Mutual Information-Enhanced Contrastive Learning with Margin for Maximal Speaker Separability

Zhe Li, *Student Member, IEEE*, Man-Wai Mak, *Senior Member, IEEE*, Mert Pilanci, *Member, IEEE*, Helen Meng, *Fellow, IEEE*

***Abstract*—Contrastive learning across various augmentations of the same utterance can enhance speaker representations' ability to distinguish new speakers. This paper introduces a supervised contrastive learning objective that optimizes a speaker embedding space using label information from training data. Besides augmenting different segments of an utterance to form a positive pair, our approach generates multiple positive pairs by augmenting various utterances from the same speaker. However, employing contrastive learning for speaker verification (SV) presents two challenges: (1) softmax loss is ineffective in reducing intra-class variation, and (2) previous research has shown that contrastive learning can share information across the augmented views of an object but could discard task-relevant nonshared information, suggesting that it is essential to keep nonshared speaker information across the augmented views when constructing a speaker representation space. To overcome the first challenge, we incorporate an additive angular margin in the contrastive loss. For the second challenge, we maximize the mutual information (MI) between the squeezed low-level features and speaker representations to extract the nonshared information. Evaluations on VoxCeleb, CN-Celeb, and CU-MARVEL validate that our new learning objective enables ECAPA-TDNN to identify an embedding space that exhibits robust speaker discrimination.**

***Index Terms*—Speaker verification; contrastive learning; mutual information; additive angular margin;**

I. Introduction

Speaker representation learning is crucial for speaker verification (SV). Its goal is to learn a feature embedding space characterized by 1) same-class compactness, ensuring that the embedding vectors of the same speaker are close; 2) different-class dispersion, where the embedding vectors belonging to different speakers are far apart. Recent years have witnessed significant advancements in this area, a result of the advancements in deep neural network (DNN) architectures [1], [2], [3], complex loss functions [4], [5], [6], [7], innovative pooling strategies [8], [9], and effective domain adaptation methods [10], [11], [12]. However, the models are still not sufficiently robust to noisy labels [13], [14] and are sensitive to input perturbation unless a notion of margin is introduced to their loss function [15], [16]. Research indicates that these shortcomings can reduce the models' generalization capabilities [17], [18], [19], [20].

Several methods have been developed to increase intra-class compactness and bolster inter-class separation in embedding spaces. Wen *et al.* [21] proposed a regularization term to penalize the gaps between features and their corresponding centers. Building upon this idea, Ranjan *et al.* [22] and Wang *et al.* [23] suggested constraining the L_2 norm of the feature representations for the softmax loss so that they lie on a hypersphere with a fixed radius, and Liu *et al.* [24] proposed optimizing the cosine similarity between the feature representations and their class centroids. These adjustments result in well-separated classes in the representation space and reduce intra-class dispersion, resulting in larger gradients during training. Furthermore, Liu *et al.* [16] argued for an enlarged classification margin, emphasizing that a more challenging learning objective can stimulate the acquisition of more discriminative features. Similarly, Liu *et al.* [25] introduced an angular distance metric. This metric evaluates the dissimilarity of objects based on their geodesic distance within a hypersphere manifold and uses an angular margin to heighten the strictness of decisions.

Contrastive learning is increasingly gaining attention in the SV community [26], [27], [28], [29]. This approach creates positive pairs using augmented samples of an utterance from the same speaker. It considers different utterances and their augmented versions as being from distinct speakers, thus forming negative pairs. The overarching goal is to draw the embeddings of the positive pairs closer while distancing the embeddings of the negative pairs. Because the supervised information for one view comes from the other view, contrastive learning can leverage the shared information across views, but often overlooks nonshared task-relevant information. Shared information corresponds to speaker features relevant to the SV task, and the features are shared across different views of the utterances. For example, a waveform after noise contamination will still contain some information about the same speaker. Nonshared information, on the other hand, refers to speaker features that are specific to the test speakers but not shared between different views during contrastive learning; Wang *et al.* [30] also theoretically proved that the nonshared information cannot be ignored; otherwise, the representation learned through contrastive learning may not be sufficient for the downstream tasks.

A speaker embedding network optimized by contrastive learning will also tend to ignore nonshared information because the network can never see the test speaker population during contrastive training. An intuitive approach to reinforce the nonshared information in the embeddings is to explicitly

Man-Wai Mak is the corresponding author. Zhe Li and Man-Wai Mak are with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong SAR. Mert Pilanci is with the Department of Electrical Engineering at Stanford University, USA. Helen Meng is with the Department of Systems Engineering & Engineering Management at The Chinese University of Hong Kong, Hong Kong SAR.

maximize the mutual information between low-level features and segment-level embeddings. This encourages the embeddings to capture speaker information preserved in lower-level representations [19], [31].

To facilitate maximal speaker separability, many investigations [32], [33], [34] have employed the NT-Xent loss, which is essentially a variant of the cross-entropy loss integrated with a softmax function. Nevertheless, recent findings [7], [35], [36] suggest that while the conventional softmax-based loss can effectively enlarge inter-class discrepancies, it is ineffective in minimizing intra-class variations. This phenomenon implies that the resulting features, although discriminative for closed-set classification, are inadequate for open-set speaker recognition. Moreover, the widespread contrastive strategies emphasize distinguishing between positive and negative pairs [37], [38], [39], with little attention to exploring optimization objectives.

To alleviate the above challenges, we designed a speaker verification framework to learn discriminative speaker representations using mutual information-enhanced contrastive learning with margin. The capability of speaker representation is enhanced by incorporating an additive angular margin into the supervised contrastive loss. Meanwhile, our framework increases the mutual information between the speaker representation and the first convolutional layer of the speaker encoder to capture more nonshared information.

This paper substantially extends our earlier work in [40], [41]. Firstly, the paper empirically verifies that the minimal sufficient representation [30] is not sufficient for speaker verification because it misses the nonshared information across the augmented views. We maximize the mutual information between the low-level features and utterance-level embeddings to enhance the preservation of useful information. Our comprehensive experiments demonstrate that incorporating squeezed frame-level phonetic information into the embedding extractor consistently improves speaker verification performance. Secondly, the paper adds comprehensive analyses to investigate the impacts of the proposed method on speaker representation. The analyses include a detailed exploration of the angular margin's influence, an investigation into the impact of varying the number of positive samples, and an analysis of alignment and uniformity. Thirdly, we have extended our experimental evaluations from the VoxCeleb1 dataset to encompass the larger and more challenging VoxCeleb2 dataset. This expansion ensures a more comprehensive validation of our proposed method across different datasets. Fourthly, we investigate the behavior of our method under low-resource scenarios using a Cantonese dataset called CU-MARVEL. The dataset was originally developed for dementia detection, and we repurposed it for speaker verification research under low-resource conditions.

II. Methodology

We aim to develop a speaker representation network based on contrastive learning using labeled audio data. The embedding vectors should cluster together for similar speakers and be distant apart for dissimilar speakers. To this end, for each training batch, we apply data augmentation to create diverse samples for each utterance in the batch. Despite various augmentations, the embedding vectors of the same instance should remain consistent. Conversely, embeddings from different samples should be distinct.

As depicted in Fig. 1, an encoder network processes the spectrograms of both the original instances and their augmented samples. This process yields a set of normalized embeddings. At the end of this process, we maximize the mutual information between the representation and the frame-level embedding from the encoder. We also compute the contrastive loss with an additive angular margin on the network's output.

A. Representation Learning Framework

Inspired by recent contrastive learning methods, our method aims to enhance representation learning. It maximizes the agreement across various augmented views of the same data via contrastive loss in the embedding space. As illustrated in Fig. 1, the framework consists of four pivotal components.

a) Data Augmentation: For each input sample, we generated one or multiple random augmentations, denoted as $\hat{\boldsymbol{x}}_i = Augmentation(\boldsymbol{x}_i)$. Each augmentation provides a unique data perspective and comprises some of the original sample's information. Following the Kaldi's recipe [42], we employed augmentation techniques such as adding noise, music, and chatter from the MUSAN dataset [43]. Additionally, we generated reverberation effects by convolving the original waveforms with room impulse responses (RIR) from the RIR dataset [44]. We also employ speed perturbation [45].

b) Encoder Network: Our primary objective involves training an encoder network using a labeled audio dataset $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$. $h_\theta(\cdot)$ denotes the first convolutional layer in the speaker encoder followed by global pooling, transforms each input audio $\boldsymbol{x}_i$ into a low-dimensional vector $\boldsymbol{h}_i = h_\theta(\boldsymbol{x}_i) \in \mathbb{R}^{T\times d}$, where T is the number of frames and d represents the dimension. Both original and augmented samples are independently fed into the same encoder, resulting in two representation vectors $\boldsymbol{h}_i$ and $\hat{\boldsymbol{h}}_i$.

c) Projection Network: The projection network, denoted as $g_\phi(\cdot)$ in Fig. 1, is a shallow network with one linear output layer responsible for transforming the encoder's output into a space where we apply the contrastive loss. We normalize the network's output to ensure the embedding vectors lie on a unit hypersphere. This normalization enables us to estimate distances in the projection space using inner products.

B. Recap of Supervised Contrastive Learning with Margin

1) Supervised Contrastive Learning: As shown in Fig. 2, we explore the supervised contrastive loss, where positive examples of a given class are contrasted with negative examples from different classes, utilizing the provided labels. We incorporated the original and augmented speaker embeddings into a supervised contrastive loss [40], [46]:

$$\mathcal{L}_{SupCon} = \sum_{i=1}^{N} \frac{-1}{|\mathcal{P}(i)|} \sum_{p\in\mathcal{P}(i)} \log \frac{\exp(sim(\boldsymbol{z}_i, \boldsymbol{z}_p)/\tau)}{\sum_{a\in\mathcal{A}(i)} \exp(sim(\boldsymbol{z}_i, \boldsymbol{z}_a)/\tau)}, \tag{1}$$

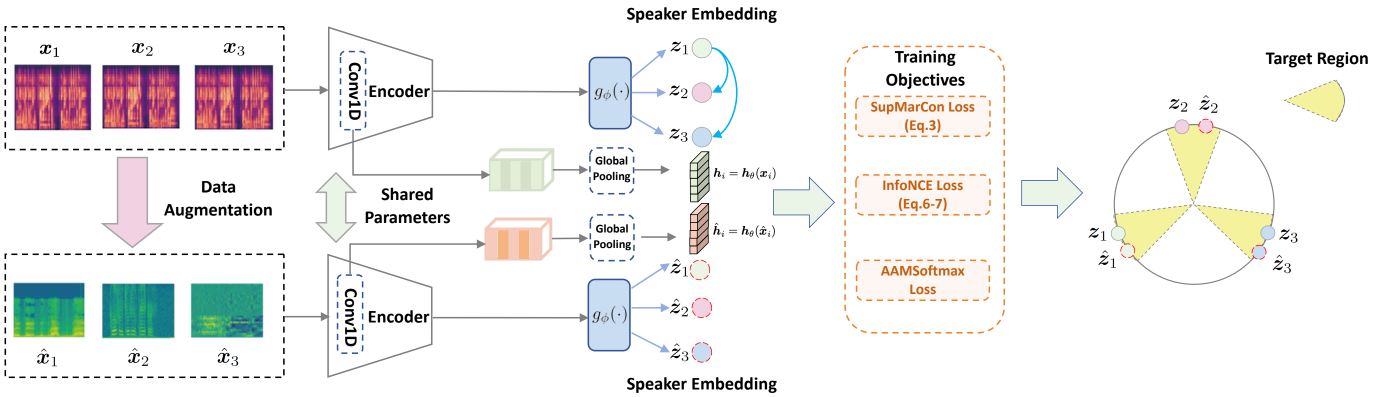


Fig. 1: Our architecture uses additive angular margin for mutual information-enhanced supervised contrastive learning. The encoder transforms acoustic features (MFCC or FBank) into normalized embedding vectors. Invariance occurs for the embeddings (e.g., z_1 and $\hat{z}_1$) whose acoustic features (x_1 and $\hat{x}_1$) come from the same speaker. On the other hand, embeddings (e.g., z_1 and z_2) whose acoustic features (x_1 and x_2) belong to different speakers are far apart.

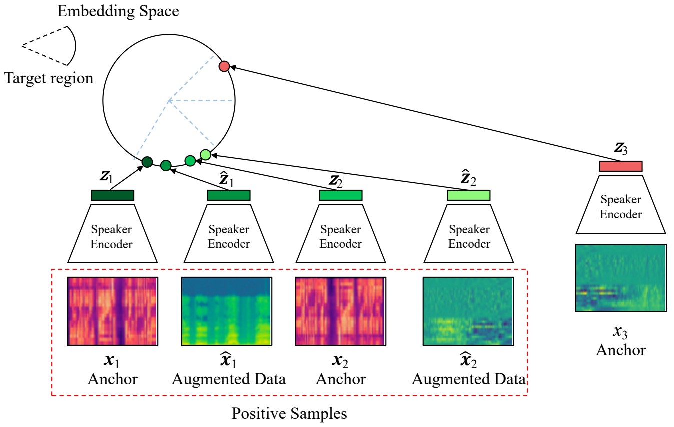


Fig. 2: Our basic idea is illustrated by contrasting all samples from the same class (positives) against those from other classes (negatives) in a batch. By incorporating class label information, we create an embedding space where similar speakers stay close to each other while dissimilar ones are far apart. In this example, x_1 and x_2 come from the same speaker, whereas x_3 comes from another speaker.

where $sim(z_i, z_p)$ is the cosine similarity. In Eq. 1, z_i is an anchor, z_a is a negative sample, $\mathcal{A}(i)$ comprises the indices of the negative samples with respect to z_i, z_p is a positive sample with respect to z_i, and $\mathcal{P}(i)$ contains the indices of positive samples in the augmented batch (original + augmentation). $\tau \in \mathcal{R}^+$ is a scalar temperature parameter.

2) Angular Margin Based Contrastive Learning: Although the training objective attempts to pull the representations of similar speakers closer together and push the representations of different speakers apart, these representations may not be sufficiently discriminative or robust against noise. Let us denote the cosine similarity as

$$\cos\theta_{i,p} = \frac{z_i^\top z_p}{\|z_i\| \, \|z_p\|}, \tag{2}$$

where $\theta_{i,p}$ is the angle between the embeddings z_i and z_p. A similar formula applies to z_i and z_a. The decision boundary of z_i for specific p and a is $\theta_{i,p} = \theta_{i,a}$, where p and a index to the positive and negative samples, respectively (Fig. 3a. A small perturbation of the embedding vectors around the decision boundary may result in an incorrect decision if no decision margin exists (Figs. 3b and 3c. To overcome this problem, we advocate adding an additive angular margin m to the decision boundary. We name the resulting objective as **sup**ervised **margin con**trastive (SupMarginCon) loss [41], which is formulated as:

$$\mathcal{L}_{SupMarginCon} = \sum_{i=1}^{N} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\cos\left(\theta_{i,p} + m\right)/\tau\right)}{\sum_{a \in A(i)} \exp\left(\cos\left(\theta_{i,a}\right)/\tau\right)}. \tag{3}$$

As shown in Fig. 3d, in this loss, the decision boundary of z_i for specific p and a is $\theta_{i,p} + m = \theta_{i,a}$. The minimization of Eq. 3 will push z_i further towards the area where $\theta_{i,p}$ decreases and $\theta_{i,a}$ increases. Therefore, adding a margin can increase the compactness of same-speaker representations and the divergences between the different-speaker representations. This aid improves alignment and uniformity – two quality measures fundamental to contrastive learning [47]. These metrics indicate how close positive-pair embeddings are to one another and how uniformly distributed the embeddings are. These properties make the SupMarginCon loss more discriminative than the conventional loss, such as the SupCon loss (Eq. 1).

C. Leveraging Nonshared Speaker Information

In contrastive learning, the augmented views provide supervision information for an anchor. For example, the input $\hat{x}_i$ in Fig. 1 and Fig. 2 provides a supervision signal to x_i because they come from the same utterance. The signal plays a similar role as class labels in supervised learning [48]. The analyses in [30] suggest that in contrastive learning, the minimal sufficient representation falls short for downstream tasks due to the missing nonshared task-related information in the representations. Additionally, contrastive learning tends to produce a minimal sufficient representation (i.e., ignoring the

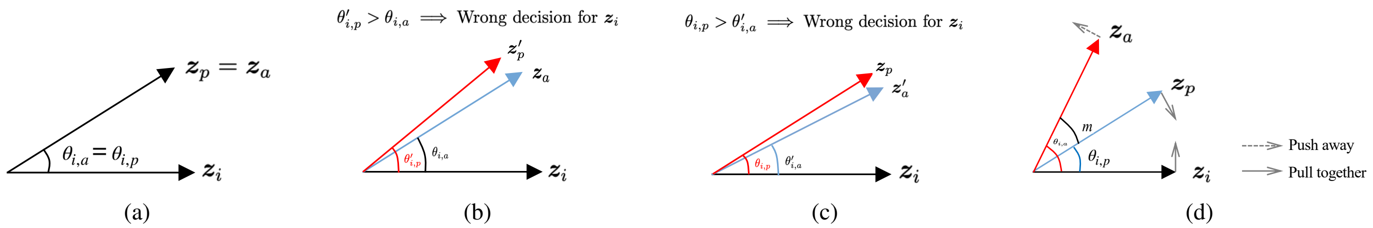


Fig. 3: (3a) Without a decision margin, the decision boundary for $\boldsymbol{z}_i$ is $\theta_{i,p} = \theta_{i,a}$. (3b) – (3c) A small perturbation on $\boldsymbol{z}_p$ or $\boldsymbol{z}_a$ but in the wrong directions can lead to incorrect decisions. (3d) *SupMarginCon* incorporates an additive angular margin m, ensuring the decision boundary for $\boldsymbol{z}_i$ satisfies $\theta_{i,p} + m = \theta_{i,a}$ for specific positive and negative samples. With the tolerance m, both $\boldsymbol{z}_p$ and $\boldsymbol{z}_a$ can be subject to a larger perturbation without causing a wrong decision for $\boldsymbol{z}_i$.

nonshared information between multiple views of the same object), thus risking overfitting the shared information across views.

Unlike contrastive learning methods that use InfoNCE [49], [50] to maximize the similarity between positive samples [51], [52], [53] or the approaches that leverage mutual information to disentangle speaker embeddings from factors such as age and domain [54], [55], [56], our method employs InfoNCE [50] to increase the mutual information between low-level features and utterance-level embeddings. We extract additional nonshared information from $\boldsymbol{h}_i$ and $\hat{\boldsymbol{h}}_i$. $\boldsymbol{h}_i$ is the output of the first convolutional layer of the speaker encoder followed by global pooling, sharing the same dimensionality as $\boldsymbol{z}_i$. $\hat{\boldsymbol{h}}_i$ is the augmented version of $\boldsymbol{h}_i$. We maximize the mutual information $I(\boldsymbol{z}_i, \boldsymbol{h}_i)$ and $I(\hat{\boldsymbol{z}}_i, \hat{\boldsymbol{h}}_i)$ to enhance the speaker information in $\boldsymbol{z}_i$ and $\hat{\boldsymbol{z}}_i$. Given the symmetry between $\boldsymbol{h}_i$ and $\hat{\boldsymbol{h}}_i$, our objective is to maximize

$$I\left(\boldsymbol{z}_i, \boldsymbol{h}_i\right) + I\left(\hat{\boldsymbol{z}}_i, \hat{\boldsymbol{h}}_i\right). \tag{4}$$

For optimizing $I\left(\boldsymbol{z}_i, \boldsymbol{h}_i\right)$ and $I\left(\hat{\boldsymbol{z}}_i, \hat{\boldsymbol{h}}_i\right)$, we choose the InfoNCE as the lower bound estimates of mutual information. Concretely, the InfoNCE lower bound is [30]

$$\hat{I}_{NCE}(\boldsymbol{z}, \boldsymbol{h}) = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \ln \frac{p\left(\boldsymbol{z}_i \mid \boldsymbol{h}_i\right)}{\frac{1}{N}\sum_{l=1}^{N} p\left(\boldsymbol{z}_l \mid \boldsymbol{h}_i\right)}\right], \tag{5}$$

where $\{\boldsymbol{z}_l\}_{l=1}^{N}$ are sampled from the conditional distribution $p(\boldsymbol{z}|\boldsymbol{h})$, with $\boldsymbol{h}_i$ drawn from the mini-batch, and N is the batch size.

To compute the InfoNCE [50] lower bound, we require a probabilistic model for $p(\boldsymbol{z}|\boldsymbol{h})$ from which N samples of $\boldsymbol{z}$, $\{\boldsymbol{z}_l\}_{l=1}^{N}$, are drawn. Following [30], [57], we employ the reparameterization trick during training. Specifically, we model $p(\boldsymbol{z}|\boldsymbol{h})$ as a Gaussian distribution $\mathcal{N}(\boldsymbol{z}; f_\theta(\boldsymbol{h}), \sigma^2\boldsymbol{I})$, where σ^2 is a pre-defined variance and $f_\theta(\boldsymbol{h})$ is a deterministic function implemented by a DNN parameterized by θ. Consequently, we have $\boldsymbol{z} = f_\theta(\boldsymbol{h}) + \sigma\boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Then, $\hat{I}_{NCE}$ is equivalent to

$$\hat{I}_{NCE}(\boldsymbol{z}, \boldsymbol{h}) = \mathbb{E}\left[-\frac{1}{N}\sum_{i=1}^{N} \ln \sum_{l=1}^{N} \exp\left(-\rho \left\|\boldsymbol{z}_l - f_\theta\left(\boldsymbol{h}_i\right)\right\|_2^2\right)\right], \tag{6}$$

where ρ is a scale factor. For estimating $I(\hat{\boldsymbol{z}}_i, \hat{\boldsymbol{h}}_i)$, the approach is the same. Therefore, the loss function for maximizing the mutual information is:

$$\mathcal{L}_{\mathit{InfoNCE}} = -\hat{I}_{NCE}(\boldsymbol{z}, \boldsymbol{h}) - \hat{I}_{NCE}(\hat{\boldsymbol{z}}_i, \hat{\boldsymbol{h}}_i). \tag{7}$$

D. Model Training

After finishing the contrastive loss minimization, the encoder's parameters are typically frozen before training a linear classification layer. However, we advocate achieving both contrastive and classification learning simultaneously. To this end, we introduce AAMSoftmax [7] to our classification task, which is optimized alongside the contrastive loss during training.

The SupMarginCon, incorporating InfoNCE loss [50], [30], can be added to the total loss as a regularization term. The combination can be implemented as follows:

$$\mathcal{L} = \mathcal{L}_{\mathit{AAMSoftmax}} + \mathcal{L}_{\mathit{SupMarginCon}} + \lambda\mathcal{L}_{\mathit{InfoNCE}}. \tag{8}$$

We aim to enhance the sufficiency of the information in the representations without compressing it. Additionally, we must avoid introducing excessive nonshared information to $\boldsymbol{z}$ from $\boldsymbol{h}$. We utilize a coefficient λ to control this.

E. Analysis and Discussion

Our proposed SupMarginCon loss function (Eq. 3) incorporated margin into SupCon (Eq. 1). The SupCon loss leads to an innovative contrastive approach that allows multiple positives for each anchor. Our proposed loss effectively leverages label information in contrastive learning and derives highly discriminative features essential for speaker verification. The proposed supervised contrastive learning-based framework has the following advantages:

- **Generalization to arbitrary positives**. Within a multiview batch, every anchor benefits from its augmented sample and other samples with the same label, contributing to the loss function's numerator. The supervised loss guides the encoder to generate representations that closely align with their respective classes, resulting in denser speaker clusters within the embedding space.
- **Enhanced contrastive capability with increased negatives:** As indicated by Eq. 3, the loss function has a sum over the negatives in the denominator. As a result,

the capability to distinguish between noise and signal is enhanced when more negative samples are added.

- **Additive margin increases discriminative power:** The additive-angular-margin supervised contrastive loss improves speaker discrimination by increasing the decision margin in the angular space.

III. Experiments and Results

A. Implementation Details

We incorporated the proposed loss function into the models in the 3D-Speaker toolkit [58] and evaluated them on the VoxCeleb [59], [60], CN-Celeb [61], [62], and CU-MARVEL [63] datasets for speaker verification. We used various architectures in the 3D-Speaker toolkit and the ERes2NetV2 architecture [64] for the encoder. We utilized 80-dimensional Fbank vector as input features. Our experiments incorporated four types of data augmentations: room impulse responses, music, background noise, and babble noise. We employ speed perturbation [45] with scaling factors of 0.9 and 1.1. We use mini-batches of size 1024, and for each utterance in a mini-batch, we randomly extracted a 3-second segment. The Adam optimizer was used. The parameter m in Eq. 3 was set to 0.2 or 0.3. The margin and scale in AAM-Softmax were set to 0.3 and 32, respectively. The contrastive learning temperature τ was set to 0.07. For Eq. 6, $\sigma = 0.1$ and $\rho = 0.05$.

B. Results and Analysis

Table I presents a comprehensive evaluation of various speaker encoders trained on VoxCeleb2 and tested on VoxCeleb1-O, VoxCeleb1-E, and VoxCeleb1-H. Multiple architectures including Res2Net, ResNet34, ECAPA-TDNN, ERes2Net, CAM++, and ERes2NetV2 were evaluated across several loss functions: Cross-Entropy, AM-Softmax, AAM-Softmax, and our proposed loss function combining AAM-Softmax, supervised contrastive learning with margin, and mutual information enhancement (Eq. 8).

Across all architectures and test sets, our proposed method consistently achieved superior performance. Specifically, the ERes2NetV2 architecture with our proposed loss achieved the best overall results, obtaining EERs of 0.53%, 0.66%, and 1.21%, and minDCFs of 0.049, 0.071, and 0.121 on VoxCeleb1-O, VoxCeleb1-E, and VoxCeleb1-H respectively. This represents notable improvements over the baseline AAM-Softmax loss with relative reductions of approximately 14.5%, 14.3%, and 17.1% in EER, highlighting the effectiveness of our approach.

The results further illustrate that margin-based losses (AM-Softmax, AAM-Softmax, and our loss) significantly outperform traditional cross-entropy across all tested architectures and datasets, reinforcing the efficacy of margin-based constraints in speaker verification. Moreover, our approach consistently outperforms standard AAM-Softmax, demonstrating the complementary advantages provided by supervised contrastive learning and mutual information enhancement, particularly in the more challenging test set VoxCeleb1-H.

Our proposed method demonstrates consistent superiority on the CN-Celeb evaluation set, achieving the best results across all speaker encoders and loss functions. As shown in Table II, the ERes2NetV2 architecture combined with our loss function achieves an EER of 5.73% and a minDCF of 0.341, outperforming the second-best AAM-Softmax baseline (6.14% EER, 0.370 minDCF). This trend holds across all architectures, with our method consistently reducing EER and minDCF over traditional loss functions. The improved performance on CN-Celeb, which includes diverse and challenging real-world scenarios, further underscores the generalization capability of our method. These results align with the findings on VoxCeleb, confirming the effectiveness of our hybrid optimization strategy across different datasets and evaluation protocols.

C. Comparing with Margin-based Contrastive-based Loss

To verify the effectiveness of our proposed loss function, we compare it against several well-known contrastive learning and margin-based methods, including AMC-loss [69], triplet loss [11], [70], angular prototypical loss (Ang-Prototy) [71], and CBRW-BCE [72]. AMC-Loss [69] combines traditional cross-entropy loss with an angular margin, explicitly minimizing geodesic distances within classes and maximizing interclass angular separations. Triplet loss [70] is closely related to supervised contrastive learning, representing a special case of contrastive loss that uses exactly one positive and one negative sample per anchor. Angular prototypical loss [71] does not require explicit speaker identities for each utterance; instead, positive pairs are sampled from within the same utterance and negative pairs from different utterances. CBRW-BCE [72] leverages a bipartite ranking method to mitigate the imbalance of trials, integrating curriculum learning that gradually selects harder negative samples, thus improving training stability and model performance.

Table III summarizes the comparison results. When ECAPA-TDNN was used as the speaker encoder, the proposed loss achieved an EER of 0.74% and minDCF of 0.096, significantly outperforming AMC-Loss (2.54% EER), triplet loss (2.30% EER), angular prototypical loss (1.19% EER), and CBRW-BCE (1.10% EER). These results demonstrate the clear advantage of our method in terms of speaker discriminative capability.

Table III also shows that under the AAM-Softmax loss, the speaker encoder CAM++ [68] and ECAPA++ [73] achieve a similar performance (0.66% and 0.65% EER, respectively) but outperform IM-ECAPA-SimAM [74] and NeXt-TDNN [75] (0.79% EER). Notably, when trained with our proposed loss, CAM++ [68] can achieve an even better performance (0.59% EER, 0.076 minDCF), surpassing both ECAPA++ [73] and NeXt-TDNN [75], despite these two encoders are more advanced. This observation indicates that our proposed loss function can significantly enhance the performance of encoders with simpler architectures, demonstrating its robustness and effectiveness across different speaker encoders.

D. Ablation Study

To further understand the contributions of each component in our proposed loss function, we conducted an ablation study using the ERes2NetV2 architecture, shown in Table IV.

TABLE I: Performance comparison of speaker encoders trained on VoxCeleb2 and evaluated on VoxCeleb1 test sets (VoxCeleb1-O, VoxCeleb1-E, and VoxCeleb1-H), using various loss functions, including Cross-Entropy, AM-Softmax, AAM-Softmax, and our proposed one (AAM-Softmax + supervised contrastive learning with margin + mutual information enhancement). The best results are highlighted in **bold**.

Speaker Encoder	Loss function	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
		EER (%)	minDCF	EER (%)	minDCF	EER (%)	minDCF
Res2Net [65]	Cross-Entropy	4.12	0.453	4.31	0.461	5.52	0.552
	AM-Softmax	1.63	0.181	1.52	0.161	2.61	0.301
	AAM-Softmax	1.56	0.151	1.42	0.149	2.48	0.231
	Ours (Eq. 8)	1.41	0.132	1.26	0.136	2.21	0.212
ResNet34 [66]	Cross-Entropy	3.95	0.431	4.16	0.441	5.21	0.531
	AM-Softmax	1.26	0.121	1.26	0.121	2.11	0.201
	AAM-Softmax	1.05	0.108	1.12	0.117	1.99	0.193
	Ours (Eq. 8)	0.93	0.099	0.99	0.106	1.79	0.176
ECAPA-TDNN [3]	Cross-Entropy	3.71	0.411	3.86	0.421	5.91	0.511
	AM-Softmax	1.06	0.121	1.16	0.121	2.01	0.201
	AAM-Softmax	0.87	0.117	0.98	0.113	1.91	0.194
	Ours (Eq. 8)	0.74	0.096	0.83	0.103	1.63	0.166
ERes2Net [67]	Cross-Entropy	4.51	0.391	3.66	0.401	4.61	0.491
	AM-Softmax	1.01	0.101	1.06	0.105	1.81	0.192
	AAM-Softmax	0.85	0.089	0.97	0.103	1.79	0.176
	Ours (Eq. 8)	0.69	0.083	0.76	0.091	1.49	0.149
CAM++ [68]	Cross-Entropy	3.31	0.371	3.46	0.381	4.31	0.471
	AM-Softmax	0.71	0.095	0.91	0.101	1.61	0.181
	AAM-Softmax	0.66	0.087	0.82	0.095	1.59	0.164
	Ours (Eq. 8)	0.59	0.076	0.71	0.086	1.36	0.136
ERes2NetV2 [64]	Cross-Entropy	3.16	0.351	3.31	0.361	4.01	0.451
	AM-Softmax	0.68	0.076	0.81	0.092	1.58	0.161
	AAM-Softmax	0.62	0.055	0.77	0.083	1.46	0.144
	Ours (Eq. 8)	**0.53**	**0.049**	**0.66**	**0.071**	**1.21**	**0.121**

TABLE II: The performance of the proposed and conventional loss functions on the CN-Celeb evaluation set using different speaker encoders. Each metric's best result is in bold.

Speaker Encoder	Loss function	CN-Celeb1-Test	
		EER (%)	minDCF
Res2Net [65]	Cross-Entropy	12.12	0.682
	AM-Softmax	8.35	0.523
	AAM-Softmax	7.96	0.452
	Ours (Eq. 8)	7.21	0.423
ResNet34 [66]	Cross-Entropy	11.95	0.663
	AM-Softmax	7.39	0.507
	AAM-Softmax	6.92	0.421
	Ours (Eq. 8)	6.48	0.395
ECAPA-TDNN [3]	Cross-Entropy	11.63	0.651
	AM-Softmax	8.08	0.442
	AAM-Softmax	8.01	0.445
	Ours (Eq. 8)	7.45	0.413
ERes2Net [67]	Cross-Entropy	10.84	0.623
	AM-Softmax	7.11	0.468
	AAM-Softmax	6.69	0.388
	Ours (Eq. 8)	5.98	0.348
CAM++ [68]	Cross-Entropy	10.51	0.602
	AM-Softmax	6.93	0.451
	AAM-Softmax	6.78	0.393
	Ours (Eq. 8)	5.95	0.353
ERes2NetV2 [64]	Cross-Entropy	10.22	0.585
	AM-Softmax	6.65	0.433
	AAM-Softmax	6.14	0.370
	Ours (Eq. 8)	**5.73**	**0.341**

Specifically, we evaluated the individual and combined effects of MI (Eq. 6), SupCon (Eq. 1), and SupMarginCon (Eq. 3).

Table IV shows that incorporating MI alone with AAM-Softmax provides a slight performance improvement, reducing the EER from 0.65% to 0.63% on VoxCeleb1 and from 6.14% to 6.08% on CN-Celeb1. The addition of supervised contrastive learning (SupCon) significantly enhanced performance, further decreasing EER to 0.57% on VoxCeleb1 and 5.90% on CN-Celeb1. When introducing the margin into supervised contrastive learning (SupMarginCon), additional gains were observed, reducing the EER to 0.54% and 5.80%, respectively. Finally, combining all three components (AAM-Softmax, SupMarginCon, and MI) results in the best overall performance, achieving an EER of 0.53% on VoxCeleb1 and 5.73% on CN-Celeb1, demonstrating the effectiveness of each component and their synergistic combination.

TABLE III: Performance comparison of the proposed loss function and existing margin-based and contrastive learning methods on VoxCeleb1-O.

Speaker Encoder	Loss Function	VoxCeleb1-O	
		EER(%)	minDCF
ECAPA-TDNN [3]	AMC-Loss [69]	2.54	0.195
	Triplet (semi-hard) [76]	2.30	0.185
	Ang-Prototy Loss [71]	1.19	0.113
	CBRW-BCE [72]	1.10	0.088
	Ours	**0.74**	**0.096**
NeXt-TDNN [75]	AAMSoftmax [7]	0.79	0.086
IM-ECAPA-SimAM [74]	AAMSoftmax [7]	0.79	0.064
ECAPA++ [73]	AAMSoftmax [7]	0.65	0.079
CAM++ [68]	AAMSoftmax [7]	0.66	0.087
CAM++ [68]	Ours	**0.59**	**0.076**

E. Effect of Maximizing Mutual Information

We selected three classic contrastive learning models—SimCLR [77], MOCO [78], and SupCon [46] (Eq. 1)—as our baselines to evaluate the impact of increasing the mutual information between frame-level features and speaker embeddings. The results on VoxCeleb1-O and CN-Celeb1-test are

TABLE IV: Ablation study of the proposed loss components on ERes2NetV2.

Loss Components	VoxCeleb1-test		CN-Celeb1-test	
	EER (%)	minDCF	EER (%)	minDCF
AAM-Softmax	0.62	0.055	6.14	0.370
AAM-Softmax + MI	0.60	0.055	6.08	0.368
AAM-Softmax + SupCon	0.57	0.053	5.90	0.350
AAM-Softmax + SupMarginCon	0.54	0.050	5.80	0.345
AAM-Softmax + SupMarginCon + MI (Ours)	**0.53**	**0.049**	**5.73**	**0.341**

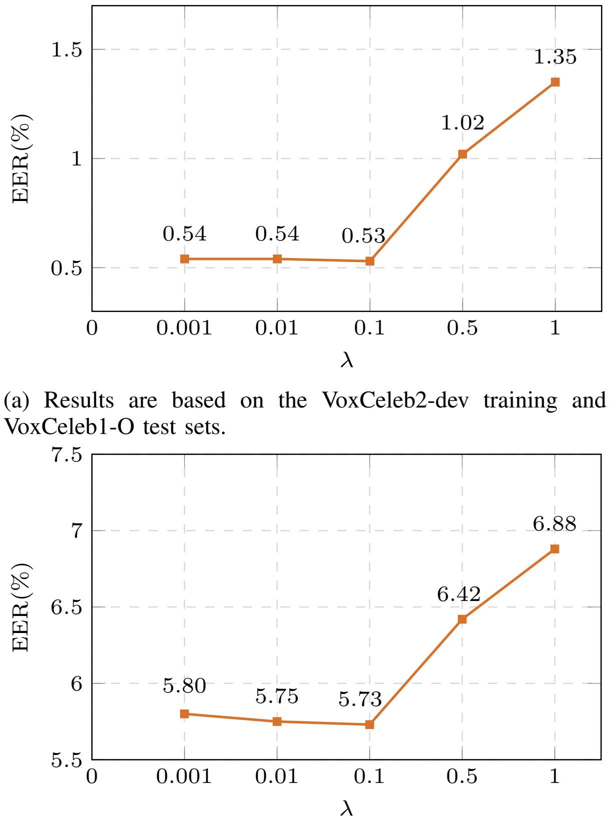
TABLE V: Effect of increasing mutual information on contrastive learning. Results are based on CN-Celeb1&2 or VoxCeleb2-dev for training and CN-Celeb1-test or VoxCeleb1-test for evaluation.

Model	VoxCeleb1-O		CN-Celeb1-test	
	EER(%)	minDCF	EER(%)	minDCF
SimCLR	6.78	0.548	15.88	0.677
SimCLR+MI	**6.63**	**0.523**	**15.47**	**0.652**
MOCO	7.46	0.627	16.17	0.706
MOCO+MI	**7.34**	**0.608**	**15.89**	**0.692**
SupCon	2.07	0.238	10.44	0.597
SupCon+MI	**2.02**	**0.231**	**10.26**	**0.583**

displayed in Table V. Maximizing mutual information between the frame-level output and the utterance-level representation introduces nonshared information, enhancing performance notably in self-supervised learning. This suggests that the shared information between views is insufficient for speaker verification, where enhanced mutual information leads to substantial improvements. Previous best practices [20] have shown that using the output of the first convolutional layer of the speaker encoder followed by global pooling as $\boldsymbol{h}$ achieves the optimal results. We follow this recipe in our approach. Furthermore, its effectiveness across different contrastive learning models indicates that our findings are broadly applicable.

F. Effect of Increasing the Role of Mutual Information

Accurately quantifying mutual information between high-dimensional variables is challenging and frequently results in imprecise estimations. We hypothesize that the hyper-parameter λ plays an important role in regularizing the non-shared information in the embedding. Specifically, a larger λ will introduce more nonshared information across views, enriching speaker information in the embeddings. To test this hypothesis, we varied λ and set it to 0.001, 0.01, 0.1, 0.5, and 1.0 and assessed the performance of the proposed loss function (Eq. 8) using ERes2NetV2 as the speaker encoder. Fig. 4 shows the EER against different values of λ. We observe a non-monotonic V-shape in EER with varying λ, suggesting that increasing mutual information consistently enhances performance in speaker verification, but excessively increasing mutual information could introduce noise along with useful information.



(a) Results are based on the VoxCeleb2-dev training and VoxCeleb1-O test sets.

(b) Results are based on the CN-Celeb1&2 training and CN-Celeb1 test sets.

Fig. 4: EER with varying hyper-parameter λ in Eq. 8.

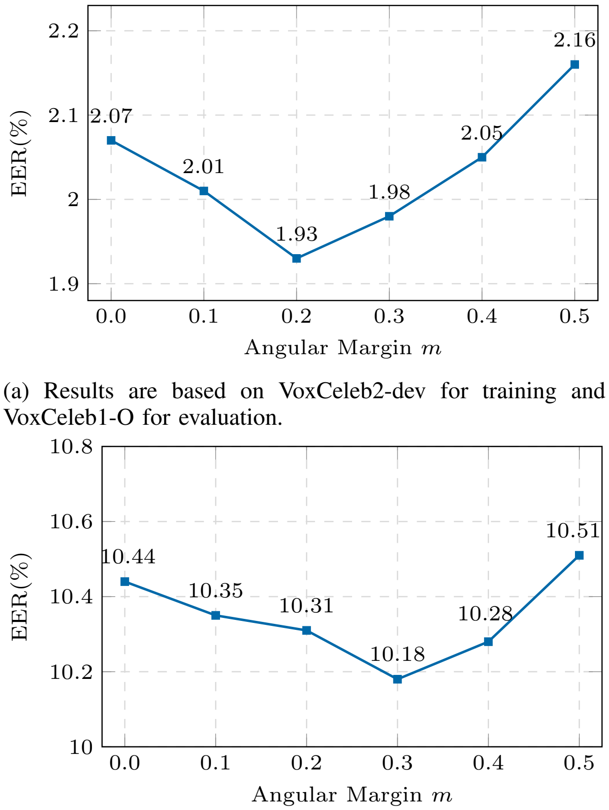
G. Effect of Angular Margin

The angular margin m in the SupMarginCon (Eq. 3) loss function affects the model's ability to discriminate. To explore this effect further, we systematically varied m from 0 to 0.5 degree in increments of 0.1 degrees. When the margin m is set to 0, SupMarginCon naturally degenerates into the SupCon (Eq. 1). The results are shown in Fig. 5.

As shown in Fig. 5a, the model achieves optimal performance on VoxCeleb1-O when $m = 0.2$. Fig. 5b shows that the best performance (EER = 10.18%) on CN-Celeb is achieved when $m = 0.3$. Any deviation from these optimal values results in a performance drop. This observation aligns with the common intuition, i.e., excessively small m causes the contrastive objective to lose discriminative power, while unnecessarily large m makes training difficult, causing suboptimal embedding networks.

H. Effect of Contrastive Learning

We further validated the capability of the proposed contrastive learning paradigm by visualizing the embedding of 20 speakers in the VoxCeleb1. Each speaker has 100 utterances.



(a) Results are based on VoxCeleb2-dev for training and VoxCeleb1-O for evaluation.

(b) Results are based on CN-Celeb1&2 for training and CN-Celeb1-test for evaluation.

Fig. 5: Effect of the angular margin m in the SupMarginCon (Eq. 3) loss on EER.

We used t-SNE to project the high-dimensional embedding vectors to a 2D space.

Fig. 6a shows the embeddings obtained from AAMSoftmax, while Figs. 6b, 6c, and 6d provide visualizations of SupCon, SupMarginCon, and our loss, respectively. Figs. 6c and 6d show that incorporating the margin makes the speaker clusters more compact. Our loss combines AAMSoftmax, SupMarginCon, and MI, leveraging both the enhanced classification capability of AAMSoftmax and the distinctive feature separation ability of SupMarginCon. This combination not only results in tighter speaker clusters but also leads to clear boundaries between different speakers. This enhanced clustering confirms the effectiveness of our model in distinguishing between different speakers and further validates the efficacy of our proposed contrastive learning approach.

I. *Effect of Number of Positives*

We investigated the effect of positive samples by incrementally increasing their number up to k per anchor. It is important to ensure that these samples do not appear in the denominator of the loss function in Eq. 3. This exclusion ensures that the model does not consider them negative.

We utilized the ECAPA-TDNN encoder and conducted experiments on the VoxCeleb1 dataset with a batch size of 1024. We trained the network for 300 epochs. Fig. 7 shows the results. A noticeable trend emerges from the result: introducing more positives consistently enhances the model's performance. Therefore, we conclude that more positive samples encourage the encoder to give closely aligned representations, resulting in compact speaker clusters in the embedding space.

J. *Sensitivity of Temperature Parameter*

The loss function in contrastive learning is typically constructed from a softmax function of feature similarities to contrast between the positive and negative pairs, with the similarity scaled by a temperature parameter τ (see Eq. 1). We observe that contrastive loss, a hardness-aware function, optimizes hard negative samples by penalizing them based on their difficulty. The temperature parameter controls the severity of penalties applied to the hard negatives. Specifically, a small temperature in contrastive loss results in stronger penalties on the hardest negative samples, promoting greater separation in their local structure and a more uniform embedding distribution. Conversely, with a large temperature, contrastive loss becomes less responsive to hard negative samples, diminishing its hardness-aware characteristics when the temperature approaches $+\infty$. The hardness-aware characteristic significantly contributes to the efficacy of softmax-based contrastive loss. As demonstrated in Fig. 8, a temperature of 0.07 leads to competitive SV performances.

K. *Alignment and Uniformity Analysis*

Alignment and uniformity are two closely related properties in contrastive learning and serve as valuable metrics for evaluating the quality of representations. Specifically, alignment refers to how an encoder generates similar representations for similar samples. It can be quantitatively defined by calculating the expected distance between the embeddings of positive pairs:

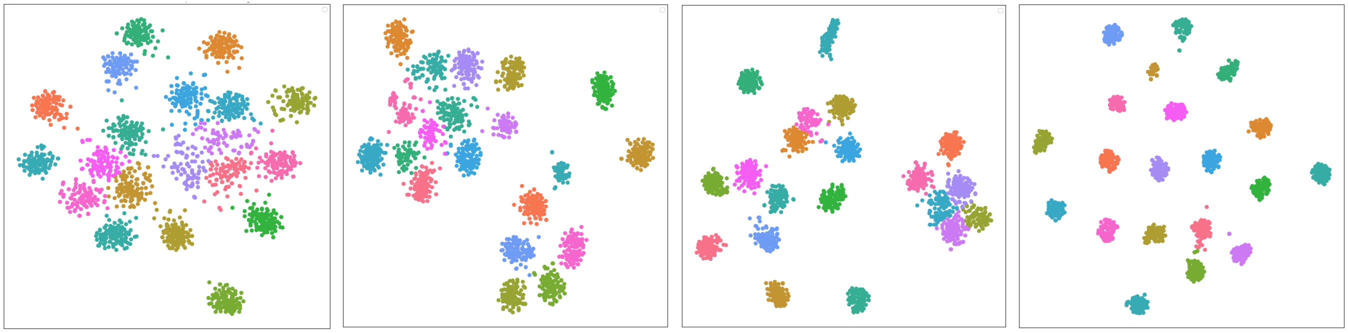
$$\ell_{\text{align}} = \underset{\boldsymbol{x},\boldsymbol{x}_p \sim p_{\text{pos}}}{\mathbb{E}} \left\| f(\boldsymbol{x}) - f\left(\boldsymbol{x}_p\right) \right\|_2^2, \tag{9}$$

where p_{pos} denotes the distribution of positive samples. Uniformity refers to how uniform the distribution of the embedding is, which helps preserve information. It is defined as

$$\ell_{\text{uniform}} = \log \underset{\boldsymbol{x},\boldsymbol{y} \overset{i.i.d}{\sim} p_{\text{data}}}{\mathbb{E}} e^{-2\|f(\boldsymbol{x})-f(\boldsymbol{y})\|_2^2}, \tag{10}$$

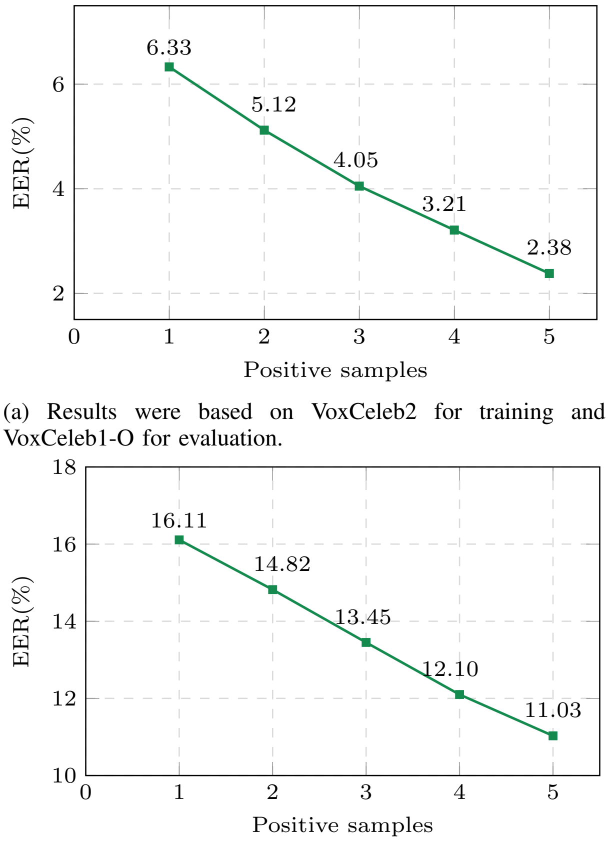
where p_{data} represents the distribution of all data.

To evaluate the alignment and uniformity of our method, we conducted an assessment using the CN-Celeb dataset. For every 10 iterations, we computed the alignment and uniformity of SupMarginCon and compared them against the alignment and uniformity of the original supervised contrastive learning. The results are presented in Fig. 9. SupMarginCon consistently enhances alignment and uniformity throughout the training process compared to supervised contrastive learning. These findings validate the intuition behind our approach and indicate that incorporating margin can significantly enhance the quality of speaker representations.



(a) AAMSoftmax (b) SupCon (Eq. 1) (c) SupMarginCon (Eq. 3) (d) Our loss (Eq. 8)

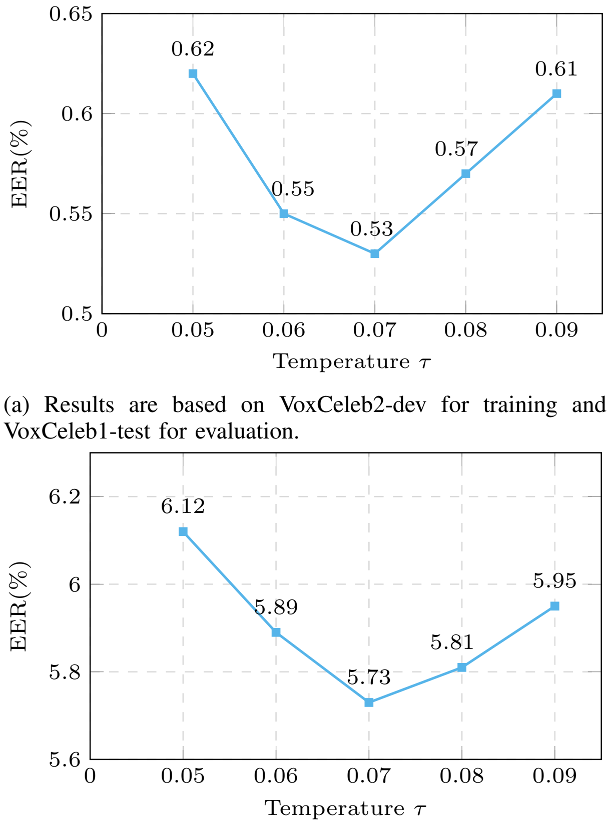
Fig. 6: t-SNE plots of the embeddings of 20 speakers in VoxCeleb1. Each color represents one speaker. The graphs show the speaker clustering effects produced by four different loss functions using the ECAPA-TDNN: (6a) AAMSoftmax (1st term of Eq. 8), (6b) SupCon (Eq. 1), (6c) SupMarginCon (2nd term of Eq. 8), and (6d) our proposed loss (Eq. 8).



(a) Results were based on VoxCeleb2 for training and VoxCeleb1-O for evaluation.

(b) Results were based on CN-Celeb1&2 for training and CN-Celeb1-test for evaluation.

Fig. 7: EER versus the maximum number of positives in $\mathcal{P}(i)$. Adding more positives reduces EER.



(a) Results are based on VoxCeleb2-dev for training and VoxCeleb1-test for evaluation.

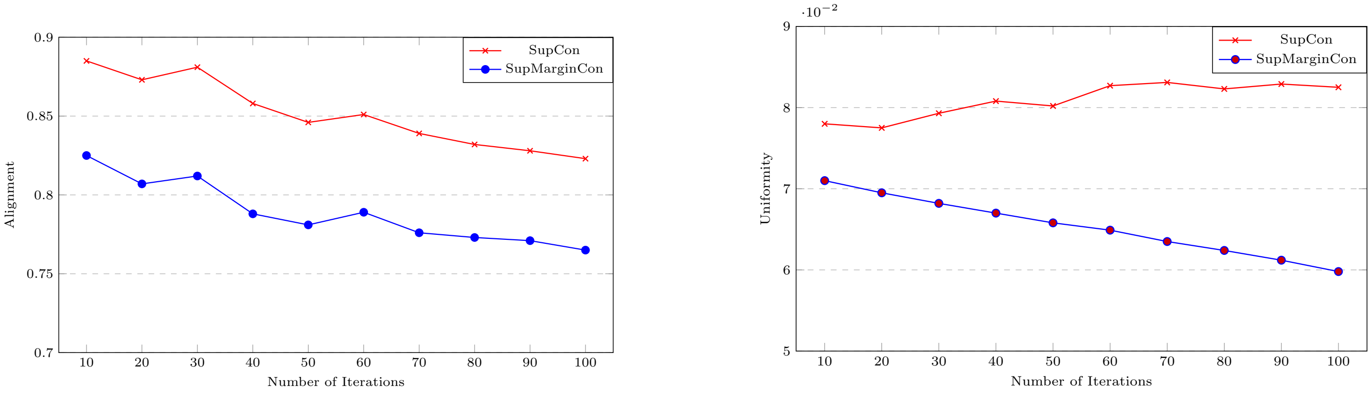
(b) Results are based on CN-Celeb1&2 for training and CN-Celeb1-test for evaluation.

Fig. 8: EER versus the temperature parameter τ in the loss function in Eq. 3. The results are based on an ERes2NetV2 speaker encoder optimized by minimizing the total loss in Eq. 8 with $\lambda = 0.1$.

L. Low-resource Scenario

We investigated the behavior of our loss under a low-resource scenario. To this end, we applied our proposed loss on an ERes2NetV encoder using the CU-MARVEL dataset, a Cantonese dataset for dementia detection [63]. It comprises 280 speakers with the majority of audio recordings shorter than 2 seconds. We repurposed the dataset for speaker verification, and the statistics of CU-MARVEL are shown in Table VI.

Table VII presents the performance and the conventional methods on CU-MARVEL version 0915. When using the ERes2NetV2 encoder with Fbank features, applying data augmentation and utilizing the AAMSoftmax loss function results in an EER of 6.80% and a minDCF of 0.74. When



(a) l_{align}

(b) $l_{\text{Uniformity}}$

Fig. 9: The alignment and uniformity of SupCon (Eq. 1) and SupMarginCon (Eq. 3). (9a) l_{align} measures the alignment between positive pairs. (9b) $l_{\text{Uniformity}}$ measures the uniformity of the embedding distribution. For both metrics, a lower value indicates better performance.

TABLE VI: Statistics of CU-MARVEL.

Data Split	# of Speakers	# of Utterances	# of Trials
Train	280	206,034	N/A
Test	53	43,319	400,000

TABLE VII: The performance of the proposed loss and conventional losses on CU-MARVEL. Fbank features were used as the input to an ERes2NetV2 speaker encoder.

Loss Function	EER(%)	minDCF
AAMSoftmax	6.80	0.74
SupCon	5.95	0.72
SupMarginCon	5.72	0.71
SupMarginCon + MI	5.63	0.70
SupMarginCon + MI + AAMSoftmax (Ours)	**4.98**	**0.67**

we employed the SupCon loss function, we observed a significant performance improvement, achieving an EER of 5.95% and a minDCF of 0.72. When we sequentially add margin and mutual information, performance improves with each addition. We noted that under low-resource conditions, contrastive learning loss outperforms classification-based loss. We attribute this performance gain to the training objectives. Speaker verification is an open-set task where a limited number of utterances are insufficient to train a robust classifier for unseen samples. However, the goal of contrastive learning is to enhance discriminative ability, which allows it to perform better on unseen test datasets under insufficient training data scenarios.

IV. Conclusions

We introduce a supervised contrastive learning framework designed to learn discriminative speaker representations. Our approach incorporates mutual information into contrastive learning, enhancing speaker-related information. We use an angular margin to improve the discriminative power of the contrastive learning loss. These enhancements improve speaker representation learning. The experimental results from CN-Celeb, VoxCeleb, and CU-MARVEL show that both techniques significantly enhance the performance of the speaker encoder in contrastive learning. On the low-resource CU-MARVEL dataset, our contrastive learning method even outperforms the classification loss.

V. Acknowledgment

This work was in part supported by the Research Grants Council of Hong Kong, Theme-based Research Scheme (Ref.: T45-407/19-N), a GRF (No. 15228223), and the Research Student Attachment Programme of HKPolyU.

References

[1] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Proc. of InterSpeech*, 2017, pp. 999–1003.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[3] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. of InterSpeech*, 2020, pp. 3830–3834.

[4] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4879–4883.

[5] J. S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," *Proc. of InterSpeech 2020*, pp. 2977–2981, 2020.

[6] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5265–5274.

[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.

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

[9] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. of InterSpeech 2018*, 2018, pp. 2252–2256.

[10] M. Sang, W. Xia, and J. H. Hansen, "Open-set short utterance forensic speaker verification using teacher-student network with explicit inductive bias," *Proc. of InterSpeech*, pp. 2262–2266, 2020.

[11] G. Bhattacharya, J. Monteiro, J. Alam, and P. Kenny, "Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6226–6230.

[12] M. Sang, W. Xia, and J. H. Hansen, "DEAAN: Disentangled embedding and adversarial adaptation network for robust speaker representation learning," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6169–6173.

[13] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Prof. of Advances in Neural Information Processing Systems (NeruIPS)*, vol. 31, 2018.

[14] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," in *Proc. of International Conference on Learning Representations (ICLR)*, 2015.

[15] G. Elsayed, D. Krishnan, H. Mobahi, K. Regan, and S. Bengio, "Large margin deep networks for classification," *Prof. of Advances in Neural Information Processing Systems (NeruIPS)*, vol. 31, 2018.

[16] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proc. of International Conference on Machine Learning (ICML)*, 2016, pp. 507–516.

[17] W.-W. Lin, M.-W. Mak, and J.-T. Chien, "Multisource i-vectors domain adaptation using maximum mean discrepancy based autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2412–2422, 2018.

[18] L. Li, M.-W. Mak, and J.-T. Chien, "Contrastive adversarial domain adaptation networks for speaker recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, pp. 2236–2245, 2022.

[19] Y. Tu, M.-W. Mak, and J.-T. Chien, "Variational domain adversarial learning with mutual information maximization for speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2013–2024, 2020.

[20] ——, "Information maximized variational domain adversarial learning for speaker verification," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6449–6453.

[21] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. European Conference on Computer Vision (ECCV)*, 2016, pp. 499–515.

[22] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," *arXiv preprint arXiv:1703.09507*, 2017.

[23] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "NormFace: L2 hypersphere embedding for face verification," in *Proc. 25th ACM International Conference on Multimedia*, 2017, pp. 1041–1049.

[24] Y. Liu, H. Li, and X. Wang, "Rethinking feature discrimination and polymerization for large-scale recognition," *arXiv preprint arXiv:1710.00870*, 2017.

[25] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 212–220.

[26] H. Zhang, Y. Zou, and H. Wang, "Contrastive self-supervised learning for text-independent speaker verification," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6713–6717.

[27] Y. Zhang, H. Zhu, Y. Wang, N. Xu, X. Li, and B. Zhao, "A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space," in *Proc. 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 4892–4903.

[28] Y. Tu, M.-W. Mak, and J.-T. Chien, "Contrastive self-supervised speaker embedding with sequential disentanglement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[29] C.-X. Gan, M.-W. Mak, W. Lin, and J.-T. Chien, "Asymmetric clean segments-guided self-supervised learning for robust speaker verification," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 081–11 085.

[30] H. Wang, X. Guo, Z.-H. Deng, and Y. Lu, "Rethinking minimal sufficient representation in contrastive learning," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 041–16 050.

[31] Y. Tu and M.-W. Mak, "Mutual information enhanced training for speaker embedding." in *Prof. of InterSpeech*, 2021, pp. 91–95.

[32] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, and W. Xu, "Consert: A contrastive framework for self-supervised sentence representation transfer," *arXiv preprint arXiv:2105.11741*, 2021.

[33] J. Giorgi, O. Nitski, B. Wang, and G. Bader, "DeCLUTR: Deep contrastive learning for unsupervised textual representations," in *Proc. of Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*, 2021, pp. 879–895.

[34] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 6894–6910.

[35] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[36] L. Li, R. Nai, and D. Wang, "Real additive margin softmax for speaker verification," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7527–7531.

[37] Z. Chen, S. Wang, and Y. Qian, "Self-supervised learning based domain adaptation for robust speaker verification," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5834–5838.

[38] M. Sang, H. Li, F. Liu, A. O. Arnold, and L. Wan, "Self-supervised speaker verification with simple siamese network and self-supervised regularization," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6127–6131.

[39] W. Xia, C. Zhang, C. Weng, M. Yu, and D. Yu, "Self-supervised text-independent speaker verification using prototypical momentum contrastive learning," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6723–6727.

[40] Z. Li and M.-W. Mak, "Speaker representation learning via contrastive loss with maximal speaker separability," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 962–967.

[41] Z. Li, M.-W. Mak, and H. M.-L. Meng, "Discriminative speaker representation via contrastive learning with class-aware attention in angular space," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[42] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5796–5800.

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

[44] 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. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[45] W. Wang, D. Cai, X. Qin, and M. Li, "The dku-dukeece systems for voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2010.12731*, 2020.

[46] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Prof. of Advances in Neural Information Processing Systems (NeruIPS)*, vol. 33, pp. 18 661–18 673, 2020.

[47] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proc. of International Conference on Machine Learning (ICML)*, 2020, pp. 9929–9939.

[48] M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata, "Learning robust representations via multi-view information bottleneck," in *Proc. of International Conference on Learning Representations (ICLR)*, 2019.

[49] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[50] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *Proc. of International Conference on Machine Learning (ICML)*, 2019, pp. 5171–5180.

[51] T. Lepage and R. Dehak, "Label-efficient self-supervised speaker verification with information maximization and contrastive learning," in *Proc. of InterSpeech*, 2022, pp. 4018–4022.

[52] C. Zhang and D. Yu, "C3-dino: Joint contrastive and non-contrastive self-supervised learning for speaker verification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1273–1283, 2022.

[53] Z. Li, W. Guo, B. Gu, S. Peng, and J. Zhang, "Contrastive learning and inter-speaker distribution alignment based unsupervised domain adaptation for robust speaker verification," in *Proc. of InterSpeech*, 2024.

[54] S. H. Mun, M. H. Han, M. Kim, D. Lee, and N. S. Kim, "Disentangled speaker representation learning via mutual information minimization,"

in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 89–96.

[55] W. H. Kang, J. Alam, and A. Fathan, "Domain generalized speaker embedding learning via mutual information minimization." in *Odyssey*, 2022, pp. 178–184.

[56] F. Zhang, W. Zhou, Y. Liu, W. Geng, Y. Shan, and C. Zhang, "Disentangling age and identity with a mutual information minimization for cross-age speaker verification," in *Proc. of InterSpeech*, 2024, pp. 3789–3793.

[57] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. of International Conference on Learning Representations*, 2014.

[58] Y. Chen, S. Zheng, H. Wang, L. Cheng, T. Zhu, R. Huang, C. Deng, Q. Chen, S. Zhang, W. Wang *et al.*, "3d-speaker-toolkit: An open-source toolkit for multimodal speaker verification and diarization," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[59] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. of InterSpeech*, pp. 2616–2620, 2017.

[60] J. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Proc. of InterSpeech*, 2018.

[61] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "CN-Celeb: A challenging chinese speaker recognition dataset," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7604–7608.

[62] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang, "CN-Celeb: multi-genre speaker recognition," *Speech Communication*, vol. 137, pp. 77–91, 2022.

[63] H. Meng, B. Mak, M.-W. Mak, H. Fung, X. Gong, T. Kwok, X. Liu, V. Mok, P. Wong, J. Woo *et al.*, "Integrated and enhanced pipeline system to support spoken language analytics for screening neurocognitive disorders," in *Proc. of InterSpeech*, 2023, pp. 1713–1717.

[64] Y. Chen, S. Zheng, H. Wang, L. Cheng, Q. Chen, S. Zhang, and J. Li, "ERes2NetV2: Boosting short-duration speaker verification performance with computational efficiency," in *Proc. of InterSpeech 2024*, 2024, pp. 3245–3249.

[65] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

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

[67] Y. Chen, S. Zheng, H. Wang, L. Cheng, Q. Chen, and J. Qi, "An enhanced res2net with local and global feature fusion for speaker verification," in *Proc. of InterSpeech 2023*, 2023, pp. 2228–2232.

[68] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "CAM++: A fast and efficient network for speaker verification using context-aware masking," in *Proc. of InterSpeech 2023*, 2023, pp. 5301–5305.

[69] H. Choi, A. Som, and P. Turaga, "AMC-Loss: Angular margin contrastive loss for improved explainability in image classification," in *Proc. of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 838–839.

[70] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[71] Z. Chen, B. Han, X. Xiang, H. Huang, B. Liu, and Y. Qian, "Sjtu-aispeech system for voxceleb speaker recognition challenge 2022," *arXiv preprint arXiv:2209.09076*, 2022.

[72] Z. Bai, J. Wang, X.-L. Zhang, and J. Chen, "End-to-end speaker verification via curriculum bipartite ranking weighted binary cross-entropy," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1330–1344, 2022.

[73] B. Liu and Y. Qian, "Ecapa++: Fine-grained deep embedding learning for tdnn based speaker verification," in *Proc. of InterSpeech 2023*, 2023, pp. 3132–3136.

[74] S.-H. Liou, P.-C. Chan, C.-P. Chen, T.-C. Lin, C.-L. Lu, Y.-H. Cheng, H.-F. Chuang, and W.-Y. Chen, "Enhancing ecapa-tdnn with feature processing module and attention mechanism for speaker verification," in *Proc. of InterSpeech*, 2024, pp. 2120–2124.

[75] H.-J. Heo, U.-H. Shin, R. Lee, Y. Cheon, and H.-M. Park, "NeXt-TDNN: Modernizing multi-scale temporal convolution backbone for speaker verification," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 186–11 190.

[76] C. Zhang, K. Koishida, and J. H. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1633–1644, 2018.

[77] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. of International Conference on Machine Learning*, 2020, pp. 1597–1607.

[78] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2020, pp. 9729–9738.

VI. Biography Section

Zhe LI (Student Member, IEEE) is a PhD candidate in the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University. He received his B.Eng. degree in computer science from Qilu University of Technology (Shandong Academy of Sciences), Jinan, China, in 2016 and his M.Sc. degree in software engineering from Xinjiang University, Urumqi, China, in 2021. He is a IEEE student member. He has led a research postgraduate student innovation project and participated in several National Key Research and Development Programs of China, NSFC projects in multilingual intelligent information processing, and projects funded by the Hong Kong Grants Council. He has received an excellent scientific and technological achievement award from the Chinese association for artificial intelligence.

Man-Wai MAK (Senior Member, IEEE) received a BEng (Hons) degree in Electronic Engineering from Newcastle Upon Tyne Polytechnic in 1989 and a Ph.D. degree in Electronic Engineering from the University of Northumbria at Newcastle (now Northumbria University) in 1993. He joined the Department of Electronic and Information Engineering (EIE) at The Hong Kong Polytechnic University in 1993, served as Interim Head of EIE from 2021 to 2023, and is presently a Professor and Associate Head of the Department of Electrical and Electronic Engineering. He has authored more than 220 technical articles and books in speaker recognition, machine learning, bioinformatics, and biomedical engineering and served as a guest editor of international journals. He has been an associate editor of IEEE Trans. on Audio, Speech and Language Processing, Journal of Signal Processing Systems, Advances in Artificial Neural Systems, and IEEE Biometrics Compendium. He is a tutorial speaker in Interspeech '16. Dr. Mak is also a co-author of the postgraduate textbook ”Biometric Authentication: A Machine Learning Approach, Prentice-Hall, 2005.”, ”Machine Learning for Protein Subcellular Localization Prediction, De Gruyter, 2015”, and ”Machine Learning for Speaker Recognition, Cambridge University Press, 2020.” He has received three Faculty of Engineering Research Grant Achievement Awards and a Faculty Award for Outstanding Performance (Research and Scholarly Activities). Dr. Mak has been an Executive Committee member of the IEEE Hong Kong Section Computer Chapter from 1995-2007 and the Chairman of the IEEE Hong Kong Section Computer Chapter from 2003-2005. He also served as a member of the IEEE Machine Learning for Signal Processing Technical Committee in 2005–2007 and as a Technical Committee Member of the IEEE Computation Intelligence Society, Intelligent Systems Applications, in 2008. Prof. Mak has served as Area Chair of Interspeech'14 and ICTAI 2016, Steering Committee Member of ISCSLP and ISCSLP16, and Program Co-Chair of ISCSLP 2018 and ISCSLP 2021. Prof. Mak's research interests include speaker recognition, machine learning, spoken language processing, biomedical engineering, and bioinformatics.

Mert Pilanci He is an assistant professor in the Department of Electrical Engineering at Stanford University. Prior to joining Stanford, he was an assistant professor of Electrical Engineering and Computer Science at the University of Michigan. In 2017, he was a Math+X postdoctoral fellow working with Emmanuel Candès at Stanford University. He received Ph.D. in Electrical Engineering and Computer Science from UC Berkeley in 2016. His studies were supported partially by a Microsoft Research PhD Fellowship. He obtained his B.S. and M.S. degrees in Electrical Engineering from Bilkent University. His research interests are in large-scale machine learning, optimization, and information theory.

Helen Meng (Fellow Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA. Helen Meng is Patrick Huen Wing Ming Professor of Systems Engineering and Engineering Management at The Chinese University of Hong Kong (CUHK). She is the Founding Director of the CUHK Ministry of Education (MoE)-Microsoft Key Laboratory for Human-Centric Computing and Interface Technologies (since 2005), Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems (since 2006), and Stanley Ho Big Data Decision Analytics Research Center (since 2013). Previously, she served as CUHK Faculty of Engineering's Associate Dean (Research), Chairman of the Department of Systems Engineering and Engineering Management, Editor-in-Chief of the IEEE Transactions on Audio, Speech and Language Processing, Member of the IEEE Signal Processing Society Board of Governors, ISCA Board Member and presently member of the ISCA International Advisory Council. She was elected APSIPA's inaugural Distinguished Lecturer 2012-2013 and ISCA Distinguished Lecturer 2015-2016. Her awards include the Ministry of Education Higher Education Outstanding Scientific Research Output Award 2009, Hong Kong Computer Society's inaugural Outstanding ICT Woman Professional Award 2015, Microsoft Research Outstanding Collaborator Award 2016 (1 in 32 worldwide), IEEE ICME 2016 Best Paper Award, IBM Faculty Award 2016, HKPWE Outstanding Women Professionals and Entrepreneurs Award 2017 (1 in 20 since 1999), Hong Kong ICT Silver Award 2018 in Smart Inclusion, CogInfoComm2018 Best Paper Award and the 2019 IEEE SPS Leo L. Beranek Meritorious Service Award for exemplary service to and leadership in the Signal Processing Society. Her research interests include AI for speech and language technologies to support multilingual and multimodal human-computer interactions, eLearning and assistive technologies, and big data decision analytics using AI. Helen has served in numerous Government appointments, including memberships in the Research Grants Council and the Steering Committee of Hong Kong's Electronic Health Record Sharing. Helen is a Fellow of HKCS, HKIE, IEEE, and ISCA.