



Automatic Labeling and Correction of Noisy Labels for Robust Self-Supervised Speaker Verification

Abderrahim Fathan¹, Jahangir Alam¹

¹Computer Research Institute of Montreal (CRIM), Montreal, Canada

abderrahim.fathan@crim.ca, jahangir.alam@crim.ca

Abstract

Supervised speaker verification relies on large labeled datasets, which are costly and labor-intensive to create. However, both manual and clustering-based labeling methods introduce label noise, degrading model generalization. To leverage unlabeled speech data, we propose a framework that automatically generates and refines pseudo speaker labels. It first generates pseudo-labels using a clustering algorithm, then trains a speaker verification system to boost the quality of pseudo-labeled data using self-supervised learning and a neural embedding extractor optimized with refined loss function. This function integrates a dynamic and adaptive label noise cleansing method, termed AdaptiveDropSC, which tracks dominant sub-centers via a dictionary table for better label correction. Experiments on VoxCeleb corpus show that our method improves pseudo-labeling accuracy across different clustering techniques, achieving state-of-the-art performance in self-supervised speaker verification.

Index Terms: speaker verification, label noise robustness, refined loss function, sub-centers, i-vector, self-supervised speaker verification

1. Introduction

Speaker verification (SV) as one of the most convenient means of biometric recognition [1], uses the voiceprint of a speaker to verify his/her identity. Based on known utterances of a speaker, the speaker verification (SV) task aims to identify whether a speaker is a legitimate user or an imposter. However, supervised SV approaches rely on large-scale labeled datasets, which are often expensive and labor-intensive to create. With the advent of big data, recently researchers in the domain of SV start to explore more affordable self-supervised learning (SSL) techniques using large noisy datasets. Indeed, since well-annotated datasets can be expensive to prepare, large-scale datasets are typically collected from the internet within automatic pipelines [2, 3]. Both manual and automatic speaker labeling methods, especially clustering-based approaches, are prone to errors, resulting in potential label noise. This noise can mislead the learning process and significantly impact the model’s generalization performance due to the memorization effects of deep models [4]. Therefore, having a reliable selection of pseudo-labels (PLs) [5] and an effective mechanism to curate these noisy annotations becomes even more crucial.

Many cleansing methods exist in the label noise and data pruning literature [6, 7, 8, 9]. They rely in general on one or multiple stages of label noise cleansing followed by a supervised training stage that uses the curated training set. Besides, iterative approaches to mitigate label noise by refining PLs such as [10, 11] can be intimidating and cumbersome as they require

more memory and computations. To build robust models, most of these methods only adopt one of the correction and filtering modes [12] or alternate between them.

In this work, we introduce a comprehensive framework for automatically generating pseudo speaker labels and refining them to mitigate label noise and enhance dataset quality. To achieve this, the proposed method trains a speaker verification system that cleans and improves pseudo-labeled data using self-supervised learning (SSL) and a neural embedding extractor optimized with a refined loss function. This loss function incorporates a dynamic and adaptive label noise cleansing technique, called AdaptiveDropSC (AdaptiveDrop [13] employing Sub-Centering mechanism), which tracks dominant sub-centers via a dictionary table for more accurate label correction. AdaptiveDropSC integrates both correction and filtering mechanisms within a single framework, allowing them to complement each other throughout training. This synergy enhances robustness against noisy labels, further improving model performance and generalization. Our approach is executed in a single-stage, end-to-end training process, eliminating the need to retrain the model from scratch on the final cleansed dataset, as required by most other methods [14, 15]. Since mislabeled samples are uncertain, an adaptive strategy is crucial. Our method dynamically adjusts filtering throughout training, recovering wrongly dropped samples in later epochs and removing inaccurate labels mistakenly included.

Our contributions in this work include:

- We introduce a general framework for the automatic generation and curation of pseudo speaker labels from unlabeled datasets to enhance the quality of pseudo-labeled data.
- We propose AdaptiveDropSC, a novel, general-purpose label noise filtering and correction method with sub-centers. It functions as a plug-and-play module compatible with any loss function, producing a refined loss for improved training.
- We conduct an extensive self-supervised SV experiments on the VoxCeleb dataset to demonstrate the effectiveness and robustness of our method under various real-world noisy labels and loss functions.

2. Related Works

A common approach to leveraging large unlabeled datasets for SV is clustering-based pseudo-labeling [16, 17, 18] or SSL-based objectives (SimCLR, MoCo) [19], followed by training the speaker embedding network discriminatively [11, 10]. However, clustering accuracy remains a bottleneck, limiting SV performance [10, 20]. Iterative clustering-classification methods [10, 21] aim to refine PLs and improve SV performance but still risk error propagation.

To address label noise, noise-robust algorithms [22, 23] can learn from noisy labels, while label-cleansing methods [9, 24]

remove or correct mislabeled samples. For example, Subcenter-ArcFace [14] improves robustness by forming dominant subclasses of clean samples, reducing intra-class constraints and mitigating label noise. In self-supervised speaker recognition, [15] proposed an audio-visual two-step label noise detection and filtering method by separating data into easy and peculiar (hard or noisy) categories. Similarly, [10] introduced a two-stage iterative loss-gated learning strategy, where embeddings are clustered to generate PLs, and high-loss samples are discarded.

3. Our Proposed Framework

Traditional supervised speaker verification (SV) approaches rely on large-scale labeled datasets, which are often expensive and labor-intensive to create. Self-supervised learning (SSL) has emerged as a powerful paradigm for speaker verification due to its ability to learn speaker representations from unlabeled data, eliminating the need for costly and time-consuming manual annotations. Moreover, these datasets, whether annotated manually or automatically, may contain label noise, which limits the model’s performance on unseen speakers and real-world scenarios. In this section, we present a framework (Figure 1) that automatically generates pseudo speaker labels and refines them to enhance dataset quality. By training a speaker verification system, this framework cleans and improves pseudo-labeled data using self-supervised learning (SSL) and a neural embedding extractor optimized with a refined loss function. This approach minimizes the need for extensive human annotations while boosting data quality and verification performance. The proposed framework consists of two main modules: the pseudo-label generation module, as shown in Figure 1a, and the speaker embedding extractor module, as illustrated in Figure 1b.

3.1. Pseudo Label Generation Module

To generate pseudo speaker labels, utterance-level representations are first extracted from the unlabeled dataset using unsupervised or self-supervised learning techniques. These embeddings are then input into a suitable clustering algorithm [16, 17, 18], which groups similar representations to assign pseudo speaker labels.

One approach to extracting embeddings is to use an i-vector extractor [25], an unsupervised method that generates fixed-dimensional speaker representations (i-vectors) from variable-length recordings. Alternatively, embeddings can be extracted by training a deep embedding extractor optimized with a self-supervised learning (SSL) objective function (e.g., SimCLR, DINO, MoCo) [19].

In the pseudo-label generation module, we adopt an i-vector extractor to extract 400-dimensional i-vectors from the unlabeled dataset. We then apply the CAMSAT algorithm [26] to generate pseudo-labels, using a predefined number of clusters from 5000, 5994, 10000. For performance comparison, we also generate pseudo-labels using other clustering algorithms, including K-means [27], Gaussian Mixture Model (GMM), Agglomerative Hierarchical Clustering (AHC) [28], CURE [29], IMSAT [30], and CIMC (Contrastive Information Maximization Clustering) [31].

3.2. Speaker Embedding Extractor Module

Given a dataset of pseudo-labeled or true-labeled training speakers, a conventional speaker embedding extractor aims to learn compact and discriminative utterance-level representa-

tions of speaker voices. It does this by minimizing intra-speaker variability while maximizing inter-speaker separation. Both manual and automatic speaker labeling methods are prone to errors, leading to potential label noise, particularly in clustering-based automatic labeling. Label noise arises from human mistakes or inaccuracies in automatic annotation, introducing incorrect labels into the dataset. This noise can degrade model performance by misleading the learning process and increasing the risk of overfitting, as the model may learn to memorize incorrect labels rather than generalizing effectively. Therefore, measures should be taken to mitigate the impact of label noise on the performance of speaker verification systems and to enhance the quality of the dataset. In our proposed framework, we incorporate a refined objective function into the speaker embedding extractor module, as depicted in Figure 1b, to enhance representation learning, improve speaker discrimination, and mitigate the effects of noisy labels. This refinement helps the model learn more robust and generalizable embeddings, ultimately boosting the performance of speaker verification in self-supervised settings. As illustrated in Figure 1b, acoustic features are first processed by the Encoder network, which extracts discriminative frame-level descriptors. The Attentive Statistical Pooling (ASP) layer then aggregates these frame-level features into a fixed-length utterance-level representation, emphasizing the most informative frames. Unlike simple averaging, ASP computes a weighted mean and standard deviation using learned attention scores, ensuring that speaker-relevant frames contribute more to the final representation. The weighted statistics are then concatenated to form the utterance-level embedding, which is subsequently projected into low-dimensional speaker embeddings through fully connected (FC) layers. These embeddings are then processed by the output layer, which employs a refined margin-based loss function to mitigate label noise and enhance the quality of the generated dataset. Except for the pseudo-label generation step, which relies on an appropriate clustering algorithm, the self-supervised speaker verification follows a similar experimental setup to the supervised speaker verification. Hence, the framework presented in Figure 1b can also be employed for supervised SV task to make it robust against label noise and to improve the quality of the manually annotated dataset.

3.2.1. Refined Loss function

In this section, we introduce our refined loss function, which integrates the AdaptiveDrop framework [13] with sub-center (denoted here as AdaptiveDropSC) into a margin-based loss function to mitigate label noise and correct mislabeled data, enhancing both training generalization and label accuracy. While we primarily integrate AdaptiveDrop into the widely used AAMSoftmax, a.k.a ArcFace [32], and BoundaryFace [33] loss functions for training, we also evaluate other margin-based losses, including AdaFace [34], AMSoftmax [35], and CosFace [36], for performance comparison.

AdaptiveDrop framework [13] utilizes Cosine Similarity (CS) as a proxy label confidence indicator to detect the mislabeled instances and set a constant CS threshold τ for label filtering to drop high-confident noisy data. Given each training sample x and its label y , the weights of currently trained embedding network are used to extract embedding ω of x on the fly at each training step. Then cosine similarity $CS(x, y) = \cos(\omega, \omega_y) = \frac{\omega \cdot \omega_y}{\|\omega\| \|\omega_y\|}$ between ω and the current learnt class center/prototype embedding ω_y corresponding to class y is computed. If $CS(x, y) \geq \tau$, sample x is retained

Algorithm 1 AdaptiveDropSC algorithm (i.e., AdaptiveDrop with sub-centers)

Inputs: Training data X , noisy pseudo-labels Y , model M , loss function l , cosine similarity threshold τ , epoch to start correcting noisy labels (ESC), epoch to start dropping out noisy labels (ESD), epoch to start tracking dominant sub-centers (ESTD), dominant sub-centers table $DSCT \in \mathbb{N}^{C \times K}$.

Output: The best model M .

Initialize: Table $DSCT$ with zeros

for epoch = 1 **to** max_epochs **do**

 Extract embeddings ω for samples of current batch x .

 Compute cosine similarities (CS) between each sample in x and its corresponding class sub-centers $\omega_{y,j}$ for $j \in \{1, \dots, K\}$.

if epoch \geq ESTD **then**

 #Update dominant sub-centers table (DSCT)

 closest_subcenter = $\arg \max_{j \in \{1, \dots, K\}} CS [y, j]$

$DSCT [y, \text{closest_subcenter}] += 1$

end if

if epoch \geq ESC **then**

 Correct pseudo-labels y of batch x .

end if

if epoch \geq ESD **then**

 #Compute dominant sub-centers

$K_{\text{dominant}} = \arg \max_{i \in \{1, \dots, K\}} DSCT [y, i]$

 Filter samples based on $CS(x, y) = \cos(\omega, \omega_{y, K_{\text{dominant}}})$.

 Set $(x_{\text{clean}}, y_{\text{clean}}) = \{(x_i, y_i) \text{ for } x_i, y_i \in (x, y) \text{ if } CS > \tau\}$

else

 Set $x_{\text{clean}} = x$ and $y_{\text{clean}} = y$

end if

 Compute margin-based loss (e.g., AAMSoftmax) with filtered x_{clean} and y_{clean} .

 Perform gradient descent to update network parameters of model M .

end for

in the training batch, otherwise it is dropped out from the batch.

Algorithm 1 shows the implementation of AdaptiveDropSC framework (i.e., AdaptiveDrop using sub-centers) that employs a sub-centers dictionary table $DSCT \in \mathbb{N}^{C \times K}$ (initially all entries are set to zero) to adaptively track the dominant ones per class from the beginning of training and all samples are reconsidered at each epoch. As a rule, the ongoing sub-centers with the highest value per class are considered dominant (see Algorithm 1). At each training step, the entry corresponding to the class sub-center assigned to each sample is increased by 1 (except a first short warmup stage of few epochs where the dictionary table is not updated until model’s predictions are reliable enough). As training gradually progresses, dominant sub-centers become clearer and the used cosine similarities become more reliable to drop out samples that are not close enough to their dominant class sub-centers.

Indeed, applying filtering after label correction helps improve correction by assuring that the corrected labels also have a high cosine similarity which is a second constraint to ensure higher quality of the corrected labels. Finally, ArcFace/BoundaryFace loss is computed using the filtered data and labels. Once training is done, embeddings are extracted from the fully connected layer closest to the pooling layer. Since cosine similarity outperforms PLDA on VoxCeleb test sets when using a margin-based loss function [37], it is used for verification scoring in this work.

4. Experimental setup

4.1. Experimental setup

We conducted a set of experiments based on the VoxCeleb dataset [3]. 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 (Vox1-O) trials lists [2]. We report results in terms of Equal Error Rates (EER) evaluation metric. We employ ECAPA-

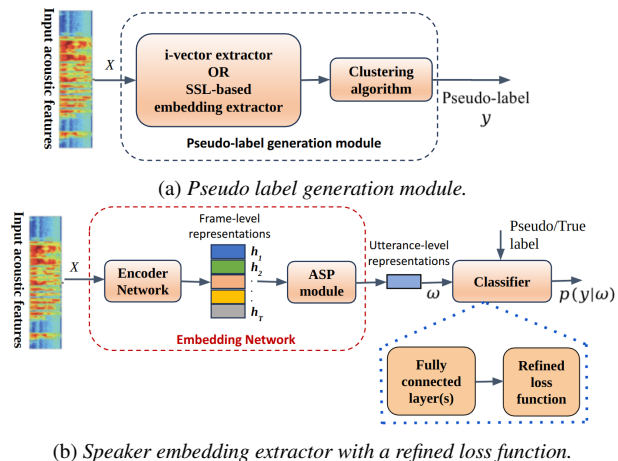


Figure 1: Schematic diagram showing different steps for the self-supervised speaker verification employing a refined loss function for label-correction.

TDNN [38] as our speaker embedding network. As acoustic features for our SV experiments, we used 40-dimensional Mel-frequency cepstral coefficients (MFCCs) extracted at every 10 ms, using a 25 ms Hamming window via Kaldi toolkit [39]. Moreover, we have used waveform-level data augmentations including additive noise and room impulse response (RIR) simulation [40] to follow other SV works. Besides, we have applied augmentation over the extracted MFCCs features, analogous to the specaugment scheme [41]. All SV experiments have been run for 150 epochs using a single A40 GPU (around 2 days training), with a batch size of 200 MFCC samples. Scale factor $s = 30$ and margin $m = 0.2$ were used across all margin-based losses. Besides, we do not use score normalization and CS was used as a backend for verification scoring between enrollment and test embeddings. We refer to our loss implementation using label correction with BoundaryFace and refer to sub-centers combined with label correction with subcenter-BoundaryFace.

Finally, we use a default $K = 3$ sub-centers when sub-centers are employed and unless specified otherwise, a default CS threshold of $\tau = 0.423$. For AdaptiveDropSC, we use $ESD = 5$, $ESC = 7$, $ESTD = 3$. Besides, to avoid training instability especially for highly noisy thresholds, we only drop out a maximum of 50% of samples from each training batch. We use the CAMSAT clustering algorithm [26] to generate PLs using predefined numbers of clusters (C) in $\{5000, 5994, 10000\}$.

4.2. Results and Discussion

Table 1 presents speaker verification results in terms of EER (%) and Relative Improvement (RI) achieved using various loss functions, with and without AdaptiveDropSC. The results show that integrating AdaptiveDropSC consistently reduces EER across all loss functions, providing significant relative improvement compared to systems without it. This improvement is attributed to AdaptiveDropSC’s ability to dynamically filter and correct mislabeled samples during training, preventing erroneous labels from degrading model performance. Additionally, the significant RI achieved across different loss functions indicates that AdaptiveDropSC is adaptable and complementary to a wide range of objective functions, making it a versatile plug-and-play module for improving speaker verification systems.

Table 2 summarizes the improvements in label accuracy, where unsupervised clustering accuracy (ACC) measures the consistency between the ground-truth labels and the initial clustering-driven pseudo labels (PLs), as well as the final rectified PLs after applying label noise correction with refined BoundaryFace loss functions—specifically when BoundaryFace is integrated with AdaptiveDropSC and AdaptiveDrop. The results show that improvements in label accuracy depend on the type of PLs used, with CAMSAT consistently achieving the highest clustering accuracies across all predefined configurations, demonstrating its effectiveness. In most cases, AdaptiveDropSC provided better final accuracy compared to AdaptiveDrop. Notably, this demonstrates that label correction can significantly enhance the performance of clustering-driven pseudo labels (PLs), highlighting its effectiveness in improving model accuracy.

Finally, Table 3 presents a comparison of our AdaptiveDropSC approach, using the best-performing configuration with Subcenter-ArcFace and CAMSAT-based pseudo labels (PLs) (refer to Table 1), against recent state-of-the-art self-supervised speaker verification (SV) methods that employ various SSL objectives, all utilizing the same ECAPA-TDNN model encoder on the Vox1-O test set. The results clearly demonstrate that our proposed adaptive filtering strategy yields significant performance improvements over all considered baselines and SOTA methods.

5. Conclusion

Noisy annotations can impair dramatically the generalization of deep learning-based speaker verification (SV) models. In this work, we proposed a comprehensive framework for self-supervised speaker verification that tackles the challenges associated with large-scale automatically and manually labeled datasets as well as label noise. Proposed approach generated pseudo speaker labels from a large-scale unlabeled dataset using CAMSAT clustering algorithm. The framework then trained a SV system to clean and improve pseudo-labeled data using self-supervised learning and a ECAPA-TDNN neural embedding extractor optimized with a refined loss function. This function integrates a dynamic and adaptive label noise cleansing method, denoted as AdaptiveDropSC (i.e., AdaptiveDrop with Sub-Centers), which tracks dominant sub-centers via a dictionary table for improved label correction. Through extensive experiments, we showed that our method consistently outperforms other loss functions, achieving significant improvements in self-supervised speaker verification across various clustering-driven pseudo-labels. Additionally, it complements several label noise-robust techniques and can be combined with various loss objectives to further enhance performance. While designed for self-supervised SV, the framework can also be applied to supervised SV tasks to enhance robustness against label noise and improve the quality of manually annotated datasets.

6. Acknowledgment

The authors wish to acknowledge the funding from the Natural Sciences and Engineering Research Council of Canada through grant RGPIN-2019-05381.

7. References

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

Table 1: *Self-supervised SV performance obtained with different loss functions with and without AdaptiveDropSC, with the relative improvement (RI). Results are reported in terms of EER (%) on the Vox1-O test set using CAMSAT-based PLs.*

Loss function	Without	With	RI (%)
BoundaryFace [33]	2.752	2.513	8.7 ↑
CosFace [36]	2.863	2.577	10.0 ↑
MV-Arc-Softmax [42]	2.884	2.519	12.7 ↑
Subcenter-ArcFace [14]	2.943	2.407	18.2 ↑
AMSoftmax [35]	2.959	2.609	11.8 ↑
Subcenter-BoundaryFace	2.959	2.672	9.7 ↑
OCSOftmax [43]	2.969	2.678	9.8 ↑
AdaFace [34]	3.059	2.688	12.1 ↑
ArcFace [32]	3.134	2.641	15.7 ↑
CurricularFace [44]	3.192	2.529	20.8 ↑

Table 2: *Final label accuracy of the rectified PLs, generated by our system after applying label noise correction using refined BoundaryFace loss functions (i.e., BoundaryFace integrated with AdaptiveDropSC & BoundaryFace integrated with AdaptiveDrop). We compare them to the initial accuracy of the PLs. Results are reported across PLs of different noise levels from various clustering algorithms.*

Pseudo-labels	No. of clusters	Initial label accuracy (%)	Final label accuracy (%)	
			AdaptiveDrop	AdaptiveDropSC
CURE [29]	5,000	15.1	29.2	29.4
KMeans [27]	5,000	30.2	31.5	31.8
GMM	5,000	45.0	54.0	54.0
AHC [28]	5,000	60.2	67.3	67.2
IMSAT [30]	5,000	57.8	60.9	61.0
	5,994	60.0	64.3	64.4
	10,000	63.9	69.5	69.7
CIMC [31]	5,000	60.2	63.7	64.2
	5,994	61.9	66.5	66.1
	10,000	63.9	69.5	69.7
CAMSAT [26]	5,000	65.5	66.8	67.3
	5,994	66.9	68.8	69.2
	10,000	70.9	73.4	73.3
True labels	5,994	100.0	100.0	100.0

Table 3: *Some recent SOTA self-supervised SV approaches compared to AdaptiveDropSC when integrated Subcenter-ArcFace loss function in EER (%) on the Vox1-O test set.*

SSL Objective	EER (%)
MoBY [19]	8.2
InfoNCE [10]	7.36
MoCo [45]	7.3
ProtoNCE [19]	7.21
PCL [19]	7.11
CA-DINO [46]	3.585
i-mix [47]	3.478
l-mix [47]	3.377
Iterative clustering [10]	3.09
CAMSAT [26]	3.065
Subcenter-BoundaryFace [48]	2.752
AdaptiveDropSC + Subcenter-ArcFace (ours)	2.407

- [2] A. Nagrani, J. S. Chung *et al.*, “Voxceleb: a large-scale speaker identification dataset,” in *INTERSPEECH*, 2017.
- [3] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” in *INTERSPEECH*, 2018.
- [4] D. Arpit, S. Jastrzebski, 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.
- [5] M. N. Rizve *et al.*, “In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning,” *arXiv preprint arXiv:2101.06329*, 2021.
- [6] B. Fréney, A. Kabán *et al.*, “A comprehensive introduction to label noise,” in *ESANN*. Citeseer, 2014.
- [7] I. Guyon, N. Matic, V. Vapnik *et al.*, “Discovering informative patterns and data cleaning,” 1996.
- [8] J.-w. Sun, F.-y. Zhao, C.-j. Wang, and S.-f. Chen, “Identifying and correcting mislabeled training instances,” in *Future generation communication and networking (FGCN 2007)*, vol. 1. IEEE, 2007, pp. 244–250.
- [9] C. E. Brodley and M. A. Friedl, “Identifying mislabeled training data,” *Journal of artificial intelligence research*, 1999.
- [10] R. Tao *et al.*, “Self-supervised speaker recognition with loss-gated learning,” in *ICASSP*. IEEE, 2022.
- [11] J. Peng *et al.*, “Progressive Contrastive Learning for Self-Supervised Text-Independent Speaker Verification,” in *Proc. of Odyssey Workshop*, 2022.
- [12] G. Jiang *et al.*, “Which is more effective in label noise cleaning, correction or filtering?” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 11, 2024, pp. 12 866–12 873.
- [13] X. Z. Abderrahim Fathan and J. Alam, “Adaptivedrop: A simple adaptive label noise filtering scheme for enhanced self-supervised speaker verification,” in *IEEE ICASSP*. IEEE, 2025.
- [14] 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.
- [15] R. Tao, K. A. Lee, Z. Shi, and H. Li, “Speaker recognition with two-step multi-modal deep cleansing,” in *IEEE ICASSP*. IEEE, 2023, pp. 1–5.
- [16] 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
- [17] 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*, vol. 16, no. 6, pp. 1263–1272, 2022.
- [18] 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.
- [19] W. Xia *et al.*, “Self-supervised text-independent speaker verification using prototypical momentum contrastive learning,” in *ICASSP*. IEEE, 2021.
- [20] B. Han, Z. Chen, and Y. Qian, “Self-supervised speaker verification using dynamic loss-gate and label correction,” *arXiv preprint arXiv:2208.01928*, 2022.
- [21] Y. Li *et al.*, “Contrastive clustering,” in *AAAI*, 2021.
- [22] M. Guan *et al.*, “Who said what: Modeling individual labelers improves classification,” in *Proc. of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
- [23] D. Rolnick, A. Veit *et al.*, “Deep learning is robust to massive label noise,” *ICLR*, 2018.
- [24] D. T. Nguyen *et al.*, “Self: Learning to filter noisy labels with self-ensembling,” *arXiv preprint arXiv:1910.01842*, 2019.
- [25] N. Dehak *et al.*, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, 2011.
- [26] 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.
- [27] J. A. Hartigan and M. A. Wong, “A k-means clustering algorithm,” *JSTOR: Applied Statistics*, vol. 28, no. 1, 1979.
- [28] W. H. E. Day *et al.*, “Efficient algorithms for agglomerative hierarchical clustering methods,” *Journal of Classification*, vol. 1, pp. 7–24, 1984.
- [29] S. Guha *et al.*, “Cure: An efficient clustering algorithm for large databases,” *SIGMOD Rec.*, vol. 27, no. 2, jun 1998. [Online]. Available: <https://doi.org/10.1145/276305.276312>
- [30] W. Hu *et al.*, “Learning discrete representations via information maximizing self-augmented training,” in *International conference on machine learning*. PMLR, 2017, pp. 1558–1567.
- [31] A. Fathan and J. Alam, “Contrastive information maximization clustering for self-supervised speaker recognition,” in *IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 2024, pp. 383–388.
- [32] J. Deng *et al.*, “Arcface: Additive angular margin loss for deep face recognition,” *IEEE TPAMI*, 2021.
- [33] 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.
- [34] 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.
- [35] F. Wang *et al.*, “Additive margin softmax for face verification,” *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.
- [36] 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.
- [37] D. Garcia-Romero, A. McCree, D. Snyder, and G. Sell, “Jhu-hltcoe system for the voxsrc speaker recognition challenge,” in *IEEE ICASSP*, 2020, pp. 7559–7563.
- [38] B. Desplanques *et al.*, “ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” in *Interspeech 2020*. ISCA.
- [39] D. Povey *et al.*, “The kaldi speech recognition toolkit,” in *In IEEE 2011 workshop*, 2011.
- [40] D. Snyder *et al.*, “X-vectors: Robust dnn embeddings for speaker recognition,” in *Proc. of IEEE ICASSP*, 2018, pp. 5329–5333.
- [41] D. S. Park *et al.*, “Specaugment: A simple data augmentation method for automatic speech recognition,” in *Interspeech*, 2019, pp. 2613–2617.
- [42] 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.
- [43] Y. Zhang *et al.*, “One-class learning towards synthetic voice spoofing detection,” *IEEE Signal Processing Letters*, 2021.
- [44] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, “Curricularface: adaptive curriculum learning loss for deep face recognition,” in *proceedings of the IEEE/CVF conference on CVPR*, 2020, pp. 5901–5910.
- [45] J. Cho *et al.*, “The jhu submission to voxsrc-21: Track 3,” *arXiv preprint arXiv:2109.13425*, 2021.
- [46] B. Han *et al.*, “Self-supervised learning with cluster-aware-dino for high-performance robust speaker verification,” *arXiv preprint arXiv:2304.05754*, 2023.
- [47] 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.
- [48] A. Fathan, X. Zhu, and J. Alam, “An investigative study of the effect of several regularization techniques on label noise robustness of self-supervised speaker verification systems,” 2024.