Int*-Match: Balancing Intra-Class Compactness and Inter-Class Discrepancy for Semi-Supervised Speaker Recognition

Xingmei Wang, Jinghan Liu, Jiaxiang Meng*, Boquan Li, Zijian Liu

College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
{wangxingmei, liujinghan, mjxwjy, liboquan, 32185246}@hrbeu.edu.cn

Abstract

Open-set speaker recognition is to identify whether the voices are from the same speaker. One challenge of speaker recognition is collecting large amounts of high-quality data. Based on the promising results of image classification, one intuitively feasible solution is *semi-supervised learning* (SSL) which uses confidence thresholds to assign pseudo labels for unlabeled data. However, we empirically demonstrated that applying SSL methods to speaker recognition is non-trivial. These methods focus solely on inter-class discrepancy as thresholds to select pseudo labels, overlooking intra-class compactness, which is particularly important for open-set speaker recognition tasks. Motivated by this, we propose **Int*-Match**, a semi-supervised speaker recognition method selecting reliable pseudo labels with intra-class compactness and inter-class discrepancy for speaker recognition. In particular, we use the inter-class discrepancy of labeled data as the threshold for pseudo-label selection and adjust the threshold based on the intra-class compactness of the pseudo labels dynamically and adaptively. Our systematic experiments demonstrate the superiority of **Int*-Match**, presenting an outstanding Equal Error Rate (EER) of 1.00% on the VoxCeleb1 original test set, which is merely 0.06% below the performance achieved by fully supervised learning.

Code — https://github.com/LiuJinghan2001/IntMatch

1 Introduction

Speaker recognition is to recognize the identity of a speaker based on voices (Kinnunen and Li 2010; Lee et al. 2011). Due to unique pronunciation organs and speaking styles, including vocal tract shapes, larynx sizes, accents, and rhythm, each speaker possesses a distinctive voice, akin to fingerprints, enabling speaker identification (Bai and Zhang 2021; Irum and Salman 2019). Based on this characteristic, speaker recognition often performs open-set recognition tasks, where the cosine similarity between two non-enrolled category samples is evaluated to determine whether they belong to the same speaker. In this paper, we focus on studying the open-set speaker recognition task.

Utilizing extensive high-quality labeled data, existing methods normalize speaker embeddings onto a hypersphere and utilize margins to explicit decision boundaries between different classes (Xiang et al. 2019; Wang et al. 2018b; Deng et al. 2019). This approach optimizes cosine distance during training, enhancing inter-class discrepancy and intra-class compactness. However, obtaining high-quality labeled data is challenging, and the scarcity of such data leads to generalization issues (Ying 2019). Therefore, the lack of high-quality labeled data is still the main challenge faced by existing efforts.

To address such issues, *semi-supervised learning* (SSL) (Zhu and Goldberg 2022) leverages abundant unlabeled data alongside a small portion of labeled ones during training, emerging as a reliable alternative to supervised learning. Specifically, existing SSL methods achieve satisfactory performance in image classification tasks, especially those state-of-the-art (SOTA) ones (Sohn et al. 2020; Zhang et al. 2021; Chen et al. 2023) assign pseudo labels (Lee et al. 2013) for unlabeled data based on confidence thresholds. The produced pseudo-label data plays the role of labeled ones to complete model training. The success of SSL indicates a prospective direction of devising effective semi-supervised speaker recognition models, towards addressing the labeled data lack issue.

We have empirically evaluated the SOTA threshold-based SSL methods (in Section 4.2) on the speaker recognition tasks, and find (1) these methods show limited success, (2) the utilization rate of pseudo labels is limited. That is, if the quality (correctness) of pseudo labels is over-focused, the quantity of the selected pseudo-label data is not enough to complete the recognition task. Moreover, in speaker recognition, enhancing both intra-class compactness and inter-class discrepancy is crucial for effectively performing open-set recognition tasks. However, we find that existing threshold-based SSL methods select pseudo labels based solely on inter-class discrepancy as thresholds, neglecting intra-class compactness. Based on such intuitions, in this work, we propose an effective SSL method (**Int*-Match**) for speaker recognition, towards selecting both high-quality and high-quantity pseudo labels for unlabeled data with intra-class compactness and inter-class discrepancy. To be specific, our contributions in this work mainly include:

- We propose **Int*-Match**, an SSL method to balance the quality and quantity of the selected pseudo labels for speaker recognition. The proposed approach takes both

*Corresponding author
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

intra-class compactness and inter-class discrepancy into consideration which offers the potential of achieving better pseudo-data quality and labeled-data efficiency.

- Our systematic experiments demonstrate the excellent speaker recognition performance of **Int*-Match**, which achieves the best EER of 1.00% on the VoxCeleb1 original test set that outperforms other baseline methods and is approximate to fully supervised learning.

2 Preliminary

In this section, we present the preparing knowledge and the motivations of this work, including the inter-class discrepancy and intra-class compactness in speaker recognition and the limitations of existing threshold-based SSL methods.

2.1 Preparing Knowledge

Let $\mathcal{D}_L = \{x_i^l, y_i\}_{i=1}^{N_L}$ and $\mathcal{D}_U = \{x_i^u\}_{i=1}^{N_U}$ denote the labeled and unlabeled datasets, respectively, where x_i^l and x_i^u are the labeled and unlabeled training samples, and y_i is the corresponding ground-truth label for labeled data. We use N_L and N_U to represent the number of training samples in $\mathcal{D}_L$ and $\mathcal{D}_U$, respectively. For the labeled data, let $z_i^l \in \mathbb{R}^d$ denote the i-th speaker embedding of the input utterance x_i^l with data augmentation. For the most widely used classification loss, the softmax loss is calculated as follows:

$$\mathcal{L}_{softmax}^l = -\log \frac{e^{(W_{y_i}^T z_i^l + b_{y_i})}}{\sum_{j=1}^N e^{(W_j^T z_i^l + b_j)}}, \tag{1}$$

where W_j is the j-th column of the last FC layer weight matrix, $W \in \mathbb{R}^{d \times N}$, and N is the number of class. $b_j \in \mathbb{R}^N$ denotes the bias term, which is simplified to 0 in most cases.

While the softmax-based cross-entropy loss is effective for closed-set classification problems like image classification, where all possible classes are known during training, it fails to produce sufficiently discriminative embeddings for open-set recognition tasks (Deng et al. 2019). To handle this challenge, speaker recognition usually adopts cosine distance to estimate the similarities between pairs of speaker embeddings.

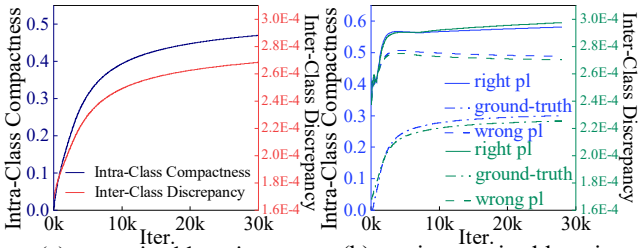
To directly optimize the cosine distance during training, W and z_i^l are l_2-normalized and rescaled to s, distributing the embeddings on a hypersphere with a radius of s (Wang et al. 2017, 2018a; Deng et al. 2019):

$$\mathcal{L}_{Norm}^l = -\log \frac{e^{s \cos_\delta \theta_{i,y_i}}}{\sum_{j=1}^N e^{s \cos_\delta \theta_{i,j}}}, \tag{2}$$

where δ denotes random augmentation, and $\cos_\delta \theta_{i,j}$ denotes the cosine similarity between the i-th speaker embedding and $W_j, j \in [1, N]$.

To enhance intra-class compactness and inter-class discrepancy, existing methods apply an additive angular margin penalty to enforce a margin between the decision boundaries of different class centers. For the classic AAM-softmax (Xiang et al. 2019), the supervised loss is defined as follows:

$$\mathcal{L}_{AAM}^l = -\log \frac{e^{s \cos_\delta (\theta_{i,y_i} + m)}}{e^{s \cos_\delta (\theta_{i,y_i} + m)} + \sum_{j=1, j \neq y_i}^N e^{s \cos_\delta \theta_{i,j}}}. \tag{3}$$



(a) supervised learning (b) semi-supervised learning

Figure 1: Intra-Class Compactness and Inter-Class Discrepancy in (a) supervised and (b) semi-supervised learning.

For unlabeled data, where the ground-truth labels are unknown, we cannot directly compute the loss using Equation (3). Instead, threshold-based SSL methods first predict pseudo labels for the unlabeled data through the cosine prediction distribution, $\cos\theta_{i,j}, j \in [1, N]$. $\hat{y_i} = \arg\max(\cos\theta)_{i,j}, j \in [1, N]$ represents the pseudo label of no-augmented unlabeled data. Subsequently, threshold-based methods select pseudo labels by applying a threshold τ. For data with pseudo labels whose predicted scores exceed the threshold, we calculate unsupervised loss between strongly-augmented unlabeled data and its predicted pseudo label:

$$\mathcal{L}_{AAM}^u = -\beta_i \log \frac{e^{s \cos_\phi (\theta_{i,\hat{y_i}} + m)}}{e^{s \cos_\phi (\theta_{i,\hat{y_i}} + m)} + \sum_{j=1, j \neq \hat{y_i}}^N e^{s \cos_\phi \theta_{i,j}}}, \tag{4}$$

where $\beta_i = \mathbb{I}(softmax(\cos\theta)_{i,\hat{y_i}} > \tau)$, and ϕ denotes strong augmentation. The SSL models ultimately optimize the sum of supervised loss and unsupervised loss to complete the training of a batch.

2.2 Intra-Class Compactness and Inter-Class Discrepancy in Speaker Recognition

In supervised learning, margin-based softmax loss can enforce both intra-class compactness and inter-class discrepancy, as shown in Figure 1 (a). However, in SSL, prediction errors in pseudo labels are common, especially during the early stages of training. This leads to unreliable inter-class discrepancy and intra-class compactness. Therefore, it motivates us to re-examine these two properties.

Intuitively, intra-class compactness represents the angle between embedding z_i and the class center W_{y_i}, which can be expressed through $\cos\theta_{i,y_i}$. For a compact intra-class distribution, z_i should be clustered tightly around W_{y_i}, exhibiting high cosine similarity. Therefore, for a batch of data, intra-class compactness can be defined as the average cosine similarity between the embeddings and their class centers:

$$\gamma_{\textbf{intra}} = \frac{1}{B} \sum_{i=1}^B \cos\theta_{i,y_i}, \tag{5}$$

where B denotes the batch size during training.

Inter-class discrepancy represents the differences in cosine similarity between z_i and different class centers W_j.

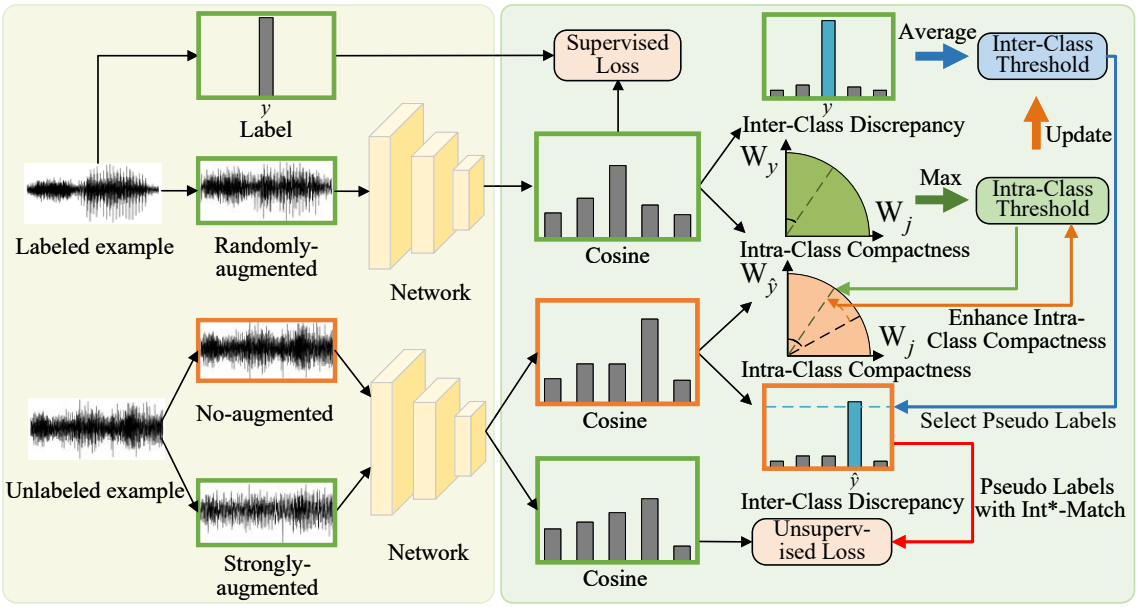


Figure 2: Overview of our method. During training, a batch of labeled data, no-augmented and strongly-augmented unlabeled data are given at the same time. Int*-Match selects reliable pseudo labels using the inter-class threshold (blue box) and restricts their intra-class compactness using the intra-class threshold (green box).

For a distribution with strong inter-class discrepancy, the cosine similarity between z_i and the corresponding class center W_{y_i} should be significantly higher than with other class centers. Based on the above analysis, the softmax function can provide a more intuitive reflection of inter-class discrepancy. Therefore, for a batch of data, the inter-class discrepancy can be defined as the average of the cosine similarities between the z_i and the corresponding class center W_{y_i}, scaled by the softmax function:

$$\gamma_{\mathbf{inter}} = \frac{1}{B}\sum_{i=1}^{B} softmax(\cos\theta)_{i,y_i}. \tag{6}$$

To obtain a more stable estimation, we aggregate $\gamma_{\mathbf{intra}}$ and $\gamma_{\mathbf{inter}}$ by employing Exponential Moving Average (EMA) with a momentum factor m over previous batches:

$$\gamma_{\mathbf{intra}_t} = m\gamma_{\mathbf{intra}_{t-1}} + (1-m)\gamma_{\mathbf{intra}}, \tag{7}$$

$$\gamma_{\mathbf{inter}_t} = m\gamma_{\mathbf{inter}_{t-1}} + (1-m)\gamma_{\mathbf{inter}}. \tag{8}$$

2.3 Limitations of Threshold-based SSL Methods

Threshold-based SSL methods use fixed or dynamic thresholds to select pseudo labels. For example, the classic FlexMatch (Zhang et al. 2021) assigns different confidence thresholds to different classes based on their learning difficulties. The threshold is applied to the predicted scores of pseudo labels, usually normalized by the softmax function. Therefore, it only acts as a filter for inter-class discrepancy.

Figure 1 (b) shows the inter-class discrepancy and intra-class compactness of pseudo labels (pl) and ground-truth labels. It is observed that in the early stages of training, regardless of the accuracy of pseudo labels, the data exhibits high inter-class discrepancy and intra-class compactness within the assigned classes. This impedes the embeddings from forming a discriminative distribution with their ground-truth class vectors, ultimately affecting the performance of open-set speaker recognition. We have visualized the embedding distributions of different methods in Section 4.2.

3 Methodology

In this section, we present our method **Int*-Match** designed to address the limitations of threshold-based SSL methods by balancing intra-class compactness and inter-class discrepancy, as shown in Figure 2. **Int*-Match** adaptively obtain inter-class and intra-class thresholds from labeled data. The inter-class threshold is used to select reliable pseudo labels, while the intra-class threshold evaluates their compactness. When their intra-class compactness exceeds the intra-class threshold, the inter-class threshold adaptively decreases to capture more pseudo labels, and the intra-class threshold adaptively increases to enhance compactness.

3.1 Inter-Class and Intra-Class Thresholds

In this section, we introduce inter-class and intra-class thresholds to balance intra-class compactness and inter-class discrepancy.

Similar to threshold-based SSL methods, the inter-class threshold $\tau_{\mathbf{inter}}$ is used to select pseudo labels. We refine the definition of $\gamma_{\mathbf{inter}}$ in Equation (6) by considering only correctly predicted labeled data to obtain a reliable $\tau_{\mathbf{inter}}$:

$$\gamma_{\mathbf{inter}}^{right} = \frac{1}{M}\sum_{i=1}^{B}\zeta_i \frac{e^{\cos_\delta \theta_{i,y_i}}}{\sum_{j=1}^{N} e^{\cos_\delta \theta_{i,j}}}, \tag{9}$$

where $\zeta_i = \mathbb{I}(\arg\max(\cos_\delta\theta)_{i,j} = y_i), j \in [1, N]$. M is the number of correctly predicted labeled data in the batch,

and M must be greater than 0 when calculating the threshold for this batch. When computing $\tau_{\mathbf{inter}}$ at iteration t, it is set to $\gamma^{right}_{\mathbf{inter}\,t}$.

The intra-class threshold $\tau_{\mathbf{intra}}$ is used to control the intra-class compactness of pseudo labels. It is defined as the average of the maximum intra-class compactness of labeled data within each class. When computing $\tau_{\mathbf{intra}}$ at iteration t, it is set to $\gamma^{max}_{\mathbf{intra}\,t}$:

$$\gamma^{max}_{\mathbf{intra}\,t} = \frac{1}{N}\sum_{j=1}^{N} \max_{(1\le i\le t\times B)} (\mathbb{I}_{(j=y_i)} \cos_\delta \theta_{i,y_i}). \quad (10)$$

3.2 Int*-Match

Int*-Match utilizes the intra-class threshold to regulate the inter-class threshold, thereby balancing the inter-class discrepancy and intra-class compactness of the pseudo labels.

To achieve this, we first set a fixed intra-class threshold τ and initially rely solely on supervised learning to obtain a reliable inter-class threshold.

Subsequently, $\tau_{\mathbf{inter}}$ is used to select pseudo labels and compute the unsupervised loss like Equation (4). When the intra-class compactness of the selected pseudo labels exceeds $\tau_{\mathbf{intra}}$, it indicates that the selected pseudo labels have achieved strong inter-class discrepancy and intra-class compactness. At this point, $\tau_{\mathbf{inter}}$ dynamically decreases to capture more pseudo labels, while $\tau_{\mathbf{intra}}$ dynamically increases to further enhance compactness:

$$\tau_{\mathbf{inter}} = \tau_{\mathbf{inter}} - (\tau_{\mathbf{inter}} - \gamma^{us}_{\mathbf{inter}\,t}) \times \alpha_t, \quad (11)$$

$$\tau_{\mathbf{intra}} = \tau_{\mathbf{intra}} + (\gamma^{max}_{\mathbf{intra}\,t} - \tau_{\mathbf{intra}}) \times \alpha_t, \quad (12)$$

where $\gamma^{us}_{\mathbf{inter}\,t}$ denotes the inter-class discrepancy of the unselected pseudo labels. α_t is an adaptive scaling parameter, defined as the maximum value between the quantity of selected pseudo labels $q_t \in [0, 1]$ and their intra-class compactness $\gamma^{s}_{\mathbf{intra}\,t}$:

$$\alpha_t = \max(q_t, \gamma^{s}_{\mathbf{intra}\,t}). \quad (13)$$

During the early stages of training, $\tau_{\mathbf{inter}}$ is used to select fewer but reliable pseudo labels, so $\gamma^{s}_{\mathbf{intra}\,t}$ is primarily responsible for updating. As $\tau_{\mathbf{intra}}$ increases and $\tau_{\mathbf{inter}}$ decreases, q_t gradually assumes the role of updating.

Through **Int*-Match**, the network can learn high-quantity and high-quality pseudo labels by balancing their inter-class discrepancy and intra-class compactness.

4 Experiment

In this section, we perform systematical experiments to evaluate our proposed **Int*-Match**. We first present the experimental setup, and further analyze our comparative experiments as well as ablation results respectively.

4.1 Experimental Setup

We start with presenting the datasets, implementations, baselines, and evaluation protocols of our experiments.

Dataset. We use the most typical VoxCeleb2 (Chung, Nagrani, and Zisserman 2018) for training, which comprises 1,092,009 utterances contributed by 5,994 speakers. On the other hand, we use the Original, Extended, and Hard VoxCeleb1 test sets (Nagrani, Chung, and Zisserman 2017; Nagrani et al. 2020) for evaluation. We follow the settings to threshold-based SSL methods (Zhang et al. 2021), selecting 2, 4, 10, and 20 utterances per class as labeled data, with the remaining data used as unlabeled data. It is worth noting that choosing 20 utterances per class represents 11% of the training dataset. Furthermore, additional experiments are conducted in Table 2 by selecting 20%, 30%, 40%, and 50% of utterances from each class proportionally, enabling comparisons with fully supervised learning. What's more, we use MUSAN (Snyder, Chen, and Povey 2015) and RIR (Ko et al. 2017) datasets for data augmentation.

Implementation. For a fair comparison, we adopt an identical training strategy across the SSL methods: we employ the popular speaker recognition model ECAPA-TDNN (Desplanques, Thienpondt, and Demuynck 2020) as the network with a channel size of 1024, and the input is an 80-dimensional logarithmic mel spectrum extracted from 2-second speech segments. Meanwhile, the output is a 192-dimensional speaker embedding. The labeled batch size is set to 150, and the unlabeled batch size is the same, with a total training step of 560k. The network parameters are optimized by Adam optimizer (Kingma and Ba 2015), where the initial learning rate is set to 0.001, which decreases 3% in every 7k iterations, roughly one epoch of unlabeled data. We use AAM-softmax loss as the loss function, with the margin as 0.2 and the scale as 30. We assess the maximum inter-class discrepancy per batch on VoxCeleb2 under fully supervised learning. The average of these values increases to 3.84×10^{-4} and stabilizes. Consequently, the confidence threshold for threshold-based SSL methods is set to 3.46×10^{-4} ($3.84 \times 10^{-4} \times 0.9$).

To improve data diversity, we apply strong and random augmentation techniques similar to those used for X-Vectors (Snyder et al. 2018), except that the unlabeled data used for predicting pseudo labels is not subjected to augmentation. The difference between strong and random augmentation is that random augmentation may apply no augmentation with some probability.

For **Int*-Match**, we set m and τ to 0.999 and 0.65, respectivly. Experiments with different τ will be shown in the ablation study (Section 4.3).

Baseline method. We first apply multiple SOTA threshold-based SSL methods in image classification as baselines, including Pseudo label (Lee et al. 2013), FixMatch (Sohn et al. 2020), FlexMatch (Zhang et al. 2021), Dash (Xu et al. 2021), FreeMatch (Wang et al. 2023), and SoftMatch (Chen et al. 2023). These methods rely on threshold-based pseudo labeling and achieve excellent performance in image classification.

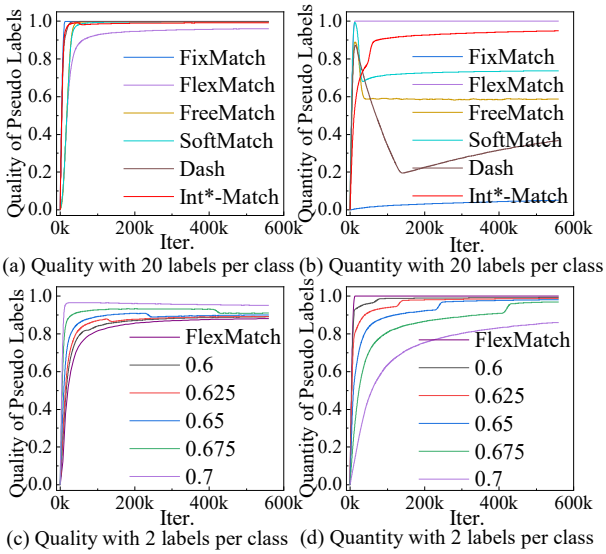
Additionally, we refer to the results of SOTA SSL methods in speaker recognition, including GCL (Inoue and Goto 2020), GCN (Tong et al. 2022) and GLL (Wang et al. 2024).

Evaluation protocol. We evaluate all methods using equal error rate (EER) and minimum detection cost function (minDCF, set $P_{target} = 0.05$), which are common metrics to evaluate the performance of speaker recognition (Kinnunen and Li 2010). Furthermore, we conduct a qualitative

	Evaluation set	VoxCeleb1-O				VoxCeleb1-E				VoxCeleb1-H			
	# Label	2	4	10	20	2	4	10	20	2	4	10	20
EER	Supervised learning	8.68	5.98	3.45	2.14	9.10	6.24	3.86	2.36	13.34	9.74	6.23	4.25
	Pseudo label (Lee et al. 2013)	14.43	10.56	10.45	4.24	14.91	11.09	10.64	4.08	19.10	14.77	13.84	6.50
	FixMatch (Sohn et al. 2020)	8.86	6.23	3.38	2.22	9.11	6.36	3.44	2.32	13.23	9.84	5.86	4.14
	FlexMatch (Zhang et al. 2021)	1.61	**1.52**	1.44	1.38	2.06	1.74	1.69	1.57	3.62	**3.15**	3.02	2.91
	Dash (Xu et al. 2021)	8.25	4.08	2.51	1.91	8.65	4.40	2.68	2.09	12.70	7.10	4.64	3.76
	FreeMatch (Wang et al. 2023)	2.55	2.49	2.20	1.79	2.90	2.69	2.27	1.86	5.06	4.74	4.06	3.36
	SoftMatch (Chen et al. 2023)	2.00	1.95	1.87	1.50	2.37	2.14	2.06	1.75	4.14	3.85	3.71	3.12
	GCL (Inoue and Goto 2020)		6.01				-				-		
	GCN (Tong et al. 2022)		**1.30**				-				-		
	GLL (Wang et al. 2024)	-	1.74	1.65	1.53		-				-		
	Int*-Match	**1.45**	1.54	**1.38**	**1.28**	**1.89**	**1.74**	**1.63**	**1.53**	**3.44**	3.16	**2.99**	**2.81**
minDCF	Supervised learning	0.476	0.364	0.227	0.150	0.500	0.377	0.244	0.153	0.613	0.496	0.339	0.245
	Pseudo label (Lee et al. 2013)	0.614	0.471	0.378	0.393	0.620	0.493	0.424	0.410	0.711	0.595	0.504	0.581
	FixMatch (Sohn et al. 2020)	0.470	0.358	0.213	0.138	0.500	0.381	0.218	0.149	0.616	0.499	0.328	0.240
	FlexMatch (Zhang et al. 2021)	0.120	0.107	0.107	0.100	0.130	0.113	0.108	0.101	**0.212**	**0.188**	0.180	0.174
	Dash (Xu et al. 2021)	0.452	0.262	0.165	0.130	0.483	0.274	0.167	0.135	0.591	0.379	0.263	0.219
	FreeMatch (Wang et al. 2023)	0.170	0.167	0.147	0.119	0.182	0.171	0.147	0.117	0.284	0.269	0.236	0.199
	SoftMatch (Chen et al. 2023)	0.134	0.120	0.125	0.102	0.148	0.137	0.131	0.112	0.234	0.227	0.215	0.186
	Int*-Match	**0.108**	**0.103**	**0.099**	**0.084**	**0.126**	**0.113**	**0.106**	**0.098**	0.214	0.189	**0.180**	**0.169**

Table 1: Performance in EER(%) and minDCF of SOTA methods and proposed Int*-Match on the VoxCeleb1 test sets. The experimental setups of GCL and GCN differ from ours: GCL selected 899 speakers (15% of VoxCeleb2) as labeled data, while GCN used the VoxCeleb1 dev set (equivalent to 14% of VoxCeleb2) as labeled data. Both methods involve more labeled data than our setting of selecting 20 samples per class (11% of VoxCeleb2).

analysis to compare the quality and quantity of pseudo labels selected by SSL methods, providing further evidence of the efficacy of our approach. The quantity of pseudo labels shows the proportion selected by the inter-class threshold relative to the total unlabeled data. The quality of pseudo labels reflects the proportion of correctly predicted pseudo labels among those selected.



(a) Quality with 20 labels per class (b) Quantity with 20 labels per class

(c) Quality with 2 labels per class (d) Quantity with 2 labels per class

Figure 3: (a) Quality and (b) Quantity of pseudo labels in SSL methods and Int*-Match. (c) Quality and (d) Quantity of pseudo labels in the ablation study of Int*-Match.

4.2 Comparative Experiment

In the following, we evaluate and compare the performance of **Int*-Match** as well as other baseline methods.

Comparison with baselines. In Table 1, we compare various SOTA methods as baselines with our proposed method. First, it is observed that **Int*-Match** almost outperforms the SOTA threshold-based SSL methods in terms of EER and minDCF on VoxCeleb1. Second, compared to the SOTA methods in semi-supervised speaker recognition, **Int*-Match** still shows better performance with fewer labels. Specifically, with 20 labels per class (3% fewer labeled data compared to GCN), **Int*-Match** achieves an EER of 1.28%, which is comparable to GCN's 1.30% on VoxCeleb1-O.

Comparison with fully supervised learning. In Table 2, we use ECAPA-TDNN as the backbone model and compare SoftMatch, FlexMatch, and **Int*-Match** to supervised learning and fully supervised learning. **Int*-Match** achieves the best performance, narrowing the average gap to 0.11%/0.6% in EER/minDCF with fully supervised learning.

Qualitative analysis. We provide a qualitative comparison of the threshold-based SSL methods with 20 labels per class, as shown in Figure 3 (a) and (b). It is first observed that **Int*-Match** consistently obtains high-quality pseudo labels with accuracy close to 100% across the training. Moreover, compared to FlexMatch, our method adjusts inter-class thresholds through intra-class compactness, which allows for a more reasonable selection of high-quality and high-quantity pseudo labels than strategies based on learning difficulty, thus enhancing the discriminative power of the speaker embeddings.

	Evaluation set	VoxCeleb1-O				VoxCeleb1-E				VoxCeleb1-H			
	# Label	20%	30%	40%	50%	20%	30%	40%	50%	20%	30%	40%	50%
EER	Supervised learning	1.64	1.42	1.40	1.33	1.78	1.61	1.56	1.55	3.21	2.93	2.84	2.76
	SoftMatch (Chen et al. 2023)	1.66	1.59	1.86	1.75	1.79	1.79	2.00	1.86	3.18	3.12	3.47	3.20
	FlexMatch (Zhang et al. 2021)	1.29	1.27	1.21	1.24	1.49	1.52	1.39	1.44	2.72	2.72	2.58	2.59
	Int*-Match	1.22	1.08	**1.00**	1.10	1.44	1.36	**1.33**	1.34	2.62	2.51	**2.44**	2.48
	Fully supervised learning			**0.94**				**1.22**				**2.29**	
minDCF	Supervised learning	0.116	0.099	0.096	0.095	0.115	0.104	0.102	0.098	0.192	0.179	0.173	0.168
	SoftMatch (Chen et al. 2023)	0.110	0.110	0.130	0.112	0.115	0.113	0.127	0.118	0.188	0.184	0.202	0.188
	FlexMatch (Zhang et al. 2021)	0.087	0.097	0.080	0.088	0.096	0.097	0.091	0.092	0.165	0.169	0.157	0.159
	Int*-Match	0.086	0.081	**0.075**	0.079	0.093	0.088	**0.085**	0.087	0.164	0.156	**0.150**	0.153
	Fully supervised learning			**0.070**				**0.080**				**0.143**	

Table 2: Performance in EER(%) and minDCF of different SSL methods and fully supervised learning on the VoxCeleb1 test sets. The performance of fully supervised learning is based on our settings and may differ slightly from the original paper.

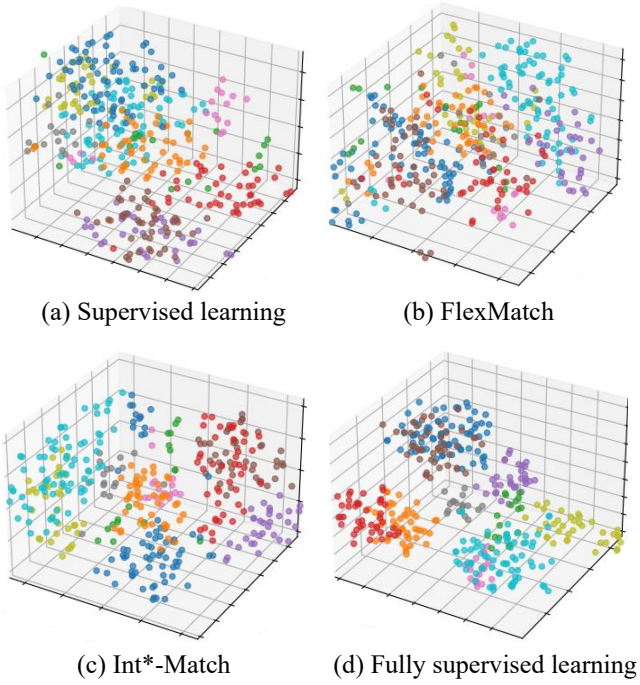
Visualization of intra-class compactness and inter-class discrepancy. To differentiate from the speakers in the training set, we use data from ten classes in VoxCeleb1 to visualize embedding distributions for different methods, as shown in Figure 4. This helps us evaluate the intra-class compactness and inter-class discrepancy of speaker embeddings for open-set speaker recognition. The figure shows that **Int*-Match** and fully supervised learning exhibit strong intra-class compactness and inter-class discrepancy for unknown classes, outperforming FlexMatch. **Int*-Match** achieves high-quality embedding distributions by balancing these two properties with limited labeled data.

τ	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
	2	10	2	10	2	10
0.6	1.55	1.47	1.93	**1.61**	3.50	**2.98**
0.625	1.56	1.40	2.00	1.65	3.62	2.99
0.65	**1.45**	**1.38**	**1.89**	1.63	**3.44**	2.99
0.675	1.55	1.48	1.93	1.73	3.53	3.17
0.7	1.71	1.60	2.05	1.85	3.72	3.32

Table 3: Performance in EER(%) on the VoxCeleb1 test sets with different τ for ablation study.

4.3 Ablation Study

We perform ablation studies to explore the effect of different initial intra-class thresholds τ. First, as shown in Table 3, the best performance in most settings is achieved when τ is set to 0.65. Performance remains relatively robust across most values of τ. Second, as shown in Figure 3 (c) and (d), as the initial intra-class threshold decreases, the inter-class threshold progressively lowers the quality of selected pseudo labels, causing the method to perform similarly to FlexMatch. Conversely, when the initial intra-class threshold is too high, the inter-class threshold does not adequately select a sufficient quantity of pseudo labels. The suitable range for selecting τ can be refined by analyzing the cosine values at the model's decision boundary.



(a) Supervised learning (b) FlexMatch

(c) Int*-Match (d) Fully supervised learning

Figure 4: Visualization of embedding distributions for 10 classes from VoxCeleb1 using t-SNE for (a) Supervised learning, (b) FlexMatch, (c) Int*-Match with 2 labels per class, and (d) Fully supervised learning.

5 Conclusion

This work proposes a novel semi-supervised speaker recognition method, **Int*-Match**, designed to leverage inter-class discrepancy and intra-class compactness to select high-quality and high-quantity pseudo labels. It utilizes the inter-class threshold to select high-quality pseudo labels while employing the intra-class threshold to enhance the compactness of selected pseudo labels and increase their quantity. Experimental results show that **Int*-Match** is superior to the SOTA methods and obtains reliable pseudo labels by balancing intra-class compactness and inter-class discrepancy.

6 Related Work

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

6.1 Speaker Recognition

Currently, deep learning has significantly advanced the field of speaker recognition by effectively extracting highly abstract embedding features from utterances (Snyder et al. 2018; Mun et al. 2020; Inoue and Goto 2020), surpassing traditional methods. We categorize deep learning-based speaker recognition research into three groups, from the perspective of labeled data requirements, i.e., (1) *supervised learning*-based, (2) *self-supervised learning*-based, and (3) *SSL*-based methods.

First, based on enough labeled data, supervised learning-based methods have consistently demonstrated improved performance in speaker recognition over the past few years. In (Variani et al. 2014), a DNN is trained at the frame level to classify speakers, and utilized to extract speaker-specific features for enrollment. Snyder et al. (Snyder et al. 2018) propose to use a time-delay neural network (TDNN) to extract the frame-level features from utterances. Then, a temporal aggregation layer is proposed to aggregate the frame-level features into fixed-length utterances-level representations. Some FC layers are next used to produce classification results. Among existing supervised learning-based methods, TDNN (Desplanques, Thienpondt, and Demuynck 2020; Liu et al. 2022; Mun et al. 2023; Thienpondt, Desplanques, and Demuynck 2021) and ResNet (Zeng et al. 2022; Zhou, Zhao, and Wu 2021) report the best performance.

Second, self-supervised learning methods utilize extensive amounts of unlabeled data, circumventing the limitation posed by the need for labeled data. Given the absence of manual annotations, a common approach in speaker recognition involves employing contrastive learning to acquire meaningful speech representations (Zhang, Zou, and Wang 2021; Zhang and Yu 2022; Sang et al. 2022). For example, Kang et al. (Kang et al. 2022) propose an Augmentation Adversarial Training (AAT) strategy, enabling the network to be speaker-discriminative while remaining invariant to the applied augmentations. Cai et al. (Cai, Wang, and Li 2021) incorporate a second-stage contrastive learning model that generates pseudo labels for unlabeled data using clustering algorithms. Tao et al. (Tao et al. 2022) present a loss-gated learning strategy within a two-stage framework, selecting reliable pseudo labels with a fixed threshold for each iteration. Han et al. (Han, Chen, and Qian 2022, 2024) utilize a Gaussian mixture model to fit data loss, which assigns dynamic thresholds to select pseudo labels.

Third, SSL methods utilize a small amount of labeled data along with a substantial pool of unlabeled data for model training. For example, Inoue et al. (Inoue and Goto 2020) introduce a contrastive learning framework utilizing Generalized Contrastive Loss (GCL) for text-independent speaker verification. Kreyssig et al. (Kreyssig and Woodland 2020) present a Cosine-Distance Virtual Adversarial Training (CD-VAT) approach to tackle speaker recognition tasks. Chen et al. (Chen, Ravichandran, and Stolcke 2021) propose an SSL approach integrating label propagation, primarily focusing on enhancing recognition performance through label inference. Tong et al. (Tong et al. 2022) perform speaker recognition tasks based on the graph convolutional network (GCN), leveraging pseudo-label clustering for unlabeled data. Wang et al. (Wang et al. 2024) propose a two-stage SSL framework, which uses a Gated Label Learning (GLL) strategy to select reliable pseudo-label data.

6.2 SSL in Image Classification

We further explore SSL, referring to its application in image classification domains. Specifically, existing work is mainly divided into three groups, i.e., consistency regularization-based, entropy-minimization-based, and holistic methods.

First, consistency regularization performs SSL by encouraging a model to provide consistent predictions for perturbed versions of the same input data, even when only a small portion of the data is labeled. For example, Laine et al. (Laine and Aila 2017) introduce two notable approaches: Π-Model and Temporal Ensembling, which utilize consistency regularization to promote consistent predictions across variations of input data. Tarvainen et al. (Tarvainen and Valpola 2017) propose Mean Teacher that minimizes the divergence between different augmented outputs as well as quickly integrates information by using a teacher model to mitigate confirmation bias. Miyato et al. (Miyato et al. 2018) propose Virtual Adversarial Training (VAT) that achieves consistency regularization by focusing on perturbations that have the most significant impact on the model's predictions.

Second, entropy minimization minimizes the entropy of the prediction function, encouraging models to provide confident predictions. For example, Lee et al. (Lee et al. 2013) propose to leverage the class with the highest prediction probability from unlabeled data as *pseudo labels*, which enhances learning by incorporating confident predictions from unlabeled datasets. Rizve et al. (Rizve et al. 2021) introduce an Uncertainty-aware Pseudo-label Selection (UPS) framework to refine pseudo labels by reducing noise, thereby improving the utilization rate of unlabeled data. Pham et al. (Pham et al. 2021) propose a meta pseudo labels approach, which employs a teacher network to guide the learning of a student network, and optimizes learning progress based on the performance feedback of the student.

Third, holistic methods that integrate consistency regularization and entropy minimization have demonstrated notable performance improvements. Sohn et al. (Sohn et al. 2020) propose FixMatch, which calculates the loss between pseudo labels of weakly-augmented samples that exceed a specified confidence threshold and their strongly augmented counterparts. Zhang et al. (Zhang et al. 2021) introduce FlexMatch, which considers variations in learning difficulties among different classes and implements a Curriculum Pseudo Label (CPL) method to encourage the selection of pseudo labels. Chen et al. (Chen et al. 2023) propose SoftMatch that employs a truncated Gaussian function to assign weights to unlabeled data, providing a balanced learning approach. Wang et al. (Wang et al. 2023) propose the self-adaptive threshold to adjust the threshold in a self-adaptive manner according to the model's learning status.

Acknowledgments

This work is supported by a grant from the Stable Supporting Fund of Acoustic Science and Technology Laboratory under Grant No.JCKYS2024604SSJS006.

References

Bai, Z.; and Zhang, X.-L. 2021. Speaker recognition based on deep learning: An overview. *Neural Networks*, 140: 65–99.

Cai, D.; Wang, W.; and Li, M. 2021. An iterative framework for self-supervised deep speaker representation learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6728–6732. IEEE.

Chen, H.; Tao, R.; Fan, Y.; Wang, Y.; Wang, J.; Schiele, B.; Xie, X.; Raj, B.; and Savvides, M. 2023. SoftMatch: Addressing the Quantity-Quality Trade-off in Semi-supervised Learning. In *International Conference on Learning Representations (ICLR)*.

Chen, L.; Ravichandran, V.; and Stolcke, A. 2021. Graph-based label propagation for semi-supervised speaker identification. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 4, 2583 – 2587.

Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. VoxceleB2: Deep speaker recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2018-September, 1086 – 1090.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Desplanques, B.; Thienpondt, J.; and Demuynck, K. 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2020-October, 3830 – 3834.

Han, B.; Chen, Z.; and Qian, Y. 2022. Self-Supervised Speaker Verification Using Dynamic Loss-Gate and Label Correction. In *Proc. Interspeech 2022*, 4780–4784.

Han, B.; Chen, Z.; and Qian, Y. 2024. Self-Supervised Learning With Cluster-Aware-DINO for High-Performance Robust Speaker Verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 529–541.

Inoue, N.; and Goto, K. 2020. Semi-supervised contrastive learning with generalized contrastive loss and its application to speaker recognition. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1641–1646. IEEE.

Irum, A.; and Salman, A. 2019. Speaker verification using deep neural networks: A. *International Journal of Machine Learning and Computing*, 9(1).

Kang, J.; Huh, J.; Heo, H. S.; and Chung, J. S. 2022. Augmentation adversarial training for self-supervised speaker representation learning. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1253–1262.

Kingma, D. P.; and Ba, J. L. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Kinnunen, T.; and Li, H. 2010. An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, 52(1): 12–40.

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

Kreyssig, F. L.; and Woodland, P. C. 2020. Cosine-distance virtual adversarial training for semi-supervised speaker-discriminative acoustic embeddings. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2020-October, 3241 – 3245.

Laine, S.; and Aila, T. 2017. Temporal ensembling for semi-supervised learning. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896. Atlanta.

Lee, K. A.; Larcher, A.; Thai, H.; Ma, B.; and Li, H. 2011. Joint application of speech and speaker recognition for automation and security in smart home. In *Annual Conference of the International Speech Communication Association*.

Liu, T.; Das, R. K.; Aik Lee, K.; and Li, H. 2022. MFA: TDNN with Multi-Scale Frequency-Channel Attention for Text-Independent Speaker Verification with Short Utterances. In *IEEE ICASSP*, 7517–7521.

Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1979–1993.

Mun, S. H.; Jung, J.-w.; Han, M. H.; and Kim, N. S. 2023. Frequency and Multi-Scale Selective Kernel Attention for Speaker Verification. In *IEEE Spoken Language Technology Workshop (SLT)*, 548–554.

Mun, S. H.; Kang, W. H.; Han, M. H.; and Kim, N. S. 2020. Unsupervised representation learning for speaker recognition via contrastive equilibrium learning. arXiv:2010.11433.

Nagrani, A.; Chung, J. S.; Xie, W.; and Zisserman, A. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60: 101027.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Proc. Interspeech 2017*, 2616–2620.

Pham, H.; Dai, Z.; Xie, Q.; and Le, Q. V. 2021. Meta pseudo labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11557–11568.

Rizve, M. N.; Duarte, K.; Rawat, Y. S.; and Shah, M. 2021. In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning. In *International Conference on Learning Representations*.

Sang, M.; Li, H.; Liu, F.; Arnold, A. O.; and Wan, L. 2022. Self-supervised speaker verification with simple siamese network and self-supervised regularization. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 6127–6131. IEEE.

Snyder, D.; Chen, G.; and Povey, D. 2015. Musan: A music, speech, and noise corpus. arXiv:1510.08484.

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5329–5333. IEEE.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.

Tao, R.; Lee, K. A.; Das, R. K.; Hautamäki, V.; and Li, H. 2022. Self-supervised speaker recognition with loss-gated learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6142–6146. IEEE.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Thienpondt, J.; Desplanques, B.; and Demuynck, K. 2021. Integrating Frequency Translational Invariance in TDNNs and Frequency Positional Information in 2D ResNets to Enhance Speaker Verification. In *Proc. Interspeech*, 2302–2306.

Tong, F.; Zheng, S.; Zhang, M.; Chen, Y.; Suo, H.; Hong, Q.; and Li, L. 2022. Graph convolutional network based semi-supervised learning on multi-speaker meeting data. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6622–6626. IEEE.

Variani, E.; Lei, X.; McDermott, E.; Moreno, I. L.; and Gonzalez-Dominguez, J. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4052–4056. IEEE.

Wang, F.; Cheng, J.; Liu, W.; and Liu, H. 2018a. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7): 926–930.

Wang, F.; Xiang, X.; Cheng, J.; and Yuille, A. L. 2017. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, 1041–1049.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018b. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5265–5274.

Wang, X.; Meng, J.; Lee, K. A.; Li, B.; and Liu, J. 2024. Two-stage Semi-supervised Speaker Recognition with Gated Label Learning. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 6495–6503. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Wang, Y.; Chen, H.; Heng, Q.; Hou, W.; Fan, Y.; ; Wu, Z.; Wang, J.; Savvides, M.; Shinozaki, T.; Raj, B.; Schiele, B.; and Xie, X. 2023. FreeMatch: Self-adaptive Thresholding for Semi-supervised Learning. In *International Conference on Learning Representations (ICLR)*.

Xiang, X.; Wang, S.; Huang, H.; Qian, Y.; and Yu, K. 2019. Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1652–1656. IEEE.

Xu, Y.; Shang, L.; Ye, J.; Qian, Q.; Li, Y.-F.; Sun, B.; Li, H.; and Jin, R. 2021. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, 11525–11536. PMLR.

Ying, X. 2019. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, 022022. IOP Publishing.

Zeng, C.; Wang, X.; Cooper, E.; Miao, X.; and Yamagishi, J. 2022. Attention Back-End for Automatic Speaker Verification with Multiple Enrollment Utterances. In *IEEE ICASSP*, 6717–6721.

Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; and Shinozaki, T. 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34: 18408–18419.

Zhang, C.; and Yu, D. 2022. C3-DINO: Joint contrastive and non-contrastive self-supervised learning for speaker verification. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1273–1283.

Zhang, H.; Zou, Y.; and Wang, H. 2021. Contrastive self-supervised learning for text-independent speaker verification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6713–6717. IEEE.

Zhou, T.; Zhao, Y.; and Wu, J. 2021. ResNeXt and Res2Net structures for speaker verification. In *IEEE Spoken Language Technology Workshop (SLT)*, 301–307.

Zhu, X.; and Goldberg, A. B. 2022. *Introduction to semi-supervised learning*. Springer Nature.