ISCA SIG-SPSC
SECURITY AND PRIVACY IN SPEECH COMMUNICATION

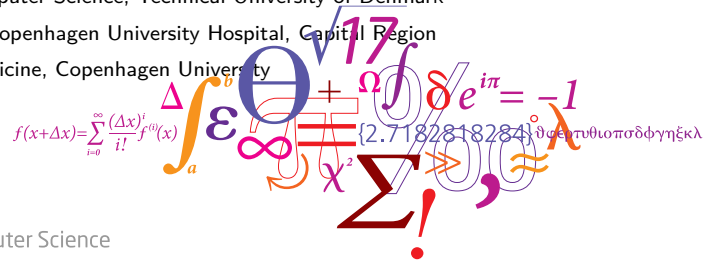
REGION H Psykiatri DTU

novo nordisk fonden

Zero-shot Cross-lingual Speech Emotion Recognition: A Study of Loss Functions and Feature Importance

Sneha Das (sned@dtu.dk), Nicole Nadine Lønfeldt[2], Nicklas Leander Lund[1],
Anne Katrine Pagsberg[2,3], Line H. Clemmensen[1]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark
[2]Child and Adolescent Mental Health Center, Copenhagen University Hospital, Capital Region
[3]Faculty of Health, Department of Clinical Medicine, Copenhagen University



DTU Compute
Department of Applied Mathematics and Computer Science

Motivation

- Speech analysis for children with OCD $\rightarrow$ Conversations in Danish
- Paralinguistic analysis $\rightarrow$ Speech emotion recognition (SER): inferring emotional state from speech signals.
- Challenges: Very little of data (Danish speech from children)
- Generalization and black-box nature of models $\rightarrow$ impacts security of systems.

Zero-shot Cross-lingual SER, with focus on loss-functions.

Research questions

1. Which model is more suitable for learning emotion representations from speech that are transferable over languages, categorical or dimensional model.
2. Does semi-supervision aid in transferring emotion representations over languages, without labelled data in the target domain?
3. How do different learning functions influence the feature attribution scores?

Methods

- Low complexity architecture, same architecture for all methods.
- Semi-supervision through loss function:
  1. Cluster-loss $\rightarrow$ Learning emotion classes
  2. Continuous metric-loss $\rightarrow$ Learning dimensional model of emotions $\rightarrow$ Activation, valence.

Evaluation

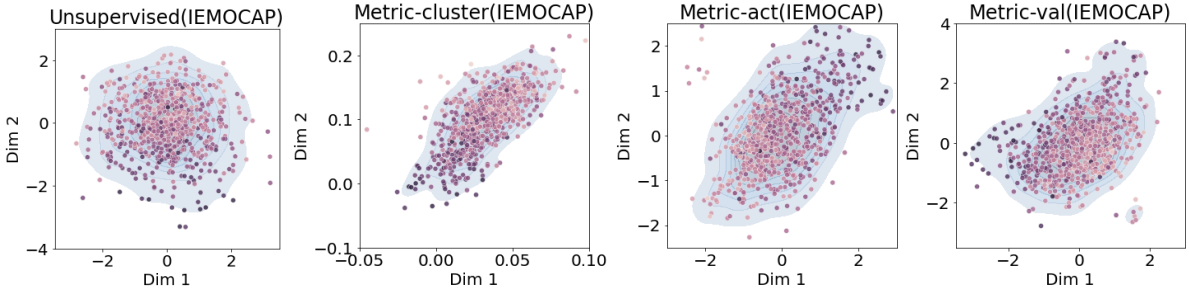
- Datasets: IEMOCAP (Training), SAVEE, Emo-DB, CaFE, URDU, AESD (Transfer)
- Input features: eGeMAPS using OpenSmile
- Preprocessing: remove outliers using z-score normalization ($-10 > z > 10$)
- 5-fold cross validation

Rank correlation analysis

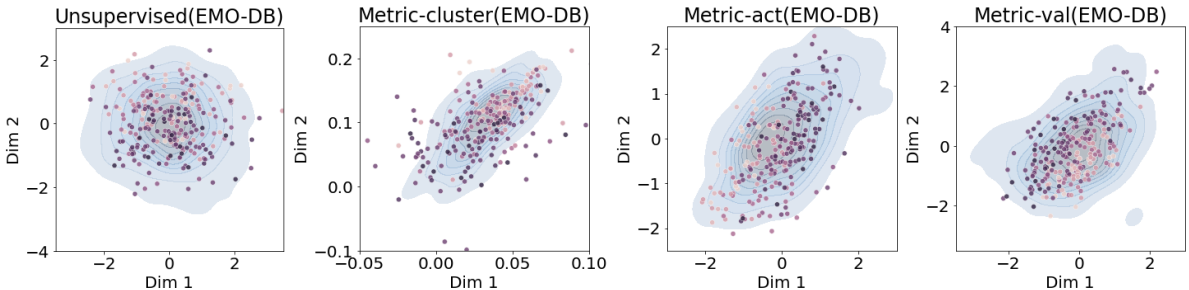
Table: Spearman's rank order correlation for the validation and transfer datasets aggregated over all model runs with different folds and random initial seeds. Higher correlation implies a larger correspondence to the ground truth labels (activation).

Method	IEMOCAP ($\mu \pm \sigma$)		EMO-DB ($\mu \pm \sigma$)		CAFE ($\mu \pm \sigma$)		URDU ($\mu \pm \sigma$)		AESD ($\mu \pm \sigma$)	
	Transfer	Supervised	Transfer	Supervised	Transfer	Supervised	Transfer	Supervised	Transfer	Supervised
Unsupervised	0.26 ± 0.17	0.26 ± 0.17	0.31 ± 0.22	0.31 ± 0.22	0.24 ± 0.14	0.24 ± 0.14	0.12 ± 0.1	0.1 ± 0.07	0.18 ± 0.11	0.16 ± 0.09
Metric-cluster	0.19 ± 0.14	0.19 ± 0.14	0.23 ± 0.16	0.28 ± 0.19	0.12 ± 0.08	0.07 ± 0.04	0.07 ± 0.06	0.09 ± 0.07	0.12 ± 0.06	0.11 ± 0.05
Metric-act	$\mathbf{0.76 \pm 0.05}$	$\mathbf{0.76 \pm 0.05}$	$\mathbf{0.53 \pm 0.08}$	$\mathbf{0.61 \pm 0.04}$	$\mathbf{0.35 \pm 0.04}$	$\mathbf{0.39 \pm 0.03}$	$\mathbf{0.38 \pm 0.05}$	$\mathbf{0.39 \pm 0.05}$	$\mathbf{0.31 \pm 0.01}$	$\mathbf{0.31 \pm 0.01}$
Metric-val	0.29 ± 0.11	0.29 ± 0.11	-0.05 ± 0.03	0.27 ± 0.24	0.31 ± 0.09	0.32 ± 0.1	0.03 ± 0.08	0.07 ± 0.1	0.01 ± 0.05	0.14 ± 0.1

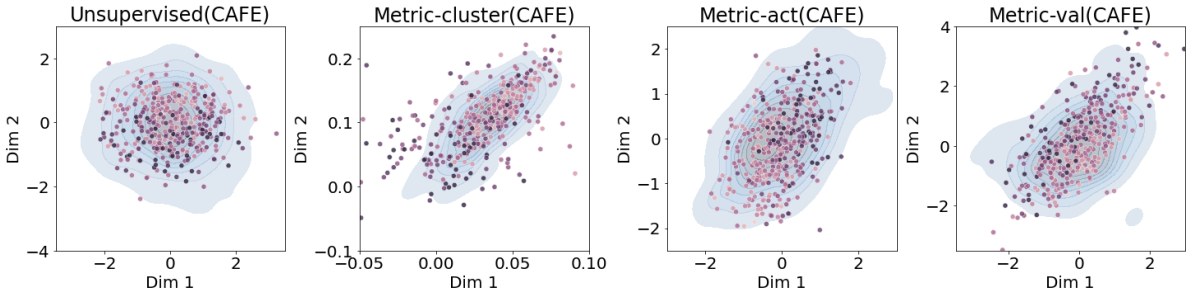
Rank correlation analysis



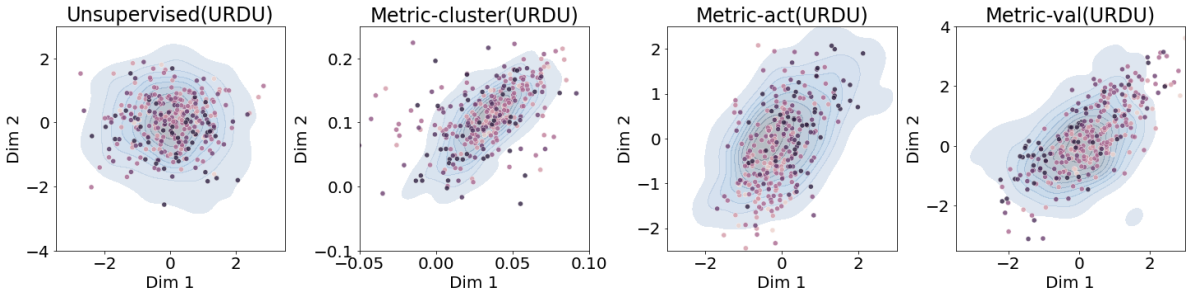
Rank correlation analysis



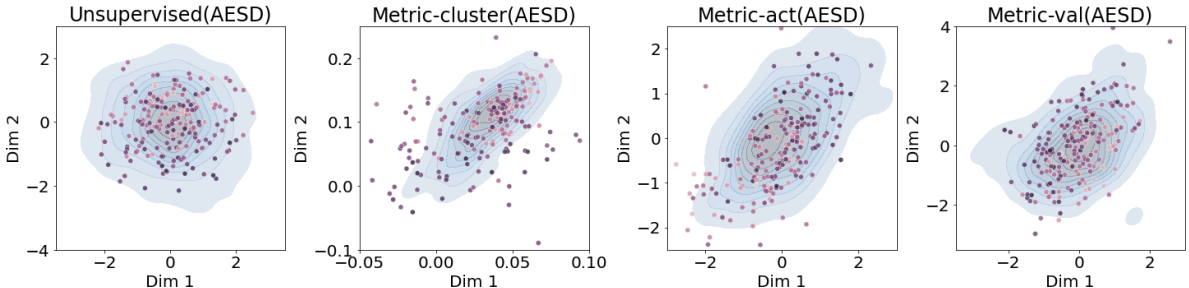
Rank correlation analysis



Rank correlation analysis



Rank correlation analysis



Rank correlation analysis

1. Metric-act consistently shows the highest Spearman's correlation coefficient; the Unsupervised model seems to perform better relative to Metric-cluster and Metric-val.
2. Distribution of the embedding in the latent space is most consistent for metric-act and closely followed by the Unsupervised model over the transfer dataset.
3. The correlation coefficient is relatively lower for CAFE, URDU and AESD.

Classification Accuracy

Table: Balanced classification accuracy over 4-class emotion classification. The model is trained on the IEMOCAP dataset only.

Method	IEMOCAP ($\mu \pm \sigma$) N-S-H-A	EMO-DB ($\mu \pm \sigma$) N-S-H-A	SAVEE ($\mu \pm \sigma$) N-S-H-A	CAFE ($\mu \pm \sigma$) N-S-H-A	URDU ($\mu \pm \sigma$) N-S-H-A	AESD ($\mu \pm \sigma$) S-H-A
Unsupervised	0.45 ± 0.01	0.39 ± 0.03	0.46 ± 0.02	0.25 ± 0.02	0.31 ± 0.02	0.17 ± 0.02
Metric-cluster	$\mathbf{0.57 \pm 0.03}$	$\mathbf{0.47 \pm 0.05}$	0.45 ± 0.05	$\mathbf{0.33 \pm 0.02}$	0.32 ± 0.03	0.31 ± 0.05
Metric-act	0.49 ± 0.01	0.44 ± 0.04	0.43 ± 0.03	0.26 ± 0.03	$\mathbf{0.33 \pm 0.03}$	0.24 ± 0.03
Metric-val	0.52 ± 0.02	0.45 ± 0.04	$\mathbf{0.47 \pm 0.03}$	0.31 ± 0.03	0.31 ± 0.03	0.25 ± 0.04

Classification accuracy

1. Metric-cluster shows highest classification performance among all the methods.
2. Cluster loss function was successful in discriminating between the classes on the training dataset. However, from the scatter points we observe that the remaining datasets have relatively larger covariances that overlap between emotions. This could indicate lower transferability of emotion representations with the metric-cluster loss function.
3. Latent samples from the Metric-act method are better contained within the kernel density plot of the training dataset. However, discrimination between the class labels are much lower relative to Metric-cluster, as can be seen from the higher overlap between datasets over the emotion classes.
4. *Anger* seems to be the most discernible from other classes, for most methods and datasets.

Feature attribution analysis

Feature attribution analysis

- Attribution scores are relatively lower for Unsupervised and Metric-cluster, in contrast to the Metric-act and Metric-val methods. Generally, the amplitude order of the mean absolute scores are Metric-val > Metric-act > Unsupervised > Metric-cluster.
- More number of feature groups for Metric-act seem to have no significant mean difference between emotion classes.
- Neutral (vs) Angry: Loudness, spectral flux and sound level are the feature groups that seem to have higher mean attribution scores for the emotion *Angry* over most of the methods considered.
- Neutral (vs) Sad: Formants and segment length show higher attribution scores.
- Metric cluster shows least changes in the attribution scores over different emotion classes.

Conclusions

- Activation is more transferable to unseen cross-lingual datasets.
- Semi-supervision helps!
- Identified feature groups with influence.

Limitations:

- Methods employed for transferability inspection increasing difficult for high dimension representations.
- Feature attribution methods here $\rightarrow$ systems with feature extractors.

References

Thankyou!

Email: sned@dtu.dk; Twitter: @dassneh