Influence of Loss Functions on the Latent Representation of Speech Emotions

Sneha Das (sned@dtu.dk)

[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

Motivation

- Speech emotion recognition (SER): inferring emotional state from speech signals.
- Emotion recognition employed in healthcare, education sector, criminal justice system.
- SER: signal processing, machine learning, deep learning.

WristAngel: Intervention and Research for OCD Treatment

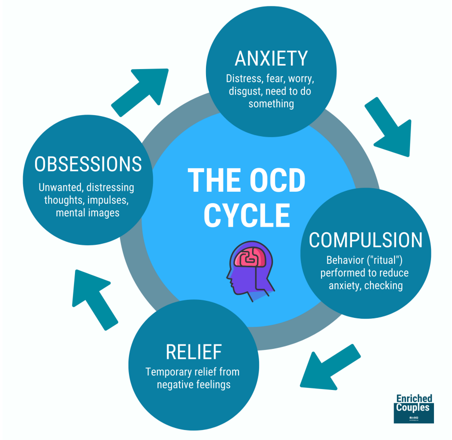
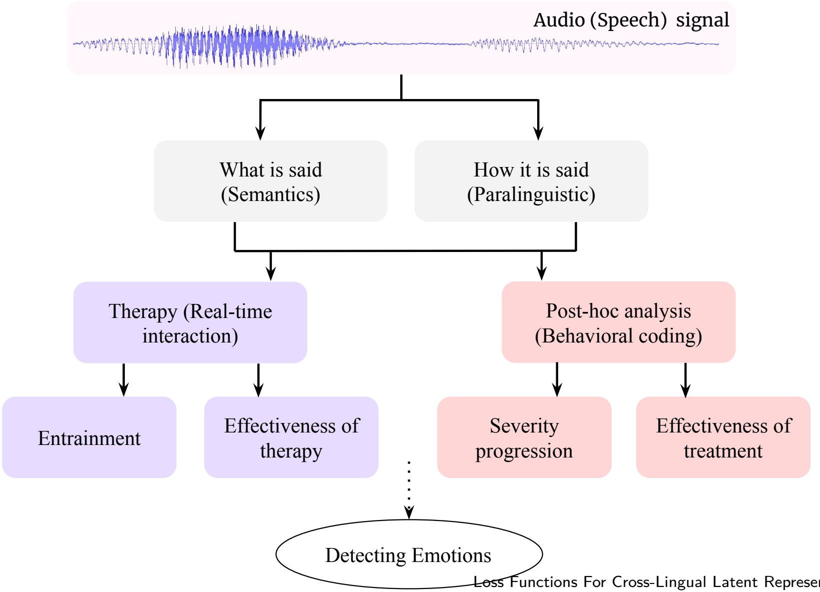


Figure: Obsessions and compulsions behave cyclically. Original image from `https://medium.com/amalgam/ocd-is-not-what-you-think-it-is-ee818028e79c`

- Mental disorder wherein ”People are caught in a cycle of obsession and compulsions".
- Obsessions → intrusive and disruptive urges, thoughts, images, etc.
- Compulsions → behavior to overcome obsessions, distress.
- In 2010, anxiety disorders - including obsessive-compulsive disorders -alone cost Europe over €74 billion (Gustavsson et al., 2011).

Role of Audio (Speech) in OCD Treatment



Speech signals and OCD

- Challenges:
  - Danish and child speech → Generalizing existing models unlikely.
  - Low resource conditions: few labels, not a lot of data (compared to input dimensions) → Training new models from scratch unlikely.
- Transferable models → Trained on open datasets and apply to Danish-speech from children.

Semi-supervision 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.

Semi-supervision with cluster-loss

Objectives and Contributions

Objectives for transferability:

1. Latent embedding with discrimination between emotion classes.
2. Latent distribution that are consistent over corpora.

Loss functions:

1. Low-complexity DAE and VAE.
2. VAE with KL-loss annealing: balancing KL-loss and reconstruction loss.
3. VAE with semi-supervision incorporating clustering in latent space.

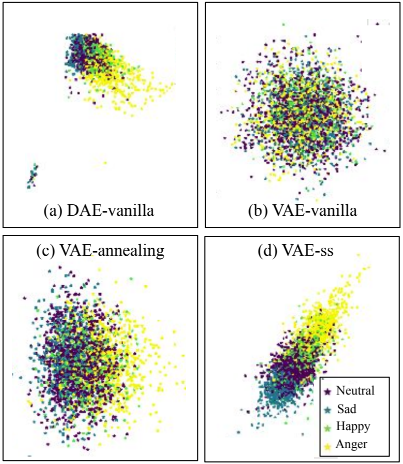
Formulation

- DAE:

$$\arg\min_{f_\theta, g_\phi} \quad \mathcal{L}_{\text{rec}} = \mathbb{E}\|\mathbf{x} - g_\phi(f_\theta(\mathbf{x_n}))\|_2^2, \tag{1}$$

- VAE:

$$\arg\min_{\theta,\phi} \quad \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}} = -\mathbb{E}_{\mathbf{z}\sim q_\theta(\mathbf{z}|\mathbf{x})} \log p_\phi(\mathbf{x}|\mathbf{z}) + D_{KL}(q_\theta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})), \tag{2}$$



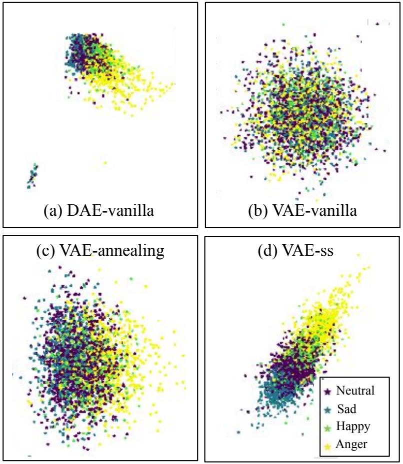
Formulation

- VAE with KL-annealing:

$$\arg\min_{\theta,\phi} \quad \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}} = -\mathbb{E}_{\mathbf{z}\sim q_\theta(\mathbf{z}|\mathbf{x})} \log p_\phi(\mathbf{x}|\mathbf{z}) + \beta_e D_{KL}(q_\theta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})), \tag{3}$$

where the standard formulation of β_e:

$$\beta_e = \begin{cases} f(\tau) = \frac{0.25}{R}\tau, & \tau \leq R \\ 0.25, & \tau > R \end{cases} \quad \text{where} \quad \tau = \frac{\text{mod}(e-1, \frac{T}{M})}{\frac{T}{M}}, \tag{4}$$

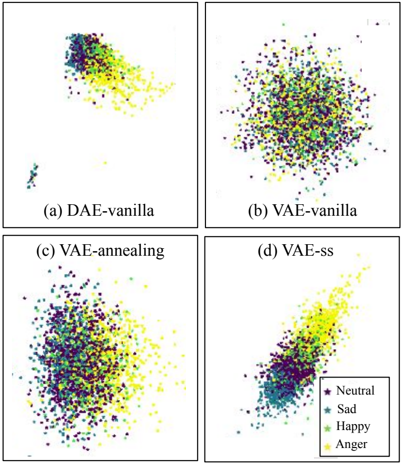


Formulation

- VAE with semi-supervision:

$$\arg\min_{\theta,\phi} \quad \mathcal{L}_{\text{rec}} + \beta_e \mathcal{L}_{\text{KL}} + \gamma \mathcal{L}_{\text{clus}},$$

$$\mathcal{L}_{\text{clus}} = \frac{D_{\text{intra}}}{D_{\text{inter}}} = \frac{\sum_{k=1}^{K} \sum_{\forall i \in k} D(\mathbf{z_i}, \overline{\mathbf{z}}^{\mathbf{k}})}{\sum_{k=1}^{K-1} \sum_{j=k+1}^{K} D(\overline{\mathbf{z}}^{\mathbf{k}}, \overline{\mathbf{z}}^{\mathbf{j}})}, \tag{5}$$



Architecture

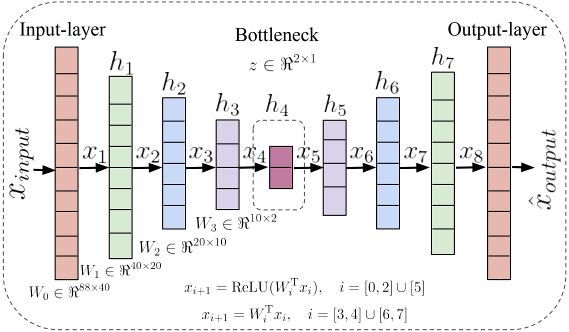


Figure: Illustration of the architecture employed for all the models explored in this work.

- Training: 50 epochs, batch size 64, Adam optimizer (learning rate: 1e-3).
- Latent embedding used as input features to a linear SVC.

Evaluation

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

Results: Classification performance

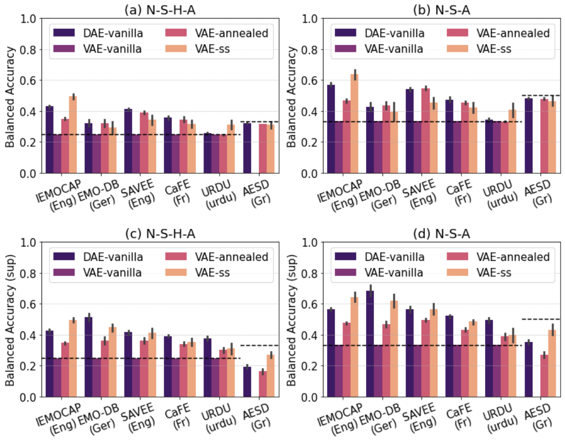


Figure: (1) Balanced accuracy on unseen transfer data sets using (a) 4 emotion classes, (b) 3 emotion classes; balanced accuracy with access to 20% of the unlabeled transfer data sets with (c) 4 emotions and (d) 3 emotion classes.

Results: Consistency of latent space

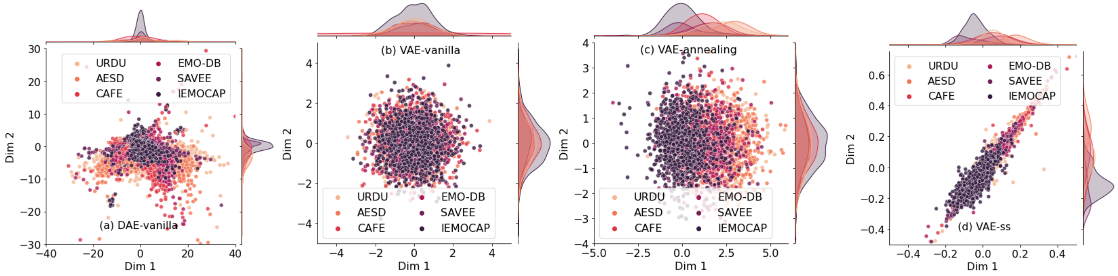
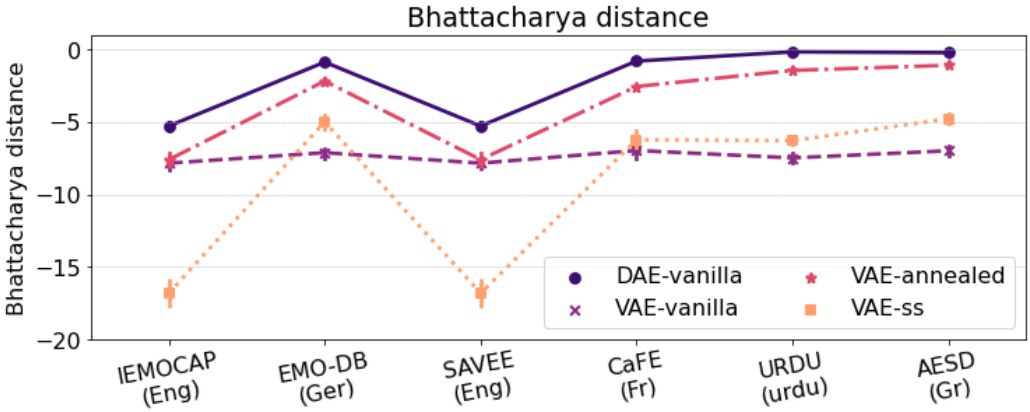


Figure: Scatter plots depicting the overlap between the latent embedding obtained from the methods investigated for all the transfer data sets.

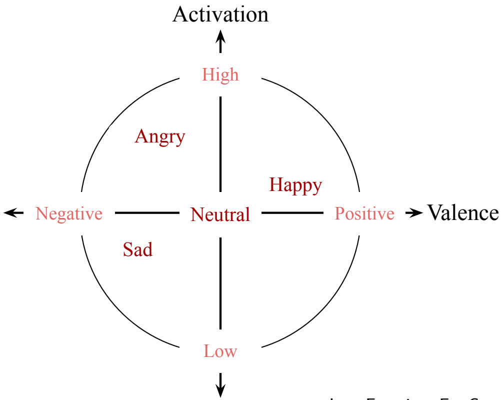
Results: Consistency of latent space



Semi-supervision with continuous metric-loss

Motivation: Dimensional model of emotions!

- Goal: Semi-supervised DAE→to shape the latent space with emotion-relevant information.
- Challenges: Labels, Continuous metric learning functions?
- Discussion: Method for continuous metric learning to order samples in latent space.



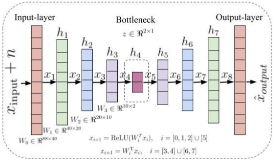
Audio-features → Feature-embedding → Emotion-recognition

Transferable feature embedding

Denoising autoencoder (DAE)

$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Rec}}$

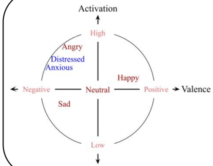
Discriminating subspace



Semi-supervision (metric learning)

$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Rec}} + \mathcal{L}_{\text{Met}}$

Emotion relevant embedding



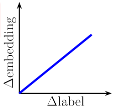
Dimensional emotion model (Activation, valence)

1. Consistent over languages, in contrast to emotion class labels
2. Allows to model emotion classes that does not exist in training dataset.

Continuous metric loss for emotion modelling

$\mathcal{L}_{\text{Met}} = \mathcal{L}_{\text{Slope}} + \mathcal{L}_{\text{Res}}$

Label: Activation, valence



Formulation

DAE:

$$\underset{f_\theta, g_\phi}{\arg\min} \quad \mathcal{L}_{\text{rec}} = \mathbb{E}\|\mathbf{x} - g_\phi(f_\theta(\mathbf{x_n}))\|_2^2, \tag{6}$$

DAE with metric-loss

$$\underset{f_\theta, g_\phi}{\arg\min} \quad \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{met}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{res}} + \mathcal{L}_{\text{sl}}, \tag{7}$$

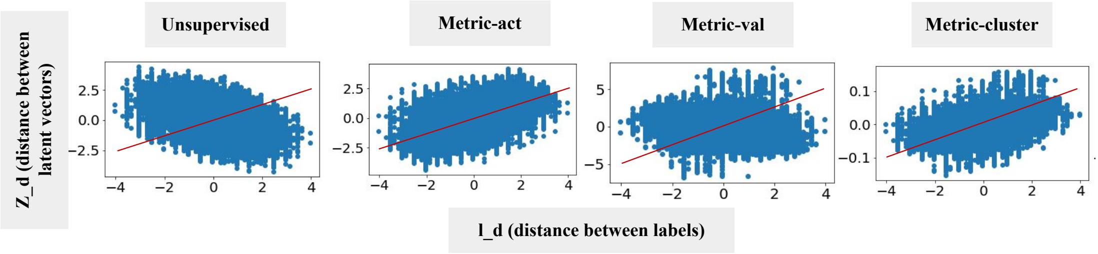
$$\mathcal{L}_{\text{res}} = \mathbb{E}\|\mathbf{z_d} - \hat{\mathbf{z}}_\mathbf{d}\|_2^2, \quad \hat{\mathbf{z}}_\mathbf{d} = p\mathbf{l_d}, \quad \mathbf{l_d} = d(l_i, l_{i+1}) \tag{8}$$

$$p = (\mathbf{l_d}^T \mathbf{l_d})^{-1} \mathbf{l_d}^T \mathbf{z_d} \tag{9}$$

$$\mathcal{L}_{\text{sl}} = \left\| \frac{\hat{\mathbf{z}}_\mathbf{d}(a_1) - \hat{\mathbf{z}}_\mathbf{d}(a_2)}{\mathbf{l_d}(a_1) - \mathbf{l_d}(a_2)} - 1 \right\|_2, \tag{10}$$

Method	R^2-Act ($\mu \pm \sigma$)	R^2-Val ($\mu \pm \sigma$)
Unsupervised	0.11±0.06	0.03±0.02
Metric-act	**0.21 ± 0.05**	**0.06 ± 0.02**
Metric-val	0.12± 0.05	0.05±0.02

Table: Adjusted squared correlation coefficient presenting the linear dependence of $\mathbf{z_d}$ on $\mathbf{l_d}$ for the three models. Mean and standard deviation over five folds are presented.



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

Reference methods

- DAE unsupervised: Correlation and classification
- Supervised SVC: Classification
- SUPERB model: Classification
- Semi-supervision with the transfer dataset labels: Correlation

Correlation analysis

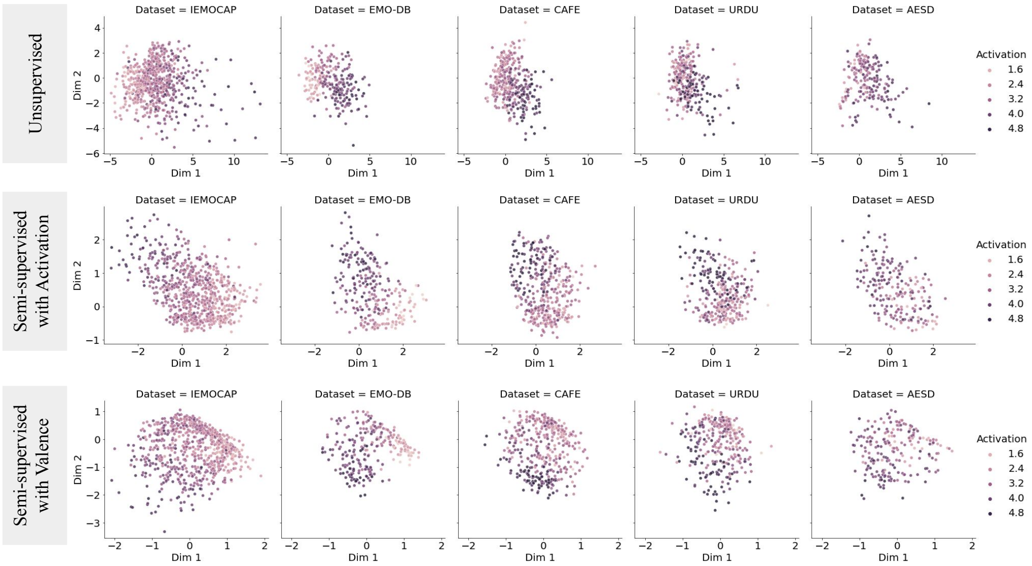
Method (DAE)	IEMOCAP		EMO-DB		CAFE		URDU		AESD	
	R^2-Act	R^2-Val	R^2-Act	R^2-Val	R^2-Act	R^2-Val	R^2-Act	R^2-Val	R^2-Act	R^2-Val
Metric-act (supervised)	NA	NA	0.38 ± 0.05	0.16 ± 0.04	0.62 ± 0.01	0.16 ± 0.01	0.34 ± 0.05	0.15 ± 0.04	0.44 ± 0.03	0.18 ± 0.01
Metric-val (supervised)	NA	NA	0.45 ± 0.03	0.21 ± 0.03	0.44 ± 0.05	0.29 ± 0.06	0.32 ± 0.06	0.16 ± 0.04	0.4 ± 0.06	0.17 ± 0.03
DAE-Unsupervised	0.41 ± 0.04	0.06 ± 0.02	$\mathbf{0.63 \pm 0.04}$	0.05 ± 0.04	0.41 ± 0.03	0.14 ± 0.02	0.28 ± 0.05	0.14 ± 0.03	0.3 ± 0.01	$-0.0 \pm 0.0^*$
DAE-Metric-act	$\mathbf{0.49 \pm 0.02}$	0.05 ± 0.01	$\mathbf{0.63 \pm 0.04}$	0.04 ± 0.02	$\mathbf{0.46 \pm 0.02}$	0.13 ± 0.03	0.32 ± 0.06	0.13 ± 0.02	$\mathbf{0.31 \pm 0.05}$	$-0.0 \pm 0.0^*$
DAE-Metric-val	0.39 ± 0.03	$\mathbf{0.11 \pm 0.01}$	0.61 ± 0.03	$\mathbf{0.1 \pm 0.04}$	0.43 ± 0.02	$\mathbf{0.15 \pm 0.01}$	$\mathbf{0.38 \pm 0.01}$	$\mathbf{0.17 \pm 0.03}$	0.27 ± 0.03	$0.01 \pm 0.01^*$

Table: Adjusted squared correlation coefficient presenting the linear dependence of l on z, the activation and valence labels for the three models. Mean and standard deviation over five folds are presented.

Classification

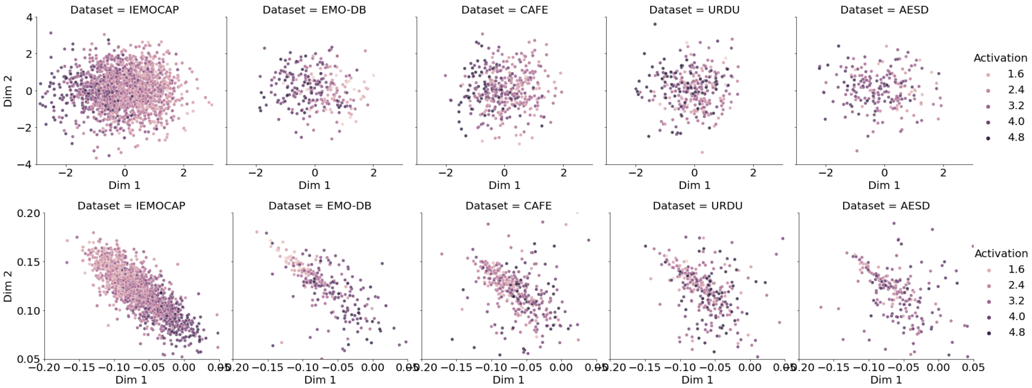
Method	IEMOCAP ($\mu \pm \sigma$)		EMO-DB ($\mu \pm \sigma$)		SAVEE ($\mu \pm \sigma$)		CAFE ($\mu \pm \sigma$)		URDU ($\mu \pm \sigma$)		AESD ($\mu \pm \sigma$)	
	N-S-A	N-S-H-A	N-S-A	N-S-H-A	N-S-A	N-S-H-A	N-S-A	N-S-H-A	N-S-A	N-S-H-A	S-A	S-H-A
SVC (supervised)	0.65 ± 0.02	0.65 ± 0.02	0.89 ± 0.03	0.68 ± 0.03	0.74 ± 0.03	0.68 ± 0.05	0.66 ± 0.03	0.51 ± 0.03	0.89 ± 0.03	0.82 ± 0.02	0.94 ± 0.03	0.7 ± 0.06
SUPERB ($> 3 \times 10^8$)	**0.79**	**0.79**	0.57	**0.66**	**0.7**	**0.68**	0.39	**0.51**	0.26	0.39	0.34	**0.53**
DAE-Unsupervised†	0.51 ± 0.02	0.51 ± 0.02	0.72 ± 0.06	0.56 ± 0.05	0.59 ± 0.02	0.49 ± 0.02	0.43 ± 0.0	0.32 ± 0.01	0.51 ± 0.05	0.38 ± 0.03	0.4 ± 0.05	0.22 ± 0.03
DAE-Metric-act‡	0.54 ± 0.02	0.54 ± 0.01	0.74 ± 0.04	0.57 ± 0.04	0.58 ± 0.02	0.46 ± 0.03	$\mathbf{0.46 \pm 0.04}$	0.33 ± 0.02	0.55 ± 0.01	0.41 ± 0.03	$\mathbf{0.44 \pm 0.02}$	0.27 ± 0.02
DAE-Metric-val‡	0.54 ± 0.01	0.54 ± 0.02	$\mathbf{0.78 \pm 0.03}$	0.61 ± 0.03	0.61 ± 0.05	0.49 ± 0.02	0.45 ± 0.01	0.34 ± 0.02	$\mathbf{0.6 \pm 0.02}$	$\mathbf{0.43 \pm 0.02}$	0.42 ± 0.02	0.25 ± 0.02
($< 4 \times 10^2$ parameters)												

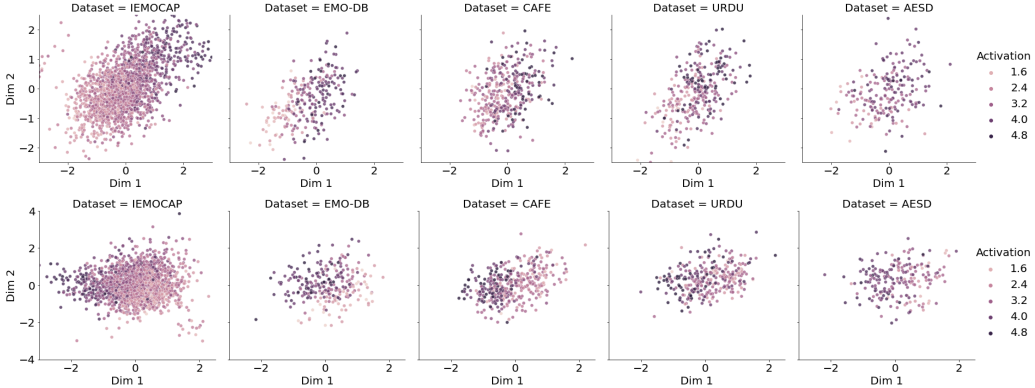
Table: Balanced classification accuracy for (a) three emotion classes (neutral, sad, anger) and (b) four emotion classes (neutral, sad, happy, anger) presented using mean and standard deviation ($\mu \pm \sigma$) computed over 5-fold cross validation. † and ‡ represents the baseline and proposed methods, respectively. Complexity of SUPERB and proposed models are presented in parentheses.



Unsupervised
Semi-supervised with Activation
Semi-supervised with Valence
Dataset = IEMOCAP
Dataset = EMO-DB
Dataset = CAFE
Dataset = URDU
Dataset = AESD
Dim 1
Dim 2
Activation
1.6
2.4
3.2
4.0
4.8

Scatter-plots of VAE latent space





Dataset = IEMOCAP
Dataset = EMO-DB
Dataset = CAFE
Dataset = URDU
Dataset = AESD
Dim 1
Dim 2
Activation
1.6
2.4
3.2
4.0
4.8

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	0.76 ± 0.05	0.76 ± 0.05	0.53 ± 0.08	0.61 ± 0.04	0.35 ± 0.04	0.39 ± 0.03	0.38 ± 0.05	0.39 ± 0.05	0.31 ± 0.01	0.31 ± 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

Table: Spearman's rank order correlation for VAEs with different losses.

Conclusions

Cluster-loss:

1. DAE: highest classification accuracy, worst distribution consistency.
2. VAE-vanilla: best consistency, classification accuracy random.

Continuous-metric loss:

1. Proposed metric loss works (Activation as self-supervision)
2. Our formulation seems to be able to model activation in the latent space $\rightarrow$ different approach necessary for valence.
3. Continuous metric loss seems better model emotion representations over language (correlation).

References

1. *Towards Transferable Speech Emotion Representation: On loss functions for cross-lingual latent representations. ICASSP, May 2022*
   Sneha Das, Nicole Nadine Lønfeldt, Anne Katrine Pagsberg, Line H. Clemmensen

2. *Continuous Metric Learning For Transferable Speech Emotions Recognition and Embedding Across Low-resource Languages. NLDL, Jan 2022*
   Sneha Das, Nicklas Leander Lund,Nicole Nadine Lønfeldt, Anne Katrine Pagsberg, and Line H. Clemmensen

Thankyou!