2024 IEEE Conference on Artificial Intelligence (CAI)

Contrastive Information Maximization Clustering for Self-Supervised Speaker Recognition

Abderrahim Fathan
Computer Research Institute of Montreal (CRIM)
Montreal, Canada
abderrahim.fathan@crim.ca

Jahangir Alam
Computer Research Institute of Montreal (CRIM)
Montreal, Canada
jahangir.alam@crim.ca

***Abstract*—Pseudo-labels (PLs) generated through clustering are extensively employed to optimize speaker embedding (SE) networks and to train self-supervised speaker verification (SV) systems. However, the effectiveness of PL-based self-supervised training is contingent on the quality of the PLs, and achieving high clustering performance often requires time-consuming and resource-intensive data augmentation regularization. In this paper, we introduce an efficient, general-purpose multi-objective clustering algorithm that outperforms all other baseline methods for clustering SEs. Our approach, named Contrastive Information Maximization Clustering (CIMC), circumvents the need for explicit data augmentation, enabling rapid training with minimal memory and computational resource usage. CIMC is founded on three key principles: (1) Self-Augmented Training, which ensures representation invariance and maximizes the information-theoretic dependency between samples and their predicted PLs (2) Virtual Mixup Training, which enforces local-Lipschitzness and upholds the cluster assumption (3) Supervised contrastive learning, which fosters the learning of more discriminative features and enhances robustness to natural corruptions by bringing together samples of the same class while separating those of different clusters. We present a comprehensive comparative analysis of our clustering method against baselines using various clustering metrics, conduct an ablation study to assess the contribution of each component, and demonstrate that our multi-objective approach provides beneficial complementary information. Furthermore, utilizing the generated PLs to train our SE system enables us to achieve high SV performance.**

***Index Terms*—Speaker verification, clustering, self-supervised speaker verification, pseudo-labels, speaker recognition**

I. Introduction

Speaker Verification (SV) involves confirming a speaker's identity based on their known utterances. In recent years, it has become a crucial technology for authenticating individuals across various applications [1]. Typically, fixed-dimensional embeddings are extracted at the utterance level from both enrollment and test speech samples. These embeddings are then fed into a scoring algorithm, such as cosine distance, to measure their similarity and determine the likelihood that they originate from the same speaker.

Traditionally, the i-vector paradigm has been one of the most prominent approaches for speaker embedding [2], [3], owing to its capacity to capture the distributive patterns of the speech in an unsupervised fashion, even with a relatively small amount of training data. This framework generates fixed-sized compact vectors (i-vectors) that encapsulate the speaker's identity in a speech utterance, regardless of its duration. Moreover, in recent years, a plethora of deep learning-based architectures and techniques have been proposed to extract embeddings [4]–[6]. These approaches have demonstrated remarkable performance when a large amount of training data from a sufficient number of speakers is available [7]. A widely adopted architecture for this purpose is ECAPA-TDNN [8], renowned for achieving state-of-the-art (SOTA) performance in text-independent speaker recognition. The ECAPA-TDNN incorporates squeeze-and-excitation (SE), utilizes channel- and context-dependent statistics pooling, multi-layer aggregation and employs self-attention pooling to obtain an utterance-level embedding.

Indeed, the majority of deep embedding models are trained under full supervision, necessitating large speaker-labeled datasets for effective training. However, creating well-annotated datasets can be a costly and time-intensive endeavor, prompting the research community to explore more affordable self-supervised learning (SSL) techniques utilizing extensive unlabeled datasets. A typical approach for addressing this issue for SV systems is to employ a one-stage "clustering-classification" scheme [5], [6], [9] by utilizing clustering algorithms (e.g., K-means, agglomerative hierarchical clustering, spectral clustering) or other self-supervised objectives to produce useful Pseudo-Labels (PLs) such as SimCLR or MoCo [10]. Subsequently, the speaker embedding network is trained using these labels in a discriminative fashion. More recently, more effective methods have emerged and gained widespread adoption in the SV domain. These frameworks utilize an iterative two-stage "clustering-classification" progressive learning process [11], [12]. Initially, SSL training (such as the InfoNCE [12] contrastive loss) is employed to train an encoder model for generating speaker embeddings. Subsequently, in the second stage, the embeddings undergo clustering to produce pseudo-labels, facilitating joint supervised training of the encoder with a classifier. This sequential process continues until further improvements are negligible.

Although these PL-based Self-Supervised SV schemes exhibit striking performance, the efficacy of clustering continues to impede all aforementioned approaches [12], [13], primarily because downstream performance heavily relies on precise PLs. However, these PLs are generally noisy and inaccurate due to the mismatch between the clustering objective and the final SV task. Additionally, despite the advantages of the iterative clustering-classification framework, the persistence of

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00077

erroneous information from incorrect PLs degrades the final downstream task performance [12], [14]. Hence, there is a demand for improved clustering algorithms to produce Pseudo-Labels (PLs) that are less noisy and more precise. Instead of relying on SOTA deep clustering models, which often require extensive domain-specific data augmentations, these methods typically utilize classical clustering algorithms such as spectral clustering or K-means. These classical algorithms are preferred due to their ease of use, faster computation, and lower resource requirements in terms of memory and GPU/CPU utilization during training.

In this paper, we introduce an efficient and general-purpose multi-objective clustering algorithm, denoted as Contrastive Information Maximization Clustering (CIMC), that outperforms all other baselines employed for clustering speaker embeddings. Our approach eliminates the need for explicit data augmentation, ensuring swift training and utilization of low memory and compute resources. The proposed approach is built upon the combination of three principles: (1) Self-Augmented Training, which enforces representation invariance and maximizes the information-theoretic dependency between samples and their predicted pseudo-labels, facilitated through the Information Maximizing Self-Augmented Training (IMSAT) clustering framework [15] (2) Virtual Mixup Training (VMT) [16], which imposes local-Lipschitzness, thereby reinforcing the cluster assumption (3) Supervised contrastive learning [17], leveraging dynamically generated pseudo-labels, to pull samples of same class closer and push samples of different clusters apart.

Rather than mixing up of inputs or using constrastive loss solely to enforce smoother model responses and compactness of embeddings, the CIMC approach effectively utilizes these predictions as supplementary supervisory signals to enhance the guidance for cluster assignment, resulting in more resilient, stable, and high-performing data clustering. The resulting algorithm shows exceptional scalability, speed, and increased robustness to data corruptions and shifts compared to IMSAT during online clustering. It is simple to implement, and adds limited computational overhead to IMSAT.

We believe the proposed CIMC clustering method can significantly enhance the optimization of current self-supervised SV frameworks by replacing the conventional clustering methods currently in use, such as k-means and spectral clustering. Moreover, proposed method holds promise for improving speaker diarization performance, where clustering is a critical module. The CIMC clustering approach is versatile and can be effectively applied to a wide range of problems and domains beyond speech or speaker verification. The contributions of this paper are as follows:

- We propose CIMC, a novel general-purpose multi-objective clustering algorithm designed for large-scale datasets and scenarios involving a high number of clusters.
- We explore several recent state-of-the-art SSL objectives for clustering, demonstrating that multi-objective clustering frequently offers valuable complementary information.
- Our proposed approach outperformed numerous clustering baselines. Furthermore, by using the generated pseudo-labels to train our SV systems, we achieved high SV performance.

II. Background and Related Work

Various designed clustering approaches have been introduced. Classical approaches include models such as K-means [18], Gaussian mixture model (GMM), BIRCH [19], CURE [20], Agglomerative Hierarchical Clustering (AHC) [21], etc. However, these methods are limited to fitting linear boundaries between data representations. Recently, the robust representational capabilities of neural networks have been employed to better model the non-linearity of complex data distributions and to scale to large datasets. As an example, Deep Embedded Clustering (DEC) [22] employs deep models to learn feature representations and cluster assignments concurrently. On the other hand, DeepCWRN [23] uses an autoencoder to learn feature representations and embeddings tailored for clustering by promoting the separation of inherent clusters within the embedding space. Additionally, other deep models have been developed based on generative models [24], [25] or dynamic architectures [26].

Although data augmentation is essential for regularizing deep neural networks in clustering and unsupervised representation learning to capture the invariance of learned representations, it also increases the size of the training set, leading to significantly longer training times, particularly for large-scale datasets and neural networks. Additionally, using blind augmentations can negatively impact speaker verification and recognition tasks, as transformations like pitch perturbation or spectral augmentation can alter a speaker's identity, sometimes creating misleading data samples. Moreover, for real-world tabular data applications [27], such as genomics and clinical data, generating additional augmented views is not straightforward and can be impractical.

III. Our proposed clustering approach

A schematic diagram of the proposed Contrastive Information Maximization Clustering (CIMC) is presented in Figure 1, which is trained via minimizing the total loss L_{total}, that integrates three different loss functions. Given a deep neural network-based clustering model f and a predefined number of clusters C, the CIMC approach constrains the predictions of the model to remain unchanged under local perturbations and implicit Virtual Mixup Training (VMT) [16] data augmentations L_{Mixup}. The model is trained in an end-to-end fashion by imposing local-Lipschitzness on the learned weights to favor the cluster assumption [28] (if samples are in the same cluster, they come from the same class), which is a critical condition for successful clustering. More explicitly, it optimizes the following L_{total} objective:

$$L_{total} = L_{IMSAT} + L_{SupCon} + L_{Mixup} \tag{1}$$

where L_{IMSAT} is the loss function for the original IMSAT clustering objective and L_{SupCon} & L_{Mixup} represent the supervised contrastive loss and the mixup loss terms, respectively. Our main focus is to harness these objectives as

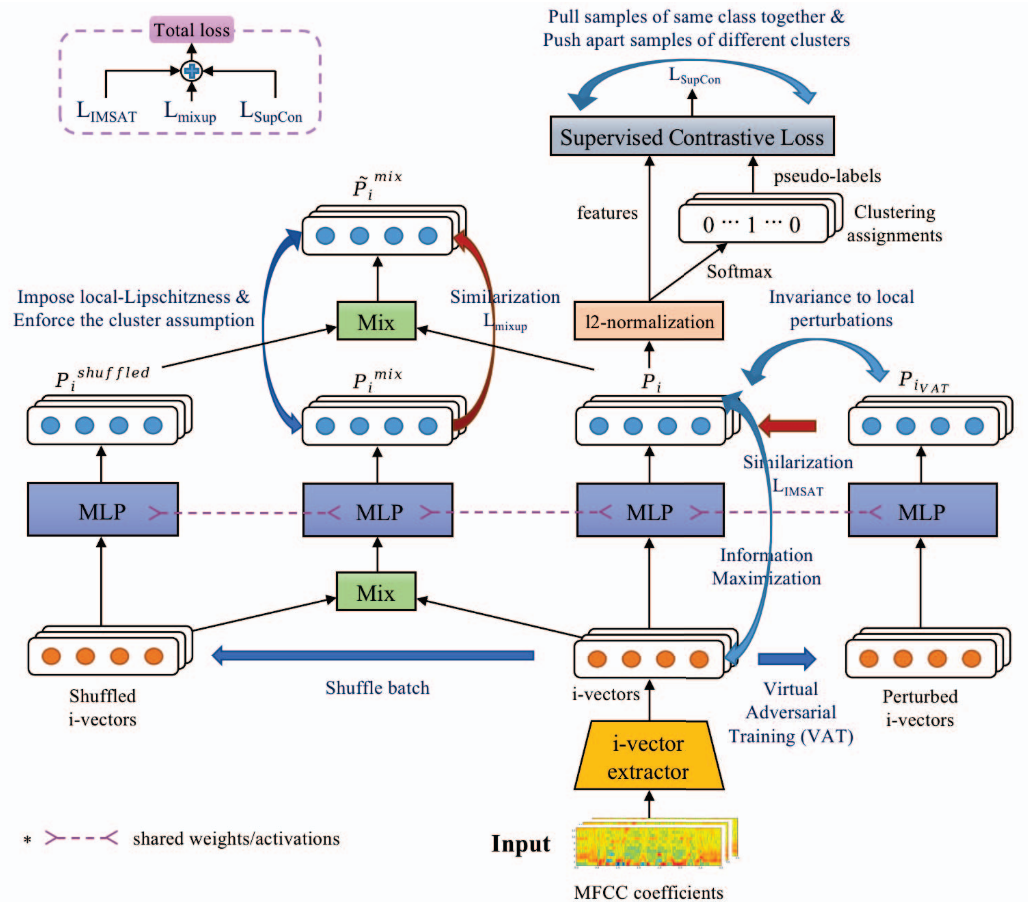

Fig. 1: The pipeline of our proposed Contrastive Information Maximization Clustering (CIMC) method depicting the data flow and the different losses employed for clustering.

additional supervisory signals to regularize the clustering model to produce consistent assignments.

Now, the original IMSAT loss L_{IMSAT} is expressed mathematically as [15]:

$$L_{IMSAT} = R_{SAT}(\theta, T_{VAT}) + \lambda(H(Y|X) - \mu H(Y)), \quad (2)$$

where $H(.)$ is the marginal entropy and $H(.|.)$ is conditional entropy. The loss term $R_{SAT}(\theta;T) = \frac{1}{N}\sum_{n=1}^{N} R_{SAT}(\theta;x_n,T(x_n))$ enables the representations of the augmented samples to be drawn closer to those of the original ones while simultaneously regularizing the network's complexity against local perturbations via Virtual Adversarial Training (VAT) [29]. $\lambda, \mu \in \mathbb{R}$ are hyper-parameters that tune the balance between the model's complexity regularization (via R_{SAT}) and the maximization of MI, and between the two entropy terms, respectively. $R_{SAT}(\theta;x,T(x)) = -\sum_{c=1}^{C}\sum_{y_c=0}^{1} p_{\hat{\theta}}(y_c|x) log p_{\theta}(y_c|T(x))$, where $p_{\hat{\theta}}(y_c|x)$ is the prediction of original data point x, and $\hat{\theta}$ are the current parameters of the neural network. $T_{VAT}(x) = x + r$ is the augmentation function using local perturbations to enforce invariance, where $r = \arg\max_{r'}\{R_{SAT}(\hat{\theta};x,x+r'); \|r'\|_2 \leq \epsilon\}$ constitutes an adversarial direction.

The difference between the marginal and conditional entropy represents the MI between sample X and its label Y that we maximize. The two entropy terms can be calculated as:

$$H(Y) = h(p_\theta(y)) = h(\frac{1}{N}\sum_{i=1}^{N} p_\theta(y|x)), \quad (3)$$

$$H(Y|X) = \frac{1}{N}\sum_{i=1}^{N} h(p_\theta(y|x_i)), \quad (4)$$

where $p_\theta(y|x)$ is our learned probabilistic classifier modeled by parameters θ of a deep network, and $h(p(y)) = -\sum_{y'} p(y') \log p(y')$ is the entropy function.

Essentially, increasing the entropy $H(Y)$ promotes uniform cluster sizes and prevents collapsing into a few clusters. Conversely, minimizing the conditional entropy $H(Y|X)$ results in less ambiguous cluster assignments and compels the classifier to be confident in our training samples [30]. For more details, please refer to [15], [29].

On the other hand, the supervised contrastive loss term L_{SupCon} helps to learn more discriminative features and has the advantage of improving robustness to natural corruptions and to out-of-distribution data [17]. Note that, the L_{SupCon} loss requires labels. But in this work, L_{SupCon} loss [17] is leveraged in an unsupervised (or self-supervised) manner employing the online generated pseudo-labels as labels and l2-normalized logits as feature embeddings. Therefore, the novelty of our usage is the use of online predictions of our clustering model as input labels, which allows us to use it in a completely unsupervised/self-supervised fashion without the need for ground-truth labels. As the performance of our clustering gradually improves, the online pseudo-labels are progressively more reliable, thus helping to generate better and more compact clusters.

The mixup loss term L_{Mixup} can be formulated as:

$$L_{Mixup} = \frac{1}{N}\sum_{i=1}^{N} KL(\alpha_i p_i + (1-\alpha_i)p_{r_i} || f(\alpha_i x_i + (1-\alpha_i)x_{r_i})). \tag{5}$$

where N is the mini-batch size of data, $r_i \in \{1, .., N\}$ is a random index, and $\alpha_i \in [0, 1]$ is the mixup interpolation coefficient. $KL(.||.)$ operator corresponds to the Kullback-Leibler divergence. $p_i = f(x_i) \in \mathbb{R}^{1xC}$, $p_{r_i} = f(x_{r_i})$ correspond to the predictions of data samples x_i and x_{r_i}, respectively.

Finally, inspired from VMT [16] regularization method which encourages the model to exhibit linearity between training points, this allows us to enforce representation smoothness during clustering and guarantee consistent predictions between the training data points and their neighboring samples. Indeed, mixup [31] which is a strategy to augment data by interpolating different data samples alongside their labels, often results in improved generalization to out-of-set samples. Mixup has also been found by [6] to enhance the generalization of self-supervised speaker verification systems when the clusters are not well distanced or not compact as it can dilute label noise and induce better class separation. In line with work [16], we opt for mixing logits instead of directly mixing probabilities in the L_{Mixup} loss. We find empirically that this step, followed by a softmax operation, enhances training effectiveness and guards against premature information loss during probability mixing. We follow the general framework in [9, Fig. 1] for training our clustering-driven self-supervised speaker embedding networks.

IV. Clustering Algorithms and Metrics

For all our clustering algorithms, we use 400-dimensional i-vectors as condensed input. These i-vectors, which serve as unsupervised representations of speakers, enable more efficient clustering and help mitigate the high dimensionality associated with MFCC acoustic features.

Additionally, to comprehensively evaluate and analyze the quality of the pseudo-labels (PLs) from multiple perspectives, we employ a set of seven supervised metrics based on both the PLs and the true labels: Unsupervised Clustering Accuracy, Normalized Mutual Information [33], Adjusted MI [34], Completeness score [35], Homogeneity score [35], Purity score, and Fowlkes-Mallows index [36].

The criteria assessed by these metrics include clustering accuracy and mutual information to evaluate the consistency between true labels and generated pseudo-labels (PLs), as well as the homogeneity, completeness, purity of clusters, and precision and recall. Additionally, we calculate three unsupervised metrics—Silhouette score [37], Calinski-Harabasz score [38], and Davies-Bouldin score [39]—based solely on the generated PLs and data samples. These metrics measure the compactness or scatter of clusters, such as intra-class dispersion, between-cluster distances, and nearest-cluster distance. We use implementations from the scikit-learn toolkit to compute these metrics. Further details and discussions can be found in the study [6], which identified a strong correlation between these metrics and speaker verification performance.

V. Results and Discussion

We assess the performance of the proposed clustering method and the resulting pseudo-labels (PLs) for self-supervised speaker verification through a series of experiments conducted on the VoxCeleb2 dataset [40]. We train the embedding networks on the development subset of VoxCeleb2, comprising 1.092 million utterances from 5,994 distinct speakers. Evaluation follows the VoxCeleb1 trials list [41], encompassing 37,720 trials with 4,874 utterances from 40 speakers. For our speaker verification (SV) system, we employ 40-dimensional Mel-frequency cepstral coefficients (MFCCs) as input features to our ECAPA-TDNN model. MFCCs are computed every 10 ms with a 25 ms Hamming window, using the Kaldi toolkit [42]. Additionally, we adopt the additive angular margin softmax (AAMSoftmax) objective [43] to enhance generalization during training of our self-supervised speaker embedding network. We set the scale factor (s) to 30 and the angular margin (m) to 0.1. Cosine similarity serves as the backend for scoring verification between embeddings of enrollment and test speech samples.

Similarly to the IMSAT setup, we adopt the MLP-based d-S-S-C architecture, with $d = 400$ and C representing the input and output dimensionality, respectively. The network has a width of $S = 20,800$ neurons. We apply RELU activation and batch normalization to all hidden layers and use softmax in the output layer. $\lambda = 0.5$ and $\mu = 3.5$. Besides, we utilize the Momentum algorithm for optimization where momentum is set to 0.9. The initial learning rate is 0.01 with an exponential rate decay of 0.996. We use a batch size of 10,240 i-vectors, normalizing the inputs independently along the sample axis to a unit l2-norm to preserve speaker information. We use $\alpha = 1$ as the coefficient of the Beta distribution used for mixup interpolation. We ran experiments for 150 epochs using 64 CPU cores for each clustering algorithm. Besides, all speaker verification experiments were run over 7 days on a single RTX2080Ti GPU, utilizing a batch size of 200 MFCC samples. All code and methods in our experiments are based on Tensorflow. Additionally, our clustering benchmarks from [6, Tab. 1] set,

TABLE I: Ablation experiments of the proposed CIMC clustering system, including various SSL-based loss objectives that do not employ data augmentation (only original data samples). C denotes the predefined number of clusters. The results are presented in terms of clustering metrics, along with the corresponding EER (%) downstream SV evaluation performance. Our examined SV system is trained from scratch by each time employing the generated clustering-based pseudo-labels.

Model	**Clustering Metrics**											**Speaker Verification**
	ACC	**AMI**	**NMI**	**No. of clusters**	**Completeness**	**Homogeneity**	**FMI**	**Purity**	**Silhouette**	**CHS**	**DBS**	**EER (%)**
L_{Mixup} (C: 10k)	0.013	0.016	0.432	10000	0.413	0.452	0.001	0.026	-0.019	1.001	17.633	9.767
L_{SupCon} (C: 10k)	0.015	0.02	0.419	10000	0.404	0.434	0.001	0.027	-0.036	1.001	19.5	20.074
L_{VICReg} [32] (C: 5k)	0.018	0.082	0.27	4496	0.309	0.239	0.004	0.021	-0.134	1.001	18.031	11.612
L_{IMSAT} (C: 10k)	0.621	0.754	0.844	9844	0.836	0.853	0.616	0.678	-0.122	0.999	16.897	4.438
L_{IMSAT} (C: 5k)	0.578	0.731	0.822	5000	0.83	0.815	0.552	0.604	-0.033	**1.002**	26.56	4.507
L_{IMSAT} (C: 5994)	0.600	0.743	0.833	5993	0.834	0.831	0.583	0.636	-0.074	0.999	23.915	4.295
$L_{Mixup} + L_{SupCon}$ (C: 10k)	0.015	0.034	0.354	9639	0.36	0.348	0.002	0.023	-0.133	0.999	**15.563**	12.54
$L_{IMSAT} + L_{Mixup} + L_{VICReg'variance} + L_{VICReg'covariance}$ (C: 5k)	0.013	0.018	0.369	5000	0.367	0.371	0.001	0.017	**-0.015**	1.0	25.571	19.952
$L_{IMSAT} + L_{Mixup} + L_{SupCon} + L_{VICReg'variance} + L_{VICReg'covariance}$ (C: 5k)	0.014	0.02	0.36	5000	0.361	0.359	0.001	0.017	-0.022	0.999	26.667	21.84
$L_{IMSAT} + L_{Mixup}$ (C: 10k)	0.628	0.764	0.852	9791	0.841	0.862	0.615	0.692	-0.149	1.0	17.297	4.321
$L_{IMSAT} + L_{SupCon}$ (C: 10k)	0.548	0.717	0.813	9585	0.823	0.803	0.361	0.589	-0.138	1.001	15.941	4.348
$L_{IMSAT} + L_{SupCon}$ (C: 5k)	0.497	0.688	0.784	4996	0.81	0.76	0.347	0.516	-0.065	0.999	24.809	4.623
$CIMC = L_{IMSAT} + L_{Mixup} + L_{SupCon}$ (C: 10k)	**0.639**	**0.776**	**0.86**	9685	**0.847**	**0.873**	**0.642**	**0.71**	-0.136	0.998	17.599	4.252
$CIMC = L_{IMSAT} + L_{Mixup} + L_{SupCon}$ (C: 5k)	0.602	0.751	0.836	4999	0.842	0.831	0.579	0.632	-0.071	0.999	26.905	**4.231**

by default, 5000 as the predefined number of clusters, which was discovered by [5] to produce the best performances.

Moreover, to follow other SV works in training the ECAPA-TDNN-based systems, we have applied data augmentation at the waveform level, such as additive noise and room impulse response (RIR) simulation, as described in [7]. Furthermore, we extended augmentation to the extracted MFCCs features, following a similar approach to the specaugment scheme [44].

In Table I, we conducted an extensive ablation study to assess the impact of each component within our CIMC system and the influence of the predefined number of clusters. We also study the VICReg method [32] which comprises a term $L_{VICReg'variance}$ that maintains the variance of each embedding dimension above a threshold and a term $L_{VICReg'covariance}$ that decorrelates each pair of variables. Results indicate that there exists complementary information among all loss terms within our proposed objective. Each term contributes to enhancing the performance of the overall clustering framework. We also observe that choosing a much higher number of clusters than ground truth leads to improved clustering performance across all studied systems. Additionally, compared to a large variety of 15 clustering benchmarks in [6, Tab. 1], we can observe that our proposed method outperforms all other baselines in terms of clustering metrics achieving 63.9% unsupervised clustering accuracy compared to 58.7% for AHC which was the best performing method (8.9% relative improvement) while having a compute time comparable to classical clustering models (3-4 days). Using our proposed system's generated PLs to train our embedding system, also enabled us to achieve a very competitive downstream SV EER performance, surpassing all other benchmarks, except the AHC PLs which lead to a slightly better performance.

Finally, Table II presents a comparison between our Self-Supervised SV (SSSV) training approach, utilizing CIMC-based PLs, and recent SOTA SSSV methods (with the ECAPA-TDNN model encoder) employing diverse self-supervised objectives. Instead of AAMSoftmax, using the margin-based OCSoftmax objective loss [48] which uses one-class learning instead of multi-class classification and which does not assume the same distribution for all speakers (which is more realistic in our case) enables us to enhance SV performance to 3.924% EER. Our findings demonstrate that our approach offers highly competitive performance compared to all baseline methods. Moreover, they indicate that enhancing the clustering modules of existing self-supervised speaker recognition systems could lead to further improvements.

TABLE II: A comparison of several SOTA Self-Supervised SV approaches to our simple SV system trained with our generated CIMC PLs. All approaches employ the same ECAPA-TDNN underlying model. Results are presented on the original VoxCeleb1 test set (Voxceleb1_O) in terms of EER (%).

SSL Objective	**EER (%)**
MoBY [10]	8.2
InfoNCE [12]	7.36
MoCo [45]	7.3
ProtoNCE [10]	7.21
PCL [10]	7.11
CA-DINO [46]	3.585
i-mix [47]	3.478
l-mix [47]	3.377
Iterative clustering [12]	3.09
Our approach (CIMC & AAMSoftmax)	$\mathbf{4.231}$
Our approach (CIMC & OCSoftmax)	$\mathbf{3.924}$

It is worth noting that although our approach slightly underperforms compared to the 2-stage iterative clustering method [12] and l-mix [47], the iterative clustering method relies on multi-stage training and requires multiple iterations for training, while l-mix incorporates an additional Variational Auto-Encoder (VAE) to generate mixup-based augmentations. Our approach, on the other hand, does not require augmentations or any additional components, making it lightweight. Moreover, our approach is faster, simpler, and has the potential to incorporate both iterative clustering and l-mix to benefit from mixup regularization and progressive clustering, further enhancing performance.

VI. Conclusion

In this paper, we introduced an efficient and general-purpose multi-objective clustering approach, denoted as Contrastive Information Maximization Clustering (CIMC). The proposed approach avoids explicit data augmentation, ensuring fast training and usage of low memory and compute resource. Clustering and speaker verification experiments on the VoxCeleb dataset demonstrated that the CIMC is robust and has better generalization capability. Additionally, we explored various recent state-of-the-art self-supervised learning objectives for clustering, demonstrating that our multi-objective approach provides beneficial complementary information. Our method outperformed all other baselines used for clustering speaker embeddings and delivered very competitive speaker verification performance compared to other benchmarks.

VII. 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.

References

[1] John H.L. Hansen and Taufiq Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] Najim Dehak et al., "Front-end factor analysis for speaker verification," *IEEE/ACM Trans. Audio Speech Lang*, 2011.

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

[4] Zhongxin Bai and Xiao-Lei 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," *JSTSP*, 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 ENLSP Workshop*, 2022.

[7] David Snyder et al., "X-vectors: Robust dnn embeddings for speaker recognition," in *IEEE-CASSP*, 2018.

[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] Bing Han, Zhengyang Chen, and Yanmin 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] Weihua Hu et al., "Learning discrete representations via information maximizing self-augmented training," PMLR, 2017.

[16] Xudong Mao et al., "Virtual mixup training for unsupervised domain adaptation," *arXiv preprint arXiv:1905.04215*, 2019.

[17] Prannay Khosla, Piotr Teterwak, et al., "Supervised contrastive learning," *NeurIPS*, 2020.

[18] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *JSTOR: Applied Statistics*, pp. 100–108, 1979.

[19] T. Zhang et al., "BIRCH: A new data clustering algorithm and its applications," *Data Mining and Knowledge Discovery*, 1997.

[20] Sudipto Guha et al., "Cure: An efficient clustering algorithm for large databases," *SIGMOD Rec.*, jun 1998.

[21] W. H. Day et al., "Efficient algorithms for agglomerative hierarchical clustering methods," *Journal of Classification*, 1984.

[22] Junyuan Xie et al., "Unsupervised deep embedding for clustering analysis," in *ICML*. PMLR, 2016, pp. 478–487.

[23] Paras Dahal, "Learning embedding space for clustering from deep representations," in *IEEE Big Data*, 2018.

[24] Nat Dilokthanakul et al., "Deep unsupervised clustering with gaussian mixture variational autoencoders," *arXiv preprint arXiv:1611.02648*, 2016.

[25] Zhuxi Jiang et al., "Variational deep embedding: A generative approach to clustering," *IJCAI*, vol. 1, 2016.

[26] Meitar Ronen et al., "Deepdpm: Deep clustering with an unknown number of clusters," in *Proceedings of IEEE/CVF*, 2022.

[27] Dara Bahri et al., "Scarf: Self-supervised contrastive learning using random feature corruption," *arXiv preprint arXiv:2106.15147*, 2021.

[28] Yves Grandvalet and Yoshua Bengio, "Semi-supervised learning by entropy minimization," *NeurIPS*, vol. 17, 2004.

[29] Takeru Miyato et al., "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *PAMI*, 2018.

[30] John Bridle et al., "Unsupervised classifiers, mutual information and'phantom targets," *NeurIPS*, vol. 4, 1991.

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

[32] Adrien Bardes, Jean Ponce, and Yann LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," *arXiv preprint arXiv:2105.04906*, 2021.

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

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

[35] Andrew Rosenberg and Julia Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proceedings of EMNLP-CoNLL*, 2007, pp. 410–420.

[36] Edward B Fowlkes and Colin L Mallows, "A method for comparing two hierarchical clusterings," *Journal of the American statistical association*, vol. 78, no. 383, pp. 553–569, 1983.

[37] Peter J Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[38] Tadeusz Caliński et al., "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, 1974.

[39] David L Davies and Donald W Bouldin, "A cluster separation measure," *IEEE transactions on PAMI*, 1979.

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

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

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

[43] Jiankang Deng et al., "Arcface: Additive angular margin loss for deep face recognition," *IEEE Transactions on PAMI*, 2021.

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

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

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

[47] 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.

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