Two-stage Semi-supervised Speaker Recognition with Gated Label Learning

Xingmei Wang1, Jiaxiang Meng1, Kong Aik Lee2, Boquan Li1,3,* and Jinghan Liu1
1 College of Computer Science and Technology, Harbin Engineering University, Harbin, China
2 Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong
3 School of Computing and Information Systems, Singapore Management University, Singapore
{wangxingmei, mjxwjq}@hrbeu.edu.cn, kong-aik.lee@polyu.edu.hk, {liboquan, liujinghan}@hrbeu.edu.cn

Abstract

Speaker recognition technologies have been successfully applied in diverse domains, benefiting from the advance of deep learning. Nevertheless, current efforts are still subject to the lack of labeled data. Such issues have been attempted in computer vision, through semi-supervised learning (SSL) that assigns pseudo labels for unlabeled data, undertaking the role of labeled ones. Through our empirical evaluations, the state-of-the-art SSL methods show unsatisfactory performance in speaker recognition tasks, due to the imbalance between the quantity and quality of pseudo labels. Therefore, in this work, we propose a two-stage SSL framework, with the aim to address the data scarcity challenge. We first construct an initial contrastive learning network, where the encoder outputs the embedding representation of utterances. Furthermore, we construct an iterative holistic semi-supervised learning network that involves a clustering strategy to assign pseudo labels, and a gated label learning (GLL) strategy to further select reliable pseudo-label data. Systematic evaluations show that our proposed framework achieves superior performance in speaker recognition than the state-of-the-art methods, matching the performance of supervised learning.

1 Introduction

Speaker recognition technologies identify speakers based on their voice features extracted from speaker utterances [Hanifa et al., 2021], which have been broadly utilized in numerous information security applications such as identity authentication and access control [Singh et al., 2012].

With the emergence of deep learning, speaker recognition is implemented as neural networks [Brecht et al., 2020], and the performance is improved into a new level [Bai and Zhang, 2021; Son et al., 2020]. Although remarkable results have been achieved, these methods require massive labeled data to perform supervised learning [Vladimir, 2017], and the challenge of annotating utterances as well as huge time expenses make them hard to be applied in practice. In response to such limitations, researchers have attempted strategies including few-shot learning [Prateek, 2020; Yanxiang et al., 2023] and transfer learning [Cunwei et al., 2018]. Based on their reported results, they still perform limited success if only limited labeled utterances are available.

Intuitively, compared to labeled utterances, the ones without labels are widely available, which inspires a potential direction, i.e., taking advantage of semi-supervised learning (SSL) [Engelen and Hoos, 2020] that trains models based on joint labeled as well as unlabeled data. To leverage unlabeled data, typical SSL technologies assign pseudo labels to them enabling such data to act as labeled ones, which has been implemented in computer vision [Xiangli et al., 2023]. Especially, those state-of-the-art holistic methods such as FixMatch [Kihyuk et al., 2020] and FlexMatch [Bowen et al., 2021] achieve promising performance matching supervised learning. In contrast, there are only limited research attempts in speaker recognition [Long et al., 2021; Nakamasa and Keita, 2020; Kreyssig and Woodland, 2020; Fuchuan et al., 2022], and their reported results are inferior to the performance in computer vision. In general, SSL in speaker recognition is still an open problem and has plenty of room for improvement.

In this work, we propose a two-stage holistic SSL framework customized for speaker recognition, and aim to achieve matching performance to supervised learning. To achieve this, the key is to understand why existing holistic methods perform excellently in computer vision but the same is not true in speaker recognition? We attribute this question to the fact that classes in utterance data greatly exceed image ones. To be specific, such huge-class data inevitably causes the imbalance between the quality and quantity of pseudo labels, i.e., SSL models focus on either assigning correct labels or selecting enough amounts of pseudo-label data. Such phenomena are analogous to the confirmation bias [Eric et al., 2020] issue, i.e., models overfit to the data assigned with incorrect pseudo labels. Based on this intuition, we aim to implement effective speaker recognition by addressing such imbalances.

In particular, our proposed framework mainly includes:

• In Stage I, we devise a contrastive learning-based network [Danwei et al., 2021] (based on unlabeled data), which provides an initial framework as well as produces speaker embedding representations for the next stage.

*Corresponding author
• In Stage II, we devise a holistic network to perform iterative semi-supervised learning (based on joint labeled and unlabeled data). First, we apply a semi-supervised clustering strategy to assign pseudo labels, rather than based on classification layers. By performing clustering based on the similarity between labeled and unlabeled embeddings, those incorrect pseudo labels are preliminarily mitigated. Second, we propose a gated label learning (GLL) network that involves flexible threshold and label verification strategies, which balance the quality and quantity of pseudo labels and further select reliable pseudo-label data.

• Based on comparative and ablation experiments, our framework (1) achieves promising speaker recognition performance (1.18%, EER) that approximates supervised learning (0.96%, EER), (2) effectively balances the quality and quantity of pseudo labels, and (3) is superior to the state-of-the-art baseline methods.

• We release our models and codes resulting from this work online\(^1\), and believe this work is pioneer to support future research around SSL-based speaker recognition.

2 Related Work

In this section, we review and compare the related work around speaker recognition and semi-supervised learning.

2.1 Speaker Recognition

Speaker recognition technologies involve identifying and verifying the identity of an individual based on their unique audio characteristics. It is a crucial branch in the domain of artificial intelligence and information security [Singh et al., 2012].

Compared to conventional approaches that perform manual feature engineering [Reynolds et al., 2000; Bai and Zhang, 2021; Noble, 2006; Billson et al., 2019; Campbell et al., 2006], advanced deep learning models adopt Deep Neural Networks (DNNs) to perform speaker recognition that extracts identification features automatically. For example, Snyder et al. [David et al., 2018] proposed a Time Delay Neural Network (TDNN) that contained segment-level as well as time-delay layers, so as to extract time-related features as x-vector for recognition. Desplanques et al. [Brecht et al., 2020] augmented TDNN with a series of Emphasized Channel Attention, Propagation, and Aggregation blocks as ECAPA-TDNN, which learned both temporal and context information for recognition. Deep learning models achieve breakthrough performance benefiting from automatic feature extraction. However, most of them follow supervised learning schemes [David et al., 2021] based on numerous labeled utterances. Expensive annotation expenses make existing work hard to be applied in practice.

In response to the labeled data absence issue, Ali et al. [Yanxiong et al., 2023] devised a few-shot learning method, which designed a feature interaction strategy to enhance the representational ability of its learned embedding. Sun et al. [Cunwei et al., 2018] proposed a Convolutional Neural Network mixed Restricted Boltzmann Machine (TLCNN-RBM) to perform transfer learning. Although these methods mitigate the data annotation issue to certain extents, based on their reported results, limited labeled utterances still make their performance inferior to supervised learning.

2.2 Semi-supervised Learning

Semi-supervised learning (SSL) technologies train deep learning models based on few labeled as well as enough unlabeled data. Especially, those typical methods produce pseudo labels for unlabeled data and thereby enable them to undertake the role of labeled ones [Yassine et al., 2020]. We review existing SSL technologies based on their application domains, i.e., computer vision and speaker recognition.

SSL in Computer Vision

Existing methods are generally categorized into three groups, i.e., consistency regularization, entropy minimization, and holistic (the former two types) methods.

First, consistency regularization methods [Mehdi et al., 2016] request models’ predictions to be consistent across unlabeled as well as augmented data, so as to learn robust and consistent features [Samuli and Timo, 2017; Takeru et al., 2019; Antti and Harri, 2017]. For example, Lee et al. [Samuli and Timo, 2017] proposed a Temporal Ensembling strategy on Pi-Model that adopted an MSE loss to obtain similar predictions between original as well as augmented inputs.

Second, entropy minimization methods [Yves and Yoshua, 2005] minimize the entropy of models’ predictions, so as to encourage models to produce confident and reliable predictions [Hyun, 2013; Qizhe et al., 2020]. For example, Lee et al. [Hyun, 2013] present pseudo labels that pick up the class with the maximum predicted probability, and can be used as if they are true labels. Xie et al. [Qizhe et al., 2020] proposed a self-training method, which assigned pseudo labels from student-teacher models.

Third, holistic methods integrate the above consistent regularization and entropy minimization strategies. Specifically, pseudo labels are assigned under weak augmentation for unlabeled data, and the labeled and pseudo-label data is then jointly utilized with cross-entropy loss [Kihyuk et al., 2020; Bowen et al., 2021]. For example, Sohn et al. [Kihyuk et al., 2020] proposed FixMatch that generated pseudo labels using a model’s predictions on weakly-augmented unlabeled data, and the model was further trained under a fixed threshold when fed the strongly-augmented data. Zhang et al. [Bowen et al., 2021] reported that FixMatch utilizes pre-defined constant thresholds for all classes to select unlabeled data, ignoring different learning statuses and difficulties of different classes. They thus proposed FlexMatch that involved a Curriculum Pseudo Labeling (CPL) strategy to obtain the flexible threshold dynamically for each class. Chen et al. [Ting et al., 2020a] proposed SimCLRv2, which adopted contrastive learning with unlabeled data, and was then fine-tuned by a few labeled ones. They found that semi-supervised learning can benefit from two-stage training strategies, especially contrastive learning.

Among the above SSL methods, the holistic ones are acknowledgedly deemed as the state-of-the-art SSL methods, and have reported promising performance close to supervised learning in computer vision tasks [Xiangli et al., 2023].

\(^1\)Our models, codes and data are available at https://github.com/aitssgll/semi-supervised-speaker-recognition
SSL in Speaker Recognition

Currently, only limited work attempts SSL in the speaker recognition domain. For example, Inoue et al. [Nakamura and Keita, 2020] proposed a framework based on Generalized Contrastive Loss (GCL), which unified losses from supervised metric learning as well as unsupervised contrastive learning. Kreyssig et al. [Kreyssig and Woodland, 2020] proposed a variant of VAT [Antti and Harri, 2017], where the loss was defined as the robustness of the speaker embedding against input perturbations, and measured by the cosine distance (termed as CD-VAT). Tong et al. [Fuchuan et al., 2022] utilized a Graph Convolutional Network (GCN) to cluster pseudo labels for unlabeled data.

In contrast to the mature SSL technologies in computer vision, existing SS-based speaker recognition methods report inferior results than supervised learning, and there is still plenty of room for improvement. Thus, our work adopts the holistic SSL strategy, given their promising performance in computer vision, and aims to achieve performance as promising as supervised learning.

3 Methodology

In this section, we present our proposed two-stage speaker recognition framework illustrated in Figure 1 in detail. As in the figure, Stage I performs contrastive learning as a pre-training task that provides an initial encoder. Stage II performs iterative holistic semi-supervised learning, where a clustering strategy is first applied on the encoder, so as to assign pseudo labels for unlabeled data based on labeled data. Further, in holistic semi-supervised learning, a cross-entropy loss is jointly utilized for supervised loss $L_i$ (for labeled data $x^l$) and unsupervised loss $L_u^{fi}$ and $L_u^{f0}$ (for unlabeled data $x^u$), and the pseudo labels are adopted as supervision signals of the unsupervised loss. Finally, we propose gated labeled learning (GLL) that involves flexible threshold and label verification strategies to further select reliable pseudo-label data. Note that these operations are iteratively conducted.

3.1 Contrastive Learning

Motivated by the conclusion reported in SimCLR-v2 [Ting et al., 2020a], contrastive learning in an upstream task is beneficial to the SSL performance in a downstream task. Such advantages are in accord with our aim to improve the performance of SSL in speaker recognition. Thus, in Stage I, we construct a contrastive learning network that provides an initial SSL framework and produces speaker embeddings.

Specifically, contrastive learning [Danwei et al., 2021; Ting et al., 2020b] trains all unlabeled data in a task-agnostic way with both positive and negative utterance pairs. Formally, the aim is to minimize the distance of the positive pairs:

$$L_{uc} = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{2} \log \frac{\exp(\cos(e_{i,1}, e_{i,2}))}{\sum_{k=1, k \neq i}^{1} \sum_{j \neq 1}^{1} \exp(\cos(e_{i,j}, e_{k,i}))},$$  \hspace{1cm} (1)

where $e_{i,j}$ is the embedding obtained by encoder $f(\cdot)$ on segment $x_{i,j}$, and $\cos(\cdot)$ is a specific cosine similarity function.

From Equation 1, to perform contrastive learning, it is necessary to obtain enough positive and negative pairs. However, meaningful information contained in positive pairs is commonly not enough [Ruijie et al., 2022]. Thus, as illustrated in Figure 1, two separate sub-segments $x_{i,1}$ and $x_{i,2}$ are cut randomly from one utterance to positive pairs, so as to enrich the quantity. In contrast, two segments from different utterances are viewed as negative pairs. Moreover, to avoid false-negative pairs, we adopt large-enough datasets and suitable batch sizes, which are validated to reduce false-negative rates [Haoran et al., 2021]. Additionally, our strategy of contrastive learning follows [Ting et al., 2020b], where an Augmentation Adversarial Training (AAT) loss [Jaesung et al., 2020] is jointly utilized with the contrastive loss, so as to maintain the classification ability of encoder $f(\cdot)$, whilst reducing its mislabeling.

3.2 Iterative Holistic Semi-supervised Learning

Based on the initial network in Stage I, we construct a further SSL framework for speaker recognition in Stage II. Following our analysis in Section 2.2, we adopt the holistic SSL strategy. It generally utilizes labeled data with ground-truth labels as well as abundant unlabeled data with pseudo labels produced under weak augmentation, and the pseudo label is a breakthrough factor to ensure reliable performance [Eric et al., 2020]. Thus, we propose two strategies to assign and select reliable pseudo labels, i.e., clustering and gated label learning.

### Semi-supervised Clustering

Based on labeled and unlabeled utterances, a constrained seed k-means [Sugato, 2002] clustering strategy is adopted to assign...
pseudo labels for the unlabeled utterance, upon the similarity between unlabeled embedding \( e^u \) and labeled embedding \( e^l \). This design preliminarily avoids the labeling of pseudo labels for unlabeled utterances [Hieu et al., 2021], instead of assigning them based on classification layers in conventional methods [Kihyuk et al., 2020; Bowen et al., 2021].

The framework is then trained based on labeled utterances with ground-truth labels as well as unlabeled ones with pseudo labels. Moreover, an additive angular softmax (AAM-softmax) loss [Jiankang et al., 2022] is augmented to the encoder \( f ( \cdot ) \). Note that we adopt an iterative learning strategy. That is, the speaker encoder is iteratively trained with multiple fixed \([Kihyuk et al., 2021; Yidong et al., 2022]\) to be balanced compared with image ones through GLL, so as to apply different guiding model training.

Gated Label Learning

Next, in Stage II, we propose an additional gated label learning (GLL) strategy that involves flexible threshold and label verification strategies, to further select reliable pseudo-label data and balance the quality and quantity of pseudo labels.

General methods select pseudo labels based on either fixed [Kihyuk et al., 2020] or flexible thresholds [Bowen et al., 2021; Yidong et al., 2022]. Although such methods achieve satisfactory performance in computer vision tasks, if high-quality (correct) pseudo labels are over-focused, it inevitably reduces the quantity of selected labels and affects the performance of SSL. In speaker recognition tasks, the quantity and quality of pseudo labels are more challenging to be balanced compared with image ones [Hao et al., 2023; Nayeem et al., 2021]. Thus, we propose GLL that provides a fusion way to filter pseudo labels.

Preliminarily, the \textit{quality} refers to the ratio of the correctly assigned pseudo labels compared with their ground-truth ones:

\[
\text{quality} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y^u_{i}^{pc} == y_u),
\]

where \( y^u_{i}^{pc} \) is the pseudo labels of the unlabeled samples through GLL, \( y_u \) represents their ground truth labels, and \( N \) is the number of unlabeled samples selected by GLL.

The \textit{quantity} refers to the ratio of the selected pseudo-label data among the total unlabeled ones:

\[
\text{quantity} = \frac{1}{N_U} \sum_{i=1}^{N_U} \mathbb{I}(B(i)),
\]

\[
B(\cdot) = \begin{cases} 
q^u_{i} > \tau & \text{GGL == flexible threshold} \\
\text{arg max} (q^u_{i}) = y^u_{i}^{pc} & \text{otherwise}
\end{cases},
\]

where \( y^u_{i}^{pc} \) is the pseudo labels assigned by clustering, \( N_U \) is the number of the total unlabeled samples, and \( \tau \) is the confidence threshold. \( q^u_{i} \) is the speaker recognition prediction on data \( x^u_{i} \), that is, \( q^u_{i} = p(y_{i} | A_{W}(x^u_{i})) \). \( p(\cdot; \cdot) \) is the output probability, where \( A_W (\cdot) \) represents a weak augmentation operation. Note that the ground truth labels of the unlabeled sample in Equation 2 are only used to perform analysis, instead of guiding model training.

First, we apply flexible thresholds [Bowen et al., 2021] into the confidence of the loss function, so as to apply different thresholds to select pseudo-label data. Formally, the flexible threshold is:

\[
\tau_t = \begin{cases} 
\frac{1}{N} & \text{t = 0} \\
\mu \tau_{t-1} + (1 - \mu) \frac{1}{\mu B} \sum_{j=1}^{\mu B} \mathbb{I}(q_{j}^u > \tau_t - 1) \cdot q_{j}^u, & \text{otherwise}
\end{cases},
\]

where \( \mu B \) is the batch size and \( \mu \) is the hyperparameter to balance the increasing speed of threshold \( \tau_t \). Here, the unsupervised loss of the flexible threshold is:

\[
L_{u_{flex}} = \frac{1}{N_U} \sum_{i=1}^{N_U} \left( \mathbb{I}(q^u_{i} > \tau_t) \cdot H \left( y^u_{i}^{pc}, p(y_{i} | A_{S}(x^u_{i})) \right) \right),
\]

where \( H (\cdot; \cdot) \) is the cross-entropy loss, and \( A_{S} (\cdot) \) represents a strong augmentation operation.

Second, we apply label verification as a decision fusion to the loss function, so as to enable pseudo labels to be assigned by clustering or classification. Such strategies provide flexible mechanisms to avoid incorrect pseudo labels caused by poor classification ability at the beginning of model training. Formally, the unsupervised loss of label verification is:

\[
L_{u_{flex}} = \frac{1}{N_U} \sum_{i=1}^{N_U} \left( \mathbb{I}(\text{arg max} (q^u_{i}) = y^u_{i}^{pc}) \cdot H \left( y^u_{i}^{pc}, p(y_{i} | A_{S}(x^u_{i})) \right) \right).
\]

Finally, the overall loss as:

\[
L = L_I + \lambda L_u,
\]

where \( L_I \) is a cross-entropy loss (for training labeled data) and \( L_u \) is the unsupervised training loss involving \( L_{u_{flex}} \) and \( L_{u_{flex}} \). Note that they are chosen alternatively to select reliable pseudo labels until the next re-clustering.

4 Experiment

In this section, we first introduce our experimental setup, and then present our comparative as well as ablation experiments.

4.1 Experimental Setup

We start with introducing the datasets, implementation, and baseline methods of our experiments.

Datasets

To train our framework, we adopt the most typical datasets, VoxCeleb2 [Son et al., 2018], which contains 1092009 utterances from 5994 speakers. In addition, we collect the testing set from VoxCeleb1 [Arsha et al., 2017], which contains 37721 utterance pairs from 40 speakers. Following the setting of typical SSL speaker recognition methods [Long et al., 2021] and the general settings of SSL, i.e., the quantity of unlabeled data should be more than labeled ones, different proportions of utterances per speaker (1% (1 sample), 2% (4 samples), 6% (10 samples), 11% (20 samples), 22% (40 samples), 33% (60 samples)) are selected as labeled data, and the remaining utterances are selected as unlabeled ones.

Implementation

In Stage I, we construct our contrastive learning network based on a Loss-gated Learning (GLL) [Ting et al., 2020b] architecture (one of the state-of-the-art self-supervised models), following the parameter settings in ECAPA-TDNN [Brecht et al., 2020], which is one of the state-of-the-art end-to-end
networks in speaker recognition [Chen et al., 2023]. Specifically, the channel size of ECAPA-TDNN is 1024, and the log mel-spectrogram dimension is 80. In Stage II, we apply strong augmentation settings in X-Vectors [David et al., 2018], and set weak augmentation as no data is augmented. The clustering component is implemented based on a faiss library [Mathilde et al., 2018]. Finally, the optimizer in both stages is Adam [Kingma and Lei, 2015] with an initial learning rate of 0.001. The learning rate is decreased by 5% each epoch in Stage I and is decreased by 5% each epoch in Stage II.

### Baseline Methods

As illustrated in the first (Method) column in Table 1, to evaluate the effectiveness of our framework, we apply multiple state-of-the-art SSL methods in both computer vision and speaker recognition domains. Computer vision methods include the holistic FixMatch [Kihyuk et al., 2020] and FlexMatch [Bowen et al., 2021], the consistency-regularization Mean Teacher [Takeru et al., 2019], the entropy-minimization Pseudo Label [Hyun, 2013], and SimCLRv2 [Ting et al., 2020a] that inspires us from applying the contrastive learning in Stage I. Speaker recognition methods include GCL [Nakamasa and Keita, 2020], CD-VAT [Kreyssig and Woodland, 2020], Mean Teacher [Takeru et al., 2019], and GCN [Fuchuan et al., 2022].

Note that the datasets utilized in speaker recognition and computer vision tasks are greatly different, especially their contained classes. For example, 5994 classes are contained in VoxCeleb2 [Son et al., 2018] and about 1000 classes are contained in ImageNet [Jia et al., 2009]. Thus, for fair comparisons, we explore the best settings for those computer vision methods (to adapt to speaker recognition datasets), by adjusting the confidence threshold.

Specifically, we explore the confidence threshold based on the state-of-the-art holistic method, FixMatch, which commonly sets the threshold to 0.9, and the setting is validated effectively in computer vision tasks. By evaluating FixMatch in our speaker recognition dataset (based on 6% labeled data), we observe that the maximum confidence will not exceed 0.0005 and the maximized mean confidence is 0.00039. We apply the obtained maximized mean confidence (0.00039) as a standard, and multiply it by 0.9 to obtain the best threshold, 0.000351. We also attempt multiple thresholds and find that 0.00051 achieves the best performance in terms of EER (6.49%), and the value is thus selected in the subsequent experiments.

### 4.2 Comparative Experiments

In the following, we first evaluate and compare our framework with other baselines. Then we analyze the quality and quantity of their pseudo labels to empirically explore the reason for their success or failure.

#### Speaker Recognition Performance

Table 1 presents our results based on the metric of Equal Error Rate (EER), which is commonly adopted to evaluate speaker recognition models [Karen and Andrew, 2015; Chen et al., 2023]. In the table, the results of SSL methods are obtained with different proportions of utilized labeled data, and the results of GCL, CD-VAT, and GCN are referred from their original literature.

First, it is observed that our frameworks achieve the best results (as the blue-highlighted values in the table) that outperform any baseline methods as well as our variants. Especially, based on 1% labeled data, we achieve an EER of 3.24%, which is worthy emphasized that our framework achieves promising results with only few labeled data. Moreover, when 33% labeled data is utilized, our framework achieves an approximate EER (1.18%) to the full supervised model (0.96%), which presents great effectiveness and advancements.

Second, SimCLRv2 outperforms others (except our framework), suggesting that the two-stage framework that inspires us is effective in speaker recognition. FixMatch and FlexMatch perform better based on the confidence threshold of 0.000351 than 0.9 in most cases. Moreover, FixMatch and FlexMatch achieve generally better performance than Mean Teacher and Pseudo Label, which demonstrates the superiority of such holistic methods.

Third, Ours w/o GLL achieves better EER even if only 1% labeled data is utilized (4.58%), which indicates our clustering strategy make contributions to the performance.

In general, our framework is effective in the speaker recognition task even if only limited labeled data is available, and is

<table>
<thead>
<tr>
<th>Method</th>
<th>The proportion of utilized labeled data</th>
<th>1%</th>
<th>2%</th>
<th>6%</th>
<th>11%</th>
<th>22%</th>
<th>33%</th>
</tr>
</thead>
<tbody>
<tr>
<td>FlexMatch (0.9)</td>
<td></td>
<td>21.22</td>
<td>10.05</td>
<td>6.68</td>
<td>8.21</td>
<td>6.83</td>
<td>6.43</td>
</tr>
<tr>
<td>FlexMatch (0.000351)</td>
<td></td>
<td>19.89</td>
<td>16.53</td>
<td>18.57</td>
<td>6.99</td>
<td>7.49</td>
<td>1.24</td>
</tr>
<tr>
<td>FixMatch (0.9)</td>
<td></td>
<td>15.62</td>
<td>10.24</td>
<td>9.07</td>
<td>8.44</td>
<td>8.04</td>
<td>7.74</td>
</tr>
<tr>
<td>FixMatch (0.000351)</td>
<td></td>
<td>13.79</td>
<td>9.33</td>
<td>5.27</td>
<td>2.82</td>
<td>2.72</td>
<td>2.53</td>
</tr>
<tr>
<td>Mean Teacher</td>
<td></td>
<td>32.62</td>
<td>10.36</td>
<td>6.35</td>
<td>3.74</td>
<td>1.95</td>
<td>1.79</td>
</tr>
<tr>
<td>Pseudo Label</td>
<td></td>
<td>17.96</td>
<td>11.51</td>
<td>7.11</td>
<td>3.28</td>
<td>6.91</td>
<td>6.52</td>
</tr>
<tr>
<td>SimCLRv2 (Ting et al., 2020a)</td>
<td></td>
<td>6.05</td>
<td>3.16</td>
<td>2.97</td>
<td>2.66</td>
<td>2.59</td>
<td>2.03</td>
</tr>
<tr>
<td>GCL [Nakamasa and Keita, 2020]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.56</td>
</tr>
<tr>
<td>CD-VAT [Kreyssig and Woodland, 2020]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.46</td>
</tr>
<tr>
<td>GCN [Fuchuan et al., 2022]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.30</td>
</tr>
<tr>
<td>Ours Stage I only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.61</td>
</tr>
<tr>
<td>Ours w/o GLL</td>
<td></td>
<td>4.58</td>
<td>2.61</td>
<td>2.15</td>
<td>1.81</td>
<td>1.64</td>
<td>1.37</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td><strong>3.24</strong></td>
<td><strong>1.74</strong></td>
<td><strong>1.65</strong></td>
<td><strong>1.53</strong></td>
<td><strong>1.41</strong></td>
<td><strong>1.18</strong></td>
</tr>
<tr>
<td>Full supervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.96</td>
</tr>
</tbody>
</table>

Table 1: EER (%) results of our framework as well as other baseline methods.
superior to the state-of-the-art methods.

Quality and Quantity of Pseudo Labels

The above experiments demonstrate our framework is superior to those baseline methods. Recall that we attribute our success to the balance of the quality and quantity of pseudo labels. Thus, in this experiment, we first take the holistic methods, FixMatch (0.000351) and FlexMatch (0.000351), as baselines to evaluate whether their pseudo labels are imbalanced, and then compare them with our framework.

Figure 2 presents the quality and quantity results on FixMatch and FlexMatch, where quality and quantity are defined in Equation 2 and Equation 3. In these figures, the quantity and quality changes are evaluated based on utilizing different amounts of labeled data, along with training epochs. First, to observe the results of FixMatch in Figure 2 (b), high-quality pseudo labels are obtained in most cases. However, in Figure 2 (a), the quantity of the selected pseudo-label data is below 0.1 in all cases. Such results confirm our intuition that FixMatch is over-focusing on assigning correct pseudo labels, which results in low quantity, i.e., the selected pseudo-label data for model training is insufficient. Second, to observe the result of FlexMatch in Figure 2 (c), the pseudo labels are high-quantity in all cases, which indicates the utilization rate of unlabeled data is satisfactory. However, in Figure 2 (d), the quality of pseudo labels is unsatisfactory in most cases, which indicates most of the labels are mislabeled. As an exception, as the yellow curve in Figure 2 (d), high-quality pseudo labels can be assigned with training iterations. This is explainable since 33% labeled data is enough for FlexMatch to assign pseudo labels based on its classification layers. In general, the evaluation results are in accordance with our intuitions, that is, such state-of-the-art SSL models can not balance the quality and quantity of pseudo labels when encountering speaker datasets, and are thus not qualified for the speaker recognition task.

Figure 3 presents the comparison results of FixMatch, FlexMatch, and our frameworks. Compared with FixMatch and FlexMatch, our frameworks (with and without GLL) have achieved both high-quality and high-quantity results. Especially, as the green curve in Figure 3 (b), the quality results achieved by our framework (with GLL) are approaching 1.0, which indicates most of the pseudo labels are correctly labeled. Moreover, as the green curve in Figure 3 (a), the success quality results (in Figure 3 (b)) are obtained based on acceptable quantity drops, which are approaching 1.0 after the 14-th epoch. In general, the results have explained the success of our framework, which has the ability to balance the two factors, and is competent for the speaker recognition task.

4.3 Ablation Experiments

Next, we perform ablation experiments to explore the effect of our designed components or strategies on the performance of speaker recognition. Specifically, we first evaluate the flexible threshold and label verification strategies in GLL, and then analyze the impact of iterative learning.

Flexible Threshold and Label Verification

We have implemented four variants of our framework as baselines, their performance is still evaluated on EER, and the quantity and quality of pseudo labels.

Table 2 presents EER results based on different proportions of labeled data. It is first observed that our framework (with GLL) performs best among these baselines in all cases, as the blue-highlighted values in the tables. Second, our framework without GLL generally performs the worst among these methods, which indicates the procedure of selecting pseudo-label data is necessary for the task. For example, based on 1% labeled data, it achieves an EER of 4.58% that is worse than the results of other methods (3.67%, 3.30%, 3.43%, and 3.24%). Third, our framework with either flexible thresholds or label verification, outperforms the fixed-threshold one. For example, based on 2% labeled data, our framework with a fixed thres-
it performs extremely low quality results. Such results are in accord with our intuition that the fixed-threshold strategy is challenging to assign pseudo labels thus affecting speaker recognition performance, and we will further analyze it in the following experiments. Fourth, the superiority of flexible thresholds and label verification is optimal as they are jointly utilized in our framework (with GLL), which demonstrates the rationality of our designed GLL. In general, the evaluation results have further confirmed the effectiveness of our proposed GLL, and each of its involved strategies is indispensable.

Figure 4 presents the quantity and quality of pseudo labels, and the results are obtained using 33% labeled data. Firstly, as the blue curve in Figure 4 (b), the framework with a fixed threshold achieves high-quality results, however, in Figure 4 (a), it performs extremely low quality results. Such results have explained their unremarkable EER performance in Table 2, i.e., it over-focuses on the correctness of pseudo labels resulting in the selected pseudo-label data being insufficient. In contrast, as the green curves in Figure 4, our framework achieves both satisfactory quality and quality results, especially after the 14-th epoch, indicating it has balanced the two factors and has selected reliable pseudo to perform SSL in speaker recognition. In general, these curves are in accordance with the EER results in Table 2, which provides further evidence to prove the necessity of our designed strategies.

Iterative Learning
Recall that our proposed iterative learning strategy iteratively performs clustering and GLL progresses for selecting optimal pseudo labels. In Table 3, we compare the performance of our framework as well as its variants in five iterations, i.e., evaluating their effectiveness with iterative learning.

It is first observed that each result in the second to the fifth iteration outperforms the first iteration, whether our framework is with or without GLL. For example, based on 1% labeled data, our framework (with GLL) achieves an EER of 4.40% in the first iteration, and the results are respectively 3.78%, 3.48%, 3.31% and 3.24% in the rest ones, which indicates our strategy of iteratively assigning and selecting pseudo labels are beneficial to the task. Moreover, we find that the best results (as the blue-highlighted values in the table) commonly appear in the middle (second to fourth) iterations. Based on 2% labeled data, the best EER of our framework (with GLL) appears in the third iteration (1.74%), which indicates the best (quality and quantity) pseudo labels can be selected within limited iterations. However, the degradation in Table 3 as iterations go on is a normal phenomenon in deep learning models, since the model inevitably overfits to labeled data. In general, the experimental results have demonstrated our proposed iterative learning strategies are effective.
References


