Interspeech 2024
1-5 September 2024, Kos, Greece

On the impact 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.

Index Terms: speaker recognition, regularization, label noise

1. Introduction

Automatic speaker verification (ASV), as one of the most convenient means of biometric recognition [1], uses the voiceprint of a speaker to verify their identity. 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] 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 [3]. They have shown great performance when large training datasets are available, particularly with a sufficient number of speakers [4]. One widely employed architecture for this purpose is ECAPA-TDNN [5], 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) [6, 7, 8], or to employ SSL-based objectives (e.g., SimCLR, MoCo [9]) to generate PLs and train the speaker embedding network using these labels in a discriminative fashion [10, 11]. Despite the impressive performance of these PL-based Self-Supervised SV schemes, clustering performance remains a bottleneck in all above approaches [11, 12] since downstream performance relies greatly on accurate PLs while 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 [11, 13]. Indeed, recent studies have shown that label noise can remarkably impact downstream performance [8]. 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 [14], label smoothing [15], employing sub-centers [16]) 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 [16] or BoundaryFace [17] for the task of SV under label noise. To generate well-performing PLs, we employed the CAMSAT clustering model [18].

Our 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 recent max-margin-based softmax losses are beneficial to mitigate memorization of label noise and outperform some widely-used losses in the domain of SV, such as the angular additive margin softmax (ArcFace) [19] loss.
- 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.
- Using CAMSAT-based PLs [18], our proposed selection of loss objectives allowed us to achieve SOTA SV performance, outperforming various benchmarks.

* Independent Researcher

10.21437/Interspeech.2024-2206

2. Background and Related Work

2.1. Noise-robust loss functions

Methods to learn from noisy data can be categorized into noise-robust algorithms [20] that directly learn from noisy labels and label-cleansing methods [21] that aim to remove or correct mislabeled data. Recent research has introduced various robust loss-based methods [22, 23, 24, 25, 26, 8, 27]. In this paper, we investigate a large variety of robust loss functions, including all mentioned losses. The full list of our results including 39 loss functions can be found in `https://github.com/fathana/label_noise_regularization`.

2.2. Maximum margin-based softmax loss objectives

To improve performance on previously unseen data and generalize to out-of-domain speech samples, various maximum margin-based softmax variants have been proposed. Indeed, softmax suffers from several drawbacks such as that (1) its computation of inter-class margin is intractable [28] and (2) the learned projections are not guaranteed equi-spaced. To solve these problems, several alternatives to softmax have been proposed. Variants like AMSoftmax [29] introduce margin constraints to enhance variance between classes and reduce variance within classes. ArcFace enhances discriminability by adding an angular margin constraint, while CosFace [30] reformulates softmax loss as cosine loss and introduces cosine margins. Besides, OCSoftmax [31] uses one-class learning instead of multi-class classification and does not assume uniform class distributions. Moreover, AdaFace [32] emphasizes misclassified samples based on speaker embedding quality. Finally, Subcenter-ArcFace relaxes intra-class constraints by introducing sub-centers for each class to learn clean dominant centers and mitigate label noise, while other robust losses like MV-Arc-Softmax [33] focus on optimizing misclassified feature vectors, and BoundaryFace [17] dynamically corrects labels by mining misclassified samples.

3. Our explored regularization techniques

To mitigate the effect of label noise in our clustering-based PLs, we investigate a variety of regularizations (mixup [14], label smoothing [15], employing sub-centers [16]) to study their effect on label noise robustness of our self-supervised SV systems. The following list provides the details of each of these regularizations 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) [34] and the latent space (l-mix) [7]. Indeed, the instance mix (i-mix) augmentation scheme [34] 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 [7] 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 [35] 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 [36] to be effective in mitigating the memorization effects of label noise and learn long enough from the simple patterns.
- **Label Smoothing**: Label Smoothing (LS) [37] regularization uses soft labels in place of one-hot labels to alleviate overfitting to noisy labels, and help mitigate label noise [38].
- **Employing Sub-centers**: [16] 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.

We incorporate the above-mentioned regularizations into our studied losses to train our SV systems for better generalization.

4. Experimental setup

We conducted a set of experiments based on the VoxCeleb2 dataset [39]. 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 [40], which consists of 37,720 trials of 4,874 utterances spoken by 40 speakers. No score normalization was applied.

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. 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 [4]. Besides, we have applied augmentation over the extracted MFCCs feature, analogous to the specaugment scheme [41]. All SV experiments have been run with a batch size of 200 MFCC samples. Scale factor $s = 30$ and margin $m = 0.2$ were used for all losses. Code and details of our hyperparameters are provided in our github repo.

4.1. Our self-supervised speaker embedding framework

Fig. 1 depicts a schematic diagram of our general clustering-based self-supervised SV process that we follow throughout the paper. 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. CAMSAT clustering-based pseudo-label generation

For clustering, we adopt the same CAMSAT clustering approach used in [18] to generate PLs. We have extracted i-vector [2] embeddings as inputs using Kaldi. I-vectors allow us here to perform clustering in a more efficient way and to avoid high dimensionality of the MFCCs acoustic features. Please refer to our github repo and [18] for details about CAMSAT architecture and training details. After training the clustering CAMSAT-based model with i-vector inputs, we selected the aligned cluster for each utterance and used the cluster-id as PL. Using these 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). Our github repo also includes

Table 1: *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	3.478	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 [32]	**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
ArcFace [19]	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 [29]	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	3.372	3.224
OCSoftmax [31]	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	3.070	3.192	**2.959**
CosFace [30]	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 [17]	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 [16]	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
MV-Arc-Softmax [33]	1.400	**2.842**	3.006	2.884	**2.937**	3.001	2.932	2.969	3.229	3.091	**2.996**	3.102	3.012	3.065	**3.112**	3.022

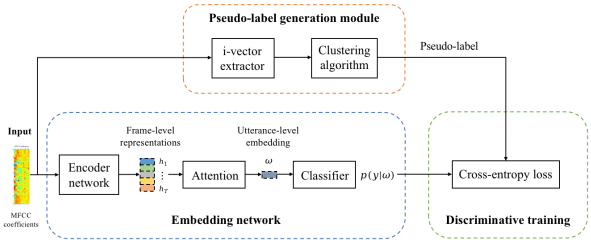

Figure 1: *General process for training our clustering generated PL-based self-supervised speaker embedding networks.*

Table 2: *Some recent SOTA Self-Supervised SV approaches in EER (%) compared to our simple SV system trained with CAMSAT-based PLs and Subcenter-BoundaryFace loss.*

SSL Objective	EER (%)
MoBY [9]	8.2
InfoNCE [11]	7.36
MoCo [42]	7.3
ProtoNCE [9]	7.21
PCL [9]	7.11
CA-DINO [43]	3.585
i-mix [44]	3.478
l-mix [44]	3.377
Iterative clustering [11]	3.09
CAMSAT [18]	3.065
Our approach (using Subcenter-BoundaryFace)	**2.752**

a Table of the clustering performance of our PLs in terms of different clustering metrics which show that our generated cluster assignments are noisy, 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.

5. Results and Discussion

As shown in our github repo, we performed a large-scale study of 39 metric learning loss functions including all the above-mentioned families of loss objectives and other recent losses using CAMSAT-based PLs. We selected our best-performing loss functions among them to further enhance SV performance by improving generalization and robustness and mitigating the memorization of label noise. In Table 1, we summarize our results employing 4 additional different regularization techniques.

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 max-margin softmax variants in Table 1 are very effective in improving the generalization of SV across all types of label noise contained in the PLs. In particular, unlike the widely used ArcFace 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 loss, perform consistently better across the 3 PLs and the ground truth labels. Indeed, ArcFace is susceptible to massive label noise [19]. This is because if a training sample is misclassified, it does not belong to the corresponding positive class. In ArcFace, this noisy sample generates a large loss value, which impairs the training. This partially explains the under-performance of ArcFace compared to other variants when using PLs for training. Besides, Fig. 2-(b) 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 compared to 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 MV-Arc-Softmax and BoundaryFace also show that sample mining and label correction are important components and can often help to mitigate label noise during training.

Finally, Table 2 shows a comparison of our approach for Self-Supervised SV training using CAMSAT-based PLs and our best-performing Subcenter-BoundaryFace loss using sub-centers regularization, compared to recent SOTA self-supervised SV approaches employing diverse SSL objectives with the same ECAPA-TDNN model encoder, on Voxceleb1_O. The results show clearly that our approach largely outperforms all baselines while being simple and fast, which suggests that the consideration of losses is still crucial and that gains in speaker recognition can still be made by simply improving these loss objectives.

5.1. Influence of regularization methods on metric learning

In Fig. 2-(c,d,e,f), we study the evolution of the downstream evaluation EER (%) performance and the training accuracy of our system trained with our selection of max-margin-based loss functions under our studied regularizations. In particular, we perform the same experiments using the original ground-truth labels to suppress the effect of label noise and simply study the impact of these losses on generalization. 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 [45] when trained

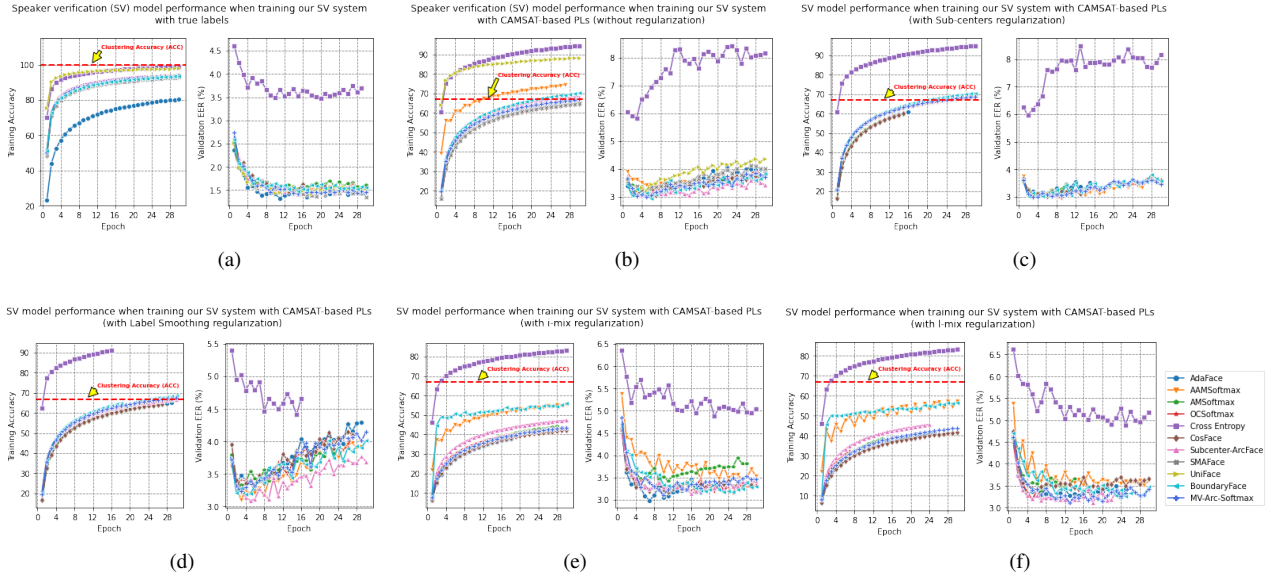

Figure 2: *Training accuracy and validation performance over time of our SV system trained under various loss functions, using different regularization techniques. We employ ground-truth labels in (a) and CAMSAT-based PLs (5994 clusters) in the rest with (b) No regularization (c) Sub-centers regularization (d) Label smoothing regularization (e) i-mix regularization and (f) l-mix regularization.*

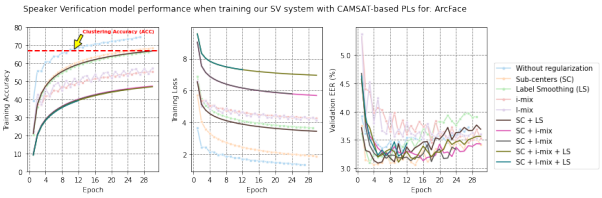

Figure 3: *Training accuracy/loss and validation EER performance over time of our speaker verification (SV) system, trained with ArcFace loss, under various combinations of regularization techniques versus without regularization.*

with noisy PLs. From the figures, we can particularly notice that the generalization to the test set starts to degrade even before reaching the clustering accuracy (ACC) of the PLs (ACC is optimal if no label correction is applied). This means that the models trained with these losses w/o regularization start to fit (memorize) the noisy labels even before taking full advantage of the clean (accurate) PLs. Despite inducing better generalization to out-of-set samples, max-margin losses and our studied regularizations do not seem to reduce sufficiently the model's ability to accommodate label noise during training. Thus, the need for effective losses that can better discover wrong labels.

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 and regularizations. 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 (see the training accuracy) but slightly underperforming sub-centers 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 always translate into better generalization, except for cross entropy in Fig. 2-(b).

Table 3: *A study of different combinations of our regularization methods for better generalization of our SV systems.*

Applied regularization methods				Loss functions					
				ArcFace			BoundaryFace		
i-mix	l-mix	Label Smoothing	Sub-centers	5,000	5,994	10,000	5,000	5,994	10,000
✗	✗	✗	✗	3.065	3.309	3.134	3.096	2.948	2.884
✓	✗	✗	✗	3.261	3.383	3.325	3.15	3.165	3.028
✗	✓	✗	✗	3.372	3.409	3.192	3.171	3.256	3.15
✗	✗	✓	✗	3.075	3.096	2.959	3.224	3.181	2.853
✗	✗	✗	✓	2.969	**3.059**	2.943	3.065	3.022	**2.752**
✓	✗	✗	✓	3.006	3.134	3.139	3.165	NA	2.927
✗	✓	✗	✓	2.959	3.118	3.033	2.974	3.038	3.086
✗	✗	✓	✓	**2.906**	3.091	**2.9**	**2.869**	**2.937**	2.943
✓	✗	✓	✓	3.091	3.107	3.128	3.383	3.139	3.006
✗	✓	✓	✓	3.123	3.213	3.165	3.187	3.091	3.155

5.2. Combination of regularization methods

In Table 3, we study the complementarities between our studied regularization methods and their effect on the performance of SV systems, using ArcFace and BoundaryFace. Results show that our combinations (e.g. label smoothing with sub-centers) often provide improvements in terms of SV performance. We can also observe in Fig. 3 using ArcFace that these combinations are beneficial in most cases compared to single regularizations by leading to better EER generalization and performance and to smoother curves that demonstrate a better ability to mitigate label noise memorization of our SV systems.

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. We investigated the effect of these

losses on the robustness of the SSSV task against label noise, and proposed a selection of regularization techniques which often lead to considerable improvements in self-supervised SV.

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] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, 2015.

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

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

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

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

[6] W. H. Kang, J. Alam, and A. Fathan, "An analytic study on clustering-based pseudo-labels for self-supervised deep speaker verification," in *SPECOM*, 2022.

[7] W. H. Kang *et al.*, "L-mix: A latent-level instance mixup regularization for robust self-supervised speaker representation learning," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[8] A. Fathan, J. Alam, and W. Kang, "On the impact of the quality of pseudo-labels on the self-supervised speaker verification task," in *NeurIPS 2022 Second ENLSP Workshop*, 2022. [Online]. Available: https://neurips2022-enlsp.github.io/papers/paper_51.pdf

[9] W. Xia *et al.*, "Self-supervised text-independent speaker verification using prototypical momentum contrastive learning," in *ICASSP*. IEEE, 2021, pp. 6723–6727.

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

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

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

[13] Y. Li *et al.*, "Contrastive clustering," in *AAAI*, 2021.

[14] H. Zhang *et al.*, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[15] C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proc. of the IEEE conference on CVPR*, 2016, pp. 2818–2826.

[16] 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.

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

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

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

[20] I. Misra *et al.*, "Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels," in *Proc. of the IEEE conference on CVPR*, 2016, pp. 2930–2939.

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

[22] A. Ghosh *et al.*, "Robust loss functions under label noise for deep neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[23] 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.

[24] Y. 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.

[25] X. Ye *et al.*, "Active negative loss functions for learning with noisy labels," in *Thirty-seventh Conference on NeurIPS*, 2023.

[26] M. Belkin *et al.*, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples." *Journal of machine learning research*, vol. 7, no. 11, 2006.

[27] T. Miyato *et al.*, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE TPAMI*, vol. 41, no. 8, pp. 1979–1993, 2018.

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

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

[30] H. Wang *et al.*, "Cosface: Large margin cosine loss for deep face recognition," in *Proc. of the IEEE conference on CVPR*, 2018, pp. 5265–5274.

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

[32] M. 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. 18 750–18 759.

[33] X. Wang, S. Zhang *et al.*, "Mis-classified vector guided softmax loss for face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 241–12 248.

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

[35] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR*, 2014. [Online]. Available: http://arxiv.org/abs/1312.6114

[36] A. Fathan and J. Alam, "An analytic study on clustering driven self-supervised speaker verification," *PRL*, 2024.

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

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

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

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

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

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

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

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

[45] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger *et al.*, "A closer look at memorization in deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 233–242.