

Rethinking auditory affective descriptors through zero-shot emotion recognition in speech

Xinzhou Xu, Jun Deng, Zixing Zhang, Xijian Fan, Li Zhao, Laurence Devillers, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Xu, Xinzhou, Jun Deng, Zixing Zhang, Xijian Fan, Li Zhao, Laurence Devillers, and Björn W. Schuller. 2022. "Rethinking auditory affective descriptors through zero-shot emotion recognition in speech." IEEE Transactions on Computational Social Systems 9 (5): 1530–41. <https://doi.org/10.1109/tcss.2021.3130401>.

Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:

Deutsches Urheberrecht

Weitere Informationen finden Sie unter: / For more information see:

<https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publiz/>



Rethinking Auditory Affective Descriptors Through Zero-Shot Emotion Recognition in Speech

Xinzhou Xu¹, Jun Deng², Zixing Zhang³, *Member, IEEE*, Xijian Fan⁴, *Member, IEEE*, Li Zhao,
Laurence Devillers, and Björn W. Schuller⁵, *Fellow, IEEE*

Abstract—Zero-shot speech emotion recognition (SER) endows machines with the ability of sensing unseen-emotional states in speech, compared with conventional SER endeavors on supervised cases. On addressing the zero-shot SER task, auditory affective descriptors (AADs) are typically employed to transfer affective knowledge from seen- to unseen-emotional states. However, it remains unknown which types of AADs can well describe emotional states in speech during the transfer. In this regard, we define and research on three types of AADs, namely, per-emotion semantic-embedding, per-emotion manually annotated, and per-sample manually annotated AADs, through zero-shot emotion recognition in speech. This leads to a systematic design including prototype- and annotation-based zero-shot SER modules, relying on the input from per-emotion and per-sample AADs, respectively. We then perform extensive experimental comparisons between human and machines’ AADs on the French emotional speech corpus CINEMO for positive-negative (PN) and within-negative (WN) tasks. The experimental results indicate that semantic-embedding prototypes from pretrained models can outperform manually annotated emotional dimensions in zero-shot SER. The results further demonstrate that it is possible for machines to understand and describe affective information in speech better than human beings, with the help of sufficient pretrained models.

Index Terms—Auditory affective descriptors (AADs), semantic-embedding prototypes, speech emotion recognition (SER), zero-shot emotion recognition.

This work was supported in part by the Natural Science Foundation of China under Grant 61801241, Grant 61802206, Grant 61902187, and Grant 62071242; in part by the Natural Science Foundation of Jiangsu under Grant BK20180746; and in part by the German Research Foundation (DFG) Reinhart Koselleck-Project AUDIONOMOUS under Grant 442218748. (*Corresponding author: Xinzhou Xu.*)

Xinzhou Xu is with the School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China, and also with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany (e-mail: xinzhou.xu@njupt.edu.cn).

Jun Deng is with Agile Robots AG, 81477 Munich, Germany (e-mail: jun.deng@tum.de).

Zixing Zhang is with the Group on Language, Audio, and Music (GLAM), Imperial College London, London SW7 2BX, U.K. (e-mail: zixing.zhang@tum.de).

Xijian Fan is with the College of Information Science and Technology, Nanjing Forestry University, Nanjing 210042, China (e-mail: xijian.fan@njfu.edu.cn).

Li Zhao is with the School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: zhaoli@seu.edu.cn).

Laurence Devillers is with CNRS-LISN, Sorbonne University, 75006 Paris, France (e-mail: devil@limsi.fr).

Björn W. Schuller is with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany, and also with the Group on Language, Audio, and Music (GLAM), Imperial College London, London SW7 2BX, U.K. (e-mail: schuller@ieee.org).

Digital Object Identifier 10.1109/TCSS.2021.3130401

I. INTRODUCTION

THE past two decades witnessed the rapid progressing in paralinguistics of auditory affective computing [1]–[3], consisting of emotion recognition in speech [4]–[6], music [7], and multimodal conditions [8]. Typically, in the research of speech emotion recognition (SER), machines learn to perceive emotional information in speech with the learning procedures appearing in certain settings [5]. Considering these settings’ differences, existing SER topics usually aim at exploring fully supervised [6], [9], [10], semisupervised [11], and data-driven transfer learning [12]–[14] as cases. Furthermore, on the basis of these works, zero-shot learning (ZSL) in SER makes it possible to recognize samples from unheard respectively “unseen” emotions, through information transfer between emotions [15]. This can help machines understand complex implicit intentions or subtle and minor emotional states hidden in social interactions and signals without seeing the corresponding samples [15], [16].

However, these endeavors mainly shed light on exploring the connection between acoustic descriptors (i.e., features extracted from speech) and fixed emotional labels, without investigating how to describe emotional states on the aspect of auditory perception [17]–[19]. A showcase is that when an SER pipeline makes use of acoustic features for categorical emotion recognition, it only has to judge the emotional states of an arbitrary speech utterance [10]. As a result, the samples from all the emotional states will be treated equally, even if some of these states tend to be much closer to each other (e.g., depression and sadness). This makes it appear promising to introduce and explore optimal auditory affective descriptors (AADs) for describing each target emotion in speech. Intuitively, it is usually impossible to evaluate high-level descriptors for each emotional state presented in speech on the aspect of auditory feeling. For example, we have to trust the annotated emotional dimensions in previous research due to the lack of evaluation methods [15], [20]–[22]. To this end, we propose to induce zero-shot SER approaches to design effective systems for investigating and evaluating the AADs.

Therefore, we focus on investigating AADs in this article, following our previous research on zero-shot SER. First, we define three types of AADs as per-emotion semantic-embedding, per-emotion manually annotated, and per-sample manually annotated AADs, considering per-emotion/per-sample and semantic-embedding/manually

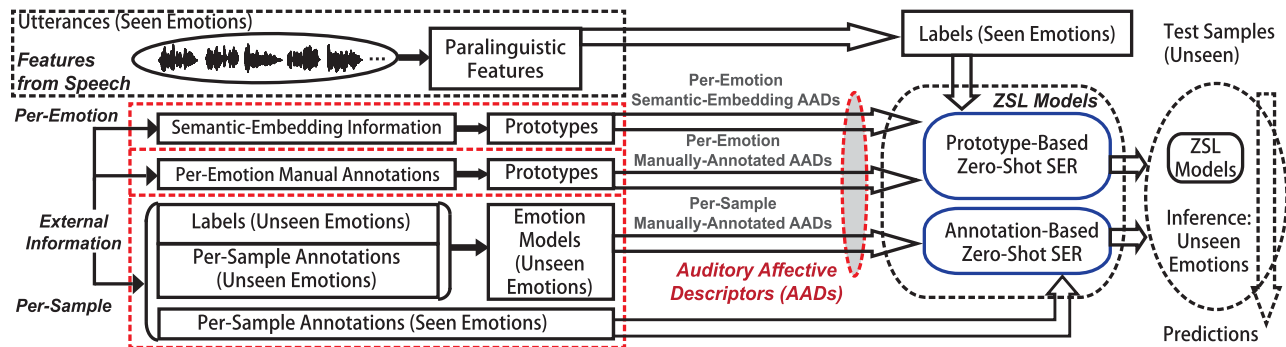


Fig. 1. Diagrammatic overview of the methodology in this work for zero-shot emotion recognition in speech, considering the prototype- and annotation-based cases.

annotated divisions, as shown in Fig. 1. The first type employs prototypes from textual semantics for each emotional state [23], while the second type sets its per-emotion prototypes through manually annotating. In contrast, the third type utilizes manual auditory annotations (i.e., emotional dimensions) on samples to describe each emotion [15]. Afterward, this work aims at answering two questions in relation to these AADs: 1) which source of AADs performs better, manually annotated or semantic embedding? and 2) can per-emotion prototypes outperform manual per-sample-annotation modeling as AADs? To answer these questions, we here resort to solving zero-shot SER tasks using the three AADs types to connect seen- and unseen-emotional states. For the first question, a prototype-based zero-shot SER model is set for the semantic-embedding or manually annotated AADs, while for the second question, we add an annotation-based model for processing the per-sample AADs. Then, these two types of models help to perform the AAD comparisons through analyzing zero-shot SER results.

This work follows our early research on exploring by further using semantic-embedding and manually annotated prototypes for comparison, in order to investigate AADs [15], [23]. Recent “conventional” SER approaches in this direction mainly focus on data-driven transfer learning [12]–[14] and few-shot learning [24], while this work is related to zero-shot information transfer between emotional states. Compared with the work of emotion embedding [25], [26], we aim to research on the affective mediation in auditory perception. In addition, the existing research of ZSL in image processing and affective computing provides different strategies to this work [27]–[30].

Our main contributions are highlighted as follows.

- 1) We define three AAD types of per-emotion semantic-embedding, per-emotion manually annotated, and per-sample manually annotated AADs for unifying emotional representations in speech.
- 2) We investigate the AADs using prototype-based zero-shot SER approaches, in order to analyze the manual and semantic-embedding emotional representations.
- 3) We further investigate the AADs using annotation-based zero-shot SER approaches, for the purpose of presenting the performance for per-emotion and per-sample AADs.

The remainder of this article is organized as follows. The basic concepts and notations are presented in Section II.

Afterward, we show the methodology for per-emotion prototypes and per-sample annotations as AADs in Section III. Then, Sections IV and V present the experimental results, experimental analysis, and the conclusions.

II. PRELIMINARIES

A. Concepts

1) *Auditory Affective Descriptors*: Intuitively, AADs have to be extracted from speech, related to feature extraction for emotions in speech [31], [32]. Unfortunately, this type of descriptors usually fails to represent emotional states due to the variety of speech data, despite the previous endeavors [6], [10]. Thus, most works resort to rating speech in emotional spaces using limited numbers of dimensions to annotate speech samples or emotional states [33]–[35]. This setup seems robust and reasonable, yet none of these spaces can completely describe every emotional state because of the complexity of emotional expression [36]–[38].

Besides learning models using per-sample annotations on emotional dimensions [15], the research of semantic embedding provides a solution for the lack of critical information when using low-dimensional spaces to describe emotions in speech [39], [40]. Typically, the usage of semantic-embedding prototypes AADs shows the feasibility on describing emotions in speech through textual representations, where the “prototype” refers to the most representative example of its corresponding category [23], [41]. We induce three types of AADs of per-emotion semantic-embedding, per-emotion manually annotated, and per-sample manually annotated ones in this article, divided by their different sources (semantic embedding or manually annotated) and forms (per emotion or per sample).

2) *Semantic-Embedding Prototypes*: The prototype of an arbitrary sample set typically refers to the representative for majority of the samples [42]–[44]. This implies that sample-wise information is not strictly required in learning procedures on a training set, possibly leading to the performance improvement through using the form of prototypes for the set. Built on mining semantic information from language, semantic-embedding prototypes aim to represent concepts (i.e., class labels), typically through learning embedding models on textual data conventionally used in natural language processing (NLP) applications [23], [45]–[47].

3) *Zero-Shot SER*: Conventional ZSL approaches make it possible to recognize the samples from unseen classes only with seen-class samples contained in training [48]–[50]. Note that the samples from the unseen classes never appear in the training set, which leads to cognitive differences between seen and unseen classes [51], [52]. The learning procedures for zero-shot SER 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 [15], [23], [29]. Within the zero-shot SER task, the seen-emotional samples refer to the speech samples from the emotional states already provided in training. In contrast, the unseen-emotional samples indicate that for some emotions (unseen-emotional states), it is unavailable to obtain any of the corresponding samples in the training procedures [23].

Our original work sheds light on the possibility to perform SER in zero-shot settings [15], focusing on constructing connections between paralinguistic features, emotional dimensions, and emotional labels. The results prove that it is feasible to perform zero-shot SER using empirical descriptions on emotional labels. Furthermore, in view of the workload for drawing the descriptions using additional annotations, we successfully include semantic-embedding prototypes for constructing the connection in zero-shot SER, relying on knowledge transfer from the textual modality [23].

B. Notations

We define the label sets from seen and unseen-emotional states as $\mathcal{D}^{(S)} = \{d_1^{(S)}, d_2^{(S)}, \dots, d_{c^{(S)}}^{(S)}\}$ and $\mathcal{D}^{(U)} = \{d_1^{(U)}, d_2^{(U)}, \dots, d_{c^{(U)}}^{(U)}\}$, containing $c^{(S)}$ and $c^{(U)}$ classes, respectively. Let $\mathcal{D}^{(S)} \cap \mathcal{D}^{(U)} = \emptyset$, implying the nongeneralized ZSL setting to simplify comparisons. Furthermore, we represent the $N^{(S)}$ seen-emotional samples as $\mathbf{X}^{(S)} = [\mathbf{x}_1^{(S)}, \mathbf{x}_2^{(S)}, \dots, \mathbf{x}_{N^{(S)}}^{(S)}] \in \mathfrak{R}^{n_F \times N^{(S)}}$ with n_F -dimensional paralinguistic features. We also set the emotional labels for the samples in $\mathbf{X}^{(S)}$ as $\mathcal{Y}^{(S)} = \{y_1^{(S)}, y_2^{(S)}, \dots, y_{N^{(S)}}^{(S)}\} \subset \mathcal{D}^{(S)}$. For an arbitrary sample $\mathbf{x}^{(U)} \in \mathfrak{R}^{n_F \times 1}$ in the unseen-emotional set, the predicted emotional label is $\hat{y}^{(U)} \in \mathcal{D}^{(U)}$.

For the case of using per-emotion prototypes, the n_A -dimensional prototypes of $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(U)}$ are $\mathbf{A}^{(S)} = [\mathbf{a}_1^{(S)}, \mathbf{a}_2^{(S)}, \dots, \mathbf{a}_{c^{(S)}}^{(S)}] \in \mathfrak{R}^{n_A \times c^{(S)}}$ and $\mathbf{A}^{(U)} = [\mathbf{a}_1^{(U)}, \mathbf{a}_2^{(U)}, \dots, \mathbf{a}_{c^{(U)}}^{(U)}] \in \mathfrak{R}^{n_A \times c^{(U)}}$, respectively. Then, we denote the samplewise prototypes for seen-emotional sample as $\mathbf{Z}^{(S)} = [\mathbf{z}_1^{(S)}, \mathbf{z}_2^{(S)}, \dots, \mathbf{z}_{N^{(S)}}^{(S)}] \in \mathfrak{R}^{n_A \times N^{(S)}}$, containing each of its column equal to the corresponding prototype from $\mathbf{A}^{(S)}$ to replace the emotional labels for $\mathcal{Y}^{(S)}$. For the case of emotion modeling using per-sample annotations, we still use $\mathbf{Z}^{(S)}$ with each of its column corresponding to a samplewise annotation vector, where n_A is the dimensionality of the per-sample annotations. Note that we employ the same n_A for these two cases since the per-emotion prototypes can also be represented as “ $\mathbf{Z}^{(S)}$ ” through the prototypes’ duplication [23]. We also set $\mathbf{Z}^{(U)} = [\mathbf{z}_1^{(U)}, \mathbf{z}_2^{(U)}, \dots, \mathbf{z}_{N^{(U)}}^{(U)}] \in \mathfrak{R}^{n_A \times N^{(U)}}$ and $\mathcal{Y}^{(U)} = \{y_1^{(U)}, y_2^{(U)}, \dots, y_{N^{(U)}}^{(U)}\}$ as the input and the target to train classifiers, instead of directly using the prototypes,

where $N^{(U)}$ is the number of unseen-emotional samples in simulating empirical procedures [15].

III. METHODOLOGY

We set two types of zero-shot SER models considering prototype- and annotation-based cases using per-emotion prototypes and per-sample annotations, respectively, as shown in Fig. 1. Both the types employ AADs and seen-emotion data in learning zero-shot SER models, in order to recognize unseen-emotional samples. Note that the prototype-based case directly employs per-emotion AADs to connect seen- and unseen-emotional states, while the annotation-based case aims to achieve the connection through modeling unseen-emotional states using the per-sample annotations.

A. Types of AADs

The AADs can be categorized on two aspects of forms and sources, leading to two binary division types of per-emotion/per-sample and semantic-embedding/manually annotated AADs. The first aspect focuses on representing each emotional state using unique or multiple representations, while the latter one considers which sources the AADs are obtained from, the automatically generated textual-data sources or the annotation sources from human annotators. This results in the per-emotion semantic-embedding, per-emotion manually annotated, and per-sample manually annotated AADs.

Following the per-emotion/per-sample division type, we design prototype-based (Section III-B) and annotation-based (Section III-C) zero-shot SER modeling types, for processing the per-emotion and per-sample AADs, respectively. For the prototype-based type, the per-emotion AADs (including the per-emotion semantic-embedding and per-emotion manually annotated AAD types) represented as “ $\mathbf{A}^{(S)}$ ” and “ $\mathbf{A}^{(U)}$ ” are defined using each of their columns as the prototype for the corresponding seen- or unseen-emotional state. Within the AADs, the prototypes can be obtained from either semantic-embedding or manually annotated sources. For the semantic-embedding source, we typically employ the presentations generated by the pretrained models from learned on textual data, while the manually annotated source comes from combining the per-sample annotations using multiple dimensions to describe one’s auditory affective feeling. In contrast, for the annotation-based type, the per-sample AADs (including the per-sample manually annotated AAD type) can be directly obtained from the auditory affective per-sample annotations, which are represented as “ $\mathbf{Z}^{(S)}$ ” and “ $\mathbf{Z}^{(U)}$ ” with their columns as the corresponding per-sample annotation vectors.

B. Prototypes as AADs

First, we focus on the framework of using prototypes $\mathbf{A}^{(S)}$ and $\mathbf{A}^{(U)}$ as AADs. Note that the prototypes include both manual and semantic-embedding sources as in Fig. 1. Within this framework, it is expected to build connection between paralinguistic features $\mathbf{X}^{(S)}$ and emotional-label information $\{\mathbf{A}^{(S)}, \mathcal{Y}^{(S)}\}$ or its equivalent form $\mathbf{Z}^{(S)}$, through training the

corresponding parameter set $\Psi_{(\text{Pro})}$ for the connection modeling. Thus, the training procedure aims at optimizing the objective function $f(\mathbf{X}^{(S)}, \mathbf{A}^{(S)}, \mathcal{Y}^{(S)}; \Psi_{(\text{Pro})})$ employing seen-emotional information with the optimal parameter set

$$\widehat{\Psi}_{(\text{Pro})} = \arg \max_{\Psi_{(\text{Pro})}} p(\mathbf{A}^{(S)}, \mathcal{Y}^{(S)} | \mathbf{X}^{(S)}; \Psi_{(\text{Pro})}). \quad (1)$$

When obtaining the optimal parameters $\widehat{\Psi}_{(\text{Pro})}$, we start the inference procedure to calculate the predicted label $\widehat{y}^{(U)}$ for an arbitrary unseen-emotional sample $\mathbf{x}^{(U)}$. To this end, the predicted emotional index \widehat{j} in $\mathcal{D}^{(U)}$ for $\mathbf{x}^{(U)}$ is denoted as

$$\widehat{j} = \arg \max_j p(d_j^{(U)}, \mathbf{A}^{(U)} | \mathbf{x}^{(U)}; \widehat{\Psi}_{(\text{Pro})}) \quad (2)$$

where $j = 1, 2, \dots, c^{(U)}$. Using the predicted index \widehat{j} , $\mathbf{x}^{(U)}$'s predicted label is $\widehat{y}^{(U)} = d_{\widehat{j}}^{(U)}$.

Within the framework, we present three ZSL strategies of embarrassingly simple zero-shot learning (ESZSL) [28], synthesized classifiers (SYNCs) [27], [53], and EXEMplar synthesis (EXEM) [27], [52], with their parameter sets Ψ s represented as $\{\mathbf{W}_{(\text{ES})}\}$, $\{\mathbf{V}_{(\text{SYNC})}\}$, and $\{\psi_{(\text{EXEM})}\}$, respectively. Then, we briefly introduce the three strategies in training and inference procedures.

1) *Training Procedure*: The ESZSL strategy makes use of combinations $\mathbf{W}_{(\text{ES})}$ to connect $\mathbf{X}^{(S)}$, $\mathbf{A}^{(S)}$, and $\mathcal{Y}^{(S)}$ in a discriminative form, where for the kernelled case $\mathbf{W}_{(\text{ES})} \in \mathfrak{R}^{N^{(S)} \times n_A}$ and the Gram matrix $\mathbf{K}(\mathbf{X}^{(S)}, \mathbf{X}^{(S)}) = \phi^T(\mathbf{X}^{(S)})\phi(\mathbf{X}^{(S)})$ using the reproducing kernel Hilbert space (RKHS) [6], [10] mapping $\phi(\cdot)$ on $\mathbf{X}^{(S)}$'s columns. Thus, the optimal $\mathbf{W}_{(\text{ES})}$ is represented as

$$\begin{aligned} \widehat{\mathbf{W}}_{(\text{ES})} &= \arg \min_{\mathbf{W}_{(\text{ES})}} f_{(\text{ES})}(\mathbf{X}^{(S)}, \mathbf{A}^{(S)}, \mathcal{Y}^{(S)}; \mathbf{W}_{(\text{ES})}) \\ &= \arg \min_{\mathbf{W}_{(\text{ES})}} (L(\mathbf{K}(\mathbf{X}^{(S)}, \mathbf{X}^{(S)})\mathbf{W}_{(\text{ES})}\mathbf{A}^{(S)}, \mathcal{Y}^{(S)}) + R(\mathbf{W}_{(\text{ES})})) \end{aligned} \quad (3)$$

jointly minimizing the regularization term $R(\mathbf{W}_{(\text{ES})})$ and the dissimilarity loss $L(\cdot, \cdot)$ with a measure of Frobenius norm distance.

The fast SYNC strategy [27], [53] aims to learn optimal linear phantom classifiers $\mathbf{V}_{(\text{SYNC})} \in \mathfrak{R}^{n_F \times c^{(P)}}$ for supervised seen-emotional samples, where $c^{(P)}$ represents the number of the phantom classifiers. $\mathbf{V}_{(\text{SYNC})}$ connects seen-emotional classifiers $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{c^{(S)}}] \in \mathfrak{R}^{n_F \times c^{(S)}}$ and unseen-emotional classifiers using similarity matrices $\mathbf{S}^{(S)} \in \mathfrak{R}^{c^{(P)} \times c^{(S)}}$ and $\mathbf{S}^{(U)} \in \mathfrak{R}^{c^{(P)} \times c^{(U)}}$, respectively. Note that the (c_P, c) element for the similarity matrices is equal to $(e^{-\text{Dis}(\mathbf{a}_c^{(S)}, \mathbf{b}_{c_P})} / \sum_{c_P=1}^{c^{(P)}} e^{-\text{Dis}(\mathbf{a}_c^{(S)}, \mathbf{b}_{c_P})})$, with $c = 1, 2, \dots, c^{(S)}$ for $\mathbf{S}^{(S)}$ while $c = 1, 2, \dots, c^{(U)}$ for $\mathbf{S}^{(U)}$, where $\text{Dis}(\mathbf{a}_c^{(S)}, \mathbf{b}_{c_P}) = \sigma^2 \|\mathbf{a}_c^{(S)} - \mathbf{b}_{c_P}\|^2$. Note that $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_{c^{(P)}}] \in \mathfrak{R}^{n_A \times c^{(P)}}$ represents the phantom prototypes. Hence, the optimal phantom classifiers

$$\begin{aligned} \widehat{\mathbf{V}}_{(\text{SYNC})} &= \arg \min_{\mathbf{V}_{(\text{SYNC})}} f_{(\text{SYNC})}(\mathbf{X}^{(S)}, \mathbf{A}^{(S)}, \mathcal{Y}^{(S)}; \mathbf{V}_{(\text{SYNC})}) \\ &= \arg \min_{\mathbf{V}_{(\text{SYNC})}} \left(J(\mathbf{X}^{(S)}, \mathcal{Y}^{(S)}, \mathbf{W}) + \frac{\tau}{2} \text{tr}(\mathbf{W}^T \mathbf{W}) \right) \\ \text{s.t. } \mathbf{W} &= \mathbf{V}_{(\text{SYNC})} \mathbf{S}^{(S)} \end{aligned} \quad (4)$$

with $\tau > 0$ representing the weight of the regularization term. We set the one-versus-other (OVO) loss as

$$\begin{aligned} J(\mathbf{X}^{(S)}, \mathcal{Y}^{(S)}, \mathbf{W}) &= \sum_{c=1}^{c^{(S)}} \sum_{i=1}^{N^{(S)}} \left(\max \left(0, 1 - \Delta \left(y_i^{(S)}, d_c^{(S)} \right) \mathbf{w}_c^T \mathbf{x}_i^{(S)} \right) \right)^2 \end{aligned} \quad (5)$$

in which $\Delta(y_i^{(S)}, d_c^{(S)})$ is equal to 1 when $y_i^{(S)} = d_c^{(S)}$ while -1 when $y_i^{(S)} \neq d_c^{(S)}$.

The EXEM strategy optimizes the relationship between exemplars $\mathbf{U}^{(S)} = [\mathbf{u}_1^{(S)}, \mathbf{u}_2^{(S)}, \dots, \mathbf{u}_{c^{(S)}}^{(S)}] \in \mathfrak{R}^{n_{\text{DR}} \times c^{(S)}}$ and prototypes $\mathbf{A}^{(S)}$, where n_{DR} represents the reduced dimensionality for principal component analysis (PCA) on seen-emotional samples. Thus, the optimal mapping $\psi_{(\text{EXEM})}(\cdot)$ can be obtained through using ν -support vector regression (ν -SVR) as

$$\begin{aligned} \widehat{\psi}_{(\text{EXEM})} &= \arg \min_{\psi_{(\text{EXEM})}} f_{(\text{EXEM})}(\mathbf{X}^{(S)}, \mathbf{A}^{(S)}, \mathcal{Y}^{(S)}; \psi_{(\text{EXEM})}) \\ &= \arg \min_{\psi_{(\text{EXEM})}} J_0(\psi_{(\text{EXEM})}(\mathbf{A}^{(S)}), \mathbf{U}^{(S)}) \\ \text{s.t. } \mathbf{u}_c &= \frac{\Omega \mathbf{X}_c^{(S)} \mathbf{e}_c}{N_c^{(S)}} \end{aligned} \quad (6)$$

where $J_0(\cdot, \cdot)$ represents the loss of ν -SVR and $\Omega \in \mathfrak{R}^{n_{\text{DR}} \times n_F}$ is the linear mapping of the PCA processing. Note that $N_c^{(S)}$ samples are contained in $d_c^{(S)}$ with $c = 1, 2, \dots, c^{(S)}$ and all the elements of $\mathbf{e}_c \in \mathfrak{R}^{N_c^{(S)} \times 1}$ are equal to 1.

2) *Inference Procedure*: Using the optimal parameters in the training procedure, we achieve the predicted index \widehat{j} for the three strategies. For ESZSL, the predicted j is

$$\widehat{j} = \arg \max_j \left(\mathbf{K}(\mathbf{X}^{(S)}, \mathbf{x}^{(U)})^T \widehat{\mathbf{W}}_{(\text{ES})} \mathbf{a}_j^{(U)} \right) \quad (7)$$

in which $\mathbf{K}(\mathbf{X}^{(S)}, \mathbf{x}^{(U)}) = \phi^T(\mathbf{X}^{(S)})\phi(\mathbf{x}^{(U)})$. For SYNC, the predicted emotional-state index

$$\widehat{j} = \arg \max_j \left(\left(\widehat{\mathbf{V}}_{(\text{SYNC})} \mathbf{s}_j^{(U)} \right)^T \mathbf{x}^{(U)} \right) \quad (8)$$

where $\mathbf{s}_j^{(U)}$ is $\mathbf{S}^{(U)}$'s j th column. Then, the predicted index for EXEM can be achieved as

$$\widehat{j} = \arg \min_j \text{Dis}(\Omega \mathbf{x}^{(U)}, \widehat{\psi}_{(\text{EXEM})}(\mathbf{a}_j^{(U)})) \quad (9)$$

where $\text{Dis}(\cdot, \cdot)$ represents a distance function.

C. Per-Sample Annotations as AADs

It is also applicable to employ per-sample annotations as AADs for modeling emotional states when providing empirical external information on describing these states [15].

1) *Training Procedure*: The training procedure, in this case, includes two learning steps, to make emotional-state decisions on paralinguistic features, through learning optimal parameters of $\Psi_{(\text{Ann})1}$ and $\Psi_{(\text{Ann})2}$.

First, we expect to learn the emotion models with parameters $\Psi_{(\text{Ann})1}$ for unseen emotions using the annotations $\mathbf{Z}^{(U)}$ and their emotional labels $\mathcal{Y}^{(U)}$ as

$$\widehat{\Psi}_{(\text{Ann})1} = \arg \max_{\Psi_{(\text{Ann})1}} p(\mathcal{Y}^{(U)} | \mathbf{Z}^{(U)}; \Psi_{(\text{Ann})1}) \quad (10)$$

TABLE I

DESCRIPTION OF THE CINEMO CORPUS FOR THE PN AND WN TASKS, INCLUDING SPEAKERS, SAMPLES, LABELS, AND ANNOTATIONS

Properties \ Tasks	PN Task	WN Task
Language & Sampling Rate	French & 16 kHz	
# Speakers	51 (21 female)	
# Samples	3992 (3591 here; 1380 female)	
# Emotional Labels & Dimensions	16 Emotions & 6 Dimensions	
Annotation Levels	{1, 2, 3} (for each dimension)	
Total Duration (hrs:min:sec)	2 : 04 : 45 here	
# Seen-Emotional Combinations	92	95
# Seen-Emotional Samples	2791	3018

for the optimization object $g_1(\mathcal{Y}^{(U)}, \mathbf{Z}^{(U)}; \Psi_{(\text{Ann})1})$. This learning step is equivalent to training emotional classifiers on the intermediated descriptors.

Then, for the seen-emotional information of $\mathbf{Z}^{(S)}$ and $\mathbf{X}^{(S)}$, the connection between paralinguistic features and annotations is represented by optimizing the parameters $\Psi_{(\text{Ann})2}$ as

$$\hat{\Psi}_{(\text{Ann})2} = \arg \max_{\Psi_{(\text{Ann})2}} p(\mathbf{Z}^{(S)} | \mathbf{X}^{(S)}; \Psi_{(\text{Ann})2}) \quad (11)$$

for the optimization object $g_2(\mathbf{Z}^{(S)}, \mathbf{X}^{(S)}; \Psi_{(\text{Ann})2})$. Note that this step can be implemented through n_A regressors on paralinguistic features.

2) *Inference Procedure*: Combining the optimized parameters in the two steps, we obtain the index of the predicted unseen-emotional label for an arbitrary sample $\mathbf{x}^{(U)}$ as

$$\hat{j} = \arg \max_j p(d_j^{(U)} | \mathbf{x}^{(U)}; \{\hat{\Psi}_{(\text{Ann})1}, \hat{\Psi}_{(\text{Ann})2}\}) \quad (12)$$

for the inference procedure, utilizing the optimal-parameter sets $\{\hat{\Psi}_{(\text{Ann})1}, \hat{\Psi}_{(\text{Ann})2}\}$ for the cascaded classifiers and regressors.

IV. EXPERIMENTS

A. Experimental Preparation

1) *Corpus and Features*: As shown in Table I, we employ the CINEMO corpus [16], [54]–[56] containing French emotional speech in the experiments, consisting of 3992 French segmentwise utterances recorded from 51 speakers (21 female) in four age groups (–15 years, 15–25 years, 25–50 years, and 50+ years), with the sampling rate of 16 kHz. The corpus considers dubbing 29 selected scenes from totally 12 French movies by the speakers. Each of these scenes could consist of one or two players at a time. The data collection paradigm involved the speakers, none of whom had professional acting experience [16]. Two persons (1 female) marked each utterance as having a major and a minor emotion label, taken from one of 16 states: “amusement (AMU),” “anger (COL),” “disappointment (DEC),” “irritation (ENE),” “anxiety (INQ),” “irony (IRO),” “joy (JOI),” “negativity (NEG),” “neutrality (NEU),” “fear (PEU),” “positivity (POS),” “satisfaction (SAT),” “seduction (SED),” “stress (STR),” “surprise (SUR),” and “sadness (TRI).” In addition, each sample was annotated in six emotional dimensions, namely, “intensity,” “activation,” “valence,” “control,” “suddenness,” and “naturalness,” using the levels from 1 to 3. The first annotator is provided the context in sequential order and manually segmented the audio

TABLE II

MEAN UAs AND THEIR STANDARD DEVIATIONS (%) FOR THE PN AND WN TASKS IN THE SPC10 CASE WHEN USING DIFFERENT AADs AND STRATEGIES FOR ZERO-SHOT SER

AADs \ Strategies	ESZSL	SYNC	EXEM
PN Task:			
Manual Descriptors	65.5 ± 2.5	59.5 ± 1.0	66.5 ± 3.1
<i>ft-crawl</i>	59.1 ± 3.8	72.7 ± 0.4	70.3 ± 0.5
<i>ft-wiki</i>	59.8 ± 3.5	71.7 ± 0.8	71.0 ± 0.2
<i>GloVe</i>	58.7 ± 2.4	72.0 ± 1.0	70.3 ± 0.3
<i>word2vec</i>	57.9 ± 1.1	71.8 ± 1.0	70.5 ± 0.9
<i>ft-crawl + SN</i>	58.3 ± 1.3	71.4 ± 1.4	69.8 ± 1.5
<i>ft-wiki + SN</i>	57.6 ± 0.7	71.9 ± 0.5	71.5 ± 1.6
<i>GloVe + SN</i>	58.4 ± 1.1	72.4 ± 1.1	69.9 ± 1.5
<i>word2vec + SN</i>	57.5 ± 0.9	72.0 ± 0.8	70.3 ± 1.7
WN Task:			
Manual Descriptors	59.9 ± 1.5	57.8 ± 2.5	63.7 ± 1.5
<i>ft-crawl</i>	70.9 ± 1.9	69.9 ± 3.1	72.3 ± 1.0
<i>ft-wiki</i>	70.6 ± 0.6	70.6 ± 2.0	72.3 ± 0.6
<i>GloVe</i>	70.6 ± 2.6	66.2 ± 2.2	68.6 ± 2.0
<i>word2vec</i>	69.9 ± 3.2	71.6 ± 1.4	71.0 ± 0.5
<i>ft-crawl + SN</i>	70.6 ± 2.0	72.0 ± 0.1	71.6 ± 0.5
<i>ft-wiki + SN</i>	70.2 ± 2.2	69.8 ± 1.7	69.8 ± 2.6
<i>GloVe + SN</i>	70.3 ± 2.6	67.8 ± 1.7	65.6 ± 1.4
<i>word2vec + SN</i>	69.9 ± 0.3	68.9 ± 0.2	71.2 ± 1.5

signals, whereas the second annotator is provided with single instances after segmentation in random order for verification [16], [56]. Note that the annotations are credible for the experiments, in view of the procedures and assessments for the annotations [54]. We choose a subset of 3591 utterances (1380 female) in accordance with [15], only considering major emotion labels. This yields 136, 373, 434, 1206, 443, 21, 140, 6, 13, 8, 41, 292, 45, 248, 51, and 134 samples corresponding to these 16 major emotions for the first annotator, while 173, 384, 370, 1234, 565, 16, 101, 12, 43, 21, 16, 218, 34, 176, 17, and 211 samples are for the second annotator.

We investigate the positive–negative (PN) and within-negative (WN) tasks from the view of valence using full-agreement major-emotion labels as in [15]. The PN task defines the positive emotions classes as “amusement” and “satisfaction,” while the negative emotions classes were “anger,” “stress,” and “sadness.” In the WN task, we aim to classify the “anger” and “disappointment” states due to existing research of social reactions [57]. Note that the remaining samples in the corpus for the two tasks are both set as seen-emotional samples, contained in the training sets.

The OPENSIMILE toolkit [58], [59] is employed in the experiments, for extracting 88-D ($n_F = 88$) extended Geneva minimalistic acoustic parameter set (eGeMAPS) (from functionals on 25 time-smoothed low-level descriptors (LLDs), temporal features, and equivalent sound level) [59] as the paralinguistic features, which has been proven effective in SER tasks when using support vector machines (SVMs). Note that we perform a min–max normalization for each feature.

2) *Per-Emotion Prototypes and Per-Sample Annotations*: For an arbitrary sample in the corpus, we choose the average-level values of the two annotators on each of the six dimensions ($n_A = 6$) as the per-sample annotations. This simulates the process of constructing emotion models and the connection between paralinguistic features and the annotations from manual knowledge. Furthermore, we regard each combination of the major emotions for the two annotators as an emotional state, without considering the order of the annotators. For each

TABLE III

MEAN UAs AND THEIR STANDARD DEVIATIONS (%) FOR THE THREE SUBTASKS WITHIN THE PN TASK AND THE SPC10 CASE, WHEN USING MANUAL AND SEMANTIC DESCRIPTORS FOR DIFFERENT STRATEGIES

AADs \ Strategies	ESZSL	SYNC	EXEM
PN Task (Bin.):			
Manual Descriptors	65.5 ± 2.5	59.5 ± 1.0	66.5 ± 3.1
Semantic (w/o SN)	58.9 ± 2.9	72.1 ± 0.9	70.5 ± 0.6
Semantic (w. SN)	58.0 ± 1.1	71.9 ± 1.0	70.4 ± 1.7
PN Task (Mul.):			
Manual Descriptors	66.0 ± 2.2	58.0 ± 2.8	68.7 ± 2.4
Semantic (w/o SN)	63.2 ± 1.2	70.5 ± 1.8	72.2 ± 0.8
Semantic (w. SN)	62.9 ± 0.8	67.6 ± 4.1	73.0 ± 1.4
Multiple Emotions:			
Manual Descriptors	41.1 ± 2.0	27.2 ± 3.2	39.2 ± 2.2
Semantic (w/o SN)	40.8 ± 1.2	42.4 ± 1.1	48.8 ± 2.0
Semantic (w. SN)	40.4 ± 0.7	44.2 ± 4.4	47.6 ± 2.9

TABLE IV

PAIRWISE COMPARISONS OF UAs WHEN USING THE SYNC STRATEGY ON THE FACTOR OF AADS USING *Post Hoc* TUKEY’S HSD, FOR PN TASK (WITH PRIOR BINARY CLASSIFICATION, NOTED AS “BIN.”) AND PN TASK (WITH LATE FUSION ON MULTIPLE EMOTIONS, NOTED AS “MUL.”) WITH SPC10 AND SPC1 CASES. WE PRESENT THE MEAN DIFFERENCE (MEAN-UA DIFFERENCE BETWEEN THE SEMANTIC-EMBEDDING AND MANUALLY ANNOTATED AADS, NOTED AS “MD”) AND THE p VALUE FOR EACH COMPARISON

Setups \ Tasks AADs & SPC	PN Task (Bin.)		PN Task (Mul.)		WN Task	
	MD	p Value	MD	p Value	MD	p Value
SPC10 Case:						
<i>ft-crawl</i>	.132	.001	.133	.001	.121	.001
<i>ft-wiki</i>	.122	.001	.134	.001	.128	.001
<i>GloVe</i>	.125	.001	.112	.001	.084	.001
<i>word2vec</i>	.123	.001	.121	.001	.138	.001
<i>ft-crawl + SN</i>	.119	.001	.138	.001	.142	.001
<i>ft-wiki + SN</i>	.124	.001	.114	.001	.120	.001
<i>GloVe + SN</i>	.129	.001	.067	.001	.100	.001
<i>word2vec + SN</i>	.125	.001	.067	.001	.111	.001
SPC1 Case:						
<i>ft-crawl</i>	.096	.001	.082	.001	.135	.001
<i>ft-wiki</i>	.090	.001	.073	.001	.142	.001
<i>GloVe</i>	.073	.001	.056	.001	.129	.001
<i>word2vec</i>	.087	.001	.044	.002	.156	.001
<i>ft-crawl + SN</i>	.078	.001	.098	.001	.104	.001
<i>ft-wiki + SN</i>	.094	.001	.055	.001	.061	.001
<i>GloVe + SN</i>	.097	.001	.072	.001	.100	.001
<i>word2vec + SN</i>	.065	.001	.047	.001	.137	.001

emotional state, we define the manual prototype of a state as the average of the annotations from its samples.

When inducing the per-emotion semantic-embedding prototypes of 300-D ($n_A = 300$) English word-vector representations in describing emotional states, we employ semantic-embedding prototypes from pretrained textual models, considering the average values of the major emotions between the two labelers. The sources of the semantic-embedding prototypes include word2vec [60], [61], GloVe [62], and fastText [63], [64] models. Note that we employ these pretrained models to generate the semantic-embedding prototypes since the models aim at producing the representations for each emotional state in speech through learning on textual data. The pretrained word2vec model utilizes Google News corpus in training with 3 million words (100 billion tokens) [30], [60], while the GloVe model considers 0.4 million vocabularies (6 billion tokens) as its training set [62]. We choose two pretrained models using Common

TABLE V

PAIRWISE COMPARISONS OF UAs WHEN USING THE EXEM STRATEGY ON THE FACTOR OF AADS USING *Post Hoc* TUKEY’S HSD, FOR PN TASK (WITH PRIOR BINARY CLASSIFICATION, NOTED AS “BIN.”) AND PN TASK (WITH LATE FUSION ON MULTIPLE EMOTIONS, NOTED AS “MUL.”) WITH SPC10 AND SPC1 CASES, WHERE THE “*” INDICATES THE INSIGNIFICANT RESULTS AT THE SIGNIFICANCE LEVEL OF 0.05. WE PRESENT THE MEAN DIFFERENCE (MEAN-UA DIFFERENCE BETWEEN THE SEMANTIC-EMBEDDING AND MANUALLY ANNOTATED AADS, NOTED AS “MD”) AND THE p VALUE FOR EACH COMPARISON

Setups \ Tasks AADs & SPC	PN Task (Bin.)		PN Task (Mul.)		WN Task	
	MD	p Value	MD	p Value	MD	p Value
SPC10 Case:						
<i>ft-crawl</i>	.038	.001	.038	.001	.086	.001
<i>ft-wiki</i>	.045	.001	.040	.001	.086	.001
<i>GloVe</i>	.038	.001	.025	.001	.049	.001
<i>word2vec</i>	.040	.001	.037	.001	.073	.001
<i>ft-crawl + SN</i>	.033	.001	.044	.001	.079	.001
<i>ft-wiki + SN</i>	.050	.001	.038	.001	.061	.001
<i>GloVe + SN</i>	.034	.001	.045	.001	.019	.103*
<i>word2vec + SN</i>	.038	.001	.044	.001	.075	.001
SPC1 Case:						
<i>ft-crawl</i>	.023	.001	.006	.900*	.138	.001
<i>ft-wiki</i>	.018	.001	.009	.710*	.143	.001
<i>GloVe</i>	.013	.039	.013	.251*	.130	.001
<i>word2vec</i>	.016	.008	.007	.900*	.134	.001
<i>ft-crawl + SN</i>	.007	.643*	.010	.596*	.158	.001
<i>ft-wiki + SN</i>	.001	.900*	.003	.900*	.154	.001
<i>GloVe + SN</i>	-.003	.900*	.013	.265*	.167	.001
<i>word2vec + SN</i>	-.007	.783*	.003	.900*	.163	.001

Crawl (noted as “ft-crawl”) and Wikipedia 2017 with UMBC webbase corpus and the statmt.org news dataset (noted as “ft-wiki”) [63], [64]. We also include the cases with and without using the neighbors in SenticNet 5 [65], [66] (noted as “SN”) to mine sentiment information for textual commonsense concepts using a long short-term memory (LSTM)-based recurrent neural network (RNN) [65]. The SenticNet 5 employs an average combination for each word vector of a corresponding emotional state’s five neighbors (if the neighboring words exist) in order to extend the word-vector models. This results in eight setups of the semantic-embedding prototypes.

3) *Parametric Setups of ZSL Strategies:* For the case of using per-sample annotations, we perform speaker-independent threefold cross validation (CV) considering individual differences of speakers [67]–[69] for the unseen-emotional samples as in [15], in order to connect the manual per-sample annotations and the emotional states. This procedure results in using two folds to learn unseen-emotional models for each validation step, considering SVMs as classifiers with the regularization parameter varying from 0.0001 to 10000 and the Gaussian-kernel scaling parameter in $\{0.01n_A, 0.1n_A, n_A, 10n_A\}$ [15]. Hence, it is guaranteed that there exists no overlapping between the steps of learning and test through employing this CV procedure. Note that for the threefold CV, we collect all the test-sample labels prior to evaluating for the purpose of fair comparisons. Then, the connection between features and annotations is characterized by SVR with its scaling parameter in $\{0.01n_F, 0.1n_F, n_F, 10n_F\}$ and the same selections of regularization parameter as in the SVMs. In addition, neural networks are employed considering 12 selections of the hidden-layer

neurons as: (32, 8), (32, 16), (64, 16), . . . , (1 024, 512), with rectified linear unit (ReLU) activation [15].

For the prototype-based cases, we utilize emotion-independent tenfold CV on the seen-emotional set using grid searching to obtain optimal parameters for the prototype-based models. Note that these tenfold CVs differ from the CV setups in the existing works on setting constraints for the fold splitting [70]–[72]. The strategies include ESZSL [28], EXEM [27], [52], and SYNC [27], [53]. The ESZSL strategy employs the weights of the regularization term as $\{10^{-3}, 10^{-2}, \dots, 10^3\}$, with the scaling parameters of the Gaussian kernels as $\{10^{-1}n_F, n_F, 10n_F\}$. The SYNC strategy directly sets the phantom classifiers $\mathbf{B} = \mathbf{A}^{(S)}$. The weight τ can be set to $\{2^{-24}, 2^{-23}, \dots, 2^{-9}\}$ and the scaling parameter σ^2 can be chosen from $\{2^{-5}, 2^{-4}, \dots, 2^{10}\}$. The EXEM strategy sets the regularization weight for ν -SVR as $\{2^{-3}, 2^{-2}, \dots, 2^{-3}\}$, the ν values as $\{2^{-8}, 2^{-7}, \dots, 2^0\}$, and the scaling parameters for Gaussian kernels as $\{2^{-4}, 2^{-3}, \dots, 2^4\}$. The dimensionality of features retained in the PCA processing is chosen as {40, 60, 80}. The distance function in EXEM is set to a 1-nearest neighbor (1NN) classifier using the Euclidean distance.

B. Experimental Results: Manual Versus Semantic-Embedding

First, we aim to employ manually annotated prototypes to simulate AADs for human affective cognition while using the semantic-embedding prototypes for knowledge representation. In order to reduce the uncertainty of small-size emotional classes, we set the lower bound for the number of per-class samples as 10 (also noted as “SPC10” for this case and “SPC1” for the original case), leading to 36 classes with 2603 samples (for the PN task) and 39 classes with 2830 samples (for the WN task), in the seen-emotional set. For each round in the tenfold CV, the unweighted accuracy (UA) measure (i.e., added recall per class divided by a number of classes to counter imbalance across classes) [2], [6] averaging through repeating for five times is used for choosing optimal parameters of the ZSL strategies, in order to reduce the influence from the random splitting for the folds. To further reduce the influence, we repeat the experiments ten times for fair comparison, in view of the training–test division in different CV rounds. Note that the PN task considers the early fusion setup to combine the multiple emotional states, that is, averaging the prototypes of the emotions for the positive and negative sets.

We present the mean UAs and their standard deviations in Table II for the two tasks considering the SPC10 case, using the ESZSL, SYNC, and EXEM strategies. It is seen from Table II that the best mean UA for manual descriptors is 66.5% and 72.7% for semantic-embedding AADs. The results indicate that semantic-embedding-based AADs perform better compared with the manual descriptors in most cases, despite using ESZSL for the PN task. Thus, we shed light on investigating the subtasks of the PN task, aiming to recognize the five unseen-emotional states contained in the task. We present the UA results with their standard deviations for three subtasks

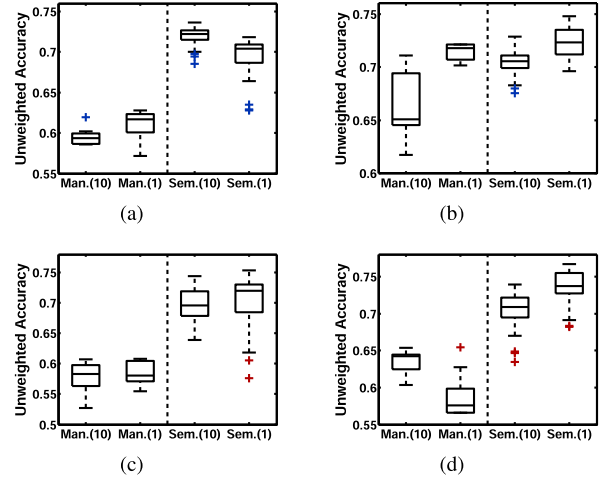


Fig. 2. Boxplots of UAs considering SPC10 [noted as “(10)”) and SPC1 [noted as “(1)”) cases with manual (noted as “Man.”) and semantic-embedding (noted as “Sem.”) prototypes for (a) SYNC and PN task, (b) EXEM and PN task, (c) SYNC and WN task, and (d) EXEM and WN task.

within the PN task in Table III for the SPC10 case, including “PN task (Bin.)” (the current PN task, defining positive and negative categories prior to zero-shot emotion recognition), “PN Task (Mul.)” (using late fusion on multiple emotions to generate positive and negative categories), and “Multiple Emotions” (recognition on the original emotions for the PN task). The results suggest that the semantic-embedding AADs can achieve at least close UA performance with the manual AADs for recognizing original emotions. In addition, the inclusion of SenticNet 5 may lead to positive or negative effects based on the comparison.

Then, in view of the better performance of the semantic-embedding prototypes, we perform a one-way ANalysis Of VAriance (ANOVA) on the SYNC and EXEM strategies within the PN task (Bin.), PN task (Mul.), and WN task, for the SPC10 and SPC1 settings. This leads to $(F(8, 81) = 199.69, p < 0.0001)$ (SYNC-SPC10), $(F(8, 81) = 8.72, p < 0.0001)$ (EXEM-SPC10), $(F(8, 81) = 31.31, p < 0.0001)$ (SYNC-SPC1), and $(F(8, 81) = 12.85, p < 0.0001)$ (EXEM-SPC1) for the PN task (Bin.), while $(F(8, 81) = 37.96, p < 0.0001)$ (SYNC-SPC10), $(F(8, 81) = 12.50, p < 0.0001)$ (EXEM-SPC10), $(F(8, 81) = 14.02, p < 0.0001)$ (SYNC-SPC1), and $(F(8, 81) = 1.52, p > 0.05)$ (EXEM-SPC1) for the PN task (Mul.). The results for the WN task are $(F(8, 81) = 51.79, p < 0.0001)$ (SYNC-SPC10), $(F(8, 81) = 43.94, p < 0.0001)$ (EXEM-SPC10), $(F(8, 81) = 37.25, p < 0.0001)$ (SYNC-SPC1), and $(F(8, 81) = 84.97, p < 0.0001)$ (EXEM-SPC1). The significance ANOVA results imply the performance gap between these types of AADs. Thus, we further focus on a *post hoc* Tukey’s Honest Significant Difference (Tukey’s HSD) test [6] with respect to the AADs’ types, comparing the semantic-embedding prototypes with the manual prototypes from dimensional annotations, as shown in Table IV (when using the SYNC strategy) and Table V (when using the EXEM strategy). It is learned from the table that the semantic-embedding prototypes perform better compared with the manual prototypes—especially for the SPC10 setting.

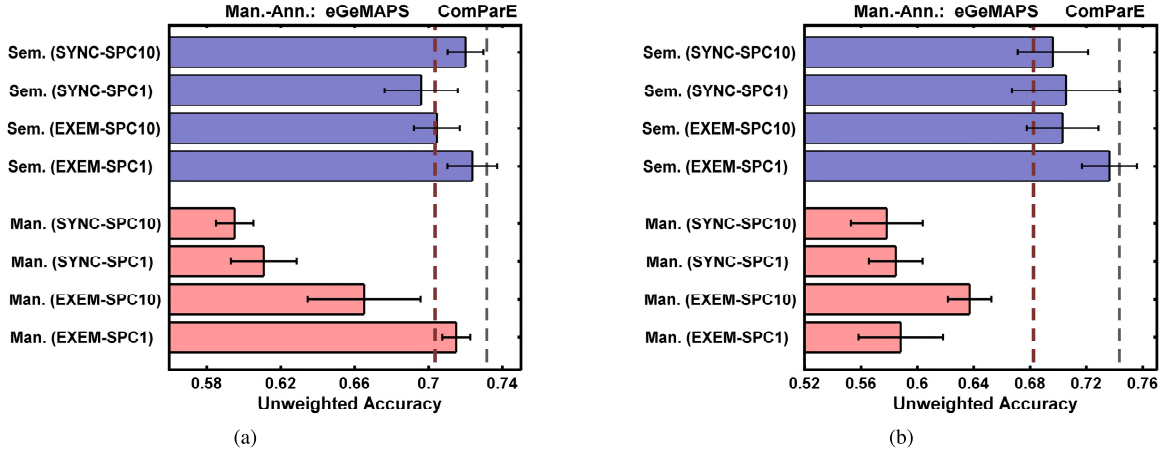


Fig. 3. Row charts of the UAs using manual (noted as “Man.”) and semantic-embedding (noted as “Sem.”) prototypes for (a) PN task and (b) WN task, with the best UA results from the eGeMAPS and ComParE feature-sets using per-sample annotations (noted as “Man.-Ann.”).

This shows the ability for machines to describe auditory emotions in the forms of AADs, exceeding human annotators using emotional dimensions. In addition, further multiple comparisons with Bonferroni correction considering SPC1/SPC10 cases and SYNC/EXEM strategies show the significance at the level of 0.05 between the semantic-embedding and manually annotated UAs for the PN (Bin.) and WN tasks.

In order to further investigate the influence from the per-class-sample limitation for seen-emotional states, we draw the boxplots for PN task (Bin.) and the WN task using the SYNC and EXEM strategies for comparison, considering the SPC10 and SPC1 settings, as shown in Fig. 2. It can be seen from the figures that semantic-embedding prototypes perform better compared with the manually annotated setups. The comparison also indicates that the SPC10 limitation may result in diverse effects on UA performance for the two tasks. This may be due to the positive or negative emotional information contained in the samples from rare seen-emotional classes.

C. Experimental Results: Prototypes Versus Annotations

Following the comparison between manual and semantic-embedding AADs, we present experiments to answer the second question on discussing the performances of per-emotion prototypes and per-sample annotations. In this regard, we make comparisons between the cases of using manual prototypes, semantic-embedding prototypes, and per-sample annotations as the AADs. Fig. 3 shows the mean UAs and their standard deviations when using the manual and semantic-embedding AADs, with the UA results when using the per-sample annotations for eGeMAPS and ComParE feature sets, respectively. The comparisons show that it is difficult for manual prototypes to outperform per-sample annotations, partially due to the loss of information in generating manual prototypes. However, semantic-embedding prototypes make it possible to approach the performance of per-sample manual annotations.

In order to further make a comparison on the performances of per-emotion prototypes and per-sample annotations, we show the best UAs for the cases of per-sample

TABLE VI
BEST UA RESULTS (%) FOR PER-SAMPLE MANUAL ANNOTATIONS (EMPLOYING eGeMAPS, GeMAPS, AND COMPARE FEATURE-SETS), MANUAL PROTOTYPES, AND SEMANTIC-EMBEDDING PROTOTYPES (USING SPC10 AND SPC1 CASES WITH SYNC AND EXEM STRATEGIES) FOR THE PN AND WN TASKS, WHERE “BINARY” REFERS TO PRIOR BINARY CLASSIFICATION AND “MULTIPLE” REFERS TO LATE FUSION ON MULTIPLE EMOTIONS

Setup	Tasks	PN Task	WN Task
Manual (Annotations)	eGeMAPS+SVR	70.4	63.4
	eGeMAPS+MLP	66.2	68.2
	GeMAPS+SVR	69.2	64.1
	GeMAPS+MLP	65.5	66.7
	ComParE [15]	72.6	65.5
	ComParE (PCA1000) [15]	72.6	64.5
	ComParE (PCA500) [15]	73.1	65.6
	ComParE (PCA100) [15]	69.1	65.0
	ComParE (PCA50) [15]	68.1	66.9
	ComParE (PCA10) [15]	68.9	74.3
Manual (Prototypes)	ComParE+DNN [15], [73]	72.3	68.9
	ComParE+DKL [15], [73]	69.7	69.7
	SYNC-SPC10 (Binary)	61.9	—
	SYNC-SPC10 (Multiple)	61.0	60.7
	EXEM-SPC10 (Binary)	71.1	—
	EXEM-SPC10 (Multiple)	71.2	65.4
	SYNC-SPC1 (Binary)	62.8	—
	SYNC-SPC1 (Multiple)	62.2	60.8
Semantic (Prototypes)	EXEM-SPC1 (Binary)	72.1	—
	EXEM-SPC1 (Multiple)	73.3	65.4
	SYNC-SPC10 (Binary)	73.6	—
	SYNC-SPC10 (Multiple)	73.5	74.5
	EXEM-SPC10 (Binary)	72.9	—
	EXEM-SPC10 (Multiple)	75.8	74.0
	SYNC-SPC1 (Binary)	71.8	—
	SYNC-SPC1 (Multiple)	74.7	75.4
Best Results	EXEM-SPC1 (Binary)	74.8	—
	EXEM-SPC1 (Multiple)	76.1	76.7

annotations, manual prototypes, and semantic-embedding prototypes in Table VI. The approaches for the manual per-sample annotations include the usage of eGeMAPS, GeMAPS, and ComParE (using the same classifiers and regressors as in [15] including deep neural network (DNN) and deep kernel learning (DKL) [73]) as the feature set, considering the same speaker-independent CV setups as in [15]. In Table VI, one finds that the semantic-embedding prototypes achieve the best

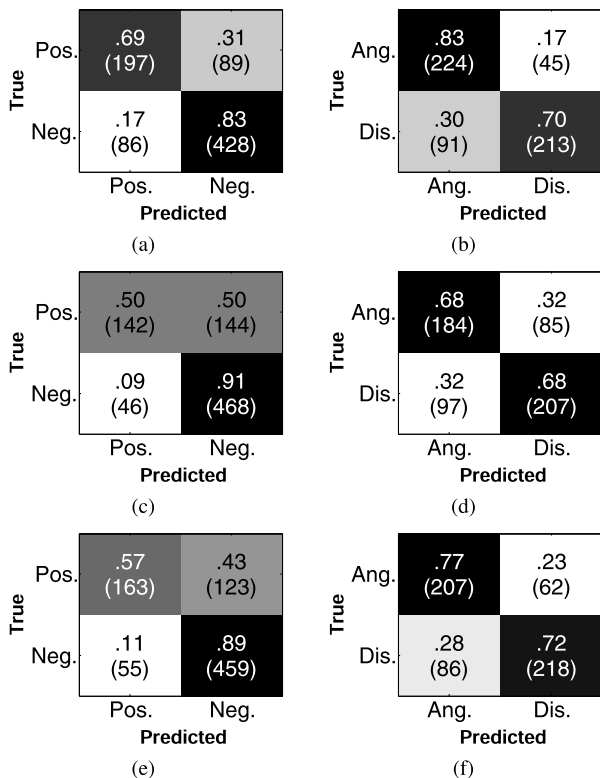


Fig. 4. Confusion matrices of the best UA results achieved by (a) and (b) semantic-embedding prototypes for the PN (positive and negative, noted as “Pos.” and “Neg.”) and WN (anger and disappointment, noted as “Ang.” and “Dis.”) tasks, respectively, (c) and (d) per-sample manual annotations using the eGeMAPS feature set for the PN and WN tasks, respectively, and (e) and (f) per-sample manual annotations using the ComParE feature set for the PN and WN tasks, respectively.

UA results as 76.1% (PN task) and 76.6% (WN task), while the UAs are 73.3% and 74.3% for the manual cases [15]. This implies that machines are able to understand emotions in speech better than human annotators, through using pretrained models from textual data as the AADs. In addition, one also observes the significance between the best semantic-embedding and the per-sample-annotation UA results for the PN and WN tasks with eGeMAPS when using a one-tailed z -test at the significance level of 0.05.

Then, we examine the confusion matrices (Fig. 4) for the best UA results achieved by the per-sample manual annotations (using eGeMAPS and ComParE feature sets) and the semantic-embedding prototypes (EXEM-SPC1 (Multiple)), when confronting the PN (positive and negative, noted as “Pos.” and “Neg.”), and WN (anger and disappointment, noted as “Ang.” and “Dis.”) tasks. The eGeMAPS feature set comparison between the per-emotion semantic-embedding and per-sample manually annotated AADs indicates that the semantic-embedding prototypes make it more effective to recognize emotional states for the PN and WN tasks, in view of the much better recalls on “Pos.” and “Ang.” for these two tasks despite the slightly unfavorable recalls on the other emotions.

To this end, we induce macro F1-score evaluation to make further accuracy-precision-balance comparisons as shown in Table VII, using the best scores as in [23]. It is observed

TABLE VII
MACRO F1-SCORES (%) CORRESPONDING TO THE BEST UA RESULTS FOR THE PER-SAMPLE MANUALLY ANNOTATED AND PER-EMOTION SEMANTIC-EMBEDDING AADs ON THE PN AND WN TASKS, CONSIDERING DIFFERENT SETUPS

AAD Setups \ Tasks	PN Task	WN Task
Per-Sample (eGeMAPS)	71.5	68.2
Per-Sample (GeMAPS)	70.3	66.7
Per-Sample (ComParE)	74.2	74.2
Per-Emotion	76.1	76.3

from the results that the per-emotion semantic-embedding prototypes achieve the best F1-score performance compared with the per-sample manually annotated AADs. This indicates the possibly better performance achieved by the per-emotion AADs compared with the per-sample ones.

V. CONCLUSION

This article focused on investigating AADs to describe emotional states in speech, through using ZSL frameworks. The investigation contained two aspects: 1) exploring the manually annotated and semantic-embedding sources of AADs and 2) exploring the performance of per-emotion prototypes and per-sample annotations as AADs. To this end, we employed zero-shot emotion recognition strategies and performed experiments on the CINEMO corpus of French emotional speech. The experimental results indicated that semantic-embedding prototypes performed better compared with manual descriptors from human annotators on both per-emotion and per-sample setups. The results also revealed the possibility for machines to replace human annotators on understanding auditory affective information in certain cases with the help from transferring external information, despite the need of further proof on other corpora.

In view of these conclusions from this work, future research may consider three topics as follows. First, one should aim to extend the current work by further exploring effective AADs from modalities other than manual and textual descriptors. Second, it is worth researching on emotional transfer from basic to complex emotional states using AADs. Furthermore, it is still unknown how to accurately model the relationship between spoken emotional expression and auditory affective perception.

ACKNOWLEDGMENT

The authors would like to thank Nicholas Cummins, Eduardo Coutinho, and the anonymous reviewers for their valuable help.

REFERENCES

- [1] B. W. Schuller, “Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends,” *Commun. ACM*, vol. 61, no. 5, pp. 90–99, 2018.
- [2] B. W. Schuller *et al.*, “The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates,” in *Proc. Interspeech*, Aug. 2021, pp. 431–435.
- [3] D. M. Schuller and B. W. Schuller, “A review on five recent and near-future developments in computational processing of emotion in the human voice,” *Emotion Rev.*, vol. 13, no. 1, pp. 44–50, Jan. 2021.

- [4] B. Schuller *et al.*, “Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge,” *Comput. Speech Lang.*, vol. 53, pp. 156–180, Jan. 2019.
- [5] B. W. Schuller *et al.*, “The INTERSPEECH 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks,” in *Proc. Interspeech*, Oct. 2020, pp. 2042–2046.
- [6] X. Xu, J. Deng, E. Coutinho, C. Wu, L. Zhao, and B. W. Schuller, “Connecting subspace learning and extreme learning machine in speech emotion recognition,” *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 795–808, Mar. 2019.
- [7] R. Panda, R. Malheiro, and R. P. Paiva, “Novel audio features for music emotion recognition,” *IEEE Trans. Affect. Comput.*, vol. 11, no. 4, pp. 614–626, Oct. 2020.
- [8] Y. Lee, S. Yoon, and K. Jung, “Multimodal speech emotion recognition using cross attention with aligned audio and text,” in *Proc. Interspeech*, Oct. 2020, pp. 2717–2721.
- [9] M. Chen, X. He, J. Yang, and H. Zhang, “3-D convolutional recurrent neural networks with attention model for speech emotion recognition,” *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018.
- [10] X. Xu *et al.*, “A two-dimensional framework of multiple kernel subspace learning for recognizing emotion in speech,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 7, pp. 1436–1449, Jul. 2017.
- [11] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, “Semisupervised autoencoders for speech emotion recognition,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 1, pp. 31–43, Jan. 2018.
- [12] 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 Commun.*, vol. 93, pp. 1–10, Oct. 2017.
- [13] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, “Universum autoencoder-based domain adaptation for speech emotion recognition,” *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 500–504, Apr. 2017.
- [14] P. Song, “Transfer linear subspace learning for cross-corpus speech emotion recognition,” *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 265–275, Apr. 2019.
- [15] X. Xu, J. Deng, N. Cummins, Z. Zhang, L. Zhao, and B. W. Schuller, “Autonomous emotion learning in speech: A view of zero-shot speech emotion recognition,” in *Proc. Interspeech*, Sep. 2019, pp. 949–953.
- [16] B. Schuller, R. Zaccarelli, N. Rollet, and L. Devillers, “CINEMO—A French spoken language resource for complex emotions: Facts and baselines,” in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*. Valletta, Malta: ELRA, 2010, pp. 1643–1647.
- [17] B. A. Erol, A. Majumdar, P. Benavidez, P. Rad, K.-K.-R. Choo, and M. Jamshidi, “Toward artificial emotional intelligence for cooperative social human–machine interaction,” *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 1, pp. 234–246, Feb. 2020.
- [18] G. B. Kempster, B. R. Gerratt, K. V. Abbott, J. Barkmeier-Kraemer, and R. E. Hillman, “Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol,” *Amer. J. Speech-Lang. Pathol.*, vol. 18, no. 2, pp. 124–132, 2009.
- [19] F. Wengler, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, “On the acoustics of emotion in audio: What speech, music, and sound have in common,” *Frontiers Psychol.*, vol. 4, pp. 1–12, May 2013.
- [20] S. D. Morgan, “Categorical and dimensional ratings of emotional speech: Behavioral findings from the Morgan emotional speech set,” *J. Speech, Lang., Hearing Res.*, vol. 62, no. 11, pp. 4015–4029, Nov. 2019.
- [21] L. Devillers, L. Vidrascu, and L. Lamel, “Challenges in real-life emotion annotation and machine learning based detection,” *Neural Netw.*, vol. 18, no. 4, pp. 407–422, May 2005.
- [22] I. Wieser, P. Barros, S. Heinrich, and S. Wermter, “Understanding auditory representations of emotional expressions with neural networks,” *Neural Comput. Appl.*, vol. 32, no. 4, pp. 1007–1022, Feb. 2020.
- [23] X. Xu, J. Deng, N. Cummins, Z. Zhang, L. Zhao, and B. W. Schuller, “Exploring zero-shot emotion recognition in speech using semantic-embedding prototypes,” *IEEE Trans. Multimedia*, early access, Jun. 9, 2021, doi: [10.1109/TMM.2021.3087098](https://doi.org/10.1109/TMM.2021.3087098).
- [24] K. Feng and T. Chaspari, “Few-shot learning in emotion recognition of spontaneous speech using a Siamese neural network with adaptive sample pair formation,” *IEEE Trans. Affect. Comput.*, early access, Sep. 3, 2021, doi: [10.1109/TAFFC.2021.3109485](https://doi.org/10.1109/TAFFC.2021.3109485).
- [25] J. Han, Z. Zhang, Z. Ren, and B. Schuller, “EmoBed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings,” *IEEE Trans. Affect. Comput.*, vol. 12, no. 3, pp. 553–564, Jul. 2021.
- [26] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, “Emotion recognition in speech using cross-modal transfer in the wild,” in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 292–301.
- [27] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, “Classifier and exemplar synthesis for zero-shot learning,” *Int. J. Comput. Vis.*, vol. 128, no. 1, pp. 166–201, 2019.
- [28] B. Romera-Paredes and P. Torr, “An embarrassingly simple approach to zero-shot learning,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, Lille, France, 2015, pp. 2152–2161.
- [29] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, “Video emotion recognition with transferred deep feature encodings,” in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2016, pp. 15–22.
- [30] 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*. Amsterdam, The Netherlands: Elsevier, 2019, pp. 349–367.
- [31] D. Luo, Y. Zou, and D. Huang, “Investigation on joint representation learning for robust feature extraction in speech emotion recognition,” in *Proc. Interspeech*, Sep. 2018, pp. 152–156.
- [32] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, “Direct modelling of speech emotion from raw speech,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2019, pp. 3920–3924.
- [33] Y. Xue, Y. Hamada, and M. Akagi, “Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space,” *Speech Commun.*, vol. 102, pp. 54–67, Sep. 2018.
- [34] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, “Speech emotion recognition with emotion-pair based framework considering emotion distribution information in dimensional emotion space,” in *Proc. Interspeech*, Aug. 2017, pp. 1238–1242.
- [35] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, “Prediction-based learning for continuous emotion recognition in speech,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5005–5009.
- [36] A. S. Cowen, H. A. Elfенbein, P. Laukka, and D. Keltner, “Mapping 24 emotions conveyed by brief human vocalization,” *Amer. Psychologist*, vol. 74, no. 6, p. 698, 2019.
- [37] S. Mohammad, “Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words,” in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2018, pp. 174–184.
- [38] K. P. Truong, D. A. van Leeuwen, and F. M. G. de Jong, “Speech-based recognition of self-reported and observed emotion in a dimensional space,” *Speech Commun.*, vol. 54, no. 9, pp. 1049–1063, Nov. 2012.
- [39] M. Norouzi *et al.*, “Zero-shot learning by convex combination of semantic embeddings,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, 2014, pp. 1–9.
- [40] Y. Ohishi, A. Kimura, T. Kawanishi, K. Kashino, D. Harwath, and J. Glass, “Pair expansion for learning multilingual semantic embeddings using disjoint visually-grounded speech audio datasets,” in *Proc. Interspeech*, Oct. 2020, pp. 1486–1490.
- [41] P. Shaver, J. Schwartz, D. Kirson, and C. O’Connor, “Emotion knowledge: Further exploration of a prototype approach,” *J. Pers. Soc. Psychol.*, vol. 52, no. 6, p. 1061, 1987.
- [42] C. Luo, Z. Li, K. Huang, J. Feng, and M. Wang, “Zero-shot learning via attribute regression and class prototype rectification,” *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 637–648, Feb. 2018.
- [43] T. Chen *et al.*, “Semantically meaningful class prototype learning for one-shot image segmentation,” *IEEE Trans. Multimedia*, early access, Feb. 24, 2021, doi: [10.1109/TMM.2021.3061816](https://doi.org/10.1109/TMM.2021.3061816).
- [44] F. Angiulli, “Prototype-based domain description for one-class classification,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1131–1144, Jun. 2012.
- [45] Z. Fu, T. A. Xiang, E. Kodirov, and S. Gong, “Zero-shot object recognition by semantic manifold distance,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2635–2644.
- [46] T. Zhang, X. Gong, and C. L. P. Chen, “BMT-Net: Broad multitask transformer network for sentiment analysis,” *IEEE Trans. Cybern.*, early access, Mar. 4, 2021, doi: [10.1109/TCYB.2021.3050508](https://doi.org/10.1109/TCYB.2021.3050508).
- [47] X. Gong, T. Zhang, C. L. P. Chen, and Z. Liu, “Research review for broad learning system: Algorithms, theory, and applications,” *IEEE Trans. Cybern.*, early access, Mar. 17, 2021, doi: [10.1109/TCYB.2021.3061094](https://doi.org/10.1109/TCYB.2021.3061094).
- [48] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, “Zero-shot learning with semantic output codes,” in *Proc. Adv. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates, 2009, pp. 1410–1418.

[49] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.

[50] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, Jul. 2018.

[51] Y. Guo, G. Ding, J. Han, S. Zhao, and B. Wang, "Implicit non-linear similarity scoring for recognizing unseen classes," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 4898–4904.

[52] S. Changpinyo, W.-L. Chao, and F. Sha, "Predicting visual exemplars of unseen classes for zero-shot learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3476–3485.

[53] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5327–5336.

[54] N. Rollet, A. Delaborde, and L. Devillers, "Protocol CINEMO: The use of fiction for collecting emotional data in naturalistic controlled oriented context," in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops*, Sep. 2009, pp. 1–6.

[55] M. Brendel, R. Zaccarelli, and L. Devillers, "Building a system for emotions detection from speech to control an affective avatar," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, Valletta, Malta: ELRA, 2010, pp. 2205–2210.

[56] B. Schuller and L. Devillers, "Incremental acoustic valence recognition: An inter-corpus perspective on features, matching, and performance in a gating paradigm," in *Proc. Interspeech*, Sep. 2010, pp. 801–804.

[57] G. Johnson and S. Connelly, "Negative emotions in informal feedback: The benefits of disappointment and drawbacks of anger," *Human Relations*, vol. 67, no. 10, pp. 1265–1290, Oct. 2014.

[58] 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, Jan. 2015.

[59] F. Eyben *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affective Comput.*, vol. 7, no. 2, pp. 190–202, Apr./Jun. 2015.

[60] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates, 2013, pp. 3111–3119.

[61] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[62] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[63] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.

[64] T. Mikolov, E. Grave, P. Bojanowski, C. Puhresch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, Miyazaki, Japan, 2018, pp. 52–55.

[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. AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, 2018, pp. 1795–1802.

[66] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok, "SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 105–114.

[67] X. Zhang *et al.*, "Fatigue detection with covariance manifolds of electroencephalography in transportation industry," *IEEE Trans. Ind. Informat.*, vol. 17, no. 5, pp. 3497–3507, May 2021.

[68] X. Zhang *et al.*, "Individual similarity guided transfer modeling for EEG-based emotion recognition," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 1156–1161.

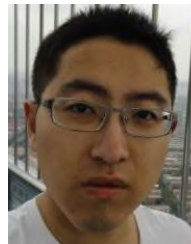
[69] Y. Saito, S. Takamichi, and H. Saruwatari, "Perceptual-similarity-aware deep speaker representation learning for multi-speaker generative modeling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1033–1048, 2021.

[70] J. Shen, S. Zhao, Y. Yao, Y. Wang, and L. Feng, "A novel depression detection method based on pervasive EEG and EEG splitting criterion," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 1879–1886.

[71] J. Shen, X. Zhang, B. Hu, G. Wang, Z. Ding, and B. Hu, "An improved empirical mode decomposition of electroencephalogram signals for depression detection," *IEEE Trans. Affect. Comput.*, early access, Jul. Aug. 14, 2020, doi: [10.1109/TAFFC.2019.2934412](https://doi.org/10.1109/TAFFC.2019.2934412).

[72] J. Shen *et al.*, "An optimal channel selection for EEG-based depression detection via kernel-target alignment," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 7, pp. 2545–2556, Jul. 2021.

[73] A. G. Wilson, Z. Hu, R. R. Salakhutdinov, and E. P. Xing, "Deep kernel learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, Cadiz, Spain: PMLR, 2016, pp. 370–378.



Xinzhou Xu received the bachelor's degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2009, and the master's and Ph.D. degrees from Southeast University, Nanjing, 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 and 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 the bachelor's degree in electronic and information engineering from Harbin Engineering University (HEU), Harbin, China, in 2009, the master's degree in information and communication engineering from the Harbin Institute of Technology (HIT), Harbin, in 2011, and the Ph.D. degree in electrical engineering and information technology from the Technical University of Munich (TUM), Munich, Germany, in 2016, with a focus on feature transfer learning for speech emotion recognition.

He was a Post-Doctoral Researcher at the Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany, from 2015 to 2017, and a Lead Researcher at audEERING, Gilching, Germany. He is currently the Head of deep learning at Agile Robots AG, Munich. His research interests are machine learning methods, such as transfer learning and deep learning with an application preference to affective computing.



Zixing Zhang (Member, IEEE) received the master's degree in physical electronics from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2010, and the Ph.D. degree in computer engineering from the Technical University of Munich (TUM), Munich, Germany, in 2015.

From 2017 to 2019, he was a Research Associate with the Department of Computing, Imperial College London (ICL), London, U.K. Before that, he was a Post-Doctoral Researcher at the University of Passau, Passau, Germany. He has authored more than 90 publications in peer-reviewed books, journals, and conference proceedings. His research mainly focuses on deep learning technologies for speaker-centered state and health computing.

Dr. Zhang has organized special sessions, such as at the IEEE 7th Affective Computing and Intelligent Interaction (ACII) Conference in 2017 and the 43rd 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 and a program committee member and an area chair for many international conferences.



Xijian Fan (Member, IEEE) received the B.Sc. degree in information and communication technology from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2009, the M.Sc. degree in computer information and science from Hohai University, Nanjing, China, in 2012, and the Ph.D. degree from the School of Engineering, University of Warwick, Coventry, U.K., in 2017.

He is currently an Associate Professor with the College of Information Science and Technology, Nanjing Forestry University, Nanjing. His research interests include affective computing, computer vision, and machine learning.



Laurence Devillers is currently a Full Professor of artificial intelligence at Sorbonne University, Paris, France, where she has been heading the team of research “Affective and Social Dimensions in Spoken Interaction With (Ro)bots: Ethical Issues” at CNRS-LISN since 2004. Since 2020, she has been heading the Interdisciplinary Chair on Artificial Intelligence HUMAINE: HUMAN-MACHINE Affective Interaction and Ethics, CNRS. She has authored the book *Les robots émotionnels* (Ed. Observatoire, March 2020) and *Des Robots et des Hommes: mythes, fantasmes et réalité* (Ed. Plon, March 2017). Her topics of research (H-index: 39) are human-machine coadaptation: from the modeling of emotions and human-robot dialog to the ethical impacts for society and the risks and benefits of artificial intelligence (AI).

Prof. Devillers is a member of the National Comity Pilot on Ethics of Numeric (CNPEN) working on conversational Agents, AI, and Ethics. She has been an expert member of the GPAI on “the future of work” since June 2020 (international group).



Li Zhao received the bachelor’s degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1982, the master’s degree from Soochow University, Suzhou, China, in 1988, and the Ph.D. degree from the Kyoto Institute of Technology, Kyoto, Japan, in 1998.

He is currently a Professor with the School of Information Science and Engineering, Southeast University, Nanjing. His research interests include spoken signal processing and affective computing.



Björn W. Schuller (Fellow, IEEE) received the Diploma and Ph.D. degrees in electrical engineering and information technology, for his study on automatic speech and emotion recognition, and the Habilitation and Adjunct Teaching Professorship in signal processing and machine intelligence from the Technical University of Munich (TUM), Munich, Germany, in 1999, 2006, and 2012, respectively.

He is currently a Professor of AI at the Group on Language, Audio, and Music (GLAM), Imperial College London, London, U.K., and a Full Professor and the Head of the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany. He is also a Guest Professor at Southeast University, Nanjing, China, among other affiliations. He has (co)authored five books and more than 1000 publications in peer-reviewed books, journals, and conference proceedings leading to more than 40000 citations (H-index of 91).

Dr. Schuller is a fellow of the International Speech Communication Association (ISCA) and the British Computer Society (BCS), a Senior Member of the Association for Computing Machinery (ACM), and the President-Emeritus of the Association for the Advancement of Affective Computing (AAAC).