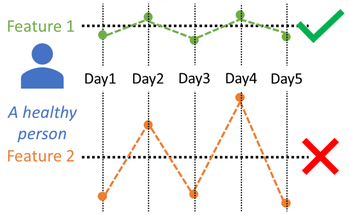
Learning Repeatable Speech Embeddings Using An Intra-class Correlation Regularizer

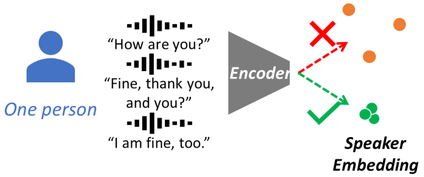
Jianwei Zhang
Suren Jayasuriya
Visar Berisha
Arizona State University

Good supervised embedding should …

①Sensitive to changes in the target class (speaker identity, clinical state)
②Remaining invariant to unrelated factors (noise, natural variations)

Embeddings that satisfy these two desired properties have ***high repeatability***





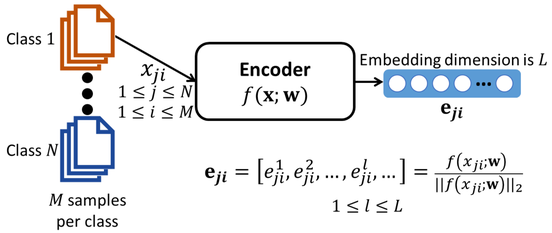
Repeatability is rarely explicitly considered during training or evaluation of embeddings. We posit that directly assessing repeatability can help improve the quality of learned latent representations.

Our solution

We propose to use the ***intra-class correlation coefficient (ICC)*** to evaluate embeddings' repeatability, and propose a novel regularizer, ***ICC regularizer***, to enforce networks to learn repeatable embeddings.

Intra-class correlation coefficient (ICC) & ICC regularizer

ICC is a statistical measure used to describe how strongly units within the same group resemble each other. It is a well-defined metric used in various areas for assessing reliability of raters or measurements and comparing variability within vs. between groups.



$$\mathbf{e_{ji}} = [e_{ji}^1, e_{ji}^2, \dots, e_{ji}^l, \dots] = \frac{f(x_{ji};\mathbf{w})}{||f(x_{ji};\mathbf{w})||_2}, \quad 1 \le l \le L$$

$$MS_B(e^l) = \frac{M \cdot \sum_j^N (\bar{e_j^l} - \bar{e^l})^2}{N-1}$$

*Inter-class (**B**etween-class) variance*

$$MS_W(e^l) = \frac{\sum_j^N M\sigma_{j,l}^2}{N \cdot (M-1)}$$

*Intra-class (**W**ithin-class) variance*

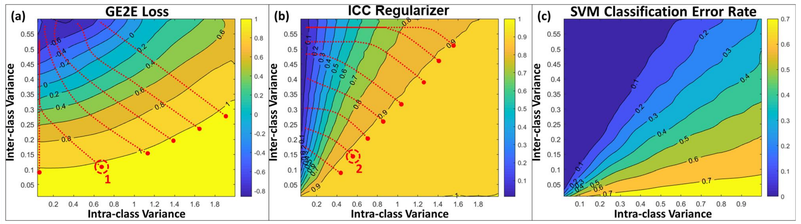
$$ICC(e^l) = \frac{MS_B(e^l) - MS_W(e^l)}{MS_B(e^l) + (M-1) \cdot MS_W(e^l)}$$

ICC value of l-th embedding dimension

$$R_{ICC} = 1 - \frac{\sum_l^L ICC(e^l)}{L}$$

ICC Regularizer

Compare ICC regularizer vs. contrastive loss



- Use Monte Carlo simulation to to study their similarity and difference.
- The total variance of the embeddings (e_{ji}) is constrained. It is beneficial to focus on the intra-class variance for increasing representation repeatability.
- GE2W, a contrastive loss, optimize intra-c and inter-class variance simultaneously.
- The ICC regularizer places greater emphasis on minimizing intra-class variance.
- **The ICC regularizer makes a better trade-off on minimizing intra-class variance and maximizing inter-class variance.**

Task 1: text-independent speaker verification

	VGG-M-40			FastResNet-34		
	EER	minDCF	ICC	EER	minDCF	ICC
GE2E	4.39%	0.2925	0.4494	2.49%	0.2133	0.7215
GE2E + ICC	3.96%	0.2778	0.5487	2.39%	0.2012	0.7366
AngleProto	4.36%	0.2809	0.4399	2.28%	0.1960	0.7501
AngleProto + ICC	4.02%	0.2790	0.5455	2.17%	0.1871	0.7627
SupCon	3.91%	0.2791	0.5693	2.30%	0.1956	0.7500
SupCon + ICC	**3.78%**	**0.2597**	**0.6661**	**2.16%**	**0.1867**	**0.7615**

We select three contrastive losses as baseline and run TI-SV task on VoxCeleb dataset. The experimental results demonstrate that the proposed ICC regularizer can improve the repeatability of embeddings. The speaker embeddings with higher repeatability achieve improved performance on the TI-SV task.

Task 2: zero-shot voice style conversion

AB Preference Result

	GE2E Loss	GE2E Loss + ICC
Select Ratio	38.10%	**61.90%**

Objective Evaluation

	Speaker Similarity Score	WER	CER
GE2E Loss	0.2231	0.5810	0.3817
GE2E Loss + ICC	0.2309	0.5109	0.3324

Evaluate AutoVC with two speaker embeddings generated in task 1 for zero-shot voice style conversion. The experimental results demonstrate that speaker embeddings with higher repeatability also result in better performance for voice style conversion task.

Task 3: dysphonic voice detection

	Dysphonic Voice Classification / [mean accuracy] (95% CI)				ICC
	SVD Train	SVD Validation	MEEI Testing	HUPA Testing	ALS
Proposed Method	0.735(.004)	**0.729(.009)**	**0.821(.004)**	**0.689(.011)**	**0.571**
Zhang et al. (2022)	0.800(.018)	0.708(.011)	0.821(.014)	0.665(.008)	0.437
Harar et al. (2017)	0.774(.017)	0.691(.009)	0.661(.024)	0.492(.008)	0.474
Verde et al. (2018)	0.891(.006)	0.627(.009)	0.704(.018)	0.598(.015)	0.018
Huckvale et al. (2021)	0.729(.019)	0.626(.012)	0.698(.044)	0.549(.014)	0.291

We implement the ICC regularizer to enhance the repeatability of embeddings for a clinical application, dysphonic voice detection. The experimental results of voice feature embeddings for dysphonic voice detection task demonstrate that the ICC regularizer can improve the repeatability of embeddings and embeddings with higher repeatability exhibit better accuracy and generalizability in dysphonic voice detection.

Conclusion

We propose to use the ICC as an evaluation metric in representation learning and use the ICC regularizer as a complementary component for contrastive loss to regularize deep-learned embeddings to be more repeatable. The experimental results demonstrate that the ICC regularizer can improve the repeatability of learned embeddings, and embeddings with higher repeatability perform better in downstream tasks.

Paper and code