019, pp. 2863–2867.

[19] N. Inoue and K. Goto, “Semi-supervised contrastive learning with generalized contrastive loss and its application to speaker recognition,” in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 1641–1646.

[20] S. H. Mun, W. H. Kang, M. H. Han, and N. S. Kim, “Unsupervised representation learning for speaker recognition via contrastive equilibrium learning,” 2020, *arXiv:2010.11433*.

[21] J. Huh, H. S. Heo, J. Kang, S. Watanabe, and J. S. Chung, “Augmentation adversarial training for self-supervised speaker recognition,” 2020, *arXiv:2007.12085*.

[22] H. Zhang, Y. Zou, and H. Wang, “Contrastive self-supervised learning for text-independent speaker verification,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6713–6717.

[23] W. Xia, C. Zhang, C. Weng, M. Yu, and D. Yu, “Self-supervised text-independent speaker verification using prototypical momentum contrastive learning,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6723–6727.

[24] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, “Deep clustering for unsupervised learning of visual features,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 132–149.

[25] D. Cai, W. Wang, and M. Li, “An iterative framework for self-supervised deep speaker representation learning,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6728–6732.

[26] J. Thienpondt, B. Desplanques, and K. Demuynck, “The idlab voxceleb speaker recognition challenge 2020 system description,” 2020, *arXiv:2010.12468*.

[27] J. Cho, J. Villalba, and N. Dehak, “The JHU submission to voxsrc-21: Track 3,” 2021, *arXiv:2109.13425*.

[28] S. H. Mun, M. H. Han, and N. S. Kim, “SNU-HIL system for the voxceleb speaker recognition challenge 2021,” VoxSRC, Tech. Rep., 2021.

[29] R. Tao, K. A. Lee, R. K. Das, V. Hautamäki, and H. Li, “Self-supervised speaker recognition with loss-gated learning,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 6142–6146.

[30] D. Cai, W. Wang, and M. Li, “Incorporating visual information in audio based self-supervised speaker recognition,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 1422–1435, 2022.

[31] M. Caron et al., “Emerging properties in self-supervised vision transformers,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9650–9660.

[32] S. Ranjan and J. H. Hansen, “Curriculum learning based approaches for noise robust speaker recognition,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 1, pp. 197–210, Jan. 2018.

[33] F. Tong, Y. Liu, S. Li, J. Wang, L. Li, and Q. Hong, “Automatic error correction for speaker embedding learning with noisy labels,” in *Proc. ISCA Interspeech*, 2021, pp. 4628–4632.

[34] L. Li, F. Tong, and Q. Hong, “When speaker recognition meets noisy labels: Optimizations for front-ends and back-ends,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 1586–1599, 2022.

[35] J. Cho, R. Pappagari, P. Zelasko, L. Moro-Velázquez, J. Villalba, and N. Dehak, “Non-contrastive self-supervised learning of utterance-level speech representations,” in *Proc. ISCA Interspeech*, 2022, pp. 4028–4032.

[36] C. Zhang and D. Yu, “C3-DINO: Joint contrastive and non-contrastive self-supervised learning for speaker verification,” *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1273–1283, Oct. 2022.

[37] J. Cho, J. Villalba, L. Moro-Velazquez, and N. Dehak, “Non-contrastive self-supervised learning for utterance-level information extraction from speech,” *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1284–1295, Oct. 2022.

[38] J. Jung, Y. J. Kim, H. Heo, B. Lee, Y. Kwon, and J. S. Chung, “Pushing the limits of raw waveform speaker recognition,” in *Proc. ISCA Interspeech*, 2022, pp. 2228–2232.

[39] H.-S. Heo et al., “Curriculum learning for self-supervised speaker verification,” in *Proc. Interspeech*, 2023, pp. 4693–4697, doi: 10.21437/Interspeech.2023-1202.

[40] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[41] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.

[42] D. Cai and M. Li, “The dku-dukeece system for the self-supervision speaker verification task of the 2021 voxceleb speaker recognition challenge,” 2021, *arXiv:2109.02853*.

[43] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised learning of visual features by contrasting cluster assignments,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 9912–9924.

[44] G. Hinton et al., “Distilling the knowledge in a neural network,” 2015, *arXiv:1503.02531*.

[45] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, “In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning,” in *Proc. Int. Conf. Learn. Representations*, 2021.

[46] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” in *Proc. ISCA Interspeech*, 2018, pp. 1086–1090.

[47] E. Arazo, D. Ortego, P. Albert, N. O’Connor, and K. McGuinness, “Unsupervised label noise modeling and loss correction,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 312–321.

[48] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, “Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition,” in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2019, pp. 1652–1656.

[49] D.-H. Lee et al., “Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2013, p. 896.

[50] K. Sohn et al., “Fixmatch: Simplifying semi-supervised learning with consistency and confidence,” *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 596–608.

[51] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: A large-scale speaker identification dataset,” in *Proc. ISCA Interspeech*, 2017, pp. 2616–2620.

[52] W. Cai, J. Chen, J. Zhang, and M. Li, “On-the-fly data loader and utterance-level aggregation for speaker and language recognition,” *IEEE/ACM Trans. Audio Speech Lang. Process..*, vol. 28, pp. 1038–1051, 2020.

[53] D. Snyder, G. Chen, and D. Povey, “Musan: A music, speech, and noise corpus,” 2015, *arXiv:1510.08484*.

[54] 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 Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5220–5224.

[55] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[56] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” in *Proc. ISCA Interspeech*, 2020, pp. 3830–3834.

[57] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, “Res2net: A new multi-scale backbone architecture,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2 , pp. 652–662, Feb. 2021.

[58] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[59] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with GPUs,” *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021.

[60] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[61] Y. Kim, W. Park, M.-C. Roh, and J. Shin, “Groupface: Learning latent groups and constructing group-based representations for face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5621–5630.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 06,2024 at 09:07:02 UTC from IEEE Xplore. Restrictions apply.

[62] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4690–4699.

[63] A. Nagrani, J. S. Chung, S. Albanie, and A. Zisserman, “Disentangled speech embeddings using cross-modal self-supervision,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6829–6833.

[64] S. Chung, H. Kang, and J. S. Chung, “Seeing voices and hearing voices: Learning discriminative embeddings using cross-modal self-supervision,” in *Proc. ISCA Interspeech*, 2020, pp. 3486–3490.

[65] P. Matějka, O. Novotný, O. Plchot, L. Burget, M. D. Sánchez, and J. Černocký, “Analysis of score normalization in multilingual speaker recognition,” in *Proc. ISCA Interspeech*, 2017, pp. 1567–1571.

[66] A. Nagrani et al., “Voxsrc 2020: The second voxceleb speaker recognition challenge,” 2020, *arXiv:2012.06867*.

[67] A. Brown, J. Huh, J. S. Chung, A. Nagrani, and A. Zisserman, “Voxsrc 2021: The third voxceleb speaker recognition challenge,” 2022, *arXiv:2201.04583*.

Bing Han (Member, IEEE) received the B.Eng. degree from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2020. He is currently working toward the Ph.D. degree with Shanghai Jiao Tong University, Shanghai, China, under the supervision of Yanmin Qian. His research mainly focuses on speaker recognition.

Zhengyang Chen (Student Member, IEEE) received the B.Eng. degree from the Department of Electronic and Information Engineering, the Huazhong University of Science and Technology, Wuhan, China, in 2019. He is currently working toward the Ph.D. degree with Shanghai Jiao Tong University, Shanghai, China, under the supervision of Yanmin Qian. His research interests include speaker recognition and speaker diarization.

Yanmin Qian (Senior Member, IEEE) received the B.S. degree from the Department of Electronic and Information Engineering, the Huazhong University of Science and Technology, Wuhan, China, in 2007 and the Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2012. Since 2013, he has been with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, where he is currently a Full Professor. From 2015 to 2016, he was an Associate Researcher with the Speech Group, Cambridge University Engineering Department, Cambridge, U.K. He has authored or coauthored more than 200 papers in peer-reviewed journals and conferences on speech and language processing, including T-ASLP, Speech Communication, ICASSP, INTERSPEECH, and ASRU. He has applied for more than 80 Chinese and American patents and was the recipient of the five championships of international challenges. His research interests include automatic speech recognition and translation, speaker and language recognition, speech separation and enhancement, music generation and understanding, speech emotion perception, multimodal information processing, natural language understanding, and deep learning and multi-media signal processing. He was the recipient of several top academic awards in China, including Chang Jiang Scholars Program of the Ministry of Education, Excellent Youth Fund of the National Natural Science Foundation of China, and the First Prize of Wu Wenjun Artificial Intelligence Science and Technology Award (First Completion). He was also the recipient of several awards from international research committee, including the Best Paper Award in Speech Communication and Best Paper Award from IEEE ASRU in 2019. He is also the Member of IEEE Signal Processing Society Speech and Language Technical Committee.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on April 06,2024 at 09:07:02 UTC from IEEE Xplore. Restrictions apply.