Exploring zero-shot emotion recognition in speech using semantic-embedding prototypes

Xinzhou Xu, Jun Deng, Nicholas Cummins, Zixing Zhang, Li Zhao, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Xu, Xinzhou, Jun Deng, Nicholas Cummins, Zixing Zhang, Li Zhao, and Björn W. Schuller. 2022. "Exploring zero-shot emotion recognition in speech using semantic-embedding prototypes." *IEEE Transactions on Multimedia* 24: 2752–65. https://doi.org/10.1109/tmm.2021.3087098.

Exploring Zero-Shot Emotion Recognition in Speech Using Semantic-Embedding Prototypes

Xinzhou Xu, Jun Deng, Nicholas Cummins, *Member, IEEE*, Zixing Zhang, *Member, IEEE*, Li Zhao, and Björn W. Schuller, *Fellow, IEEE*

***Abstract*—Speech Emotion Recognition (SER) makes it possible for machines to perceive affective information. Our previous research differed from conventional SER endeavours in that it focused on recognising unseen emotions in speech autonomously through machine learning. Such a step would enable the automatic leaning of unknown emerging emotional states. This type of learning framework, however, still relied on manual annotations to obtain multiple samples of each emotion. In order to reduce this additional workload, herein, we propose a zero-shot SER framework employing a per-emotion semantic-embedding paradigm to describe emotions in zero-shot SER, instead of using the sample-wise descriptors. Aiming to optimise the relationship between emotions, prototypes, and speech samples, this framework includes two types of learning strategies: Sample-wise learning and emotion-wise learning. These strategies apply a novel learning process to speech samples and emotions, respectively, via specifically designed semantic-embedding prototypes. We verify the utility of these approaches by performing an extensive experimental evaluation on two corpora on three aspects, namely the influence of different types of learning strategies, emotional-pair comparison, and the selections of semantic-embedding prototypes and paralinguistic features. The experimental results indicate that it is applicable to use semantic-embedding prototypes for zero-shot emotion recognition in speech, despite the influence of choosing optimal strategies and prototypes.**

***Index Terms*—Speech emotion recognition, paralinguistics, zero-shot learning, semantic-embedding prototypes**

I. Introduction

PARALINGUISTICS include the field of *Speech Emotion Recognition* (SER) for affective computing regarding speech signals [1]–[3]. Current SER research focuses, mainly, on the search for suitable features, approaches, and models, which can result in close to ideal performances [4]–[7]. Apart from SER's basic learning object, various other learning cases have been investigated. These include semi-supervised learning [8], cross-domain transfer [9]–[12] based on domain adaptation [13], [14], cross-modal transfer [15], and multi-task learning [16]–[19].

What these techniques have in common is that they all fail to learn an emotion which is unfamiliar or even never seen in training examples. Examples for unfamiliar or unknown emotions are not uncommon; they may appear in mining mixed emotions [20], defining new emotions [21], and learning minor emotions [22], where it is difficult to provide well-coordinated examples for these sorts of emotions (referred to as 'unseen emotions'). As a typical example in human-machine interaction, a machine may expect to decide on whether a speaker is trustable, friendly, aggressive, or violent when receiving an utterance. However, it will be unable to perform this task if it has not been taught how to estimate the speaker's complex implicit intention. A small number of existing works have made attempts to address this, e. g., [23].

In order to deal with recognising unseen-emotional speech samples, we propose an autonomous learning strategy based on zero-shot emotion recognition [24]. *Zero-Shot Learning* (ZSL) has demonstrated high utility in image processing [25]–[28] and affective computing [29]–[31]. We have also demonstrated a zero-shot framework for SER [24]. However, this framework required a high number of fully labelled emotional descriptors or attributes (i.e., emotional dimensions) which construct the label-learning models for seen-emotional samples. This type of per-sample annotation often leads to heavy workload and high cost for annotators to make the emotional descriptors [26]. Further, when employing the annotations to learn emotional labels, it requires additional computational steps to construct the associated label-learning models.

In overcoming these limitations, our framework employs per-emotion semantic-embedding prototypes, instead of the standard per-sample annotation for emotional descriptors. Our framework includes joint learning for acoustic features, emotional labels (for seen-emotional samples), and semantic-embedding prototypes (for seen and unseen emotions) [32]–[34].

Within the proposed framework using semantic-embedding prototypes, the ZSL strategies are categorised into two types of sample-wise learning and emotion-wise learning. The sample-wise learning opts to perform learning on each sample using their paralinguistic features and the duplication of the corresponding per-emotion prototypes. In contrast, emotion-wise learning transforms the samples into emotion-wise forms and constructs the relationship between the form and the prototype on each emotional state.

This work was supported by the Natural Science Foundation of China under Grants No. 61801241, No. 61673108, and No. 61801236, the Natural Science Foundation for Jiangsu Higher Education Institutions under Grant 18KJB510029, the Natural Science Foundation of Jiangsu under Grant BK20180746, the NUPTSF under Grant NY217149, and the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE).

X. Xu is with the School of Internet of Things, Nanjing University of Posts and Telecommunications, P.R. China, and also with the Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany. (e-mail: xinzhou.xu@njupt.edu.cn).

J. Deng is with Agile Robots AG, Germany. (e-mail: jun.deng@tum.de).

N. Cummins is with the Department of Biostatistics and Health informatics, Institute of Psychiatry, Psychology and Neuroscience, Kings College London, London, UK, and with the Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany. (e-mail: nicholas.cummins@ieee.org).

Z. Zhang is with GLAM–the Group on Language Audio & Music, Imperial College London, UK. (e-mail: zixing.zhang@tum.de).

L. Zhao is with School of Information Science and Engineering, Southeast University, P.R. China. (e-mail: zhaoli@seu.edu.cn).

B. W. Schuller is with the Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany, and also with GLAM–the Group on Language Audio & Music, Imperial College London, UK. (e-mail: schuller@ieee.org).

On defining the prototypes, the semantic-embedding sources make it more convenient to extract latent sentiment representations from textual data for emotional words. This semantic-embedding prototype-based setup should be efficient in generating connections between seen and unseen emotional states through replacing the sample-wise manual annotation into automatic emotion-to-vector transformation by learning the corresponding semantic-embedding prototypes.

Our main contributions can be summarised as follows:

- We propose a framework of using semantic-embedding prototypes in zero-shot emotion recognition in speech, and demonstrate the effectiveness of the framework.
- We divide the learning strategies in the framework into sample-wise learning and emotion-wise learning, and explore feasible strategies for zero-shot SER.
- We also explore the influence of different learning strategies, inter-emotion information transfer, semantic-embedding prototypes, and paralinguistic features.

The remainder of this paper is organised as follows. Section II presents brief theoretical preliminaries for this paper. Then, the zero-shot speech emotion recognition framework is shown in Section III. Finally, Section IV and V present the experiments, the analysis, and the summarisation.

II. Preliminaries

Within this section, we present preliminary knowledge, including related works in previous research and basic notations.

A. Related works

Zero-Shot Learning [26], [28], [35]: ZSL aims to recognise unseen-class samples, using only seen-class samples in learning. In this case, unseen classes refer to the classes without any samples in the training material [36], [37]. The learning procedures rely on transferring related information from seen to unseen emotions, both through the representations of features from samples and the latent description from their labels or prototypes in different modalities. We investigate the typical setting of ZSL in this paper, where the test set only includes unseen-emotional samples [28]. Noting that *Generalised ZSL* (GZSL) is a more difficult recognition task, as the test set includes more labels for decision [38]–[40]. Considering the intersection between emotions [20], we perform non-generalised ZSL in this paper [28], [41]. This is due to different tasks on perceiving basic and complex unseen emotions in application [42]. In addition, it is also applicable to simulate the generalised case based on making an assumption of multi-label activation for seen and unseen emotions [43].

Semantic-Embedding Prototypes: A prototype refers to the most typical and representative example of its corresponding category. Prototypes can be used instead of learning a group of samples from this category [44]. Each semantic-embedding prototype, in this case, represents a class (an emotional state) in the form of a vector lying in a semantic embedding space [35], [45], through learning latent information on cross-modality sources [46]–[48]. The semantic-embedding prototype in this paper employs a semantic source generated through textual embedding transformed from its original one-hot representations [47], [49].

Zero-Shot Speech Emotion Recognition: A basic framework for zero-shot learning in SER, containing two phases of attribute learning and label learning, is presented in [24]. The attribute-learning phase constructs the relationship between paralinguistic features and emotional descriptors or attributes, through the procedure of regression on seen-emotional samples using *Support Vector Regression* (SVR) or *Deep Neural Networks* (DNNs). Then, employing empirical knowledge of attributes, the label-learning phase trains classifiers which link each emotion to its corresponding attribute samples. These two phases provide a possibility to recognise an arbitrary unseen-emotional speech sample without knowing any sample from its domain. As mentioned in the introduction, this framework relies on manual annotations for the descriptors on seen-emotional samples and the empirical judgements on the descriptors for each seen and unseen emotion.

We extend the previous research [24] by using semantic-embedding prototypes [35], [45] instead of the per-sample emotional descriptors. In SER, conventional approaches focus on raising accuracies in fully supervised or semi-supervised learning [6], [7], [50], [51], without considering the zero-shot cases. The current works of transfer learning [10], [12] and domain adaptation [9], [11], [13], [14] in SER investigate the inter-corpus information transfer, differing from our emotion-transfer research. Multi-task learning in SER is designed using auxiliary information to improve accuracies [16]–[19], while this work makes use of the semantic-embedding information to achieve the zero-shot SER. ZSL works in affective computing of image emotion [29], [31] and affective video recognition [30] shed light on zero-shot emotion recognition for visual cases. The research on typical ZSL approaches can provide various basic algorithms for our zero-shot SER research [32], [33], [41], [47], [52]–[54]. On learning strategies, this work also proposes a novel framework including sample-wise and emotion-wise learning compared with the existing research [26], [28], [55].

B. Notations

For the non-generalised ZSL, the seen-emotional and unseen-emotional label sets are denoted as $\mathcal{D}^{(S)} = \{d_1^{(S)}, d_2^{(S)}, \dots, d_{c^{(S)}}^{(S)}\}$ and $\mathcal{D}^{(U)} = \{d_1^{(U)}, d_2^{(U)}, \dots, d_{c^{(U)}}^{(U)}\}$. They contain $c^{(S)}$ and $c^{(U)}$ emotions respectively, where $\mathcal{D}^{(S)} \cap \mathcal{D}^{(U)} = \emptyset$. The semantic-embedding prototypes of $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(U)}$ are denoted as $A^{(S)} = [a_1^{(S)}, a_2^{(S)}, \dots, a_{c^{(S)}}^{(S)}] \in \Re^{n_A \times c^{(S)}}$ and $A^{(U)} = [a_1^{(U)}, a_2^{(U)}, \dots, a_{c^{(U)}}^{(U)}] \in \Re^{n_A \times c^{(U)}}$ respectively, where n_A represents the dimensionality of the prototypes. The $N^{(S)}$ seen-emotional samples with n_F-dimensional paralinguistic features are denoted as $X^{(S)} = [x_1^{(S)}, x_2^{(S)}, \dots, x_{N^{(S)}}^{(S)}] \in \Re^{n_F \times N^{(S)}}$ with their corresponding emotional labels as $\mathcal{Y}^{(S)} = \{y_1^{(S)}, y_2^{(S)}, \dots, y_{N^{(S)}}^{(S)}\}$ and the corresponding sample-wise prototypes $Z^{(S)} = [z_1^{(S)}, z_2^{(S)}, \dots, z_{N^{(S)}}^{(S)}] \in \Re^{n_A \times N^{(S)}}$ with each column equal to the prototypes in the $A^{(S)}$ corresponding to the sample's label, while for an arbitrary unseen-emotional sample, the features are $x^{(U)} \in \Re^{n_F \times 1}$, with the sample's predicted label $\widehat{y}^{(U)}$.

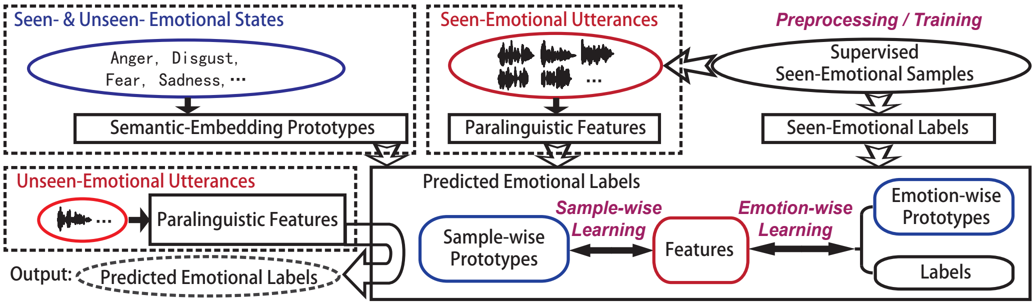


Figure 1: A diagrammatic overview of the proposed zero-shot SER framework using semantic-embedding prototypes and paralinguistic features. Note that the learning step appears in constructing the zero-shot emotion recognition model, while the prediction step consists of using the unseen-emotional utterances to achieve their predicted emotional labels.

The emotional labels $y_i^{(S)} \in \mathcal{D}^{(S)}$ and $\widehat{y}^{(U)} \in \mathcal{D}^{(U)}$, where $i = 1, 2, \ldots N^{(S)}$.

III. Methodology

A. Proposed framework

The proposed framework contains two steps of learning and prediction (inference). The learning step trains a zero-shot emotion recognition model, using semantic-embedding prototypes, emotional labels (from seen emotions), and paralinguistic features of speech utterances (from seen emotions), as presented in Figure 1. The prediction step employs the learnt model to predict the emotional state of an arbitrary sample of an emotion representation not present in the training step, herein referred to as an unseen emotional state.

First, semantic-embedding prototypes of seen and unseen emotional states are generated as $A^{(S)}$ and $A^{(U)}$, respectively. This generation is achieved via learnt text-to-vector models on the emotional label sets $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(U)}$. The raw speech utterances are also processed to obtain paralinguistic features $X^{(S)}$ for the supervised samples from seen-emotional states.

Then, we learn the zero-shot emotion recognition model using the semantic-embedding prototypes $A^{(S)}$, the seen-emotional paralinguistic features $X^{(S)}$, and the emotional labels $\mathcal{Y}^{(S)}$ of the samples, through using either of two types of learning strategies. We define the two types as sample-wise learning and emotion-wise learning, making use of $A^{(S)}$, $X^{(S)}$, and $\mathcal{Y}^{(S)}$ in different forms, as shown in Figure 2; this division differs from similar previous related research [28], [55]. We introduce these two types of learning in the learning step using the objective function $f(\cdot)$ as follows.

Sample-wise Learning: The sample-wise learning strategies perform learning on sample-wise pairs of prototypical duplication and seen-emotional samples. This step initialises with the set of $\{A^{(S)}, \mathcal{Y}^{(S)}\}$, leading to the sample-wise intermediation seen in Figure 2 which provides intermediate variables for the further learning on seen-emotional samples. The intermediation consists of the sample-wise duplication of prototypes as $\{Z^{(S)}\}$, which also includes the label information $\mathcal{Y}^{(S)}$ implicitly.

Thus, the learning step of sample-wise learning can be represented as estimating $f(\cdot)$'s optimal parameter set:

$$\widehat{\Psi} = \arg\max_{\Psi} P\left(A^{(S)}, \mathcal{Y}^{(S)} | X^{(S)}; \Psi\right) = \arg\max_{\Psi} P\left(Z^{(S)} | X^{(S)}; \Psi\right), \tag{1}$$

with the optimisation objective function $f(X^{(S)}, \Phi^{(S)}; \Psi)$, in which the $\Phi^{(S)}$ can be $\{A^{(S)}, \mathcal{Y}^{(S)}\}$, or equally be $\{Z^{(S)}\}$. The representation of the sample-wise $Z^{(S)}$ is equivalent to using the emotion-wise prototypes $A^{(S)}$ and the emotional labels $\mathcal{Y}^{(S)}$. We employ the representation of $f(X^{(S)}, Z^{(S)}; \Psi)$ in introducing the sample-wise learning strategies since one of the most attractive characteristics for this type is the sample-wise intermediation of $Z^{(S)}$.

Emotion-wise Learning: The emotion-wise learning strategies focus on learning between the ground-truth or generated prototypes $A^{(S)}$ and predicted emotional models for the seen-emotional states. These transformed models can be generalised to the form of emotion-wise intermediation, as seen in Figure 2. This intermediation can be learnt through discriminative or generative emotional classifiers, using the supervised information of $\{X^{(S)}, \mathcal{Y}^{(S)}\}$ for seen-emotional samples. Further, this type of learning can be iterative or reversible, indicating that the learning on emotions can be performed in advance of generating the intermediation. It is also applicable to include an additional $Z^{(S)}$ in generating the emotion-wise intermediation implicitly, without affecting the essence of emotion-wise learning.

In emotion-wise learning, we represent the optimised parameter set of the object $f(\cdot)$ as

$$\widehat{\Psi} = \arg\max_{\Psi} P\left(X^{(S)}, \mathcal{Y}^{(S)} | A^{(S)}; \Psi\right), \tag{2}$$

with the objective function $f(X^{(S)}, A^{(S)}, \mathcal{Y}^{(S)}; \Psi)$, specifically for emotion-wise learning due to its respective learning on the the seen-emotional samples' labels $\mathcal{Y}^{(S)}$ and the seen-emotional prototypes $A^{(S)}$.

In the prediction step, we input the paralinguistic features extracted from unseen-emotional samples into the model to predict their labels $\widehat{y}^{(U)}$s. The predicted index of emotions in the label set can be

$$\widehat{j} = \arg\max_{j} P\left(d_j^{(U)}, A^{(U)} | x^{(U)}; \widehat{\Psi}\right) \quad s.t.\ j = 1, 2, \ldots, c^{(U)}, \tag{3}$$

from which the predicted unseen-emotional label $\widehat{y}^{(U)} = d_{\widehat{j}}^{(U)}$.

Section III-B introduces the strategies of sample-wise learning, which directly employ the $Z^{(S)}$ in the fitting of the seen-emotional samples $X^{(S)}$ and the duplicated sample-wise prototypes $Z^{(S)}$. In Section III-C, we investigate the emotion-wise learning strategies, employing the information of $A^{(S)}$

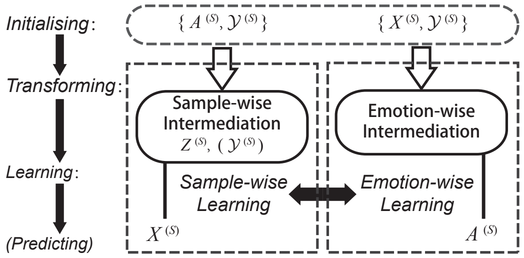

Initialising:
Transforming:
Learning:
(Predicting)
$\{A^{(S)}, \mathcal{Y}^{(S)}\}$
$\{X^{(S)}, \mathcal{Y}^{(S)}\}$
Sample-wise Intermediation
$Z^{(S)}, (\mathcal{Y}^{(S)})$
Emotion-wise Intermediation
Sample-wise Learning
Emotion-wise Learning
$X^{(S)}$
$A^{(S)}$

Figure 2: A block diagram of the proposed strategic categories of sample-wise and emotion-wise learning.

and $\mathcal{Y}^{(S)}$ separately through intra-emotion generative or inter-emotion discriminative modelling. Finally in Section III-D, we highlight applicable selections of semantic-embedding prototypes appearing in Figure 1.

B. Applicable strategies: Sample-wise learning

The sample-wise learning strategies are characterised by the direct use of per-sample prototypes $Z^{(S)}$, leading to optimising the form of objective function as $f(X^{(S)}, Z^{(S)}; \Psi)$.

SAE: The *Semantic AutoEncoder* (SAE) [33] represents the sample-wise prototypes as the encoded layer in an autoencoder, with the parameter set $\Psi = \{W_{\text{(SAE)}}\}$ representing the linear encoder. Thus, the optimal encoder

$$\begin{aligned}\widehat{W}_{\text{(SAE)}} &= \arg\min_{W_{\text{(SAE)}}} f_{\text{(SAE)}}(X^{(S)}, Z^{(S)}; W_{\text{(SAE)}}) \\ &= \arg\min_{W_{\text{(SAE)}}} ||X^{(S)} - W_{\text{(SAE)}}^T Z^{(S)}||_F^2 + \lambda ||W_{\text{(SAE)}} X^{(S)} - Z^{(S)}||_F^2,\end{aligned} \tag{4}$$

which is solved by the Bartels-Stewart algorithm [33], where the $W_{\text{(SAE)}} \in \Re^{n_A \times n_F}$. The parameter $\lambda > 0$ refers to the weight between the terms of data reconstruction and sample-wise prototype fitting. The estimated label index of $x^{(U)}$ can be drawn through the encoder or decoder as

$$\widehat{j} = \begin{cases} \arg\min_j Dis(\widehat{W}_{\text{(SAE)}} x^{(U)}, a_j^{(U)}), \\ \arg\min_j Dis(x^{(U)}, \widehat{W}_{\text{(SAE)}}^T a_j^{(U)}), \end{cases} \tag{5}$$

through calculating the distance operator $Dis(\cdot, \cdot)$.

DEM: The *Deep Embedding Model* (DEM) [47] strategy employs deep structures in the regression on the paralinguistic features of training samples with the sample-wise prototypes as the target in fitting. This leads to optimising the parameters $\Psi = \{W_1, W_2\}$ of the basic deep structures as

$$\begin{aligned}&\widehat{W}_{\text{(DEM)},1}, \widehat{W}_{\text{(DEM)},2} = \arg\min_{W_1, W_2} f_{\text{(DEM)}}(X^{(S)}, Z^{(S)}; W_1, W_2) \\ &= \arg\min_{W_1, W_2} \frac{1}{N^{(S)}} \sum_{i=1}^{N^{(S)}} \left|\left| x_i^{(S)} - g(W_2 g(W_1 z_i^{(S)})) \right|\right|^2 + R(W_1, W_2),\end{aligned} \tag{6}$$

where the connections $W_1 \in \Re^{n_H \times n_A}$ and $W_2 \in \Re^{n_F \times n_H}$ with the n_H-size hidden layer. The $g(\cdot)$ represents the nonlinearity mapping using a *Rectified Linear Unit* (ReLU), and the regularisation term of $R(W_1, W_2)$ can be $\lambda(||W_1||_F^2 + ||W_2||_F^2)$ with the weight constant λ. The estimated unseen-emotional label index is

$$\widehat{j} = \arg\min_j Dis\left(x^{(U)}, g(\widehat{W}_{\text{(DEM)},2}\, g(\widehat{W}_{\text{(DEM)},1}\, a_j^{(U)})\right). \tag{7}$$

LatEm: The *Latent Embeddings* (LatEm) [54] strategy utilises the optimisation between the columns of $X^{(S)}$ and $Z^{(S)}$, by calculating the cost function for each pair of $x_i^{(S)}$ and $z_l^{(S)}$, using the parameter set $\Psi = \mathcal{W}_{\text{(LE)}}$, where $i, l = 1, 2, \ldots, N^{(S)}$. The optimal parameters

$$\begin{aligned}&\widehat{\mathcal{W}}_{\text{(LE)}} = \arg\min_{\mathcal{W}_{\text{(LE)}}} f_{\text{(LE)}}(X^{(S)}, Z^{(S)}; \mathcal{W}_{\text{(LE)}}) = \arg\min_{\mathcal{W}_{\text{(LE)}}} \sum_{i=1}^{N^{(S)}} \frac{L(x_i^{(S)}, z_i^{(S)})}{N^{(S)}} \\ &s.t.\ L(x_i^{(S)}, z_i^{(S)}) = \sum_{l=1}^{N^{(S)}} \max\{0, G(x_i^{(S)}, z_i^{(S)}, z_l^{(S)})\}, \\ &G(x_i^{(S)}, z_i^{(S)}, z_l^{(S)}) = \Delta(z_i^{(S)}, z_l^{(S)}) + F(x_i^{(S)}, z_l^{(S)}) - F(x_i^{(S)}, z_i^{(S)}), \\ &F(x_i^{(S)}, z_l^{(S)}) = \max_{m=1,2,\ldots,M} (x_i^{(S)T} W_{\text{(LE)},m} z_l^{(S)}),\end{aligned} \tag{8}$$

where $\widehat{\mathcal{W}}_{\text{(LE)}} = \{\widehat{W}_{\text{(LE)},1}, \widehat{W}_{\text{(LE)},2}, \ldots, \widehat{W}_{\text{(LE)},M}\}$ and $\mathcal{W}_{\text{(LE)}} = \{W_{\text{(LE)},1}, W_{\text{(LE)},2}, \ldots, W_{\text{(LE)},M}\} \subset \Re^{n_F \times n_A}$. This optimisation form includes a discriminative term of $\Delta(z_i, z_l)$, equal to 0 for $z_i = z_l$ and 1 for $z_i \neq z_l$. This makes LatEm be a mixture of the sample-wise and emotion-wise learning implicitly, as the discriminative term only requires the labels $\mathcal{Y}^{(S)}$. The optimisation can be solved using *Stochastic Gradient Descent* (SGD) [54], obtaining $x^{(U)}$'s predicted index

$$\widehat{j} = \arg\max_j \left(\max_{m=1,2,\ldots,M} (x^{(U)T} \widehat{W}_{\text{(LE)},m} a_j^{(U)}) \right). \tag{9}$$

C. Applicable strategies: Emotion-wise learning

The emotion-wise learning strategies make use of $\mathcal{Y}^{(S)}\}$ and $A^{(S)}$ separately, with the objective function represented by $f(X^{(S)}, A^{(S)}, \mathcal{Y}^{(S)}; \Psi)$.

ESZSL: The *Embarrassingly Simple Zero-Shot Learning* (ESZSL) [52] methodology induces the emotion-wise prototypes $A^{(S)}$ utilising linear discrimination for the seen-emotional samples, in order to transfer knowledge from seen to unseen emotions. We define the parameter set as $\Psi = \{W_{\text{(LES)}}\}$ for the linear case, while $\Psi = \{W_{\text{(KES)}}\}$ for the corresponding kernelised case. The optimal $W_{\text{(LES)}}$ is

$$\begin{aligned}\widehat{W}_{\text{(LES)}} &= \arg\min_{W_{\text{(LES)}}} f_{\text{(LES)}}(X^{(S)}, A^{(S)}, \mathcal{Y}^{(S)}; W_{\text{(LES)}}) \\ &= \arg\min_{W_{\text{(LES)}}} \left(L_0(X^{(S)T} W_{\text{(LES)}} A^{(S)}, \mathcal{Y}^{(S)}) + R_1(W_{\text{(LES)}}) \right),\end{aligned} \tag{10}$$

where the linear mapping matrix $W_{\text{(LES)}} \in \Re^{n_F \times n_A}$ and $L_0(\cdot, \cdot)$ is a loss function to measure the dissimilarity, using Frobenius norm between the transformed features and the labels $\mathcal{Y}^{(S)}$ in the one-hot form as a matrix. $R_1(W_{\text{(LES)}})$ represents the regularisation for $W_{\text{(LES)}}$. Thus, the predicted index

$$\widehat{j} = \arg\max_j \left(x^{(U)T} \widehat{W}_{\text{(LES)}} a_j^{(U)} \right). \tag{11}$$

The kerenlised form of ESZSL includes learning linear combination of *Reproducing Kernel Hilbert Space* (RKHS) [6], [7] vectors from $W_{\text{(KES)}} \in \Re^{N^{(S)} \times n_A}$, as

$$\begin{aligned}\widehat{W}_{\text{(KES)}} &= \arg\min_{W_{\text{(KES)}}} f_{\text{(KES)}}(X^{(S)}, A^{(S)}, \mathcal{Y}^{(S)}; W_{\text{(KES)}}) \\ &= \arg\min_{W_{\text{(KES)}}} \left(L_0(K(X^{(S)}, X^{(S)}) W_{\text{(KES)}} A^{(S)}, \mathcal{Y}^{(S)}) + R_2(W_{\text{(KES)}}) \right),\end{aligned} \tag{12}$$

where $K(X^{(S)}, X^{(S)}) = \phi^T(X^{(S)}) \phi(X^{(S)})$ represents the inner products between RKHS vectors, in which $\phi(\cdot)$ is the RKHS mapping for each column of $X^{(S)}$. $R_2(W_{\text{(KES)}})$ represents the regularisation for $W_{\text{(KES)}}$. The predicted index

$$\widehat{j} = \arg\max_j \left(K(X^{(S)}, x^{(U)})^T \widehat{W}_{\text{(KES)}} a_j^{(U)} \right), \tag{13}$$

where $K(X^{(S)}, x^{(U)}) = \phi^T(X^{(S)})\phi(x^{(U)})$.

SYNC: The fast *SYNthesised Classifiers* (SYNC) [53], [56] strategy makes use of seen-emotional labels $\mathcal{Y}^{(S)}$ to generate linear discriminative mappings, while $A^{(S)}$ is employed to generate $c^{(P)}$ phantom classifiers connecting seen and unseen emotions. Defining the parameters as $\Psi = \{V_{\text{(SYNC)}}\}$, we have the optimal parameters

$$\begin{aligned} \widehat{V}_{\text{(SYNC)}} &= \arg\min_{V_{\text{(SYNC)}}} f_{\text{(SYNC)}}(X^{(S)}, A^{(S)}, \mathcal{Y}^{(S)}; V_{\text{(SYNC)}}) \\ &= \arg\min_{V_{\text{(SYNC)}}} \left(J(X^{(S)}, \mathcal{Y}^{(S)}, W) + \frac{\tau}{2} tr(W^T W) \right) \quad s.t.\ W = V_{\text{(SYNC)}} S, \end{aligned} \tag{14}$$

where the weight $\tau > 0$. The phantom linear classifiers $V_{\text{(SYNC)}} \in \Re^{n_F \times c^{(P)}}$ and the seen classifiers $W = [w_1, w_2, \dots, w_{c^{(S)}}] \in \Re^{n_F \times c^{(S)}}$. The loss function of J aims to optimise the discrimination between $X^{(S)}$ and $\mathcal{Y}^{(S)}$. For the *One-Versus-Other* (OVO) case (noted as 'SYNC-OVO'),

$$J(X^{(S)}, \mathcal{Y}^{(S)}, W) = \sum_{c=1}^{c^{(S)}} \sum_{i=1}^{N^{(S)}} \left(\max(0, 1 - \Delta_0(y_i^{(S)}, d_c^{(S)}) w_c^T x_i^{(S)}) \right)^2, \tag{15}$$

where $\Delta_0(y_i^{(S)}, d_c^{(S)})$ is equal to 1 for $y_i^{(S)} = d_c^{(S)}$ and -1 for $y_i^{(S)} \neq d_c^{(S)}$. The similarity matrix $S \in \Re^{c^{(P)} \times c^{(S)}}$ contains its c_Pth-row and c_Sth-column element as $S_{c_P c_S} = \frac{e^{-Dis(a_{c_S}^{(S)}, b_{c_P})}}{\sum_{c_P=1}^{c^{(P)}} e^{-Dis(a_{c_S}^{(S)}, b_{c_P})}}$, using the phantom prototypes $B = [b_1, b_2, \dots, b_{c^{(P)}}] \in \Re^{n_A \times c^{(P)}}$. The simplified form of the distance operator $Dis(a_{c_S}^{(S)}, b_{c_P}) = \sigma^2 ||a_{c_S}^{(S)} - b_{c_P}||^2$.

We can also yield the J to standard *Crammer-Singer* (CS) multi-class *Support Vector Machine* (SVM) loss (noted as 'SYNC-CS') or the CS loss with l_2 distance of prototypes (noted as 'SYNC-struct'). The l_2-distance CS loss implies a mixture of the sample-wise learning, due to the inclusion of $Z^{(S)}$. With the s_j representing the jth column of S, the predicted emotional-state index

$$\widehat{j} = \arg\max_j \left((\widehat{V}_{\text{(SYNC)}} s_j)^T x^{(U)} \right). \tag{16}$$

EXEM: The *EXEMplar synthesis* (EXEM) [37], [53] strategy utilises a one dimensionality-reduced exemplar for each emotional state, optimising the alignment as the parameter set $\Psi = \{\psi_{\text{(EXEM)}}\}$, between each exemplar and its corresponding emotional prototype. This leads to the optimal mapping for the alignment as

$$\begin{aligned} \widehat{\psi}_{\text{(EXEM)}} &= \arg\min_{\psi_{\text{(EXEM)}}} f_{\text{(EXEM)}}(X^{(S)}, A^{(S)}, \mathcal{Y}^{(S)}; \psi_{\text{(EXEM)}}) \\ &= \arg\min_{\psi_{\text{(EXEM)}}} Sim(\psi_{\text{(EXEM)}}(A^{(S)}), U) \quad s.t.\ u_c = \frac{\Omega X_c^{(S)} \Gamma_c}{N_c^{(S)}}, \end{aligned} \tag{17}$$

where $Sim(\cdot, \cdot)$ is a similarity measurement. The exemplar matrix $U = [u_1, u_2, \dots, u_{c^{(S)}}] \in \Re^{n_{DR} \times c^{(S)}}$ with the reduced dimensionality equal to n_{DR} using the linear mapping matrix $\Omega \in \Re^{n_{DR} \times n_F}$ of the *Principal Component Analysis* (PCA). In order to draw the mapped prototypes $\psi_{\text{(EXEM)}}(A^{(S)}) \in \Re^{n_{DR} \times c^{(S)}}$, ν- *Support Vector Regression* (ν-SVR) can be employed in this optimisation. $N_c^{(S)}$ represents the number of samples belonging to $d_c^{(S)}$. All of the elements of $\Gamma_c \in \Re^{N_c^{(S)} \times 1}$ are equal to 1. Then, we have $x^{(U)}$'s predicted index

$$\widehat{j} = \arg\min_j Dis(\Omega x^{(U)}, \widehat{\psi}_{\text{(EXEM)}}(a_j^{(U)})). \tag{18}$$

Table I: The best UAs (%; corresponding to the corpora 'GEMEP / DEMoS') among multiple prototypes averaging on each pair of unseen emotions for different learning strategies.

Strategies \ Data Forms	$Z^{(S)}$	$A^{(S)}$	$\mathcal{Y}^{(S)}$
SAE [33]	✓		
DEM [47]	✓		
LatEm [54]	✓		✓ (*)
SSE [32]	✓	✓	✓
ESZSL [52]		✓	✓
EXEM (1NNS) [37], [53]		✓	✓
SYNC (-OVO, -CS) [53], [56]		✓	✓
SYNC (-struct) [53], [56]	✓	✓	✓
GNN-ZSL (GCNZ [57], DGP [58])		✓	✓

* *Implicitly using the discriminative label information for seen emotions.*

Similar to SYNC-struct, some strategies use sample-wise prototypes $Z^{(S)}$, jointly with $A^{(S)}$ and $\mathcal{Y}^{(S)}$ during learning. We, therefore, consider these a special case of emotion-wise learning, such as the SSE strategy, since they include learning procedures for $A^{(S)}$.

SSE: The *Semantic Similarity Embedding* (SSE) [32] strategy projects both of features and prototypes into an emotional-contribution space represented by seen-emotional prototypes, optimising the parameter set $\Psi = \{w_{\text{(SSE)}}, \mathcal{V}_{\text{(SSE)}}, \varphi_{\text{(SSE)}}\}$. First, the optimal mapping $\varphi_{\text{(SSE)}}(\cdot)$ for each prototype from the prototype domain to contribution weights can be drawn through sparse coding as

$$\widehat{\varphi}_{\text{(SSE)}} = \arg\min_{\varphi_{\text{(SSE)}}} f_{\text{(SSE)},1}(A^{(S)}; \varphi_{\text{(SSE)}}), \tag{19}$$

where the $f_{\text{(SSE)},1}$ refers to a reconstruction function for seen-emotional prototypes. The linear mapping $w_{\text{(SSE)}} \in \Re^{n_F \times 1}$ and the emotion-wise nonlinearisation parameter set $\mathcal{V}_{\text{(SSE)}} = \{v_{\text{(SSE)},1}, v_{\text{(SSE)},2}, \dots, v_{\text{(SSE)},c^{(S)}}\} \subset \Re^{n_F \times 1}$. Thus, the optimal $w_{\text{(SSE)}}$ and $\mathcal{V}_{\text{(SSE)}}$ are obtained through a discriminative form with the $\widehat{\varphi}_{\text{(SSE)}}$, as

$$\begin{aligned} \widehat{w}_{\text{(SSE)}}, \widehat{\mathcal{V}}_{\text{(SSE)}} &= \arg\min_{w_{\text{(SSE)}}, \mathcal{V}_{\text{(SSE)}}} f_{\text{(SSE)},2}(X^{(S)}, \mathcal{Y}^{(S)}, Z^{(S)}; w_{\text{(SSE)}}, \mathcal{V}_{\text{(SSE)}}) \\ &= \arg\min_{w_{\text{(SSE)}}, \mathcal{V}_{\text{(SSE)}}} f_{MM}(\{w_{\text{(SSE)}}^T\ \psi_{\text{(SSE)}}(x_i^{(S)}, \mathcal{V}_{\text{(SSE)}})\ \widehat{\varphi}_{\text{(SSE)}}(z_i^{(S)})\}), \end{aligned} \tag{20}$$

considering the minimisation of the max-margin formulation f_{MM} [32], with arbitrary $i = 1, 2, \dots, N^{(S)}$, which can be solved alternatively using SVM and *ConCave-Convex Procedure* (CCCP) [32]. The nonlinearisation mapping $\psi_{\text{(SSE)}}(x_i^{(S)}, \mathcal{V}_{\text{(SSE)}}) \in \Re^{n_F \times c^{(S)}}$ can be *INTersction* (INT) or ReLU functions [32], while the $\widehat{\varphi}_{\text{(SSE)}}(z_i^{(S)}) \in \Re^{c^{(S)} \times 1}$ represents the corresponding sparsely coded unseen-emotional prototype. The predicted label index

$$\widehat{j} = \arg\max_j \left(\widehat{w}_{\text{(SSE)}}^T\ \psi_{\text{(SSE)}}(x^{(U)}, \widehat{\mathcal{V}}_{\text{(SSE)}})\ \widehat{\varphi}_{\text{(SSE)}}(a_j^{(U)}) \right). \tag{21}$$

Note that the *Graph Neural Network* (GNN) based ZSL strategies, including GCNZ [57] and *Dense Graph Propagation* (DGP) [58] can also be sorted as the emotion-wise learning by replacing the prototypes $A^{(S)}$ and $a^{(U)}$ into

Table II: The description of the GEMEP and DEMoS corpora, including the dimensional *arousal-valence* polarity for the emotional states.

Properties \ Corpora		GEMEP (12 emotions)	DEMoS (excluding *neutral*)
Polarity	*Arousal* (Pos.) & *Valence* (Pos.)	*amusement*, *pride*, *elation*	*happiness*, *surprise*
	Arousal (Neg.) & *Valence* (Pos.)	*relief*, *interest*, *pleasure*	-
	Arousal (Pos.) & *Valence* (Neg.)	*hot anger* (*rage*), *panic fear*, *despair*	*anger*, *sadness*, *fear*, *disgust*, *guilt*
	Arousal (Neg.) & *Valence* (Neg.)	*cold anger* (*irritation*), *anxiety*, *sadness* (*depression*)	-
Languages / # Speakers / # Samples		French / 10 (5 female) / 1 080	Italian / 68 (23 female) / 9 697

learning on knowledge graphs consisting of concept entities. This step, however, requires well-designed concept structures for the specific case of affective computing. We summarise different usages on the data forms of $Z^{(S)}$, $A^{(S)}$, and $\mathcal{Y}^{(S)}$ for the applicable strategies in Table I. Thus for the **sample-wise learning** strategies, we unify the forms of optimisation as $f(X^{(S)}, Z^{(S)}; \Psi)$, while the **emotion-wise learning** strategies employ the form of $f(X^{(S)}, A^{(S)}, \mathcal{Y}^{(S)}; \Psi)$. Note that the boundary of the two categories of the learning strategies is not clear since the data forms are interconvertible.

D. Semantic-embedding prototypes for emotions

A common method for prototype selection when performing sentiment analysis is to employ a pre-trained word-vector model to build the semantic-prototype generators. Applicable generators include *word2vec* [49], [59], *GloVe* [60], and *fastText* [61], [62].

Based on the continuous skip-gram model, the *word2vec* model generates word representations by using several extensions on sub-sampling of frequent words and a simplified variant of *Noise Contrastive Estimation* (NCE) [31], [49]. The *GloVe* model differs from *word2vec* in that it is built using a weighted least-squares model, trained on global word-word co-occurrence counts [60]. It is also possible to use pre-trained word vectors based on multiple text data. Similar to *word2vec*, the *fastText* generator extends the continuous skip-gram model, through considering subword information [61].

Built on the research of *SenticNet* [63], [64], *SenticNet 5* [65] aims to identify sentiment information from textual commonsense concepts. It employs a *Long Short-Term Memory* (LSTM) based *Recurrent Neural Network* (RNN) [65] to aid the construction of a *SenticNet* concept's values in multiple sentiment dimensions, related moods and polarities, and 5 highly related semantic words or phrases.

In our framework, we propose two categories of prototypes. The first category employs the raw word vectors to form the semantic-embedding prototypes directly, while the other category utilises the *SenticNet 5* to obtain an emotional state's neighbouring textual phrases. We also replace the raw word vectors into the averaging word vectors of the corresponding concepts' neighbours, which may reduce the error probability compared with a single word's representations.

IV. Experiments

A. Preparation

Table III: Brief introduction of the features of the eGeMAPS set (with *Mel-Frequency Cepstral Coefficients* (MFCC)). '$\langle\cdot\rangle$' represents the extensive features compared with the GeMAPS set.

Feature Types and Description of eGeMAPS (with # dimensions)
Within Voiced Regions ($42\langle+14\rangle$ dim.):
• **Frequency** ($18\langle+4\rangle$ dim.):
Pitch (10 func.‡); Jitter (2 func.†);
Formant 1 to 3 frequency & Formant 1 bandwidth (2 func.†);
⟨Formant 2 & 3 bandwidth (2 func.†)⟩;
• **Energy/Amplitude** (4 dim.):
Shimmer (2 func.†); *Harmonics-to-Noise Ratio* (HNR) (2 func.†);
• **Spectral** ($18\langle+10\rangle$ dim.):
Alpha Ratio (2 func.†); Hammarberg Index (2 func.†);
Spectral Slope 0-500 Hz & 500-1500 Hz (2 func.†);
Formant 1 to 3 relative energy (2 func.†);
Harmonic Difference H1-H2 and H1-A3 (2 func.†);
⟨Spectral Flux (2 func.†)⟩; ⟨MFCC 1 to 4 (2 func.†)⟩;
Within Unvoiced Regions ($4\langle+1\rangle$ dim.):
Alpha Ratio (1 func.⋆); Hammarberg Index (1 func.⋆);
Spectral Slope 0-500 Hz & 500-1500 Hz (1 func.⋆);
⟨Spectral Flux (1 func.⋆)⟩;
Within Global Regions (Voiced & Unvoiced) ($16\langle+11\rangle$ dim.):
Loudness (10 func.‡); rate of loudness peaks;
continuous voiced regions per second;
mean length & standard deviation of continuous voiced regions;
mean length & standard deviation of continuous unvoiced regions;
⟨Spectral Flux (2 func.†)⟩; ⟨MFCC 1-4 (2 func.†)⟩;
⟨equivalent sound level⟩.

⋆ 1 *func.: Only applying the functional of arithmetic mean.*
† 2 *func.: Applying the functionals of '1 func.' and coefficient of variation.*
‡ 10 *func.: Applying the functionals of* $\{20, 50, 80\}$*th percentile, the range of* 20*th to* 80*th percentile, and the mean and standard deviation of the slope of rising & falling parts, in addition to '2 func.'.*

Corpora and Features

We use the *GEneva Multimodal Emotion Portrayals* (GEMEP) [6], [66], [67] and *Database of Elicited Mood in Speech* (DEMoS) [68] corpora in our experiments. For both corpora, we set the learning-data partition as *Leave-Two-Emotions-Out* (LTEO). This strategy results in leaving the samples from two emotion categories as the test set, while using the other samples in training and validating. This step enables us to investigate the knowledge transfer between emotional samples in the zero-shot SER.

GEMEP: The GEMEP corpus includes 1 260 French utterance samples with a sampling rate of 44.1 kHz, belonging to 18 emotions from 10 speakers (5 female). As in [6], [7], we choose 1 080 samples from 12 emotions (90 samples per emotion), without considering the additional emotions. This leads to the emotional inclusion of *amusement*, *anxiety*, *cold anger* (or *irritation*), *despair*, *elation*, *hot anger* (or *rage*), *interest*, *panic fear*, *pleasure*, *pride*, *relief*, and *sadness* (or *depression*).

DEMoS: In order to investigate the small-size emotion cases with more utterance samples, we further employ the DEMoS corpus, which includes 9 697 Italian utterances from 8 categories (including 7 emotional states and the neutral state). The categories of *anger*, *sadness*, *happiness*, *fear*, *disgust*, *guilt*, *surprise*, and *neutral* contain 1 447, 1 530, 1 395, 1 156, 1 678, 1 129, 1 000, and 332 samples respectively, with the sampling rate of 44.1 kHz. Produced by 68 native Italian speakers (23 female), the utterances are elicited by combinations of *Mood Induction Procedures* (MIP). Table II provides a

Table IV: Unweighted Accuracies (UAs; %), corresponding to the corpora of 'GEMEP / DEMoS', using the eGeMAPS feature set and different strategies, when employing the 8 sorts of semantic-embedding prototypes. We omit the insignificant results on UAs compared with the chance level.

GEMEP / DEMoS Corpora:	w/o *SenticNet 5*				w/ *SenticNet 5*			
Strategies \ Prototypes	*word2vec*	*GloVe*	*fastText*-crawl	*fastText*-wiki	*word2vec*	*GloVe*	*fastText*-crawl	*fastText*-wiki
SAE	51.5 / –	52.6 / –	51.9 / 50.7	53.9 / –	51.1 / 50.5	53.0 / –	52.5 / –	52.7 / 50.6
DEM	51.9 / –	54.5 / –	53.1 / –	54.0 / –	– / –	– / 51.3	– / –	– / 50.7
LatEm ($M = 2$)	56.5 / 51.5	55.3 / 51.1	54.7 / 51.5	57.5 / 51.0	51.5 / 53.8	– / 51.3	51.1 / 52.0	– / 52.6
LatEm ($M = 4$)	55.8 / 50.8	53.5 / 51.6	55.7 / 51.0	54.6 / 50.7	51.4 / 52.0	51.4 / 52.8	52.2 / 50.9	52.1 / 51.3
SSE (INT)	– / 50.5	51.7 / –	54.0 / –	53.5 / –	– / –	– / –	53.9 / –	– / –
SSE (ReLU)	51.3 / –	52.8 / 51.1	54.4 / –	54.6 / –	– / 51.9	51.7 / 52.1	55.0 / 51.8	– / 52.2
Linear ESZSL	58.2 / –	56.5 / –	59.2 / –	58.2 / –	51.5 / –	54.1 / –	54.5 / –	**53.9** / –
Kernel ESZSL	**59.3** / –	**57.6** / –	**59.3** / 52.0	**59.5** / –	52.9 / 53.8	**54.2** / 53.2	54.9 / 52.0	53.4 / 52.4
EXEM (1NNS)	57.3 / **53.1**	52.3 / 52.1	55.9 / 53.4	56.1 / **52.7**	**54.0** / 56.2	52.6 / 55.4	**56.5** / 54.3	53.5 / 54.7
SYNC-OVO	59.2 / 51.3	57.2 / 51.2	58.5 / 52.0	59.0 / 52.3	53.0 / 56.6	53.8 / 54.9	56.0 / **54.6**	52.8 / **55.3**
SYNC-CS	59.0 / 52.7	55.6 / **52.5**	57.4 / **53.8**	56.9 / 52.5	50.9 / **56.8**	52.8 / **56.1**	55.1 / 53.0	53.8 / 53.7
SYNC-struct	57.0 / 50.6	55.6 / 51.9	56.9 / 52.1	57.0 / 51.5	51.9 / 55.8	53.2 / 55.2	55.0 / 54.1	52.9 / 54.8

Table V: Pair-wise comparisons of UAs (mean difference and significance, noted as 'MD' and 'Signif.' respectively) on the factor of *strategy* using post-hoc *Tukey*'s HSD, between the 3 strategies (including Kernel ESZSL, EXEM (1NNS), and SYNC-OVO) and the 7 strategies (including SAE, DEM, LatEm, SSE (ReLU), in addition to the 3 ones).

Strategies \ Corpora		GEMEP corpus		DEMoS corpus	
Str. 1	Str. 2	MD (Str. 1-2)	Signif. (p value)	MD (Str. 1-2)	Signif. (p value)
Kernel ESZSL	SAE	0.0397	< .005*	0.0142	< .005*
	DEM	0.0582	< .005*	0.0163	< .005*
	LatEm	0.0307	< .005*	−0.0016	> .05
	SSE (ReLU)	0.0475	< .005*	0.0041	> .05
	EXEM (1NNS)	0.0161	> .05	−0.0229	< .005*
	SYNC-OVO	0.0022	> .05	−0.0183	< .005*
EXEM (1NNS)	SAE	0.0236	> .05	0.0371	< .005*
	DEM	0.0421	< .005*	0.0392	< .005*
	LatEm	0.0146	> .05	0.0213	< .005*
	SSE (ReLU)	0.0314	< .005*	0.0270	< .005*
	Kernel ESZSL	−0.0161	> .05	0.0229	< .005*
	SYNC-OVO	−0.0139	> .05	0.0046	> .05
SYNC -OVO	SAE	0.0375	< .005*	0.0325	< .005*
	DEM	0.0560	< .005*	0.0346	< .005*
	LatEm	0.0285	< .01*	0.0167	< .005*
	SSE (ReLU)	0.0453	< .005*	0.0224	< .005*
	Kernel ESZSL	−0.0022	> .05	0.0183	< .005*
	EXEM (1NNS)	0.0139	> .05	−0.0046	> .05

* *Significant at the level of* 0.05 *for post-hoc* Tukey*'s HSD.*

brief introduction for the GEMEP (12 emotions) and DEMoS corpora, including the dimensional *arousal*-*valence* polarity for the emotional states.

The openSMILE toolkit is utilised for extracting paralinguistic features [69], [70]. We employ the 88-dimensional *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) in the experiments (from functionals on 25 time-smoothed *Low-Level Descriptors* (LLDs), temporal features, and equivalent sound level) [70], which has been proven effective in SER when using SVMs. In supplementary experiments, we also use GeMAPS (from functionals on 18 LLDs and temporal features) [70] and ComParE (from functionals on 65 LLDs and temporal features) [2], [3], [7] sets, containing 62 and 6 373 features respectively. Min-max normalisation is employed for feature pre-processing. We present a brief introduction for the mainly used eGeMAPS feature set (including GeMAPS) in Table III, for the extraction within voiced, unvoiced, and global regions in an arbitrary speech sample [70].

Semantic-Embedding Prototypes

The experiments employed 300-dimensional English word-vector representations as the selections of semantic-embedding prototypes, based on pre-trained models from *word2vec* (1 model), *GloVe* (1 model), and *fastText* (2 models), all with and without using *SenticNet 5*. The *SenticNet 5* extends the 4 models through presenting an average combination for each word vector of a corresponding emotion's 5 neighbours (if their word vectors exist), leading to 8 selections of prototypes. We utilise the conventional *word2vec* model trained on the *Google News* corpus containing 3 million words (100 billion tokens) [31], [49]. The *GloVe* model in the experiments employs *Wikipedia 2014* plus *Gigaword 5* as its training data, containing 0.4 million vocabularies (6 billion tokens) [60]. For the *fastText* models, we employ the training data of the pre-trained representations respectively including: 1) 2 million word vectors (600 billion tokens) trained on Common Crawl (noted as '*fastText*-crawl'); 2) 1 million word vectors trained on *Wikipedia 2017*, *UMBC webbase* corpus and the *statmt.org news* dataset (noted as '*fastText*-wiki') [61], [62].

The 12 emotional categories in the GEMEP corpus can be used to generate semantic-embedding prototypes into two folds. For the emotions of *anxiety*, *cold anger*, *hot anger*, *panic fear*, and *sadness*, the prototype generator utilises the emotions' average word-vector representations of emotional pairs as *anxiety*-*worry*, *irritation*-*anger*, *rage*-*anger*, *panic*-*fear*, and *sadness*-*depression* respectively, in accordance with the emotional-label description for the GEMEP corpus [66]. The emotions in DEMoS and the other emotions in GEMEP are used to obtain their semantic-embedding prototypes directly through the word-vector models.

Experimental Setup for Learning Strategies

As mentioned above, the LTEO data partition implies pair-wise emotion recognition in the experiments. To select optimal parameters for each strategies, we utilise emotion-independent 5-fold and 3-fold *Cross-Validation* (CV) in grid-searching on the GEMEP and DEMoS corpora, respectively, which makes the validation set include samples from two emotions in each CV round. The measurement of *Unweighted Accuracy* (UA) is chosen in the CV rounds as the standard, calculated through averaging recalls [7], [16], [71].

In the proposed zero-shot SER framework, we employ the ZSL approaches of SAE [33], DEM [47], LatEm [54], SSE [32], ESZSL [52], EXEM [37], [53], and SYNC [53], [56] in the experiments. For the SAE strategy, the parameters $\lambda = \{10^{-1}, 10^{0}, 10^{1}, 10^{3}, 10^{5}, 10^{7}, \ldots, 10^{27}\}$. The DEM employs the learning rates of $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$, the weight decays of $\{10^{-3}, 10^{-2}\}$, and the intermediate-layer node numbers of $\{100, 300, 500, 700, 900\}$, within $3\,000$ epochs using the *Adaptive moment estimation* (Adam) optimiser [47]. The LatEm considers two cases of Ms for the numbers of the linear mapping matrices, equal to 2 and 4, while the learning rates of SGD are $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$, within 150 epochs. The SSE retains the experimental setup as in [32], considering both of INT and ReLU cases. For the linear and kernelised ESZSL strategies, we choose both of the regulariser weights for samples and prototypes (cf. [52]) as $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$, while for the kernelised ESZSL, we employ Gaussian kernels with the scaling parameters of $\{10^{-1}n_F, n_F, 10n_F\}$. The EXEM considers the ν-SVR parameters of the regularisation coefficients, the ν values, and the kernel-scaling parameters as $\{2^{-3}, 2^{-2}, \ldots, 2^{3}\}$, $\{2^{-8}, 2^{-7}, \ldots, 2^{0}\}$, and $\{2^{-4}, 2^{-3}, \ldots, 2^{4}\}$ respectively, while the reduced dimensions in PCA are $\{40, 60, 80\}$. We set the distance function in EXEM as a 1-nearest neighbour classifier with standard Euclidean distance (noted as EXEM (1NNS)). The SYNC strategies (SYNC-OVO, SYNC-CS, and SYNC-struct) utilise semantic-embedding prototypes of seen emotions as the phantom prototypes, using the regularisation weights τs and the distance-scaling parameters σ^2s chosen from $\{2^{-24}, 2^{-23}, \ldots, 2^{-9}\}$ and $\{2^{-5}, 2^{-4}, \ldots, 2^{5}\}$, respectively.

B. Experimental results: Strategic comparison

Our experiments contain three modules. For the first module of **strategic comparison** (Section IV-B), we aim to verify the applicability of the proposed framework and make a comparison between the learning strategies. The module of **emotional pair-wise analysis** (Section IV-C) allows to investigate emotion transfer between different types of emotional states due to the needs of application. Afterwards, the **influence of semantic-embedding prototypes and paralinguistic features** (Section IV-D) focuses on exploring the influence from different critical settings and parameters.

First, we perform the experiments on both the GEMEP and DEMoS corpora comparing the different semantic-embedding prototypes *word2vec*, *GloVe*, and *fastText*, with and without *SenticNet 5*, using UA as the performance-evaluation metric [7]. As shown in Table IV, we calculate the average value of all the pair-wise emotion recognition results (UAs) in different selections of specific semantic-embedding prototypes and learning strategies, where we omit the insignificant UA results in the table compared to the chance level, using a one-tailed *z-test* at the significance level of 0.05 [6], [72]. Thus, the UAs present imply that it is feasible to use semantic-embedding prototypes in zero-shot SER instead of the sample-wise annotation as used in [24], at least for recognising some specific pairs of emotions.

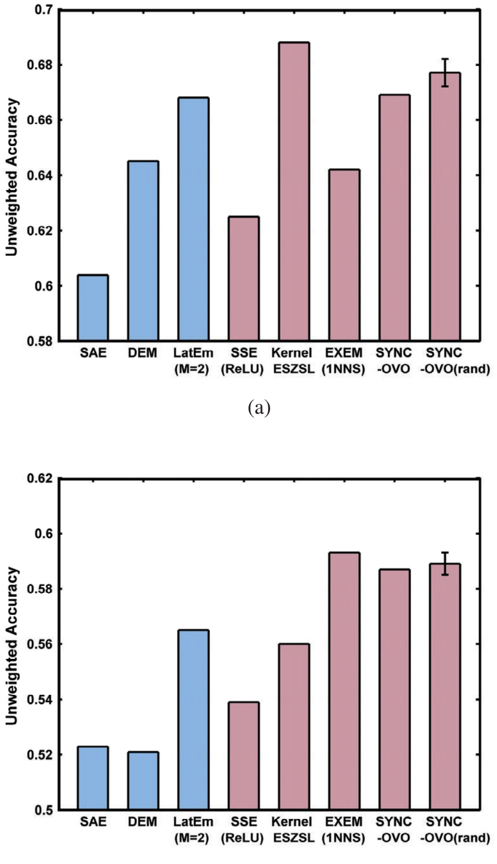
As observed in Table IV, different learning strategies and semantic-embedding prototypes can result in diverse performances on UA and F1. Most of the best results appears among the learning strategies of ESZSL, SYNC, and EXEM, all included in the form of emotion-wise learning. We can infer that the fitting procedures in sample-wise learning may make it difficult to coordinate valuable information for emotion recognition due to the limited number of speech samples. Further, for the strategies of kernelised ESZSL, EXEM, and SYNC-OVO, the highest average UAs (among the 8 categories of prototypes) for the GEMEP corpus outperform the results on the DEMoS corpus. This may be due to the small number of emotional states in the DEMoS corpus, which provides insufficient information transfer from seen to unseen emotions.

Table VI: The best UAs (%; corresponding to the corpora of 'GEMEP / DEMoS') among multiple prototypes averaging on each pair of unseen emotions for different learning strategies.

Strategies \ Prototypes	w/o *SenticNet 5*	w/ *SenticNet 5*	both
SAE	57.2 / 51.1	57.0 / 51.7	60.4 / 52.3
DEM	59.3 / 50.3	54.4 / 52.1	64.5 / 52.1
LatEm ($M = 2$)	64.2 / 54.2	58.0 / 55.5	66.8 / 56.5
LatEm ($M = 4$)	63.5 / 54.2	59.3 / 54.7	66.6 / 56.0
SSE (INT)	58.3 / 50.8	56.9 / 50.1	61.6 / 51.6
SSE (ReLU)	58.6 / 51.4	57.8 / 53.6	62.5 / 53.9
Linear ESZSL	63.2 / 50.0	60.2 / 50.6	66.9 / 50.6
Kernel ESZSL	64.6 / 52.8	**61.4** / 55.6	**68.8** / 56.0
EXEM (1NNS)	62.3 / **55.6**	60.3 / **58.7**	64.2 / **59.3**
SYNC-OVO	64.4 / 54.7	59.7 / 58.3	66.9 / 58.7
SYNC-CS	63.4 / 54.9	58.8 / 57.8	66.2 / 58.4
SYNC-struct	62.9 / 53.8	58.8 / 57.4	66.0 / 57.7
SYNC-OVO (rand)	**65.0** $\pm$ 0.9 / 55.2 $\pm$ 0.4	60.9 $\pm$ 1.1 / 58.2 $\pm$ 0.5	67.7 $\pm$ 0.5 / 58.9 $\pm$ 0.4

In order to compare inter-strategy performances, we employ a two-way *ANalysis Of VAriance* (ANOVA) [7] for the UAs across all the emotional pairs (66 for GEMEP corpus and 28 for DEMoS corpus) on the two corpora respectively, with its factors *strategy* (including 7 categories of SAE, DEM, LatEm ($M = 2$), SSE (ReLU), Kernel ESZSL, EXEM (1NNS), and SYNC-OVO) and *prototype* (including 8 categories as in Section IV-A). For the *strategy* factor, we obtain its significant effect with $(F(6, 3640) = 15.59,\ p < 0.0001)$ (on the GEMEP corpus) and $(F(6, 1512) = 31.52,\ p < 0.0001)$ (on the DEMoS corpus). We also perform a post-hoc *Tukey's Honest Significant Difference* (*Tukey*'s HSD) test [7] with respect to the *strategy*, focusing on comparing the strategies of Kernel ESZSL, EXEM (1NNS), and SYNC-OVO, with the remaining ones among the 7 categories for *strategy*, as shown in Table V. The results in Table V show the best UA performance appearing in employing Kernel ESZSL and SYNC-OVO for the GEMEP corpus, while the DEMoS corpus makes EXEM (1NNS) and SYNC-OVO perform significantly better. This implies that using SYNC strategies results in the most balanced performance, in accordance with the experimental results of conventional ZSL research [28].

Afterwards, we investigate the best UAs for each emotional pair among the prototypes (Table VI). This analysis includes the best UA results among the 4 prototypes with and without *SenticNet 5* respectively, and all the 8 prototypes, averaging on each pair of unseen emotions. In view of the most balanced UA results for SYNC-OVO presented above, we propose the 'SYNC-OVO (rand)' strategy based on SYNC-OVO using randomised phantom prototypes, with the elements ranging from 0 to 1 following a uniform distribution. We set the number of the phantom prototypes $c^{(P)}$ as $1\,000$, repeating



(a)

(b)

Figure 3: Bar charts of the average UAs among all the 8 sorts of prototypes for each pair of emotions with eGeMAPS paralinguistic features, when using the algorithms SAE, DEM, LatEm ($M = 2$), SSE (ReLU), Kernel ESZSL, EXEM (1NNS), SYNC-OVO, and SYNC-OVO (rand), on the (a) GEMEP and (b) DEMoS corpora.

the experiments for 15 times. The results indicate that Kernel ESZSL and SYNC-OVO / SYNC-OVO (rand) perform well on the GEMEP corpus, while EXEM (1NNS) and SYNC-OVO / SYNC-OVO (rand) achieve better performance on the DEMoS corpus (Figure 3). As an additional investigation, we present the UA and macro F1-score (calculated through averaging recalls and precision-recall integration) [71], [73] results of the strategies of Kernel ESZSL, EXEM (1NNS), SYNC-OVO, and SYNC-OVO (rand). The F1-score results obey the tendency of the UAs (Table VII), indicating that the analysis on the UAs in this work is possible to be transferred onto the joint precision-accuracy measurement.

C. Experimental results: Emotional pair-wise analysis

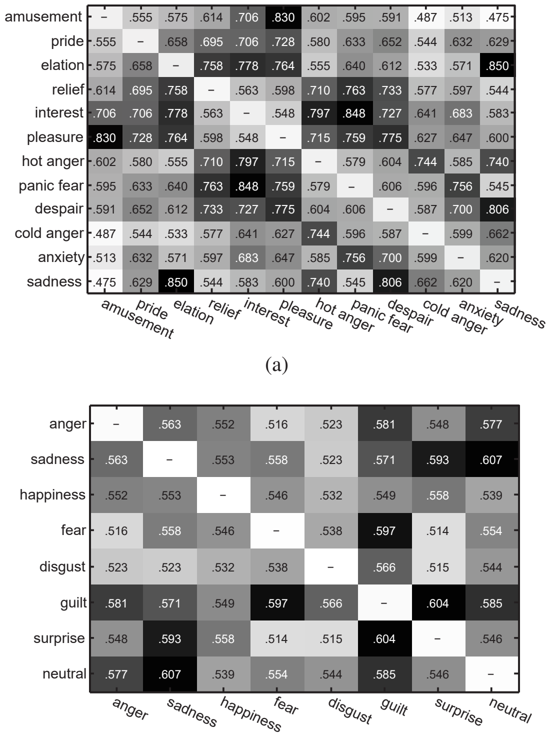
We start an analysis between emotional states as only it is required to recognise very few numbers of emotional states in most application scenarios. We present the pair-wise emotional matrices on the corpora GEMEP (Figure 4a) and DEMoS (Figure 4b), using the average UA results for the SAE, DEM, LatEm ($M = 2$), SSE (ReLU), Kernel ESZSL, EXEM (1NNS), and SYNC-OVO strategies, considering the best results among the 8 prototypes. Each row or column with quite different UAs in Figure 4 implies that the information transfer for zero-shot SER jointly depends on the source domain for seen emotions and the target domain for unseen emotions. It is also learnt from Figure 4 that the pair-wise zero-shot recognition on the GEMEP corpus performs better compared to the DEMoS corpus. We set a one-way ANOVA on these UAs between the two corpora [7], showing significantly better performance for the GEMEP corpus. This may be partially due to the variety of seen emotions in the target domain.

Table VII: The best UAs and F1-scores (%) on the GEMEP and DEMoS corpora among the 8 prototypes averaging on each pair of unseen emotions for different learning strategies. We consider the best and average results for SYNC-OVO (rand).

Strategies \ Measurements	GEMEP Corpus		DEMoS Corpus	
	UA	F1-score	UA	F1-score
Kernel ESZSL	68.8	68.0	56.0	54.8
EXEM (1NNS)	64.2	63.6	59.3	57.1
SYNC-OVO	66.9	66.1	58.7	56.7
SYNC-OVO (rand) Avg.	67.7	66.8	58.9	56.4
SYNC-OVO (rand) Best	68.3	67.5	59.5	57.2

On the GEMEP corpus in Figure 4a, the UA results ($> 75.0\%$) of emotional pairs indicate that the dimension of *arousal* is a key factor in zero-shot SER, since all the top UAs appear in the emotional pairs with different polarities of *arousal*. This is in accordance with the previous SER research [18], [74], [75], showing better performance on *arousal*-polarity separation. We also jointly investigate the influence from the dimensions *arousal* and *valence*, in view that the emotional states of the GEMEP corpus were chosen evenly based on the level of *arousal* and *valence*. Figure 5 presents the average UAs of the emotional pairs with the same *arousal*-*valence* polarity (noted as 'Same Polarity'), with the different *arousal* polarity (noted as 'Diff. Arousal'), with the different *valence* polarity (noted as 'Diff. Valence'), and with different polarities on both dimensions (noted as 'Diff. Ar.-Val.'). We calculate the average UAs of 7 strategies (noted as '7 Str.'; including the strategies in Figure 3) and 3 strategies (noted as '3 Str.'; including Kernel ESZSL, EXEM (1NNS), and SYNC-OVO).

The results in Figure 5 indicate that the inter-polarity zero-shot recognition outperforms the intra-emotion setups (see the UAs between 'Same Polarity' and other bars in Figure 5). Thus, we can infer that the dimension *valence* still plays an effective role in zero-shot SER, despite of the dominant effect from *arousal*. However, the UAs of the different-*arousal*-*valence* cases are lower than the different-*arousal* cases (see the UAs between 'Diff. Ar.-Val.' and 'Diff. Arousal'), which are 67.0% / 68.0% and 70.2% / 71.1% for the '7 Str. / 3 Str.' UAs. This is different from the experimental analysis in [6], where for conventional SER, the emotions with different polarities on both of *arousal* and *valence* lead to a better recognition performance compared with the other cases. In view of this difference, we analyse the average UAs for the two cases of positive-*arousal* / positive-*valence* versus negative-*arousal* / negative-*valence* and positive-*arousal* / negative-*valence* versus negative-*arousal* / positive-*valence*, which are 58.1% / 60.5% and 75.9% / 75.4% respectively (in the form of '7 Str. / 3 Str.'). Thus, the zero-shot-SER case of positive-



(a)

(b)

Figure 4: Average UA matrices of emotional pairs considering the strategies of SAE, DEM, LatEm ($M = 2$), SSE (ReLU), Kernel ESZSL, EXEM (1NNS), and SYNC-OVO, on the corpora (a) GEMEP and (b) DEMoS, choosing the best UAs among 8 sorts of prototypes using eGeMAPS features.

arousal / negative-*valence* versus negative-*arousal* / positive-*valence* is in line with the analysis of conventional SER methods, while it is difficult to classify the emotional pairs for the case of positive-*arousal* / positive-*valence* versus negative-*arousal* / negative-*valence*.

On the DEMoS corpus in Figure 4b, we are interested in the performance on recognising the categories of *guilt* and *neutral*, since the investigation of the two categories results from the rarely-appeared emotion of *guilt* [68] in the research on emotion detection in speech [76]. The results are shown in Table VIII by presenting the UAs between the two categories and the remaining emotions when considering the '7 Str.' and '3 Str.' averaging UAs. The left column of Table VIII provides the UAs of zero-shot recognition for the emotion of *guilt*. The best UAs appear in classifying *guilt* from *surprise* (positive *valence*) and *fear* (negative *valence*), which implies that the *valence* polarity does not determine the recognition performance for *guilt*, directly. The right column for recognising *neutral* achieves the best UAs when classifying *sadness*. This may provide evidence and motivation for detecting depressive factors in speech [77], [78], for the zero-shot case without providing the speech samples of depression.

In view of the possibility in solving zero-shot depression detection, we then perform comparisons between *sadness* and the other emotional states on DEMoS corpus using different strategies, as shown in Table IX. In addition to conventional ZSL strategies, we employ the *Feature Generating Network* (FGN) strategy based on *Generative Adversarial Networks* (GANs) [79], [80], using 600 synthesised samples for each unseen emotion with the classification-loss weight of 0.01. It is observed from Table IX that the strategies achieve their best UAs on detecting depression from different emotions, where FGN performs better for *happiness* and *disgust*.

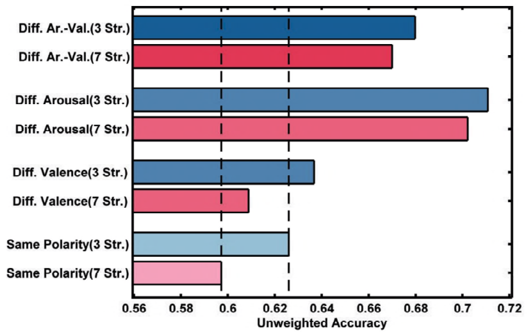


Figure 5: Row charts of the UAs among all the 8 sorts of prototypes for recognising the emotional pairs with the same *arousal-valence* polarity (noted as 'Same Polarity'), with the different *arousal* polarity (noted as 'Diff. Arousal'), with the different *valence* polarity (noted as 'Diff. Valence'), and with different polarities on both dimensions (noted as 'Diff. Ar.-Val.'), using the averaging of '7 Str.' and '3 Str.'.

Table VIII: The best UAs (%) among all the 8 sorts of prototypes, averaging across the cases of '7 Str.' and '3 Str.', for recognising *guilt* and *neutral* from the other emotional categories on the DEMoS corpus.

Target Emo. (Val.):	*guilt* (Neg.)		*neutral* (-)	
Emo. (Val.)\Avg.	7 Str.	3 Str.	7 Str.	3 Str.
happiness (Pos.)	54.9	58.8	53.9	56.2
surprise (Pos.)	**60.4**	64.1	54.6	55.3
anger (Neg.)	58.1	62.2	57.7	60.3
sadness (Neg.)	57.1	61.7	**60.7**	**65.2**
fear (Neg.)	59.7	**65.5**	55.4	55.6
disgust (Neg.)	56.6	57.0	54.4	54.9
neutral (-)	58.5	62.4	–	–

D. Experimental results: Influence of semantic-embedding prototypes and paralinguistic features

When investigating the influence of semantic-prototype selections, it can be observed that the best performance corresponds to different choices of semantic-embedding prototypes as in Table VI. Thus, we continue to analyse the factor of *prototype* with 8 categories, for the two-way ANOVA in Section IV-B.

We investigate the factor of *prototype*, obtaining a significant effect with $(F(7, 3640) = 12.45,\ p < 0.0001)$ (on the GEMEP corpus) and $(F(7, 1512) = 12.85,\ p < 0.0001)$ (on the DEMoS corpus), which indicates the factor of *prototype* affects the UAs significantly. Afterwards, a post-hoc *Tukey*'s HSD is performed to further examine the pair-wise comparison between prototypes. First, we present the pair-wise comparisons of UAs (calculating mean difference and significance) on the factor of *prototype*, between the cases of with and without employing *SenticNet 5* on the prototypes, when using the 4 models of *word2vec*, *GloVe*, *fastText-crawl*, and *fastText-wiki* respectively, as shown in Table X. The results of Table X show the significance between with and without *SenticNet*

Table IX: The best UAs (%) among the prototypes on DEMoS corpus, for recognising '*sadness*' from the other emotions using different strategies.

Emo. (Val.) \ Strategies	Kernel ESZSL	EXEM (1NNS)	SYNC -OVO	FGN (GANs)
happiness (Pos.)	52.8	60.4	55.2	**62.9**
surprise (Pos.)	60.9	**66.5**	62.8	65.2
guilt (Neg.)	61.8	61.0	**62.4**	59.3
disgust (Neg.)	52.2	51.4	53.5	**60.0**
fear (Neg.)	57.8	**61.1**	58.9	58.2
anger (Neg.)	59.3	58.9	58.5	**59.4**
neutral (-)	54.7	**71.7**	69.3	65.8

Table X: Pair-wise comparisons of UAs (mean difference and significance, noted as 'MD' and 'Signif.' respectively) on the factor of *prototype* using *Tukey*'s HSD, between the cases of with and without employing *SenticNet 5* (noted as 'SN') on the prototypes (noted as 'Prot.'), when using *word2vec*, *GloVe*, *fastText-crawl*, and *fastText-wiki* respectively.

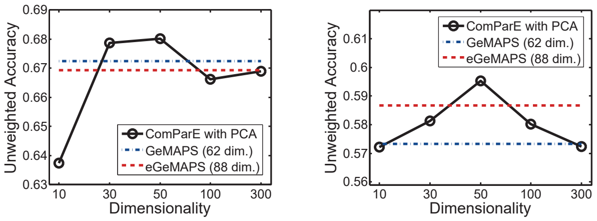
Prototypes\Corpora		GEMEP corpus		DEMoS corpus	
Prot. 1 (w/o *SN*)	Prot. 2 (w/ *SN*)	MD (Prot.1-2)	Signif. (p value)	MD (Prot.1-2)	Signif. (p value)
word2vec		0.0488	$< .005$*	-0.0257	$< .005$*
GloVe		0.0269	$< .05$*	-0.0196	$< .005$*
fastText-crawl		0.0194	$> .05$	-0.0083	$> .05$
fastText-wiki		0.0488	$< .005$*	-0.0177	$< .005$*

* *Significant at the level of* 0.05 *for post-hoc* Tukey*'s HSD.*

5 cases, when using the models of *word2vec*, *GloVe*, and *fastText-wiki*. This verifies that the usage of *SenticNet 5* can affect the performance of zero-shot SER, where the DEMoS corpus achieves better UAs, while the GEMEP corpus prefers to perform zero-shot SER without the tool of *SenticNet 5*. In addition, we perform the post-hoc comparison within the 4 models for the GEMEP corpus (without *SenticNet 5*) and the DEMoS corpus (with *SenticNet 5*), due to the significance above on these two corpora in relation to *SenticNet 5*. The results reveal insignificant differences between these models, which suggests that different word-vector models do not affect UA performance significantly. However, different processing on these models may lead to distinct results.

Finally, we explore the performance for the proposed selections of paralinguistic features. Considering the performance of SYNC-OVO, we illustrate, in Figure 6, the UAs for SYNC-OVO when using the state-of-the-art paralinguistic feature sets GeMAPS (62 dimensions), eGeMAPS (88 dimensions), and ComParE ($\{10, 30, 50, 100, 300\}$ dimensions). We choose the ComParE features with these dimensions through a PCA trained on the training / validation set from the seen-emotional samples in ZSL, in order to present a fair comparison using similar numbers of features. The results suggest that UAs for different feature sets depended on setups of data sets (see the contrast between GeMAPS and eGeMAPS in Figure 6). Furthermore, the ComParE set achieves the best performance in a low dimensionality of 50, which might be caused by the highly related samples from seen and unseen domains.

V. Conclusions



(a) (b)

Figure 6: Comparison of the best UAs among the 8 sorts of prototypes for each emotional pair of the SYNC-OVO strategy, when using the paralinguistic features of ComParE (with PCA dimensionality of 10, 30, 50, 100, and 300 components), GeMAPS, and eGeMAPS, on the corpora (a) GEMEP and (b) DEMoS.

Within this paper, an exploration into zero-shot learning for emotion recognition in speech using semantic per-emotion prototypes was presented. First, we analysed the applicable approaches for zero-shot emotion recognition in speech, considering the two types of learning strategies: sample-wise and emotion-wise learning. Then, experiments on the corpora GEMEP and DEMoS allowed us to draw three conclusions: 1) It was applicable to employ the per-sample semantic-embedding prototypes in recognising zero-shot emotional states, typically for some target emotions with application background; 2) Different learning strategies might lead to different performances, where the strategies of *Embarrassingly Simple Zero-Shot Learning* (ESZSL), *EXEMplar synthesis* (EXEM), and *SYNthesized Classifiers* (SYNC) could result in better performance; 3) The target emotions, selected prototypes, and paralinguistic features were able to affect the zero-shot recognition performance, which could be specifically designed.

Despite of these conclusions on performing *Zero-Shot Learning* (ZSL) when processing emotional factors in audio signals, it remains some challenges.

First, one should investigate the relationship between artificially annotated attributes and the semantic-embedding prototypes, and further research on the modalities of the prototypes. It would also be helpful to investigate emotional-speech augmentation for unseen emotions using improved *Generative Adversarial Networks* (GANs) based ZSL strategies [38], [40], [80]. Furthermore, the research on *Graph Neural Networks* (GNNs) [57] can be a powerful tool in addition to the semantic-embedding prototypes, in order to better represent the prototypes. It appears also applicable to specifically design the emotional states for the training and test sets in zero-shot emotion recognition in speech. Finally, *Generalised ZSL* (GZSL) can be further investigated in order to adapt to complex cases.

Acknowledgement

The authors would like to thank Jing Han, Eduardo Coutinho, Francesco Barone, and the anonymous reviewers for their valuable help.

References

[1] B. Schuller, F. Weninger, Y. Zhang, F. Ringeval, A. Batliner, S. Steidl, F. Eyben, E. Marchi, A. Vinciarelli, and K. Scherer, "Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge," *Computer Speech & Language*, vol. 53, pp. 156–180, 2019.

[2] B. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny *et al.*, "The INTERSPEECH 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Hyderabad, India: ISCA, 2018, pp. 122–126.

[3] B. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The INTERSPEECH 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Shanghai, China: ISCA, 2020, p. no pagination.

[4] B. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[5] Z. Zhang, J. Han, E. Coutinho, and B. Schuller, "Dynamic difficulty awareness training for continuous emotion prediction," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1289–1301, 2018.

[6] X. Xu, J. Deng, N. Cummins, Z. Zhang, C. Wu, L. Zhao, and B. Schuller, "A two-dimensional framework of multiple kernel subspace learning for recognizing emotion in speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1436–1449, 2017.

[7] X. Xu, J. Deng, E. Coutinho, C. Wu, L. Zhao, and B. Schuller, "Connecting subspace learning and extreme learning machine in speech emotion recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 3, pp. 795–808, 2019.

[8] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 31–43, 2018.

[9] Q. Mao, G. Xu, W. Xue, J. Gou, and Y. Zhan, "Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition," *Speech Communication*, vol. 93, pp. 1–10, 2017.

[10] P. Song, "Transfer linear subspace learning for cross-corpus speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 265–275, 2019.

[11] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 500–504, 2017.

[12] W. Zhang and P. Song, "Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 307–318, 2020.

[13] J. Li, E. Chen, Z. Ding, L. Zhu, K. Lu, and H. T. Shen, "Maximum density divergence for domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear.

[14] J. Li, M. Jing, K. Lu, L. Zhu, and H. T. Shen, "Locality preserving joint transfer for domain adaptation," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 6103–6115, 2019.

[15] J. Han, Z. Zhang, Z. Ren, and B. Schuller, "EmoBed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings," *IEEE Transactions on Affective Computing*, to appear.

[16] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Hyderabad, India: ISCA, 2018, pp. 951–955.

[17] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 85–99, 2017.

[18] D. Le, Z. Aldeneh, and E. M. Provost, "Discretized continuous speech emotion recognition with multi-task deep recurrent neural network." in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Stockholm, Sweden, 2017, pp. 1108–1112.

[19] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, 2017.

[20] D. Watson and K. Stanton, "Emotion blends and mixed emotions in the hierarchical structure of affect," *Emotion Review*, vol. 9, no. 2, pp. 99–104, 2017.

[21] N. L. Nelson, E. Nowicki, M. C. Diemer, K. Sangster, C. Cheng, and J. A. Russell, "Children can create a new emotion category through a process of elimination," *Cognitive Development*, vol. 47, pp. 117–123, 2018.

[22] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[23] Y. Zhang, F. Weninger, Z. Ren, and B. Schuller, "Sincerity and deception in speech: Two sides of the same coin? a transfer-and multi-task learning perspective." in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, San Francisco, CA, 2016, pp. 2041–2045.

[24] X. Xu, J. Deng, N. Cummins, Z. Zhang, L. Zhao, and B. Schuller, "Autonomous emotion learning in speech: A view of zero-shot speech emotion recognition," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Graz, Austria: ISCA, 2019, pp. 949–953.

[25] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013, pp. 935–943.

[26] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2014.

[27] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 7, pp. 1425–1438, 2015.

[28] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2251–2265, 2018.

[29] C. Zhan, D. She, S. Zhao, M.-M. Cheng, and J. Yang, "Zero-shot emotion recognition via affective structural embedding," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA: IEEE, 2019, pp. 1151–1160.

[30] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Video emotion recognition with transferred deep feature encodings," in *Proc. ACM International Conference on Multimedia Retrieval*. New York, NY: ACM, 2016, pp. 15–22.

[31] V. Campos, X. Giro-i Nieto, B. Jou, J. Torres, and S.-F. Chang, "Sentiment concept embedding for visual affect recognition," in *Multimodal Behavior Analysis in the Wild*. Elsevier, 2019, pp. 349–367.

[32] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *Proc. the IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: IEEE Computer Society, 2015, pp. 4166–4174.

[33] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE Computer Society, 2017, pp. 3174–3183.

[34] Y. Annadani and S. Biswas, "Preserving semantic relations for zero-shot learning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT: IEEE, 2018, pp. 7603–7612.

[35] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Advances in Neural Information Processing Systems 22*. Curran Associates, Inc., 2009, pp. 1410–1418.

[36] Y. Guo, G. Ding, J. Han, S. Zhao, and B. Wang, "Implicit non-linear similarity scoring for recognizing unseen classes." in *Proc. International Joint Conferences on Artificial Intelligence (IJCAI)*, Stockholm, Sweden, 2018, pp. 4898–4904.

[37] S. Changpinyo, W.-L. Chao, and F. Sha, "Predicting visual exemplars of unseen classes for zero-shot learning," in *Proc. the IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy: IEEE, 2017, pp. 3476–3485.

[38] J. Li, M. Jing, L. Zhu, Z. Ding, K. Lu, and Y. Yang, "Learning modality-invariant latent representations for generalized zero-shot learning," in *Proc. ACM International Conference on Multimedia*, Seattle, WA, 2020, pp. 1348–1356.

[39] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized zero-and few-shot learning via aligned variational autoencoders," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, 2019, pp. 8247–8255.

[40] J. Li, M. Jing, K. Lu, L. Zhu, Y. Yang, and Z. Huang, "Alleviating feature confusion for generative zero-shot learning," in *Proc. ACM International Conference on Multimedia*, Nice, France, 2019, pp. 1587–1595.

[41] Y. Fu, T. Xiang, Y.-G. Jiang, X. Xue, L. Sigal, and S. Gong, "Recent advances in zero-shot recognition: Toward data-efficient understanding

of visual content," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 112–125, 2018.

[42] P. Ekman, "Are there basic emotions?" *Psychological Review*, vol. 99, no. 3, pp. 550–553, 1992.

[43] Y. Liu, X. Gao, Q. Gao, J. Han, and L. Shao, "Label-activating framework for zero-shot learning," *Neural Networks*, vol. 121, pp. 1–9, 2020.

[44] P. Shaver, J. Schwartz, D. Kirson, and C. O'Connor, "Emotion knowledge: Further exploration of a prototype approach." *Journal of Personality and Social Psychology*, vol. 52, no. 6, p. 1061, 1987.

[45] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," in *Proc. International Conference on Learning Representations (ICLR)*, Banff, AB, Canada, 2014, p. no pagination.

[46] Z. Fu, T. Xiang, E. Kodirov, and S. Gong, "Zero-shot learning on semantic class prototype graph," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 2009–2022, 2017.

[47] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 2021–2030.

[48] C. Luo, Z. Li, K. Huang, J. Feng, and M. Wang, "Zero-shot learning via attribute regression and class prototype rectification," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 637–648, 2017.

[49] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013, pp. 3111–3119.

[50] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[51] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2017.

[52] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *Proc. International Conference on Machine Learning (ICML)*, Lille, France, 2015, pp. 2152–2161.

[53] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Classifier and exemplar synthesis for zero-shot learning," *International Journal of Computer Vision*, pp. 1–36, 2019.

[54] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, "Latent embeddings for zero-shot classification," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV: IEEE, 2016, pp. 69–77.

[55] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–37, 2019.

[56] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 5327–5336.

[57] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT: IEEE, 2018, pp. 6857–6866.

[58] M. Kampffmeyer, Y. Chen, X. Liang, H. Wang, Y. Zhang, and E. P. Xing, "Rethinking knowledge graph propagation for zero-shot learning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, 2019, pp. 11 487–11 496.

[59] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[60] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1532–1543.

[61] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[62] T. Mikolov, É. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proc. the International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan, 2018, pp. 52–55.

[63] E. Cambria and A. Hussain, "SenticNet," in *Sentic Computing*. Springer, 2015, pp. 23–71.

[64] E. Cambria, S. Poria, R. Bajpai, and B. Schuller, "SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives," in *Proc. the International Conference on Computational Linguistics (COLING)*, Osaka, Japan, 2016, pp. 2666–2677.

[65] E. Cambria, S. Poria, D. Hazarika, and K. Kwok, "SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings," in *Proc. the AAAI Conference on Artificial Intelligence*, New Orleans, LA, 2018, pp. 1795–1802.

[66] T. Bänziger and K. R. Scherer, "Introducing the Geneva multimodal emotion portrayal (GEMEP) corpus," *Blueprint for affective computing: A sourcebook*, pp. 271–294, 2010.

[67] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, and E. Marchi, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Lyon, France: ISCA, 2013, pp. 148–152.

[68] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, and B. Schuller, "DEMoS: An Italian emotional speech corpus," *Language Resources and Evaluation*, pp. 1–43, 2019.

[69] F. Eyben and B. Schuller, "openSMILE:) The Munich open-source large-scale multimedia feature extractor," *ACM SIGMultimedia Records*, vol. 6, no. 4, pp. 4–13, 2015.

[70] F. Eyben, K. R. Scherer, K. P. Truong, B. Schuller *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[71] A. Schnall and M. Heckmann, "Integrating sequence information in the audio-visual detection of word prominence in a human-machine interaction scenario," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Singapore: ISCA, 2014, pp. 2640–2644.

[72] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Attention-enhanced connectionist temporal classification for discrete speech emotion recognition," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Graz, Austria: ISCA, 2019, pp. 206–210.

[73] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.

[74] P. Tzirakis, J. Zhang, and B. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB, Canada: IEEE, 2018, pp. 5089–5093.

[75] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB, Canada: IEEE, 2018, pp. 5084–5088.

[76] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, "Iterative feature normalization scheme for automatic emotion detection from speech," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 386–397, 2013.

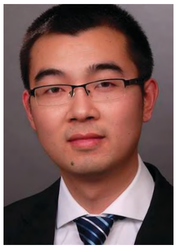
[77] A. Mallol-Ragolta, Z. Zhao, L. Stappen, N. Cummins, and B. Schuller, "A hierarchical attention network-based approach for depression detection from transcribed clinical interviews," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Graz, Austria, 2019, pp. 221–225.

[78] N. Cummins, V. Sethu, J. Epps, J. W. Williamson, T. F. Quatieri, and J. Krajewski, "Generalized two-stage rank regression framework for depression score prediction from speech," *IEEE Transactions on Affective Computing*, to appear.

[79] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, 2018, pp. 5542–5551.

[80] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, and Z. Huang, "Leveraging the invariant side of generative zero-shot learning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, 2019, pp. 7402–7411.

Xinzhou Xu received the bachelor's degree from Nanjing University of Posts and Telecommunications, Nanjing/China, in 2009, the master's and the PhD degree from Southeast University, Nanjing/China, in 2012 and 2017, respectively. He is currently a Lecturer with the School of Internet of Things, Nanjing University of Posts and Telecommunications. Previously, he was with the Machine Intelligence & Signal Processing group, MMK, Technical University of Munich (TUM), Munich/Germany (from 2014 to 2016), and the Chair of Complex and Intelligent Systems, University of Passau, Passau/Germany (from 2015 to 2016). His research interests include audio signal processing, pattern recognition, machine learning, and affective computing.



Jun Deng received his bachelor degree (2009) in electronic and information engineering from Harbin Engineering University (HEU) and his master degree (2011) in information and communication engineering from Harbin Institute of Technology (HIT), Heilongjiang/China, and his doctoral degree (2016) for his study on Feature Transfer Learning for Speech Emotion Recognition, in electrical engineering and information technology from Technical University of Munich (TUM), Germany. He was a postdoctoral researcher from 2015 to 2017 at the Chair of Complex and Intelligent Systems at the University of Passau in Passau/Germany, and was also a Leader Researcher at audEERING, Germany. Currently, he is the Head of deep learning at Agile Robots AG, Germany. His interests are machine learning methods such as transfer learning and deep learning with an application preference to affective computing.

Nicholas (Nick) Cummins is a lecturer in AI for speech analysis for health at the Department of Biostatistics and Health Informatics at Kings College London. Nicks current research interests include speech processing, affective computing and multisensory signal analysis. He is fascinated by the application of machine learning techniques to improve our understanding of different health conditions and mental health disorders in particular. Nick is actively involved in RADAR-CNS project in which he assists in the management of Work Package 8: Data Analysis & Biosignatures. Nick was awarded his PhD in electrical engineering from UNSW Australia in February 2016 for his thesis Automatic assessment of depression from speech: paralinguistic analysis, modelling and machine learning. After completing his PhD, he was a postdoctoral researcher at the Chair of Complex and Intelligent Systems at the University of Passau, Germany. Most recently, he was a habilitation candidate at the Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg, also in Germany. During his time in Germany, he was involved in the DE-ENIGMA, RADAR-CNS, TAPAS and sustAGE Horizon 2020 projects. He also wrote and delivered courses in speech pathology, deep learning and intelligent signal analysis in medicine. Nick is an external advisor on the National Science Foundation of China (NSFC) funded project, Diagnosis of Depression by Speech Signals (grant No.31860285). He has (co-)authored over 100 conference and journal papers leading to over 1900 citations (h-index: 23). He is a frequent reviewer for IEEE, ACM and ISCA journals and conferences as well as serving on program and organisational committees. He is a member of ACM, ISCA, IEEE and a full member of the IET.

Zixing Zhang (M'15) received his master degree in physical electronics from the Beijing University of Posts and Telecommunications (BUPT), China, in 2010, and his PhD degree in computer engineering from Technical University of Munich (TUM), Germany, in 2015. From 2017 to 2019, he was a research associate with the Department of Computing at the Imperial College London (ICL), UK. Before that, he was a postdoctoral researcher at the University of Passau, Germany. To date, he has authored more than ninty publications in peer-reviewed books, journals, and conference proceedings. His research mainly focuses on deep learning technologies for speaker-centred state and health computing. He has organised special sessions, such as at the IEEE 7th Affective Computing and Intelligent Interaction (ACII) conference in 2017 and at the 43nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in 2018. Moreover, he serves as a reviewer for numerous leading-in-their fields journals and conferences, a programme committee member and an area chair for many international conferences.

Li Zhao received the bachelor's degree from Nanjing University of Aeronautics and Astronautics, Nanjing/China, in 1982, the master's degree from Soochow University, Suzhou/China, in 1988, and the PhD degree from Kyoto Institute of Technology, Kyoto/Japan, in 1998. He is currently a Professor with the School of Information Science and Engineering, Southeast University, Nanjing/China. His research interests include spoken signal processing and affective computing.

Björn Schuller (M'05-SM'15-F'18) received his diploma in 1999, his doctoral degree for his study on Automatic Speech and Emotion Recognition in 2006, and his habilitation and Adjunct Teaching Professorship in the subject area of Signal Processing and Machine Intelligence in 2012, all in electrical engineering and information technology from TUM in Munich/Germany. He is Professor with GLAM– the Group on Language Audio and Music in the Department of Computing at the Imperial College London/UK, Full Professor and head of the Chair of Embedded Intelligence for Health Care and Wellbeing at Augsburg University/Germany. Dr. Schuller is President-Emeritus of the Association for the Advancement of Affective Computing (AAAC), Fellow of the IEEE, ISCA, and BCS, and Senior Member of the ACM and (co-)authored 5 books and more than 1 000 publications in peer reviewed books, journals, and conference proceedings leading to more than 37 000 citations (h-index = 87).