The Speaker and Language Recognition Workshop (Odyssey 2024)
18-21 June 2024, Quebec City, Canada

An investigative study of the effect of several regularization techniques on label noise robustness of self-supervised speaker verification systems

Abderrahim Fathan, Xiaolin Zhu, Jahangir Alam*

Computer Research Institute of Montreal (CRIM)
Montreal, Canada

abderrahim.fathan@crim.ca, alice.zhuxl@gmail.com, jahangir.alam@crim.ca

Abstract

Clustering-based Pseudo-Labels (PLs) are widely used to optimize Speaker Embedding (SE) networks and train Self-Supervised (SS) Speaker Verification (SV) systems. However, this SS training scheme relies on highly accurate PLs. In this paper, we perform a large investigative study of the effect of several regularization techniques (mixup, label smoothing, employing sub-centers) on the label noise robustness of SSSV systems. We study these techniques and apply them on various recent metric learning loss functions for better generalization of SSSV systems. In particular, we investigate the effect of these losses and regularizations on the robustness of the self-supervised SV task against label noise using the CAMSAT clustering model to generate PLs. We provide a thorough comparative analysis of the performance of these techniques using different numbers of clusters and show that some of them are effective against label noise and lead to considerable improvements in SV performance.

1. Introduction

Automatic speaker verification (ASV) consists of using the voiceprint of a speaker to verify their identity. ASV is one of the most convenient means of biometric recognition [1]. Based on a speaker's known utterances, the speaker verification (SV) task consists of confirming that the identity of a speaker is who they purport to be.

Typically, utterance-level fixed-dimensional embedding vectors are extracted from the enrollment and test speech samples and then fed into a scoring algorithm (e.g., cosine distance) to measure their likelihood of being from the same speaker. Classically, the i-vector framework has been one of the most dominant approaches for speech embedding [2, 3] thanks to its ability to summarize the distributive patterns of speech in an unsupervised manner and with relatively small training datasets. It generates fixed-sized compact vectors that represent the speaker's identity in a speech utterance regardless of its length. Besides, in the past years, various deep learning-based architectures and techniques have been proposed to extract embeddings [4, 5, 6]. They have shown great performance when large training datasets are available, particularly with a sufficient number of speakers [7]. One widely employed architecture for this purpose is ECAPA-TDNN [8], which has achieved state-of-the-art (SOTA) performance in text-independent speaker recognition. The latter uses squeeze-and-excitation (SE), employs channel- and context-dependent statistics pooling & multi-layer aggregation and applies self-attention pooling to obtain an utterance-level embedding vector.

Indeed, most of the deep embedding models are trained in a fully supervised manner and require large speaker-labeled datasets for training. However, well-annotated datasets can be expensive and time-consuming to prepare, which has led the research community to explore more affordable self-supervised learning (SSL) techniques using larger unlabeled datasets. One common way to solve this issue for SV systems is to use clustering models to generate Pseudo-Labels (PLs) [9, 5, 6], or to employ SSL-based objectives (e.g., SimCLR, MoCo [10]) to generate PLs and train the speaker embedding network using these labels in a discriminative fashion [11, 12]. Despite the impressive performance of these PL-based Self-Supervised SV schemes, clustering performance remains a bottleneck in all above approaches [12, 13] as downstream performance relies greatly on accurate PLs since these are in general noisy and inaccurate due to the discrepancy between the clustering objective(s) and the final SV task. Besides, even with iterative clustering-classification paradigms, the erroneous information from the wrong PLs keeps propagating iteratively, which degrades the final performance [12, 14]. Indeed, recent studies have shown that label noise can remarkably impact downstream performance [6]. Thus, the need for better-performing SV approaches that are robust to label noise to mitigate its negative effect on generalization. In this paper, we investigate several regularization techniques (mixup [15], label smoothing [16], employing sub-centers [17]) to incorporate into our SV systems, jointly with our loss functions to study their effect on the label noise robustness of self-supervised SV systems. To this aim, we explore a variety of metric learning loss functions, including maximum margin-based softmax losses (e.g. CosFace, AdaFace), symmetric losses, normalized losses, and noise-robust loss functions such as Subcenter-ArcFace [17] or BoundaryFace [18] for the task of SV under label noise. To generate well-performing PLs, we employed the CAMSAT clustering model [19]. We used three different predefined numbers of clusters {5000, 5994, 10000} during clustering to study the generalization and behavior of self-supervised SV systems under various types of real-world label noise.

We propose a curated selection of loss objectives (see Table 1) that we experimentally found to be effective against label noise and enhance the generalization of self-supervised SV systems to out-of-set samples, beyond discrepancies in the PLs. The contributions of this paper are as follows:

- We propose the first large-scale investigative study of different regularization techniques, using various recent state-of-the-art loss objectives, for the task of speaker verification (SV). Several of these losses and regularizations we apply for the first time in the domain of SV.
- We show that maximum-margin -based softmax losses are beneficial to mitigate the memorization effects of label noise during training.

* Independent Researcher

10.21437/odyssey.2024-7

- We show that the mixup regularization strategy and using sub-centers are effective against label noise memorization and lead to better robustness and generalization.

- To our knowledge, we are the first to generalize the regularization idea of using sub-centers of classes, introduced in subcenter-ArcFace, to other types of losses.

- We demonstrate that several recent maximum-margin softmax variants provide a great advantage in terms of generalization and noise-robustness over some widely-used losses in the domain of SV, such as the angular additive margin softmax (AAMSoftmax) [20] loss.

- Using CAMSAT-based PLs [19], our proposed selection of loss objectives allowed us to achieve SOTA SV performance, outperforming various benchmarks.

2. Background and Related Work

2.1. Noise-robust loss functions

We can generally group the methods to learn from noisy data into two categories: approaches focusing on creating noise-robust algorithms to learn directly from noisy labels [21, 22, 23, 24, 25, 26], and label-cleansing approaches that aim to remove or correct mislabeled data [27, 28, 29, 30]. In recent years, various robust loss-based methods were proposed to learn with noisy labels. [31] proved theoretically that symmetric loss functions, such as Mean Absolute Error (MAE), are robust to label noise, while other losses like commonly used Cross Entropy (CE) are not. Besides, [32] introduced Generalized Cross Entropy (GCE), a generalized mixture of CE and MAE. [33] proposed Symmetric Cross Entropy (SCE) which is a combination of CE and scaled MAE. Reverse Cross Entropy (RCE) was also suggested to learn more distinguished feature representations for detecting adversarial examples. Additionally, [34] suggested a state-of-the-art Active Passive Loss (APL) to create fully robust loss functions. It showed that any loss function can be made robust to noisy labels by a simple normalization operation that makes loss functions symmetric. On the other hand, recently [35] found that APL still struggles with MAE and suffers from a problem of underfitting. For this reason, they suggested a new class of passive loss functions that are different from MAE, called Negative Loss Functions (NLFs), and proposed a new class of theoretically robust passive loss functions, called Normalized Negative Loss Functions (NNLFs). By replacing the MAE in APL with NNLF, they proposed an additional Active Negative Loss (ANL), a robust loss function framework with stronger fitting ability. In this paper, we investigate several robust loss functions created by the APL framework and NLFs, including the proposed normalization operation.

Moreover, in the domain of SV, [6] found that Mixup regularization is effective against label noise memorization [36], and leads to better generalization of self-supervised SV systems since Mixup can dilute the label noise and create synthetic samples around the borders that lead to smoothing the data manifold and better class separation. In the same line of work, [37] also proposed an effective noise-robust self-supervised Multi-task learning framework based on various mixup variants to make use of the variety of complementary information that can potentially be gained through the combination of the different tasks to improve the performance and robustness of SV systems.

2.2. Maximum margin-based softmax loss objectives

The goal of Metric Learning is to learn representation functions that map objects into an embedded space. The aim is to simplify the comparison function of speaker utterances all the way down to the most simple distance function (e.g. cosine distance) by delegating the hard task of generating speaker representations to the trained embedding network which should ensure intra-class compactness and inter-class separability.

To improve performance on previously unseen data and generalize to out-of-domain speech samples, various maximum margin-based softmax variants based on different objectives have been proposed. Indeed, softmax suffers from several drawbacks such as that (1) its computation of inter-class margin is intractable [38] and (2) the learned projections are not guaranteed equi-spaced. Indeed, the projection vectors for majority classes occupy more angular space compared to minority classes [39]. To solve these problems, several alternatives to softmax have been proposed [20, 40, 41, 42, 43]. For instance, AMSoftmax [40] loss applies an additive margin constraint in the angular space to the softmax loss for maximizing inter-class variance and minimizing intra-class variance. To provide a clear geometric interpretation of data samples and enhance the discriminative power of deep models, AAMSoftmax (angular additive margin softmax) [20] objective introduces an additive angular margin to the target angle (between the given features and the target center). Due to the exact correspondence between the angle and arc in the normalized hypersphere, AAMSoftmax can directly optimize the geodesic distance margin, thus its other name ArcFace.

Additionally, CosFace (large margin cosine loss) [43] reformulates the softmax loss as a cosine loss by L2 normalizing both features and weight vectors to remove radial variations, based on which a cosine margin term is introduced to further maximize the decision margin in the angular space. On the other hand, OCSoftmax [41] uses one-class learning instead of multi-class classification and does not assume the same distribution for all classes/speakers. More recently, AdaFace [42] loss has been proposed which emphasizes misclassified samples according to the quality of speaker embeddings (via feature norms). As an improvement, SMAFace was also introduced for low-quality face recognition images by incorporating sample mining into conventional margin-based methods. At its core, SMAFace focuses on prioritizing information-dense samples, namely hard samples or easy samples, which present more distinctive features. To this aim, it employs a probability-driven mining strategy, enabling the model to adeptly navigate hard samples, thereby bolstering its robustness and adaptability. Besides, as softmax has no unified threshold to separate positive sample-to-class pairs from negative sample-to-class pairs, a Unified Cross Entropy (UniFace) [44] loss for face recognition model training was designed on the vital constraint that all the positive sample-to-class similarities shall be larger than the negative ones. Additionally, as sample-to-class loss-based models can not fully explore the cross-sample relationship among large amounts of samples, UniTSFace [45] proposed a unified threshold integrated sample-to-sample based loss (USS), which features an explicit unified threshold for distinguishing positive from negative pairs. Furthermore, to incorporate additional sample-to-sample comparisons during training, [46] proposed Variational Prototype Learning (VPL), which represents every class as a distribution instead of a point in the latent space. Identifying the slow feature drift phenomenon, authors directly injected memorized features into prototypes to approximate variational prototype sampling. Finally, as above methods are susceptible to label noise, Subcenter-ArcFace [17] relaxes

Table 1: A study of a wide variety of metric learning loss functions. Results are reported in terms of the EER (%) downstream SV evaluation performance. We used the CAMSAT algorithm to generate PLs using different predefined numbers of clusters.

Loss function	No. of clusters			Loss function	No. of clusters		
	5,000	**5,994**	**10,000**		**5,000**	**5,994**	**10,000**
MV-Arc-Softmax	**2.842**	3.006	2.884	Agent Center loss	13.34	13.393	12.508
OCSoftmax	2.964	3.134	2.969	Focal loss	13.001	13.340	12.561
Subcenter-ArcFace	2.969	3.059	2.943	Generalized Cross Entropy	13.351	13.277	13.966
ArcFace-VPL	2.996	3.059	2.996	Revserse Cross Entropy	14.252	14.687	14.555
SMAFace	3.049	3.112	3.171	Softmax	14.486	14.507	15.085
AMSoftmax	3.054	3.224	2.959	AGCE loss	14.464	14.390	14.608
AdaFace	3.059	3.112	3.059	AExp loss	14.565	14.973	14.756
AAMSoftmax	3.065	3.309	3.134	Mean Absolute Error	14.613	15.021	14.570
CosFace-VPL	3.075	3.022	2.948	AUE loss	14.666	14.947	14.772
CosFace	3.096	3.043	**2.863**	Normalized Cross Entropy	18.664	19.692	20.594
BoundaryFace	3.096	**2.948**	2.884	MagFace	8.499	8.409	3.139
Normalized Softmax loss	3.134	3.118	3.028	Normalized Focal loss	18.754	19.565	20.700
Unified Cross Entropy (UniFace) loss	3.15	3.208	3.16	Normalized Negative Focal loss	22.969	24.146	25.779
Normalized BCE loss	3.213	3.181	3.192	Hard Gumbel-Softmax	23.096	47.397	22.778
CurricularFace	3.229	3.256	3.192	Normalized Negative Cross Entropy	23.261	26.156	27.45
Cross Entropy	5.477	5.827	5.546	Soft Gumbel-Softmax	25.774	43.871	22.683
AS-Softmax	5.748	6.272	6.607	Center loss	27.126	29.173	27.625
DropMax	7.137	6.601	8.006	Unified Threshold Integrated Sample-to-Sample (UniTSFace) loss	36.49	36.437	36.946
Symmetric Cross Entropy	12.773	13.266	13.091	Sparsemax	42.179	42.54	46.124

Table 2: A study of different regularization methods incorporated into whether our metric learning loss functions directly or our ECAPA-TDNN model for better overall generalization of our SV system. Results are reported in terms of the EER (%) downstream SV evaluation performance. We used the CAMSAT algorithm to generate PLs using different predefined numbers of clusters.

Loss function \ Regularization method	True labels	No regularization			Sub-centers			Label Smoothing			i-mix			l-mix		
	5,994	**5,000**	**5,994**	**10,000**	**5,000**	**5,994**	**10,000**	**5,000**	**5,994**	**10,000**	**5,000**	**5,994**	**10,000**	**5,000**	**5,994**	**10,000**
Cross Entropy	3.489	5.477	5.827	5.546	5.795	5.97	6.177	4.369	4.412	4.592	4.73	4.883	4.798	5.095	4.883	4.989
AdaFace [42]	**1.326**	3.059	3.112	3.059	3.134	3.134	2.937	3.325	3.24	2.98	3.128	**2.985**	**2.916**	3.224	3.224	3.171
AAMSoftmax [20]	1.437	3.065	3.309	3.134	**2.969**	3.059	2.943	3.075	3.096	2.959	3.261	3.383	3.325	3.372	3.409	3.192
AMSoftmax [40]	1.522	3.054	3.224	2.959	2.996	3.049	2.996	3.128	3.33	3.017	3.213	3.425	3.372	3.409	3.499	3.224
OCSoftmax [41]	1.416	**2.964**	3.134	2.969	3.028	**2.948**	2.985	**2.906**	3.309	2.99	3.118	3.139	3.059	3.123	3.219	**2.959**
CosFace [43]	1.463	3.096	3.043	**2.863**	2.974	3.006	2.847	2.996	3.272	3.171	3.208	3.176	3.181	3.081	3.425	3.065
BoundaryFace [18]	1.479	3.096	**2.948**	2.884	3.065	3.022	**2.752**	3.224	3.181	**2.853**	3.150	3.165	3.028	3.171	3.256	3.150
Subcenter-ArcFace [17]	1.400	2.969	3.059	2.943	NA	NA	NA	**2.906**	**3.091**	2.9	**3.006**	3.134	3.139	**2.959**	**3.118**	3.033

the intra-class constraint of ArcFace by designing K sub-centers for each class to improve the robustness to label noise. In this case, the training sample only needs to be close to any of the K positive sub-centers instead of the only one positive center.

Other robust losses are MV-Arc-Softmax [47] which adaptively concentrates on optimizing the mis-classified (hard) feature vectors, as they are more crucial to enhance feature discriminability, to guide the discriminative feature learning. This loss combines the advantages of feature margin and feature mining into a unified loss function. Additionally, BoundaryFace [18] which, starting from the perspective of decision boundary, employs a novel mining framework that focuses on the relationship between a sample's ground truth class center and its nearest negative class center. Specifically, a noise label self-correction module is put forward to emphasize hard sample features that are between the ground truth class center and the nearest negative class center. If a sample is misclassified, there is a high probability that it is distributed within the nearest negative class's decision boundary, and the nearest negative class is likely to be the ground truth class of this misclassified sample. Based on this idea, BoundaryFace employs a module that automatically discovers misclassified samples during training and dynamically corrects their labels.

3. Our explored regularization techniques

In order to mitigate the effect of label noise in our clustering-based pseudo-labels, we investigate a variety of regularization techniques to incorporate into our SV systems, jointly with our loss functions (mixup [15], label smoothing [16], employing sub-centers [17]) to study their effect on the label noise robustness of self-supervised SV systems. The following list provides the details of each of these regularization techniques that we adopt with our best-performing loss variants:

- **Mixup augmentation**: We study two variants of mixup at both the instance input-level (i-mix) [48] and the latent space (l-mix) [5]. Indeed, the instance mix (i-mix) augmentation scheme [48] performs interpolation on the training samples and their PLs. As a result, the i-mix strategy can be applied to self-supervised learning tasks where no actual class labels are provided, and has shown potential in a number of self-supervised tasks including image classification and voice command recognition. On the other hand, the l-mix [5] strategy that applies i-mix on the latent space, instead of the raw data domain, may yield more diverse synthetic samples. To apply i-mix on the latent space of the speech, l-mix incorporates a variational autoencoder (VAE) encoder [49] to extract the latent variable of the given acoustic features. The resulting mixed latent variable is then fed into the VAE decoder to generate a new synthetic sample, with different patterns than the standard i-mix generated samples. As it favors the smoothness of the output distribution, the mixup strategy has been shown in [50] to be effective in mitigating the memorization effects of label noise, and help to slow down the memorization of noisy labels and learn long enough from the simple patterns available.

  In our experiments, we train our maximum-margin loss-based SV systems jointly with i-mix or l-mix augmentations to regularize the model weights.

- **Label Smoothing**: Label Smoothing (LS) [16, 51] regularization uses soft labels in place of one-hot labels to alleviate overfitting to noisy labels, and help mitigate label noise [52, 53]. We incorporate this regularization directly into our studied losses.
- **Employing Sub-centers**: [17] introduced a novel loss function called Subcenter-ArcFace which relaxes the intra-class constraint (force all samples close to the corresponding positive center) of ArcFace to improve the robustness to label noise. More specifically, the authors designed K sub-centers for each class and a training sample only needs to be close to any of the K positive sub-centers instead of only one positive center as employed in usual metric learning losses. Very importantly, since the intra-class constraint enforces a training sample to be close to one of the multiple positive sub-classes but not all of them, the proposed subcenter-ArcFace encourages one dominant sub-class that contains the majority of clean samples and non-dominant sub-classes that include hard or noisy samples. As a consequence, the noise is likely to form a non-dominant sub-class and will not be enforced into the dominant sub-class. Therefore, subcenter-ArcFace is more robust to label noise. In this paper, we incorporate this idea directly into other losses to improve their generalizability and study their behavior.

4. Experimental setup

As input to our CAMSAT clustering algorithm, we employ 400-dim i-vectors. The compact i-vectors, which are unsupervised speaker representations, allow us here to perform clustering in a more efficient way and to avoid high dimensionality of the MFCC acoustic features.

In order to evaluate the performance of our proposed self-supervised approach for SV, we conducted a set of experiments based on the VoxCeleb2 dataset [54]. To train the embedding networks, we used the development subset of the VoxCeleb2 dataset, which consists of 1,092,009 utterances collected from 5,994 speakers. The evaluation was performed according to the original VoxCeleb1 trials list [55], which consists of 37,720 trials of 4,874 utterances spoken by 40 speakers.

For our ECAPA-TDNN-based SV system, the acoustic features used in the experiments were 40-dimensional Mel-frequency cepstral coefficients (MFCCs) extracted at every 10 ms, using a 25 ms Hamming window via Kaldi toolkit [56]. Moreover, to follow other SV works in training the ECAPA-TDNN-based systems, we have used waveform-level data augmentations including additive noise and room impulse response (RIR) simulation [7]. In addition to the waveform-level augmentations, we have also applied augmentation over the extracted MFCCs feature, analogous to the specaugment scheme [57].

All speaker verification experiments have been run for 7 days using a single A40 GPU (or RTX2080Ti in some cases), with a batch size of 200 MFCC samples. All margin-based losses are run with scale factor $s = 30$ and angular margin $m = 0.2$. Cosine similarity was used as a backend for verification scoring between enrollment and test embeddings.

i-mix and l-mix regularization strategies are used with $\alpha = 0.5$, where $\lambda \sim Beta(\alpha, \alpha)$ is the mixing coefficient from the Beta distribution to interpolate inputs or latent embeddings, respectively. We use a smoothing coefficient of 0.15 when applying a weighted average between the uniform distribution and the provided PLs during label smoothing. Finally, we use $K = 3$ sub-centers wherever sub-centers are employed.

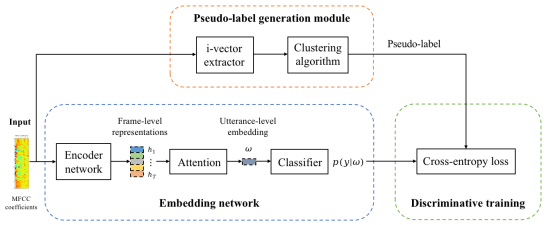

Figure 1: General process for training our clustering generated pseudo-label-based self-supervised speaker embedding networks.

4.1. Our clustering-based self-supervised speaker embedding framework

Figure 1 depicts a schematic diagram of our general clustering-based self-supervised SV process that we follow throughout the paper. During our work, we explore various loss functions and regularization methods and conduct different analyses of their impact on the robustness of SV performance. We employ ECAPA-TDNN as our speaker embedding network and use our adopted loss objectives to train this system using PLs generated by the CAMSAT clustering algorithm.

4.2. Clustering-based pseudo-label generation

For clustering, we have extracted i-vector [2, 3] embeddings using the Kaldi toolkit [56], which is a statistical unsupervised fixed-dimensional representation from each training utterance and performed clustering on top of them. After training the clustering CAMSAT-based model, we selected the aligned cluster for each utterance and used the cluster-id as PL. With the clustering-based PLs, we can train the speaker embedding network via our metric learning loss objectives, analogous to supervised learning.

For a thorough comparison, we have set the number of clusters to be in {5000, 5994, 10000} to study the influence of the predefined number of clusters on the downstream speaker verification performance (5994 is the ground truth number).

4.3. Clustering performance of our pseudo-labels

Table 3 shows the clustering performance of our employed clustering-based PLs using CAMSAT to generate these PLs. From the unsupervised accuracy and mutual information scores, we can observe that our obtained cluster assignments are noisy and not pure, hence the existence of discrepancies between the PLs and the speaker-identity ground truths. As a result, in several cases, our SV performance was degraded from overfitting this label noise.

Table 3: The clustering performance of our CAMSAT-based pseudo-labels using different numbers of predefined clusters.

No. of predefined clusters	No. of discovered clusters	ACC	NMI	AMI
5,000	4,596	0.655	0.874	0.812
5,994	5,194	0.669	0.878	0.816
10,000	6,364	0.709	0.889	0.830

4.4. Clustering Evaluation Metrics

Following the commonly used evaluation metrics for clustering, we assess the quality of the generated pseudo-labels using the following three supervised clustering metrics:

- **Unsupervised Clustering Accuracy (ACC)**: measures the consistency between the true labels and the generated PLs. $ACC = \max_{m} \frac{\sum_{i=1}^{N} \mathbb{1}\{y_i = m(c_i)\}}{N}$ where y_i is the true label, c_i is the generated PL assignment, and m is a mapping function which ranges over all possible one-to-one mappings between true labels and assignments. The optimal mapping can be efficiently computed using the Hungarian algorithm [58].
- **Normalized Mutual Information (NMI)** [59]: $NMI(Y, C) = \frac{I(Y,C)}{\frac{1}{2}[H(Y)+H(C)]}$ where Y and C denote the ground-truth labels and the clustering assignments, respectively. H is the entropy function and I denotes the MI metric. NMI is the harmonic mean between below homogeneity and completeness scores.
- **Adjusted MI (AMI)** [60]: Since the NMI measure is not adjusted for chance, including the adjusted MI score might be preferred for comparison in some of our cases.

4.5. CAMSAT clustering algorithm

For clustering, we adopt the same CAMSAT clustering approach used in [19] to generate pseudo-labels. CAMSAT is based on augmentation mix and self-augmented training. The goal is to impose invariance to data augmentation on the output predictions of deep models in an end-to-end fashion while maximizing the information-theoretic dependency between samples and their predicted discrete representations (cluster assignments). It provided both state-of-the-art speaker clustering and SV performance. In this paper, we try to investigate several metric learning loss functions to enhance the generalization performance of self-supervised speaker embedding systems and to mitigate the negative effect of heavy noise in the generated pseudo-labels (PLs) used to train these systems. Please refer to [19] for details about CAMSAT architecture and training details.

5. Results and Discussion

In Table 1, we performed a large-scale study of 39 metric learning loss functions including all the above-mentioned families of loss objectives and other widely used losses using CAMSAT-based pseudo-labels.

Besides, in an attempt to further enhance SV performance by improving generalization and robustness and mitigating the memorization of label noise, in Table 2 we summarize our results using 4 additional different regularization techniques (with different predefined numbers of clusters) employed to train our SV model using the selection of our best-performing loss functions in Table 1.

Throughout our experiments, we can observe that incorporating a margin can easily enhance the performance of our metric learning loss functions, often outperforming supervised training with cross entropy using the true labels. Results show clearly that our selection of maximum-margin softmax variants in Table 2 are very effective in improving the generalization of our SV systems across all types of label noise contained in the PLs. In particular, unlike the widely used AAMSoftmax loss in SV, to our knowledge, our results indicate for the first time that variants such as OCSoftmax using one-class learning instead of multi-class classification and not assuming the same distribution for all speakers (which is more realistic in our case), or the recent AdaFace and SMAFace losses, perform consistently better across the 3 PLs and the ground truth labels. Indeed, AAMSoftmax is susceptible to massive label noise [20]. This is because if a training sample is noisy (misclassified), it does not belong to the corresponding positive class. In AAMSoftmax, this noisy sample generates a large wrong loss value, which impairs the model training. This partially explains the under-performance of AAMSoftmax compared to other variants when using pseudo-labels for training. Figure 2 also shows clearly this overfitting phenomenon which affects the majority of loss functions, and consequently the dramatic degradation of the downstream validation EER performance over epochs due to memorization of noisy labels. Interestingly, thanks to its design to be robust to label noise, we can also observe the good performance of Subcenter-ArcFace, which often outperforms all other losses across our various studied PLs. This can be explained by using sub-centers which make the final dominant vector centers (the clean ones) more compact and well distant from each other. The high-performing BoundaryFace also shows that label correction is an important component and can often help to mitigate label noise during training.

Besides, in our experiments on the VoxCeleb1-O test set, sample-to-sample loss functions and other losses such as MagFace, BroadFace, DropMax, Center loss, Softmax, Gumbel-Softmax and Sparsemax performed poorly and seem to suffer from serious problems of convergence, numerical instability, or sensitivity to hyperparameters. On the other hand, we can observe that the normalization operation to make our losses symmetric helped us to improve performance in the case of Softmax and Binary CE (BCE). Finally, we found, as shown in Table 1, that recently proposed NLFs and NNLFs losses both performed poorly in our case compared to our suggested maximum-margin softmax-based variants.

Moreover, using different predefined numbers of clusters including the ground truth number of clusters, we can see that the final downstream SV evaluation performance depends more on the quality of the PLs, and that the consideration of the predefined number of clusters is less important.

Table 4: Some recent SOTA Self-Supervised SV approaches in EER (%) compared to our simple SV system trained with CAMSAT-based PLs and Subcenter-BoundaryFace loss. All models are based on ECAPA-TDNN. Results are reported on the original VoxCeleb1 test set (Voxceleb1_O).

SSL Objective	EER (%)
MoBY [10]	8.2
InfoNCE [12]	7.36
MoCo [61]	7.3
ProtoNCE [10]	7.21
PCL [10]	7.11
CA-DINO [62]	3.585
i-mix [63]	3.478
l-mix [63]	3.377
Iterative clustering [12]	3.09
CAMSAT [19]	3.065
Our approach (using Subcenter-BoundaryFace)	**2.752**

Finally, Table 4 shows a comparison of our approach for Self-Supervised SV training using CAMSAT-based PLs and our best-performing Subcenter-BoundaryFace loss using sub-center regularization, compared to recent SOTA self-supervised SV approaches employing diverse SSL objectives with the same ECAPA-TDNN model encoder. The results show clearly that our approach largely outperforms all the baselines while being

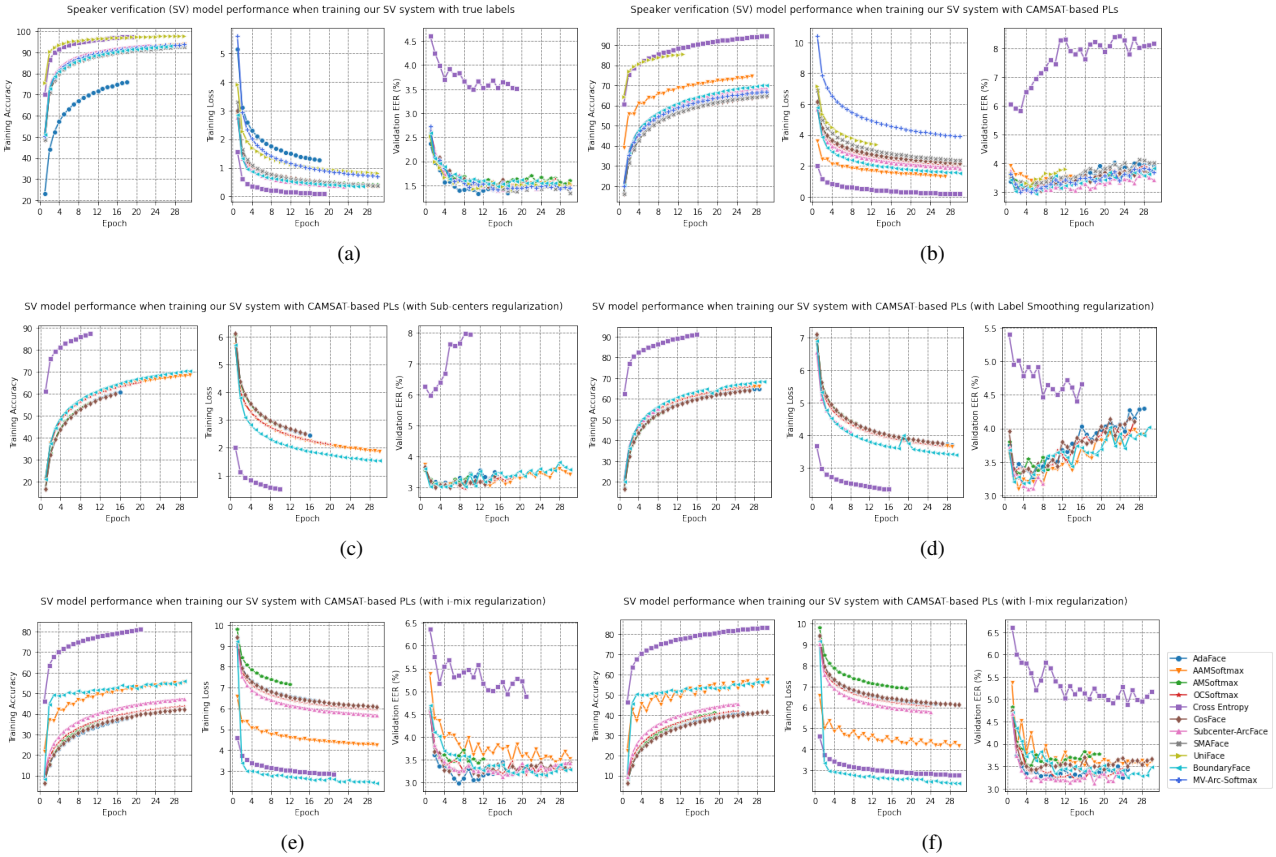

Figure 2: Training accuracy/loss and validation performance over time of our speaker verification (SV) system trained under various loss functions, using different regularization techniques. We employ ground-truth labels in (a) and CAMSAT-based PLs in the rest with (b) No regularization (c) Sub-centers regularization (d) Label smoothing regularization (e) i-mix regularization and (f) l-mix regularization.

simple and fast, which suggests that the consideration of loss functions is still crucial and that further gains can still be made by simply improving the loss objectives of current self-supervised speaker recognition systems.

5.1. Influence of regularization methods on metric learning losses over epochs

In Figure 2, we study the evolution of the downstream evaluation EER (%) performance and the training accuracy and loss of our system trained with our selection of maximum-margin-based loss functions under our studied regularization (we use our 5994-based CAMSAT PLs). In particular, we perform the same experiments using the original ground-truth labels to suppress the effect of label noise and study its impact on the generalization and training of SV systems. First of all, despite the good generalization of our SV systems, we can observe that these metric learning losses still suffer from overfitting and from the phenomenon of label noise memorization [36] when trained with our noisy pseudo-labels. This demonstrates that producing compact cluster assignments (compact probabilities) with more discriminative ability does not necessarily help to mitigate memorization of label-noise. Despite inducing better generalization to out-of-set samples, maximum-margin softmax losses do not seem to reduce sufficiently the model's ability to accommodate random noise during training.

Indeed, due to the memorization effects [36], deep models (in particular, overparameterized networks), tend to fit easy (clean) patterns in the pseudo-labels first, and then overfit the hard and complex (noisy) patterns gradually. This leads to overfitting the noise and corruptions in the training pseudo-labels and eventually, the validation curve starts to drop gradually. This highlights the importance of having highly accurate PLs for good generalization of self-supervised SV systems. Very interestingly, on the contrary to other losses where validation performance starts to degrade after only the first few epochs, we can observe in figures 2-(b) and (c) that using sub-centers is more robust to label noise and does suffer the least from overfitting compared to other losses. It is worth mentioning, however, that using sub-centers remains much slower than other methods due to its use of a much bigger matrix of sub-centers. Besides, we can also observe that mixup regularization via both i-mix and l-mix in Fig. 2-(e) and (f) are really beneficial to prevent overfitting through time, with a strong regularization effect than using sub-centers but slightly underperforming sub-center regularization overall. As far as label smoothing is concerned, we could observe a lighter regularization effect that prevents the training loss from overfitting strongly the PLs, which can be explained by the model becoming less overconfident about its predictions. However, this effect does not necessarily translate into better generalization, except for the cross entropy loss in Figure 2-(b).

Finally, although we could observe a slight underfitting phe-

nomenon when it comes to using our studied loss functions (except the cross entropy loss which always overfits easily), in particular the robustly-designed variants such as Subcenter-ArcFace, BoundaryFace and MV-Arc-Softmax when trained with the ground-truth labels in 2-(a). On the contrary to other papers such as [34] that underscore this as a negative side effect, we find this behavior to remain less of an issue and more of a feature to delay overfitting and allow the model to learn long enough from the simple patterns available. Importantly, this induces better generalization as this can be observed in the same figure. Additionally, this result can be confirmed in Figure 2-(b) when using the CAMSAT-based PLs where we can observe that, instead of overfitting the PLs, the models are converging slowly and steadily towards 65 to 70% training accuracy, which is around the unsupervised clustering accuracy (ACC) of our CAMSAT PLs. This can also point to an interesting ability of these losses to somehow assess how reliable is each PL and to remain focused on the most relevant (accurate) ones. We believe this is an important result that needs further investigation and some theoretical analysis to explain.

6. Conclusion

In this work, we performed a comparative study of a wide range of recent metric learning loss functions and 4 regularization techniques for better generalization of Self-Supervised Speaker Verification (SSSV) systems. In particular, we investigated the effect of these losses on the robustness of the SSSV task against label noise, and proposed a selection of loss functions combining with our proposed regularization techniques against label noise that often lead to considerable improvements in self-supervised SV performance.

7. Acknowledgment

The authors wish to acknowledge the funding from the Government of Canada's New Frontiers in Research Fund through grant NFRFR-2021-00338 and Natural Sciences and Engineering Research Council of Canada through grant RGPIN-2019-05381.

8. References

[1] John H.L. Hansen and Taufiq Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, 2015.

[2] Najim Dehak et al., "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, 2011.

[3] Patrick Kenny, "A Small Footprint I-vector Extractor," in *Odyssey*, 2012, pp. 1–6.

[4] Z. Bai and X. L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, 2021.

[5] Woo Hyun Kang, Jahangir Alam, and Abderrahim Fathan, "l-mix: a latent-level instance mixup regularization for robust self-supervised speaker representation learning," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[6] Abderrahim Fathan, Jahangir Alam, and Woohyun Kang, "On the impact of the quality of pseudo-labels on the self-supervised speaker verification task," in *NeurIPS 2022 Second ENLSP Workshop*, 2022.

[7] David Snyder et al., "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. of IEEE ICASSP*, 2018, pp. 5329–5333.

[8] Brecht Desplanques et al., "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*. ISCA.

[9] Woo Hyun Kang, Jahangir Alam, and Abderrahim Fathan, "An analytic study on clustering-based pseudo-labels for self-supervised deep speaker verification," in *SPECOM*, 2022.

[10] Wei Xia et al., "Self-supervised text-independent speaker verification using prototypical momentum contrastive learning," in *ICASSP*. IEEE, 2021.

[11] Junyi Peng et al., "Progressive Contrastive Learning for Self-Supervised Text-Independent Speaker Verification," in *Proc. of Odyssey Workshop*, 2022.

[12] Ruijie Tao et al., "Self-supervised speaker recognition with loss-gated learning," in *ICASSP*. IEEE, 2022.

[13] B. Han, Z. Chen, and Y. Qian, "Self-supervised speaker verification using dynamic loss-gate and label correction," *arXiv preprint arXiv:2208.01928*, 2022.

[14] Yunfan Li et al., "Contrastive clustering," in *AAAI*, 2021.

[15] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[16] Christian Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[17] J. Deng, J. Guo, et al., "Sub-center arcface: Boosting face recognition by large-scale noisy web faces," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, Proceedings, Part XI 16*. Springer, 2020, pp. 741–757.

[18] Shijie Wu and Xun Gong, "Boundaryface: A mining framework with noise label self-correction for face recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 91–106.

[19] Abderrahim Fathan and Jahangir Alam, "Camsat: Augmentation mix and self-augmented training clustering for self-supervised speaker recognition," in *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, 2023.

[20] Jiankang Deng et al., "Arcface: Additive angular margin loss for deep face recognition," *IEEE TPAMI*, 2021.

[21] Eyal Beigman and Beata Beigman Klebanov, "Learning with annotation noise," in *Proc. of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 2009, pp. 280–287.

[22] Melody Guan et al., "Who said what: Modeling individual labelers improves classification," in *Proc. of the AAAI conference on artificial intelligence*, 2018, vol. 32.

[23] David Rolnick, Andreas Veit, et al., "Deep learning is robust to massive label noise," *ICLR*, 2018.

[24] Armand Joulin, Laurens van der Maaten, et al., "Learning visual features from large weakly supervised data," in *European Conference on Computer Vision*. Springer, 2016, pp. 67–84.

[25] Ishan Misra et al., "Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2930–2939.

[26] Davood Karimi et al., "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical Image Analysis*, vol. 65, 2020.

[27] Carla E Brodley and Mark A Friedl, "Identifying mislabeled training data," *Journal of artificial intelligence research*, 1999.

[28] Sainbayar Sukhbaatar et al., "Training convolutional networks with noisy labels," *arXiv preprint arXiv:1406.2080*, 2014.

[29] Andreas Veit et al., "Learning from noisy large-scale datasets with minimal supervision," in *Proc. of the IEEE conference on CVPR*, 2017.

[30] Duc Tam Nguyen et al., "Self: Learning to filter noisy labels with self-ensembling," *arXiv preprint arXiv:1910.01842*, 2019.

[31] Aritra Ghosh et al., "Robust loss functions under label noise for deep neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, 2017, vol. 31.

[32] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[33] Yisen Wang et al., "Symmetric cross entropy for robust learning with noisy labels," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 322–330.

[34] X. Ma, H. Huang, et al., "Normalized loss functions for deep learning with noisy labels," in *International conference on machine learning*. PMLR, 2020, pp. 6543–6553.

[35] Xichen Ye et al., "Active negative loss functions for learning with noisy labels," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[36] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, et al., "A closer look at memorization in deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 233–242.

[37] Abderrahim Fathan, Jahangir Alam, and Xiaolin Zhu, "Multi-task learning over mixup variants for the speaker verification task," in *International Conference on Speech and Computer*. Springer, 2023, pp. 446–460.

[38] G. F. Elsayed et al., "Large margin deep networks for classification," 2018.

[39] W. Liu, Y. Wen, et al., "Large-margin softmax loss for convolutional neural networks.," in *ICML*, 2016, vol. 2.

[40] F. Wang et al., "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[41] You Zhang et al., "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, 2021.

[42] Minchul Kim et al., "Adaface: Quality adaptive margin for face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18750–18759.

[43] Hao Wang et al., "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

[44] J. Zhou, X. Jia, Q. Li, et al., "Uniface: Unified cross-entropy loss for deep face recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20730–20739.

[45] Qiufu Li, Xi Jia, Jiancan Zhou, et al., "Unitsface: Unified threshold integrated sample-to-sample loss for face recognition," *arXiv preprint arXiv:2311.02523*, 2023.

[46] Jiankang Deng, Jia Guo, et al., "Variational prototype learning for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11906–11915.

[47] Xiaobo Wang, Shifeng Zhang, et al., "Mis-classified vector guided softmax loss for face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 12241–12248.

[48] Kibok Lee et al., "i-mix: A domain-agnostic strategy for contrastive representation learning," in *ICLR*, 2021.

[49] Diederik P. Kingma and Max Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR*, 2014.

[50] Abderrahim Fathan and Jahangir Alam, "An analytic study on clustering driven self-supervised speaker verification," *Pattern Recognition Letters*, 2024.

[51] Gabriel Pereyra, George Tucker, et al., "Regularizing neural networks by penalizing confident output distributions," *arXiv preprint arXiv:1701.06548*, 2017.

[52] Michal Lukasik et al., "Does label smoothing mitigate label noise?," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6448–6458.

[53] Blair Chen, Liu Ziyin, et al., "An investigation of how label smoothing affects generalization," *arXiv preprint arXiv:2010.12648*, 2020.

[54] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[55] A. Nagrani, J. S. Chung, et al., "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[56] Daniel Povey et al., "The kaldi speech recognition toolkit," in *In IEEE 2011 workshop*, 2011.

[57] Daniel S. Park et al., "Specaugment: A simple data augmentation method for automatic speech recognition.," in *Interspeech*, 2019, pp. 2613–2617.

[58] Harold W Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics (NRL)*, vol. 52, no. 1, pp. 7–21, 2005.

[59] Pablo A Estévez et al., "Normalized mutual information feature selection," *IEEE Transactions on neural networks*, 2009.

[60] Nguyen Xuan et al., "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," 2010.

[61] Jejin Cho et al., "The jhu submission to voxsrc-21: Track 3," *arXiv preprint arXiv:2109.13425*, 2021.

[62] Bing Han et al., "Self-supervised learning with cluster-aware-dino for high-performance robust speaker verification," *arXiv preprint arXiv:2304.05754*, 2023.

[63] Abderrahim Fathan and Jahangir Alam, "On the influence of the quality of pseudo-labels on the self-supervised speaker verification task: a thorough analysis," in *IWBF*. IEEE, 2023.