Continuous Metric Learning For Transferable Speech Emotion Recognition and Embedding Across Low-resource Languages

Sneha Das[1], Nicklas Leander Lund[1],
Nicole Nadine Lønfeldt[2], Anne Katrine Pagsberg[2,3], and 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

Abstract

Speech emotion recognition (SER) refers to the technique of inferring the emotional state of an individual from speech signals. SERs continue to garner interest due to their wide applicability. Although the domain is mainly founded on signal processing, machine learning, and deep learning, generalizing over languages continues to remain a challenge. However, developing generalizable and transferable models are critical due to a lack of sufficient resources in terms of data and labels for languages beyond the most commonly spoken ones. To improve performance over languages, we propose a denoising autoencoder with semi-supervision using a continuous metric loss based on either activation or valence. The novelty of this work lies in our proposal of continuous metric learning, which is among the first proposals on the topic to the best of our knowledge. Furthermore, to address the lack of activation and valence labels in the transfer datasets, we annotate the signal samples with activation and valence levels corresponding to a dimensional model of emotions, which were then used to evaluate the quality of the embedding over the transfer datasets[1]. We show that the proposed semi-supervised model consistently outperforms the baseline unsupervised method, which is a conventional denoising autoencoder, in terms of emotion classification accuracy as well as correlation with respect to the dimensional variables. Further evaluation of classification accuracy with respect to the reference, a BERT based speech representation model, shows that the proposed method is comparable to the reference method in classifying specific emotion classes at a much lower complexity.

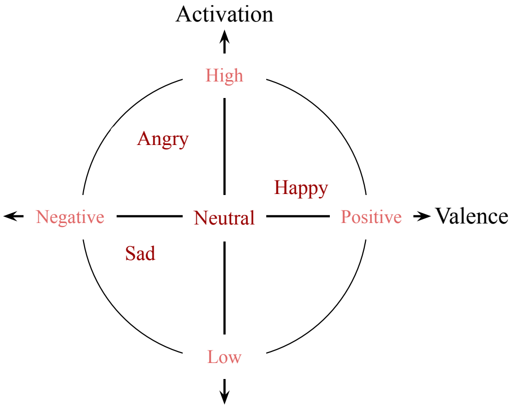

Figure 1: Illustration of the circumplex dimensional model of emotion [25].

1 Introduction

Speech emotion recognition (SER) is the process of inferring the emotional state from speech signals. The domain has found applicability in diverse fields, in healthcare in the detection of disorders, risk assessment in criminal justice system, monitoring the attentiveness of students in school, etc.

Methods for SER have evolved from solely signal processing and conventional machine learning methods to using more advanced deep learning. Notable methods employed the use of hidden Markov models and Gaussian mixture models for emotion classification. Support vector machines have been a widely used tool for SER,

[1] The labels are available at:https://bit.ly/3rg6VsA

https://doi.org/10.7557/18.6300

either as a standalone tool or in tandem with other methods to improve classification [26]. Furthermore, advanced machine learning methods including deep learning models have recently shown promise for SER. This includes the use of CNNs, RNNs and LSTMs [30].

An enduring struggle of SER models is their ability to generalize over languages. This is addressed by using supervision when the languages are supported with sufficient resources, both in data and labels. However, supervised learning is inapplicable to languages beyond the commonly spoken ones, due to insufficient resources in terms of small data sets and few-to-no labels. Therefore, to cater to low-resource languages, it is necessary to develop methods that generalize better over languages. The associated challenge is the subjectivity in emotion perception and classification. For instance, the perception of speech signals demonstrating a neutral emotion can vary between languages, owing to phonetic and cultural differences. Despite this, most models use class labels to train models, instead of employing a more universal and continuous metric like activation and valence from the dimensional model of emotions [25].

Latent representation methods, like the autoencoder (AE), can compress the input features to a smaller and ideally more target relevant latent embedding and are commonly employed in various applications, for instance biosignal processing, computer-vision, and speech-processing [33]. Furthermore, these methods are relatively more interpretable. This is a favourable characteristic as the decision making process of the model is more transparent thereby enabling smoother deployment of such systems in medical and clinical setup. Therefore, latent representation methods can also be useful in the modelling of emotions from speech signals, such that the models retain only the emotion-relevant paralinguistic content from the speech signal. This may also lead to better knowledge transfer between data sets by transferring only generic emotion representations to unseen languages and corpora and not the syntactical variations between languages. Transferable emotion representations aid in addressing the acute label shortage for SER tasks in under-resourced languages.

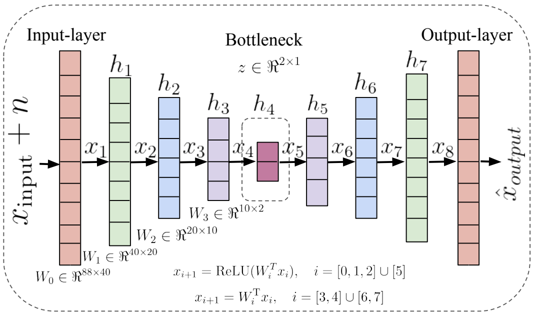
Input-layer
Bottleneck
Output-layer
$z \in \Re^{2\times 1}$
$x_{\text{input}} + n$
h_1
h_2
h_3
h_4
h_5
h_6
h_7
x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
$\hat{x}_{output}$
$W_0 \in \Re^{88\times 40}$
$W_1 \in \Re^{40\times 20}$
$W_2 \in \Re^{20\times 10}$
$W_3 \in \Re^{10\times 2}$
$x_{i+1} = \text{ReLU}(W_i^T x_i), \quad i = [0,1,2] \cup [5]$
$x_{i+1} = W_i^{\text{T}} x_i, \quad i = [3,4] \cup [6,7]$

Figure 2: Illustration of the (denoising) autoencoder architecture.

In this paper, we use the denoising autoencoder (DAE) to obtain a highly compressed emotion embedding that is more consistent over different languages. Additionally, we strive to keep the system relatively simple and interpretable as the target application of this work is clinical psychiatry. While a conventional DAE will learn a compressed latent space representative of the input features, it is not necessary that the model learns to discriminate between the desired emotions or emotion representation. Therefore, we propose to use semi-supervision to direct the DAE to learn factors relevant to emotion discrimination like activation and valence, wherein the semi-supervision is provided by the activation and valence labels. We address the following questions in the process: 1. How to provide semi-supervision to preserve the label distances between data samples in the latent space, where the labels span a continuous space. In contrast, most works on metric learning, like triplet loss and contrastive loss are discrete in nature. 2. How to validate the embedding quality in the absence of dimensional labels (activation and valence) for the transfer datasets considered in this paper. Therefore, the contributions of this work are: 1. We propose a method for continuous metric learning to order the samples in the latent space. To the best of our knowledge, this is one of the few methods addressing continuous metric learning. 2. We annotate the transfer datasets with the activation and valence labels that we use to validate the effectiveness of the proposed methods over languages. The labels will be shared for research purposes.

2 Relation To Prior Work

Autoencoders and variants: DAE was one of the earliest deep learning based unsupervised learning techniques for SER [31]. This was followed by the use of sparse AE for feature transfer [7] and for SER on spontaneous data sets [8]. Furthermore, end-to-end representation learning for affect recognition from speech was proposed and showed performance comparable to existing methods [13]. In recent years, techniques like variational and adversarial AEs and adversarial variational Bayes have been exploited to learn the latent representations of speech emotions with input features ranging from the raw signals to hand crafted features [19, 24, 9, 23].

Metric learning in SER: In a recent paper on SER, the authors proposed a convolutional autoencoder that employs a pre-trained autoencoder and a convolutional neural network, and a triplet loss in the autoencoder is used for metric learning [12]. Further on, a contrastive loss was used for metric learning in a Siamese network [21]. On similar lines, a class specific triplet-loss based LSTM neural network was proposed for SER [16]. A combination of the centre-loss, that clusters members of a class together, and the cross-entropy loss was investigated to better cluster features [5, 22]. The authors proposed multiple f-similarity preserving losses for metric learning using soft labels and tested it on classification [34].

3 Methodology

Dimensional Model of Emotion: We consider the circumplex model of emotion, wherein Russell et al proposed that emotions can be represented in a circular space, wherein the x-axis corresponds to valence, and y-axis corresponds to the activation [25], as illustrated in Fig. 1. Activation refers to how arousing an emotion is and valence refers to the level of positivity in the emotion.

3.1 Denoising autoencoder with continuous metric learning

We represent the learning function in a denoising autoencoder as:

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

where $\mathbf{x}$ and $\mathbf{x_n}$ are the clean and partially noisy feature vectors, f_θ represents the encoder, g_ϕ the decoder, and $\mathcal{L}_{\text{rec}}$ is the reconstruction loss [28]. It is known to be more robust than the AE, since it is designed to learn a subspace of the clean feature vectors from a noisy input, therefore learning the characteristics of the desired input.

Let $\mathbf{z} \in \mathbb{R}^{b\times 2}$ be the latent space embedding, and $\mathbf{l} \in \mathbb{R}^{b\times 1}$ be the label we are modelling. Then $\mathbf{z_d} = d(z_i, z_{i+1})$, such that $\mathbf{z_d} \in \mathbb{R}^{(b-1)\times 1}$, where $d(a_1, a_2)$ is the Euclidean distance between the vectors a_1 and a_2, and $\mathbf{z_d}$ is the distance between two latent samples. Similarly, $\mathbf{l_d} = d(l_i, l_{i+1}) \in \mathbb{R}^{(b-1)\times 1}$, is the distance between the labels of the data samples, and $i = 1,2,...,(b-1)$. We assume that $\mathbf{z_d}$ is a linear function of $\mathbf{l_d}$ such that $\hat{\mathbf{z}}_\mathbf{d} = p\mathbf{l_d}$; we obtain the optimum p by minimizing the squared error $\|\mathbf{z_d} - \hat{\mathbf{z}}_\mathbf{d}\|_2^2$, which yields:

$$p = (\mathbf{l_d}^T \mathbf{l_d})^{-1} \mathbf{l_d}^T \mathbf{z_d} \tag{2}$$

As motivated in the Sec. 1, our goal is to obtain a latent space through training, wherein the distance between the embedding of two samples is close to the distance between the labels of the corresponding samples. We can achieve this by 1. obtaining a slope that is close to one, and 2. minimizing the residual between $\mathbf{z_d}$ and $\hat{\mathbf{z}}_\mathbf{d}$. Therefore, the loss factor corresponding to the slope between $\mathbf{z_d}$ and $\mathbf{l_d}$ is given as:

$$\mathcal{L}_{\text{sl}} = \left\| \frac{\hat{\mathbf{z}}_\mathbf{d}(a_1) - \hat{\mathbf{z}}_\mathbf{d}(a_2)}{\mathbf{l_d}(a_1) - \mathbf{l_d}(a_2)} - 1 \right\|_2, \tag{3}$$

where a_1, a_2 are two arbitrary data instances. Furthermore, the residual component is the standard mean squared error:

$$\mathcal{L}_{\text{res}} = \mathbb{E}\|\mathbf{z_d} - \hat{\mathbf{z}}_\mathbf{d}\|_2^2. \tag{4}$$

Therefore, our proposed total metric loss is composed of:

$$\mathcal{L}_{\text{met}} = \mathcal{L}_{\text{res}} + \mathcal{L}_{\text{sl}} \tag{5}$$

The final equation for the DAE with metric learning is given as:

$$\underset{f_\theta, g_\phi}{\arg\min} \quad \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{met}}, \tag{6}$$

where $\mathcal{L}_{\text{rec}}$ has the same formulation as in Eq. 1

3.2 Pair-wise distance preservation

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 1: 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.

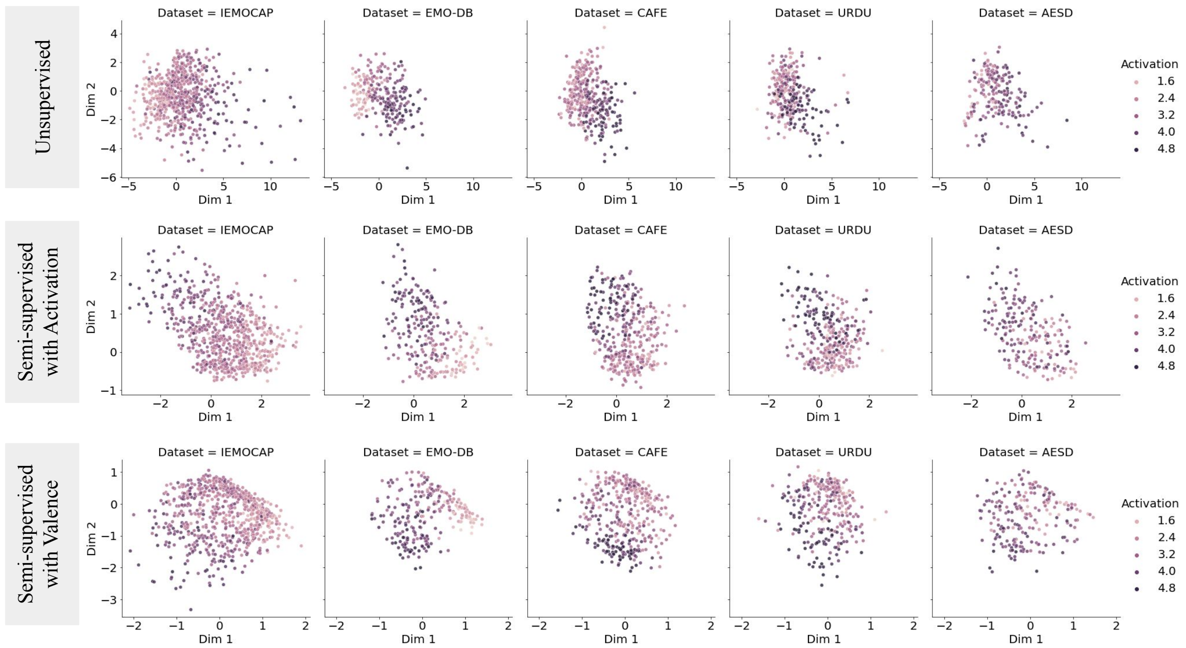

Figure 3: Latent embedding for DAE-unsupervised (Top), DAE-metric-act (Middle) and DAE-metric-val (Lower) over five datasets, color-coded by activation levels. The models have been trained on IEMOCAP only.

To obtain an insight on the effectiveness of the proposed metric loss, we study the correlation between $\mathbf{z_d}$ and $\mathbf{l_d}$ by using ordinary least squares (OLS) to model $\mathbf{l_d}$ as a function of $\mathbf{z_d}$ as follows: $\mathbf{l_d} = c + \beta_1 \mathbf{z_{d_1}} + \beta_2 \mathbf{z_{d_2}}$. The assumption of a linear relation between the pair-wise point distances in the latent space and the labels is motivated from our proposed metric loss. The R^2 adjusted is presented in Table 1.

4 Experimental setup

In this section, we describe the dataset and the input feature space, the architecture of the models, the pre-processing steps followed by the description of the methods of comparison and our evaluation setup.

Datasets and input features: IEMOCAP, an audio-visual affect data set, is used to train and validate the models [4]. The data set comprises of annotations representing both the categorical and dimensional emotional model [2]. The models are trained with data from the emotional categories *neutral (N), sad (S), happy (H), angry (A)*. We use the extended Geneva minimalistic acoustic parameter set (eGeMAPS) [11], specifically the functionals of lower-level features. Each speech sample yields a feature vector comprising of 88 features and we use the OpenSmile toolkit to extract the features [10]. To study how the latent representations are transferred between corpora and languages, we test classification accuracy over the following *transfer datasets* : 1. the Surrey Audio-Visual Expresses Emotion (SAVEE) database that is primarily English and consists of male speakers only, 2. the Berlin Database of Emotional Speech (Emo-DB) recorded in German and, 3. the Canadian French Emotional (CAFE) speech database comprising of French audio samples [17, 3, 14], 4. the URDU-dataset consisting of Urdu speech, and 5. the Acted Emotional Speech Dynamic (AESD) Database comprising of Greek speech [18, 29]. Note that AESD does not comprise of speech samples from the neutral class, whereby evaluation includes AESD speech samples from three and two classes only. For correlation analysis, due to the lack of activation and valence labels, we annotate the following datasets: 1. EMO-DB, 2. URDU 3. AESD and 4. CAFE. These labels are then used to test the baseline DAE-unsupervised, and the proposed DAE-metric-act and DAE-metric-val models.

System architecture: Past works using AEs and variants have mostly addressed novel network architectures for better classification accuracy [8, 19, 23, 24, 31]. However, since the focus of this work is to investigate the potential of metric learning in obtaining a more transferable embedding space for emotion recognition, we employ a simple architecture for the DAE baseline with performance that is comparable to existing methods. The proposed methods share the same architecture as the baseline model, illustrated in Fig. 2. The size of the input features is 88 and we compress the latent space to two dimensions. As described above, the input feature vector is comprised of the descriptive statistics of each speech signal, without temporal correlation. Therefore, in the encoder and decoder we employ fully connected NNs, instead of convolutional or recurrent NNs. The fully connected layers are followed by rectified linear units (ReLU) to incorporate non-linearity within the model.

Preprocessing: Prior to using the data sets for training and testing, we remove the outliers by computing the z-score and eliminating the data samples that have a z-score, $-10 > z > 10$. We chose a threshold of 10 instead of the standard value of 3 because the goal of this work is to understand the behavior of the models for both typical and atypical rendition of emotions in speech. Therefore, we only remove the extreme outliers. For run-time evaluation over the transfer datasets, we employ statistics from 20% of the transfer data set to standardize the remaining 80%.

Proposed methods: We train and validate the following models: 1. DAE-Unsupervised is used as a baseline, 2. DAE-Metric-Act that utilizes the proposed semi-supervision via the activation labels from Sec. 3.1, and 3. DAE-Metric-Val utilizes the proposed semi-supervision via the valence labels (Sec. 3.1). The input features to the DAE is corrupted by a noise component, $x_{\text{input}} = x_{\text{true}} + N$ and $N \in \mathcal{N}(0,1)$. We use the mean squared error (MSE) to optimize the baseline model as shown in Eq: 1. To study the consistency of the results, 5-fold cross-validation is used on the IEMOCAP database while the transfer data sets are identical over the iterations and are used for model testing only. The models were trained over 50 epochs with a batch size of 64, and we used the Adam optimizer with the learning rate set to 1e-3. Additional methods used for reference are listed in the following section.

Reference methods: In addition to comparing the performance of the DAE models with the proposed metric loss to the unsupervised DAE, we employ the following as reference methods to gauge the relative performance of the models developed in this work:

1. Since the DAE-unsupervised model is employed as a baseline to evaluate the effectiveness of the metric-loss, we compared its classification performance with similar methods from literature as presented in Table 2.

2. We implemented the support vector classifier, SVC, trained separately on all data sets towards the downstream task of classifying the input eGeMAPS features into target classes. The results are shown in Table 4. We use the supervised SVC to study the upper-bound of the performance limits of the SER task. However, the work in this paper is focused on enabling reliable SER for languages with *few or no* labelled data, and supervised learning is therefore inapplicable within this scope.

3. We used the SUPERB: Speech processing Universal PERformance Benchmark for emotion classification on the transfer datasets [32]. The method employs the HuBERT (Hidden-Unit BERT), specifically hubert-large-ll60k model, as the base model to first extract speech representations [15]. The speech representations are then given as input to a linear downstream task to classify the speech signal into emotions. The base model is trained on English data sets and the linear downstream model for emotion classification is trained on the IEMOCAP dataset. We employ the model for classification only as the model is specifically trained for that. The results are shown in Table 4. Relative to the complexity of the models developed in this paper ($< 4 \times 10^2$ parameters), note that the considered model is highly complex ($> 3 \times 10^8$ parameters).

4. Lastly, besides comparing the correlation between the embedding and the labels for the proposed models and the unsupervised model, we employ supervision to train models with the proposed metric losses. In other words, we use the labels from the transfer data sets in the metric loss during training. We do this to obtain insights on the performance limits of the proposed method and the results are presented in Table 3 as metric-act (supervised) and metric-val (supervised).

Experiments for evaluation: To investigate the efficacy of the proposed methods, in the following parts we investigate the quality of the latent embedding by using them as the input features for classification and correlation downstream tasks. Towards that, we evaluate the models in terms of 1. the correlation coefficient, and 2. the classification accuracy.

1. Correlation analysis : We study the correlation between the latent dimensions and dimensional variables (activation, valence). With the main focus to study the effectiveness of the proposed continuous metric learning method using the dimensional variables, we evaluate how large a proportion of the labels can be explained by the latent dimensions. Therefore, as described in Sec. 3.2, we compute the correlation coefficients between $\mathbf{z}$ and l, via the adjusted R^2. The mean and standard deviation over 5-folds for R^2 for the models are shown in Table 3; Note that the R^2, where the p-value of the F-statistic > 0.05, are indicated by an asterisk. Models with larger R^2 are more effective in preserving the distance between samples in the embedding space with respect to the dimensional variables.

2. Classification of emotion classes : We inspect the performance of the methods by classifying the speech samples into emotional categories using the support vector classifier (SVC) with a linear kernel. For evaluation, we use balanced accuracy for the 4-class (N-S-H-A) and 3-class (N-S-A) scenarios to account for imbalanced classes. Furthermore, supervised training on eGeMAPS using SVC and the SUPERB model are employed as references, as described above. The results are presented in Table 4.

5 Results and discussion

Method	Features+Dataset	classes	Accuracy
GAN [20]	eGeMAPS [11]+EMO-DB	2	66% (UAR)
FLUDA [1]	IS10 [27]+IEMOCAP(+)	4	50% (UA)
VAE+LSTM [19]	LogMel+IEMOCAP	4	56.08% (UA)
AE+LSTM [19]	LogMel+IEMOCAP	4	55.42% (UA)
Stacked-AE+BLSTM-RNN [13]	COVAREP+IEMOCAP [6]	4	50.26% (UA)
DAE+Linear-SVM **(baseline)**	eGeMAPS+IEMOCAP	4	52.09% (UA)

Table 2: Performance in terms of unweighted accuracy (UA) and unweighted average recall (UAR) of the reference methods (from cited papers) and the unsupervised baseline model developed in this paper.

In Table. 2, we list current models similar to the DAE-unsupervised baseline in terms of the architecture, network size and input-output format. We observe that the performance of the developed unsupervised model is comparable to state of the art. Therefore, we consider the DAE+Linear-SVM model a reasonable baseline to gauge the performance of the proposed metric-loss in this paper.

Correlation analysis: From Table 3, we observe that the mean and standard deviation of R^2 is consistently higher for the proposed methods relative to the unsupervised baseline, over the transfer data sets. However, DAE-unsupervised is as good as DAE-metric-act in terms of the correlation between $\mathbf{z}$ and the activation for EMO-DB. With supervision, as anticipated metric-act and metric-val models seem to have a higher R^2 relative to the unsupervised and semi-supervised models, specifically for the activation variable.
To summarize the observations, the proposed methods show higher adjusted R^2 between the modeled $\mathbf{z}$ and the label considered. However, we observe that R^2 corresponding to the valence variable is lower than the activation variable. An unaccounted non-linear relation between valence variable and the latent space could be a reason for the observation. Nevertheless, both metric-act and metric-val seem to effectively order and arrange data samples in the latent space, relative to the activation label. This is also evident from Fig. 3, wherein we can observe that the samples in the latent space are better distributed for metric-act and metric-val then for the unsupervised model.

Classification of emotion classes: We observe that for both 3- and 4-class scenarios, the proposed methods have higher accuracy than the baseline unsupervised method. Furthermore, while SUPERB outperforms all the methods for IEMOCAP and SAVEE, it seems to have lower accuracy than the baseline and the proposed models for the remaining data sets, specifically for the 3-class classification task. Lastly, although classification accuracy of the supervised SVC is superior and implies that there is class-discriminating information in the data sets, how much of that information corresponds to the paralinguistic aspects of speech and emotion is worth investigating.
The balanced accuracy scores (Table 4) for metric-val and metric-act is consistently better than the unsupervised baseline over the transfer datasets, indicating that incorporating a distance preservation metric in the loss function aids in shaping the distribution of features in the latent space that is more consistent over languages. It is also interesting to note that while metric-val did not show a large difference in R^2 with respect to the baseline (Table 1), in terms of classification accuracy

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	**0.63±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±0.0*
DAE-Metric-act	**0.49±0.02**	0.05±0.01	**0.63±0.04**	0.04±0.02	**0.46±0.02**	0.13±0.03	0.32±0.06	0.13±0.02	**0.31±0.05**	−0.0±0.0*
DAE-Metric-val	0.39±0.03	**0.11±0.01**	0.61±0.03	**0.1±0.04**	0.43±0.02	**0.15±0.01**	**0.38±0.01**	**0.17±0.03**	0.27±0.03	0.01±0.01*

Table 3: 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.

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 (>3×10^8)	**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	**0.46±0.04**	0.33±0.02	0.55±0.01	0.41±0.03	**0.44±0.02**	0.27±0.02
DAE-Metric-val‡	0.54±0.01	0.54±0.02	**0.78±0.03**	0.61±0.03	0.61±0.05	0.49±0.02	0.45±0.01	0.34±0.02	**0.6±0.02**	**0.43±0.02**	0.42±0.02	0.25±0.02
(<4×10^2 parameters)												

Table 4: 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.

it is often better than metric-act. Furthermore, the proposed models are unable to effectively differentiate between anger and happy, whereby its performance drops in 4-class classification. A similar trend was observed for the reference SUPERB model (lower 3-class classification accuracy), despite its much higher complexity. This indicates that more complex models do not necessary lead to learning meaningful representations of the more subtle aspects of speech.

Effectiveness of metric loss function: From Table 1 we observe that metric-act, i.e., the DAE trained with semi-supervision from the activation labels of the IEMOCAP dataset, shows the highest adjusted R^2 value between $\mathbf{l_d}$ and $\mathbf{z_d}$, $\mathbf{l_d}$ corresponding to the activation labels. This suggests that the proposed formulation of the continuous metric loss (Eq. 5) is effective when activation labels are employed for semi-supervision. In contrast, for metric-val wherein $\mathbf{l_d}$ corresponds to valence, the R^2 value is similar to the baseline. A potential reason for this could be that the relation between valence and the embedding is inherently non-linear. This implies that different approaches are necessary to model valence and activation variables. Investigating and including aspects of the true relation between valence, activation and the embedding will be addressed in future work. In conclusion, we can state that using the dimensional variables, activation and valence, to learn emotion representations yield embeddings that are more transferable to unseen datasets and new languages. Furthermore, the proposed continuous metric loss with semi-supervision enables us to incorporate information on the dimensional variables within the model, hence aiding transferability.

6 Conclusion

In this work, we proposed a method for continuous metric learning, such that the difference between two feature points is continuous. We apply the method for speech emotion recognition, wherein we model a DAE that is semi-supervised using the activation and valence labels and the continuous metric loss. Our results show that the embedding from the proposed method is generally more consistent and thereby more transferable to different languages. The proposed methods are evaluated in terms of classification performance, and the proposed models outperform the baseline method on all the transfer datasets. Furthermore, to investigate the correspondence of the latent space to activation and valence variables, we compute the adjusted R^2 that indicates how much variation in the labels can be explained by a linear combination of the latent space. While the R^2 with respect to valence is generally lower than that for activation, between the methods, the proposed models outperform the baseline method for the datasets. Addressing the lack of labels corresponding to activation and valence variables in the transfer dataset, the annotated transfer dataset is the second contribution of this work.

References

[1] Y. Ahn, S. J. Lee, and J. W. Shin. Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation. *IEEE Signal Processing Letters*, 2021. doi: 10.1109/lsp.2021.3086395.

[2] I. Bakker, T. Van Der Voordt, P. Vink, and J. De Boon. Pleasure, arousal, dominance: Mehrabian and russell revisited. *Current Psychology*, 33(3):405–421, 2014. doi: 10.1007/s12144-014-9219-4.

[3] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005. doi: 10.21437/interspeech.2005-446.

[4] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008. doi: 10.1007/s10579-008-9076-6.

[5] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng. Learning discriminative features from spectrograms using center loss for speech emotion recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7405–7409. IEEE, 2019. doi: 10.1109/icassp.2019.8683765.

[6] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE, 2014. doi: icassp.2014.6853739.

[7] J. Deng, Z. Zhang, E. Marchi, and B. Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 humaine association conference on affective computing and intelligent interaction*, pages 511–516. IEEE, 2013. doi: 10.1109/acii.2013.90.

[8] V. Dissanayake, H. Zhang, M. Billinghurst, and S. Nanayakkara. Speech emotion recognition 'in the wild'using an autoencoder. *Proc. Interspeech 2020*, pages 526–530, 2020. doi: 10.21437/interspeech.2020-1356.

[9] S. E. Eskimez, Z. Duan, and W. Heinzelman. Unsupervised learning approach to feature analysis for automatic speech emotion recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5099–5103. IEEE, 2018. doi: 10.1109/icassp.2018.8462685.

[10] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010. doi: 10.1145/1873951.1874246.

[11] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015. doi: 10.1109/taffc.2015.2457417.

[12] Y. Gao, J. Liu, L. Wang, and J. Dang. Metric learning based feature representation with gated fusion model for speech emotion recognition. *Proc. Interspeech 2021*, pages 4503–4507, 2021. doi: 10.21437/interspeech.2021-1133.

[13] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer. Representation learning for speech emotion recognition. In *Interspeech*, pages 3603–3607, 2016. doi: 10.21437/interspeech.2016-692.

[14] P. Gournay, O. Lahaie, and R. Lefebvre. A canadian french emotional speech dataset. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 399–402, 2018. doi: 10.1145/3204949.3208121.

[15] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

[16] J. Huang, Y. Li, J. Tao, Z. Lian, et al. Speech emotion recognition from variable-length inputs with

triplet loss function. In *Interspeech*, pages 3673–3677, 2018. doi: 10.21437/interspeech.2018-1432.

[17] P. Jackson and S. Haq. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*, 2014.

[18] S. Latif, A. Qayyum, M. Usman, and J. Qadir. Cross lingual speech emotion recognition: Urdu vs. western languages. In *2018 International Conference on Frontiers of Information Technology (FIT)*, pages 88–93. IEEE, 2018. doi: 10.1109/fit.2018.00023.

[19] S. Latif, R. Rana, J. Qadir, and J. Epps. Variational autoencoders for learning latent representations of speech emotion: a preliminary study. *Interspeech 2018: Proceedings*, pages 3107–3111, 2018. doi: 10.21437/interspeech.2018-1568.

[20] S. Latif, J. Qadir, and M. Bilal. Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 732–737. IEEE, 2019. doi: 10.1109/acii.2019.8925513.

[21] Z. Lian, Y. Li, J. Tao, and J. Huang. Speech emotion recognition via contrastive loss under siamese networks. In *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data*, pages 21–26, 2018. doi: 10.1145/3267935.3267946.

[22] B. Mocanu, R. Tapu, and T. Zaharia. Utterance level feature aggregation with deep metric learning for speech emotion recognition. *Sensors*, 21(12):4233, 2021. doi: 10.3390/s21124233.

[23] M. Neumann and N. T. Vu. Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7390–7394. IEEE, 2019. doi: 10.1109/icassp.2019.8682541.

[24] S. Parthasarathy, V. Rozgic, M. Sun, and C. Wang. Improving emotion classification through variational inference of latent variables. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7410–7414. IEEE, 2019. doi: 10.1109/icassp.2019.8682823.

[25] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39 (6):1161, 1980.

[26] B. Schuller, G. Rigoll, and M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *2004 IEEE international conference on acoustics, speech, and signal processing*, volume 1, pages I–577. IEEE, 2004. doi: 10.1109/icassp.2004.1326051.

[27] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan. The interspeech 2010 paralinguistic challenge. In *Proc. INTERSPEECH 2010, Makuhari, Japan*, pages 2794–2797, 2010. doi: 10.21437/interspeech.2010-739.

[28] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. doi: 10.1145/1390156.1390294.

[29] N. Vryzas, R. Kotsakis, A. Liatsou, C. A. Dimoulas, and G. Kalliris. Speech emotion recognition for performance interaction. *Journal of the Audio Engineering Society*, 66(6):457–467, 2018. doi: 10.17743/jaes.2018.0036.

[30] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *IEEE Journal of selected topics in signal processing*, 4(5):867–881, 2010. doi: 10.1109/jstsp.2010.2057200.

[31] R. Xia and Y. Liu. Using denoising autoencoder for emotion recognition. In *Interspeech*, pages 2886–2889, 2013. doi: 10.21437/interspeech.2013-256.

[32] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, et al. Superb: Speech processing universal performance benchmark.

Proc. Interspeech 2021, pages 1194–1198, 2021. doi: 10.21437/interspeech.2021-1775.

[33] O. Yildirim, R. San Tan, and U. R. Acharya. An efficient compression of ecg signals using deep convolutional autoencoders. *Cognitive Systems Research*, 52:198–211, 2018. doi: 10.1016/j.cogsys.2018.07.004.

[34] B. Zhang, Y. Kong, G. Essl, and E. M. Provost. f-similarity preservation loss for soft labels: A demonstration on cross-corpus speech emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5725–5732, 2019. doi: 10.1609/aaai.v33i01.33015725.