Towards Transferable Speech Emotion Representation: On Loss Functions For Cross-Lingual Latent Representations

Sneha Das\textsuperscript{1}, Nicole Nadine Lønfeldt\textsuperscript{2}, Anne Katrine Pagsberg\textsuperscript{2,3}, Line H. Clemmensen\textsuperscript{1}

\textsuperscript{1}Department of Applied Mathematics and Computer Science, Technical University of Denmark
\textsuperscript{2}Child and Adolescent Mental Health Center, Copenhagen University Hospital, Capital Region
\textsuperscript{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.
- Existing challenges: Generalizing over languages, corpora, recording condition (under low-resource conditions).
Objectives and Contributions

Objectives for transferability:

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

Contributions:

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:

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

• VAE:

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

• VAE with KL-annealing:

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

where the standard formulation of \( \beta_e \):

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

(a) DAE-vanilla
(b) VAE-vanilla
(c) VAE-annealing
(d) VAE-ss
Formulation

- VAE with semi-supervision:

\[ \arg \min_{\theta, \phi} \mathcal{L}_{\text{rec}} + \beta e \mathcal{L}_{\text{KL}} + \gamma \mathcal{L}_{\text{clus}}, \]

\[ \mathcal{L}_{\text{clus}} = \frac{D_{\text{intra}}}{D_{\text{inter}}} = \frac{\sum_{k=1}^{K} \sum_{i \in k} D(z_i, z^k)}{\frac{K-1}{K} \sum_{k=1}^{K} \sum_{j=k+1}^{K} D(z^k, z^j)}, \quad (5) \]
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
Conclusions

1. DAE: highest classification accuracy, worst distribution consistency.
2. VAE-vanilla: best consistency, classification accuracy random.
3. VAE-ss: Classification accuracy similar to DAE and distribution consistency similar to VAE-vanilla.