April 14~19, 2024 (ICASSP 2024)

Self-supervised Speaker Verification Employing a Novel Clustering Algorithm

Abderrahim Fathan, Jahangir Alam

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

Summary

- We propose a **new general-purpose clustering algorithm** called **CAMSAT**.
- CAMSAT combines the benefits of **mutual information (MI) maximization** of IMSAT clustering framework & the regularization benefit of AUGMIX (**mix of augmentations** at the predictions level).
- Using a thorough comparative analysis via clustering metrics, **CAMSAT allows us to outperform all other clustering algorithms** for speaker clustering.
- We achieve **better speaker verification performance (SV)** than all **SOTA SV** baselines.
- Benefits: better **generalizability**, **robustness**, and **stability** of clustering under **data shift** for **large-scale datasets** or/and a **high number of clusters**.
- We perform an ablation study to **analyze the contribution of the different components** of our proposed framework.

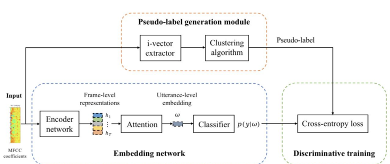
Introduction & Motivation

- **Labeled datasets** can be **expensive** and **time-consuming** to obtain.
- **Label noise** can significantly impact performance of self-supervised learning-based models.
- There are generally 2 groups of methods to learn from noisy data:
  - Noise-robust algorithms: learn directly from noisy labels
  - Label-cleansing approaches: remove or correct mislabeled data.
- How can we **improve the performance of deep clustering models**?
- How can we **improve the estimation of the ground truth number of clusters**?

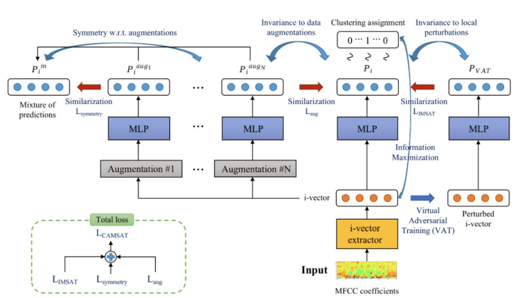
Proposed Approach

❖ Clustering-based Self-supervised Systems

- The general process for training our clustering generated pseudo-label-based self-supervised speaker embedding networks:



❖ Proposed CAMSAT Approach



❖ CAMSAT | Math Equations

- **We minimize the following objective:**

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

where: $$L_{aug} = \frac{1}{N}\sum_{i=1}^{N} KL(p_i^{aug_{r_i}}||p_i)$$

$$p_i^m = \frac{1}{|J|+1}\left[\sum_{j\in\{1,\dots,|J|\}} \alpha_j p_i^{aug_j} + p_i\right]$$

$$L_{symmetry} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|J|}\sum_{j\in\{1,\dots,|J|\}} KL(p_i^{aug_j}||p_i^m)$$

$$R_{SAT}(\theta; T) = \frac{1}{N}\sum_{n=1}^{N} R_{SAT}(\theta; x_n, T(x_n))$$

$$R_{SAT}(\theta; x, T(x)) = -\sum_{c=1}^{C}\sum_{y_c=0}^{1} p_{\hat{\theta}}(y_c|x) \log p_\theta(y_c|T(x))$$

- L_{aug} forces the predicted representations of augmented samples to be close to those of the original data points.
- R_{SAT} loss term allows the following:
  - Representations of the augmented samples are pushed close to those of the original samples.
  - Regularize the complexity of the network against local perturbations using Virtual Adversarial Training (VAT).
- $L_{symmetry}$ allows the following:
  - Produce **consistent feature representations** that are labeled identically irrespective of transformations, generalize better, and generate compact clusters.
  - Mixing augmentations generates further **diverse transformations** at the latent predictions level.
    - Induce robustness and reduce the memorization of used augmentations.
  - Enforce **representation smoothness** and symmetry w.r.t data augmentations.
  - Bootstrapping mixed predictions can be considered a **majority vote clustering** method.
  - Conduct **entropy minimization** implicitly.

Experiments & Results

❖ Experimental Setup

- **VoxCeleb Corpus**
  - Trained using the VoxCeleb2 development set: 5994 speakers (1,092,009 utterances in total)
  - Evaluated on the VoxCeleb1-O test set: 40 speakers (4,874 utterances in total)
- **Acoustic features:** 40-dimensional MFCCs.
- **Data Augmentation**
  - Waveform-level data augmentations: **additive noise** and **room impulse response** (RIR) simulation.
  - **SpecAugment** is applied on the fly.
  - Both **time** and **frequency maskings** are performed.
  - Additional augmentation for clustering (on i-vectors):
    - **Gaussian noise**
    - **Input Masking** (5-10% of input)
- **Clustering Evaluation Metrics**
  - **7 Supervised metrics**: Unsupervised clustering accuracy (ACC), Normalized Mutual Information (NMI), Adjusted MI (AMI), Completeness, Homogeneity, Purity, Fowlkes-Mallows index (FMI).
  - **3 Unsupervised metrics**: Silhouette score, Calinski-Harabasz score (CHS), Davies-Bouldin score (DBS).

❖ Results

- **Results | Clustering & SV performances**

Model	Clustering Metrics											Speaker Verification
	ACC	AMI	NMI	No. of clusters	Completeness	Homogeneity	FMI	Purity	Silhouette	CHS	DBS	EER (%)
Supervised (True Labels)	**1.0**	**1.0**	**1.0**	5994	**1.0**	**1.0**	**1.0**	**1.0**	**−0.006**	31.708	4.692	**1.437**
GMM (Full cov.)	0.45	0.631	0.747	5000	0.767	0.728	0.312	0.566	-0.015	39.266	4.673	5.143
GMM (Full cov., l2Norm)	0.504	0.678	0.789	5000	0.792	0.785	0.415	0.633	-0.015	41.568	5.114	5.429
GMM (Spherical cov.)	0.427	0.587	0.711	5000	0.739	0.685	0.22	0.539	-0.037	38.665	4.864	5.265
GMM (Diagonal cov.)	0.425	0.6	0.721	5000	0.748	0.696	0.23	0.539	-0.033	38.455	4.874	5.451
GMM (Tied cov.)	0.457	0.66	0.767	5000	0.788	0.747	0.317	0.574	-0.016	38.922	4.726	5.164
Bayesian GMM (γ=1e-5, μ=1)	0.45	0.629	0.746	5000	0.766	0.727	0.312	0.566	-0.015	39.257	4.673	5.143
Bayesian GMM 1 (l2Norm, γ=1e-5, μ=1)	0.504	0.678	0.789	5000	0.792	0.785	0.415	0.633	-0.015	**41.57**	5.115	5.159
Bayesian GMM 2 (γ=100, μ=0.01)	0.449	0.63	0.746	5000	0.766	0.727	0.311	0.566	-0.015	39.258	4.675	4.958
Divisive HC	0.097	0.204	0.477	5000	0.479	0.474	0.035	0.132	-0.06	18.044	9.068	13.531
KMeans	0.302	0.468	0.591	5000	0.645	0.546	0.194	0.311	-0.114	24.936	**2.714**	6.978
CURE	0.151	0.218	0.393	5000	0.466	0.34	0.011	0.216	-0.052	17.77	5.372	6.994
BIRCH	0.299	0.374	0.54	5000	0.725	0.43	0.013	0.353	-0.027	24.348	4.901	5.642
DEC	0.029	0.122	0.365	4911	0.386	0.345	0.007	0.036	-0.084	8.734	7.266	11.957
SOM	0.025	0.088	0.402	5041	0.404	0.4	0.01	0.037	-0.041	10.148	18.402	15.806
DeepCWRN	0.003	0.006	0.15	1008	0.179	0.129	0.001	0.003	-0.217	3.841	41.521	38.171
IMSAT	0.393	0.491	0.649	4987	0.668	0.63	0.297	0.426	-0.044	22.887	6.668	5.912
AHC	0.587	0.74	0.825	5000	0.841	0.81	0.311	0.684	**−0.01**	39.561	4.991	3.685
AHC (l2Norm)	0.602	0.756	0.838	5000	0.849	0.827	0.375	0.693	-0.034	39.638	5.147	3.621
CAMSAT (S: 30k)	0.614	0.746	0.829	4993	0.843	0.816	0.557	0.636	-0.033	1.001	25.239	3.812
CAMSAT (l2Norm, S: 20k)	0.655	0.812	0.874	4596	0.888	0.86	0.641	0.675	-0.105	0.999	26.561	**3.065**
CAMSAT (l2Norm, S: 20k, C: 10k)	**0.709**	**0.83**	**0.889**	6364	**0.892**	**0.886**	**0.708**	**0.745**	-0.141	1.0	21.656	3.134

- **Ablation study of all components**

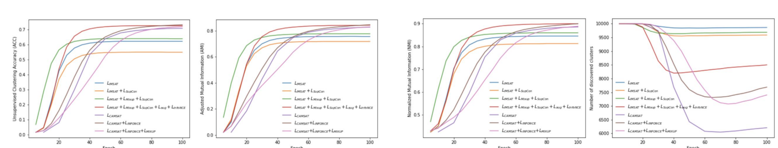
Model	Clustering Metrics											Speaker Verification
	ACC	AMI	NMI	No. of clusters	Completeness	Homogeneity	FMI	Purity	Silhouette	CHS	DBS	EER (%)
IMSAT (no L_{aug}, no $L_{symmetry}$)	0.393	0.491	0.649	4987	0.668	0.63	0.297	0.426	-0.044	**22.887**	**6.668**	5.912
no $L_{symmetry}$	0.466	0.586	0.713	4967	0.736	0.693	0.316	0.529	-0.096	1.001	24.407	5.276
no $L_{symmetry}$, no R_{SAT}	0.401	0.585	0.733	5000	0.73	0.736	0.34	0.488	-0.016	1.0	26.5	6.188
no $L_{symmetry}$, no R_{SAT}, C: 10k	0.319	0.544	0.732	9998	0.705	0.761	0.275	0.519	-0.059	1.0	18.814	6.792
no $L_{symmetry}$, no R_{SAT}, C: 15k	0.266	0.497	0.715	14798	0.68	0.754	0.22	0.502	-0.13	1.0	15.194	7.015
no L_{aug}	0.498	0.646	0.765	4999	0.773	0.757	0.454	0.555	-0.051	1.001	25.671	5.329
no L_{aug}, no mix of inputs augmentation	0.492	0.644	0.763	5000	0.772	0.755	0.446	0.551	-0.083	0.999	25.866	5.027
no L_{aug}, dropout	0.473	0.664	0.783	5000	0.781	0.785	0.462	0.542	**−0.015**	1.0	25.558	5.488
no L_{aug}, no BatchNorm	0.491	0.674	0.787	5000	0.789	0.785	0.487	0.552	-0.018	1.001	25.856	5.313
no L_{aug}, only $KL(p_i\|\|p_i^m)$ term in $L_{symmetry}$	0.378	0.471	0.621	4957	0.654	0.591	0.197	0.417	-0.11	1.0	22.795	6.161
no L_{aug}, mixture in the input space alone	0.414	0.496	0.652	4989	0.671	0.633	0.301	0.451	-0.1	1.001	23.924	5.488
no L_{aug}, add another $L_{symmetry}$ version at the input level	0.428	0.624	0.757	5000	0.755	0.759	0.373	0.519	-0.017	0.999	25.697	5.541
CAMSAT	0.52	0.704	0.806	5000	0.809	0.803	0.48	0.602	-0.021	1.0	25.238	4.825
CAMSAT (S: 15k, C: 5994)	0.556	0.724	0.821	5994	0.821	0.822	0.524	0.644	-0.049	1.001	23.174	4.518
CAMSAT (S: 20k)	0.589	0.741	0.828	5000	0.836	0.821	0.559	0.639	-0.029	1.0	25.455	4.3
CAMSAT (l2Norm, Noise)	0.583	0.754	0.835	4923	0.848	0.823	0.552	0.636	-0.128	1.002	24.968	3.95
CAMSAT (l2Norm, S: 15k)	0.654	0.815	0.875	4664	**0.891**	0.86	0.643	0.673	-0.115	1.0	25.834	3.25
CAMSAT (l2Norm, S: 15k, C: 5994)	0.67	0.821	0.88	5238	**0.893**	0.867	0.662	0.692	-0.126	1.0	23.589	3.081
CAMSAT (Noise, Masking)	0.534	0.716	0.814	5000	0.817	0.81	0.493	0.614	-0.021	1.002	25.308	4.613
CAMSAT (Noise, Masking, S: 15k)	0.585	0.743	0.831	5000	0.835	0.826	0.559	0.642	-0.024	1.001	25.651	4.178
CAMSAT (Noise, Masking, S: 30k)	0.614	0.746	0.829	4993	0.843	0.816	0.557	0.636	-0.033	1.001	25.239	3.812
CAMSAT (l2Norm, Masking)	0.581	0.755	0.835	4907	0.849	0.822	0.544	0.632	-0.093	1.0	24.836	4.067
CAMSAT (l2Norm, Noise, Masking)	0.576	0.76	0.84	4993	0.85	0.83	0.544	0.633	-0.075	1.0	25.034	4.04
CAMSAT (l2Norm, Noise, Masking, S: 15k)	0.651	0.814	0.874	4665	**0.89**	0.859	0.639	0.67	-0.116	1.0	25.82	3.43
CAMSAT (l2Norm, Noise, Masking, S: 20k)	0.655	0.812	0.874	4596	**0.888**	0.86	0.641	0.675	-0.105	0.999	26.561	3.065
CAMSAT (l2Norm, Noise, Masking, S: 20k, C: 10k)	**0.709**	**0.83**	**0.889**	6364	**0.892**	**0.886**	**0.708**	**0.745**	-0.141	1.0	21.656	3.134
CAMSAT (l2Norm, Noise, Masking, S: 20k, C: 10k, 2xBatch size)	0.704	0.828	0.888	6424	**0.891**	**0.884**	**0.703**	**0.74**	-0.132	1.0	21.13	3.224
CAMSAT (l2Norm, Noise, Masking, S: 20k, C: 10k, Label smoothing)	0.707	**0.83**	**0.889**	6462	**0.892**	**0.886**	**0.708**	**0.743**	-0.141	0.999	21.24	3.203
CAMSAT (l2Norm, Noise, Masking, S: 20k, C: 5994)	0.669	0.816	0.878	5194	**0.888**	0.868	0.659	0.694	-0.122	1.0	24.627	3.309
CAMSAT (l2Norm, Noise, Masking, S: 30k)	0.656	0.801	0.867	4909	0.878	0.857	0.646	0.678	-0.1	0.999	25.923	3.33
CAMSAT (l2Norm, Noise, Masking, S: 45k)	0.638	0.776	0.851	4999	0.861	0.842	0.614	0.662	-0.09	0.999	26.241	3.462

- **Distance metrics & Comparison to SOTA self-supervised SV approaches**

Distance Metric	Clustering Metrics											Speaker Verification
	ACC	AMI	NMI	No. of clusters	Completeness	Homogeneity	FMI	Purity	Silhouette	CHS	DBS	EER (%)
KL divergence	**0.528**	**0.702**	**0.805**	5000	**0.809**	**0.801**	**0.485**	**0.608**	-0.026	1.001	25.465	**4.608**
JS divergence	0.376	0.541	0.706	4994	0.702	0.71	0.317	0.441	**−0.014**	1.001	25.1	6.267
Cosine distance	0.308	0.411	0.612	4897	0.616	0.609	0.259	0.335	-0.095	**1.002**	25.198	6.4
L2 loss	0.251	0.385	0.564	4895	0.59	0.541	0.134	0.275	-0.132	0.998	**23.428**	8.754
AAMSoftmax [13]	0.314	0.524	0.676	4729	0.691	0.661	0.251	0.361	-0.119	1.001	24.811	9.046
Squared Earth Mover's distance [30]	0.051	0.127	0.407	4996	0.416	0.399	0.014	0.063	-0.06	0.998	25.638	18.627
Sinkhorn distance [8]	0.018	0.052	0.331	4995	0.346	0.317	0.002	0.023	-0.083	1.001	23.661	12.275

SSL Objective	EER (%)
MoBY [62]	8.2
InfoNCE [55]	7.36
MoCo [6]	7.3
ProtoNCE [62]	7.21
PCL [62]	7.11
CA-DINO [26]	3.585
i-mix [18]	3.478
l-mix [18]	3.377
Iterative clustering [54]	3.09
CAMSAT (Ours)	**3.065**

- **CAMSAT | Clustering metrics over time**



Advantages of our approach

- Our Mutual Information-based approach is **rigorously grounded in information theory**
  - Effortlessly **avoid degenerate solutions**.
  - Ensembling classifier predictions improves **prediction calibration**.
- Our clustering objective is easily scalable:
  - **No dependence on large batches** of **negative samples** or **large similarity matrices**.
  - Is more suitable for **large-scale datasets** thanks to its robustness and stability against **corruptions** and **data shift**.
- Adding Gaussian noise to inputs helps the **exploration of different clustering configurations**.
- CAMSAT is **not sensitive to the predefined number of clusters.**
- CAMSAT recovers the number of clusters almost perfectly.

Key Findings

- We demonstrate the importance of our mixtures and the relevance of the notion of symmetry within augmentations.
- Setting a higher predefined number of clustering in CAMSAT leads to improved performance.
- Beyond some clustering performance, the marginal improvement in final downstream SV performance becomes minimal and requires much more accurate PLs.