INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Autonomous Emotion Learning in Speech: A View of Zero-Shot Speech Emotion Recognition

Xinzhou Xu[1], *Jun Deng*[2], *Nicholas Cummins*[3], *Zixing Zhang*[4], *Li Zhao*[5], *Björn Schuller*[3,4]

[1] College of Internet of Things, Nanjing University of Posts and Telecommunications, P.R. China
[2] Agile Robots AG, Germany
[3] ZD.B Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
[4] Group on Language Audio and Music, Imperial College London, UK
[5] School of Information Science and Engineering, Southeast University, P.R. China

xinzhou.xu@njupt.edu.cn, {nicholas.cummins,schuller}@ieee.org

Abstract

Conventionally, speech emotion recognition is achieved using passive learning approaches. Differing from such approaches, we herein propose and develop a dynamic method of autonomous emotion learning based on zero-shot learning. The proposed methodology employs emotional dimensions as the attributes in the zero-shot learning paradigm, resulting in two phases of learning, namely attribute learning and label learning. Attribute learning connects the paralinguistic features and attributes utilising speech with known emotional labels, while label learning aims at defining unseen emotions through the attributes. The experimental results achieved on the CINEMO corpus indicate that zero-shot learning is a useful technique for autonomous speech-based emotion learning, achieving accuracies considerably better than chance level and an attribute-based gold-standard setup. Furthermore, different emotion recognition tasks, emotional attributes, and employed approaches strongly influence system performance.

Index Terms: Autonomous emotion learning, speech emotion recognition, zero-shot learning, emotional attributes

1. Introduction

Speech Emotion Recognition (SER) has been investigated comprehensively during the past decade [1,2]. Prominent directions concerning SER have focused on diverse topics such as data collection [3], data enrichment [4], deep learning [5], feature enhancement [6], and transfer learning [7]. Nevertheless, most current research on SER is, arguably, focused on the passive learning of emotional states, which need fully, or at least partially labelled training samples to learn reasonable models [8]. Furthermore, passive approaches are unsuitable for labelling samples which do not have an adequate amount of matched training data. In the worst-case scenario, if no training samples are provided, the corresponding target emotional states cannot be recognised. Hence it is natural to allow for autonomous emotion learning in speech, to model unknown emotional states.

Zero-Shot Learning (ZSL) provides a solution for such operating conditions, enabling systems to perform classification when there is a lack of adequately labelled training data [9–12]. Results in the literature indicate the suitability of ZSL for tasks such as object detection [10, 11], gesture detection [13], and even for video based emotion detection [14]. Still, very few of the existing ZSL methods provide a reasonable framework for zero-shot emotion learning in speech. This is due, in part, to the latent emotional descriptors in paralinguistics [6, 15] and complicated forms of expression of emotion [16, 17].

Thus, in this paper, we propose a zero-shot SER paradigm to recognise emotional speech samples which come from 'unseen' emotional states. We use attribute learning to associate each emotional descriptor, noted as 'attribute' (i. e., emotional dimensions), with paralinguistic features extracted from speech with known emotions. We then use label learning to leverage these attributes to model unseen emotional states using empirical annotation. The CINEMO corpus, which consists of spoken utterances and corresponding emotional labels represented in six dimensions [18–20], is used to demonstrate the effectiveness of the proposed approach. To the best of the authors' knowledge, this proposed approach is novel in the SER literature. Emotional dimensions have been used as auxiliary information to improve performance in multi-task SER [21, 22]. Similarly, recent works have focused on the relationship between vocal features and affective descriptors in a cross-language setting [23]. None of these works, however, employed a zero-shot paradigm. Moreover, popular ZSL approaches [11] focused on fixed label definition using attributes, while our method considers automatic emotional-state definition.

2. Methodology

2.1. Zero-Shot Learning

Conventional existing ZSL approaches focus on transferring knowledge between modalities to provide additional information to recognise unseen samples, usually confronted with visual learning topics [9–12]. However, it is difficult to utilise these approaches directly in SER due to two reasons. First, emotion lies in a latent layer in speech, resulting in difficulties in inferring emotional descriptors. Second, different affective expressions may also lead to diverse descriptors. Thus, we propose a zero-shot SER methodology employing emotional attributes to connect paralinguistic features and emotional states. The methodology is divided into two phases, *attribute learning* and *label learning* (Fig. 1). The attribute learning makes use of paralinguistic features to learn emotional attributes, e. g. , the emotional description for utterances such as level of relaxation or naturalness. The label learning, on the other hand, utilises the attributes to model unseen emotional states empirically.

2.2. Attribute Learning

The attribute-learning phase aims at fitting attribute values (known emotion labels) to speech samples using corresponding paralinguistic features. This phase models the relationship between n_F features and n_A emotional attributes (Fig. 1).

http://dx.doi.org/10.21437/Interspeech.2019-2406

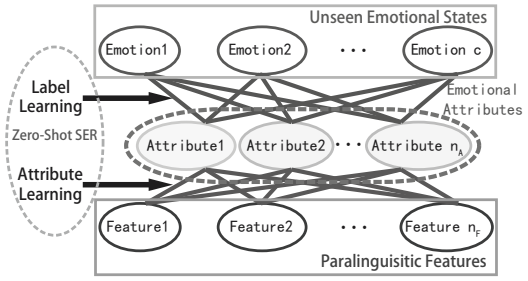


Figure 1: *A diagrammatic overview of the zero-shot SER methodology, including attribute learning and label learning.*

The $N^{(S)}$ utterance samples belonging to seen (known) emotional states used in attribute learning are denoted as

$$X^{(S)} = [x_1^{(S)}, x_2^{(S)}, \ldots, x_{N^{(S)}}^{(S)}]^T \in \Re^{N^{(S)} \times n_F}, \quad (1)$$

with the dimensionality of the features being n_F. The corresponding attributes $A^{(S)} \in \Re^{N^{(S)} \times n_A}$ of $X^{(S)}$ are denoted as

$$A^{(S)} = [a_1^{(S)}, a_2^{(S)}, \ldots, a_{N^{(S)}}^{(S)}]^T = [\alpha_1^{(S)}, \alpha_2^{(S)}, \ldots, \alpha_{n_A}^{(S)}]. \quad (2)$$

Thus, the task of attribute learning is to learn the relationship between $X^{(S)}$ and each attribute. Through defining the mapping of attribute i (the ith column of $A^{(S)}$) from $X^{(S)}$ as

$$f_i(X^{(S)}) = [f_i(x_1^{(S)}), f_i(x_2^{(S)}), \ldots, f_i(x_{N^{(S)}}^{(S)})]^T, \quad (3)$$

we have predicted mapping $\widehat{f}_i(\cdot)$ represented as

$$\arg\max_{f_i} Sim\left(f_i(X^{(S)}), \alpha_i^{(S)}\right) \quad s.t.\ \psi^{(S)}(f_i) \in \Omega^{(S)}, \quad (4)$$

in which $f_i(\cdot)$ represents the feature mapping of the ith attribute, where $i = 1, 2, \ldots, n_A$. The $Sim(\cdot, \cdot)$ represents a similarity measurement between two vectors. $\psi^{(S)}(f_i)$ is a regularisation term subjected to a condition set $\Omega^{(S)}$.

Assuming that the seen and unseen domains share similar mappings from features to attributes, we obtain the n_A-dimensional predicted attributes of one unseen sample $x^{(U)}$ as

$$\widehat{a}^{(U)} = [\widehat{f}_1(x^{(U)}), \widehat{f}_2(x^{(U)}), \ldots, \widehat{f}_{n_A}(x^{(U)})]^T. \quad (5)$$

2.3. Label Learning

For the label-learning phase, we assume that labellers are able to estimate multiple sets of n_A emotional attributes empirically for each of c unseen emotional states. We take this assumption as the first-hand learning for emotions that usually comes from observing physiological or behavioural data [24], indicating that it is natural for different human beings to vocally express one emotional state in various ways [16, 17].

This assumption allows us to model the unseen emotional states using the attributes from the samples. We define the empirical values of the attributes of the c unseen classes as being

$$A^{(E)} = [a_1^{(E)}, a_2^{(E)}, \ldots, a_{N^{(E)}}^{(E)}]^T \in \Re^{N^{(E)} \times n_A}, \quad (6)$$

with the corresponding emotional labels for the jth row as

$$d_j^{(E)} \in \mathscr{L} = \{l_1, l_2, \ldots, l_c\}, \quad (7)$$

where $\mathscr{L}$ represents the label set with the labels of the c classes from l_1 to l_c. Thus the emotional-state labels of the $N^{(E)}$ empirical samples (the rows of $A^{(E)}$) are defined as $\mathscr{D}^{(E)} = \{d_1^{(E)}, d_2^{(E)}, \ldots, d_{N^{(E)}}^{(E)}\}$ respectively.

Table 1: *Key information of the employed CINEMO corpus.*

Characteristics		Description
Speech Data	Language	French
	# Utterances	3 992 (3 591 here)
	# Speakers	51 (21 female, 30 male)
Annotates	# Dimensions	6 (A1 to A6)
	# Emotions	256 (16 major, 16 minor)

Therefore from the side of probability, the optimal label-learning model of $A^{(E)}$ can be represented as $\widehat{g}$ equal to

$$\arg\max_{g} p\left(\mathscr{D}^{(E)} | A^{(E)}, g\right) \quad s.t.\ \psi^{(E)}(g) \in \Omega^{(E)}, \quad (8)$$

where g indicates the label-learning model for the $N^{(E)}$ empirical emotional attributes, aiming at maximising the probability to obtain the $\mathscr{D}^{(E)}$. $\psi^{(E)}(g)$ is a regularisation term subjected to a condition set $\Omega^{(E)}$. Using the predicted feature mappings $f_i(\cdot)$s in Eq. (5), we assume that the $A^{(E)}$ and $a^{(U)}$ share the same estimated label-learning model $\widehat{g}$, drawing the prediction of the unseen-emotional label of $x^{(U)}$ as

$$\widehat{d}^{(U)} = \arg\max_{d^{(U)} \in \mathscr{L}} p\left(d^{(U)} | \widehat{a}^{(U)}, \widehat{g}\right), \quad (9)$$

employing the feature mappings $\widehat{f}_i(\cdot)$s to transfer attribute-learning models from seen to unseen emotional states, while using the predicted model $\widehat{g}$ to learn the unseen states from empirical definition of speech emotions.

3. Corpus

The experimental results presented in this paper were obtained using the CINEMO corpus [18–20], consisting of 3 992 French utterances recorded from 51 speakers with a total length of 2:13:59, with the sampling rate of 16kHz. The data collection paradigm involved the speakers, none of whom had professional acting experience, repeating lines from 12 well known French movies [19]. Table 1 presents key information of the corpus.

Emotional states were labelled by two annotators (1 female, 1 male) using two different labelling strategies. The first strategy involved marking each utterance as having a major and a minor emotion label, taken from one of sixteen 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)'. This process resulted in 256 combinations for each annotator. In the second strategy, each sample was labelled for strength in six emotional dimensions, namely '*intensity*', '*activation*', '*valence*', '*control*', '*suddenness*', and '*naturalness*', herein denoted as 'A1' to 'A6'.

In our experiments, due to data quality issues, we considered only a subset of 3 591 utterances (1 380 female; 2 210 male). Further, to ensure a higher quality of labels, we considered only those samples that had matching major emotion labels from both annotators as the ZSL candidates. This filtering procedure left a total of 2 628 samples with a total length of 1:32:38 for the use of ZSL in our experiments.

4. Experimental Setup

4.1. Data Setup

We partitioned the data from the CINEMO corpus into two sets, namely a ZSL set for label-learning and a second partition for

Table 2: *Three attribute combinations (marked as 'AC1', 'AC2', and 'AC3' respectively) of the emotional descriptors.*

ACs	# Attributes	Attributes
AC1	3	*intensity*, *activation*, *valence*
AC2	3	*control*, *suddenness*, *naturalness*
AC3	6	*intensity*, *activation*, *valence*, *control*, *suddenness*, *naturalness*

the attribute-learning set. The ZSL set included part of the full-agreement samples (the ZSL candidates) with their annotations in order to process label learning and recognition test. For this partition, we employed a three-fold speaker-independent *Cross-Validation* (CV) setup (Speaker ID 1-20, 22-36, and 37-51), in accordance with [19]. We used two folds for label learning to simulate the procedure of empirical attribute estimation, with this data only providing the corresponding attributes and emotional label of each sample, while we used the other fold as a test set for unseen-emotional-state recognition only using paralinguistic features extracted from the utterances.

The rest of the samples in the corpus, the attribute-learning set, were used in a regression analysis paradigm which attempts to bridge the emotional attributes and paralinguistic features. This process assumes that any sample not belonging to the emotional states in the ZSL set could be included, since the sample belongs to a different emotional state to be precise. We only consider each sample's attributes and series (or paralinguistic features). Then, for any given sample, we aim to minimise the gap between each attribute and the mapping of features; to this end we adapt several regression strategies (Sec. 4.3).

4.2. Features and Attributes

In our experiments, we use the *INTERSPEECH Computational Paralinguistics Challenge* (ComParE) [1] feature set, including 6 373 static features of functionals of 65 *Low-Level Descriptors* (LLDs) [25]), which, due to its high dimensionality, contains comprehensive paralingustic information. The features were extracted using the OPENSMILE toolkit [26].

As attributes, we use the average values of the six emotional dimensions (Sec. 3). Note that due to the gap between the two annotators in labelling the corpus, the average emotional-dimension rating scores between the annotators were calculated as the attributes and unified to vary from 0 to 1. Three *Attribute Combinations* (ACs) were also designed according to *Cohen's Kappa* (κ) [19], denoted as 'AC1', 'AC2', and 'AC3' respectively (Table 2). The 'AC1' utilised the attributes with relatively higher κs, while the 'AC2' included the other three attributes.

4.3. Learning Approaches

In the attribute-learning phase, we consider three approaches for $f_i(\cdot)$s, namely, shallow-structure *Multi-Layer Perceptron* (MLP) networks [27], *Support Vector Regression* (SVR) [28], and *Ridge Regression* (RR) [15]. The MLP consists of a two-hidden-layer structure, with 12 selections of the hidden-layer neurons as: $(32, 8), (32, 16), (64, 16), \ldots, (1\,024, 512)$. For both of the SVR and RR, we test both linear and kernelised (Gaussian kernels) representation, with the regularisation parameter C. For the SVRs, the C is varied from 0.0001 to $10\,000$, while the Gaussian-kernel SVRs σs is varied in $\{0.01n_F, 0.1n_F, n_F, 10n_F\}$. For the RRs, we test Cs from 0.0001 to 0.1, while the range of Gaussian-kernel parameter σs of the kernelised RRs is $\{0.1n_F, n_F\}$.

For the label-learning model g, we consider three classifiers: k-*Nearest Neighbour* (kNN), *Naive Bayesian* (NB) using Gaussian distribution, and *Support Vector Machines* (SVMs) with linear and Gaussian kernels, as these approaches are widely used in SER applications [1, 6, 29]. The selections of k in kNN are $\{1, 5, 10, 20, 30\}$. The SVMs include the Cs as in the SVRs. The Gaussian-kernel parameter σs for the kernelised forms are set as $\{0.01n_A, 0.1n_A, n_A, 10n_A\}$.

5. Experimental Results

5.1. Positive-Negative Experiments

First, from the view of valence, we investigated zero-shot SER on recognising samples belonging to unseen *positive and negative* (PN) emotional classes, as a common task in analysing human emotion [30]. Jointly considering data balance and the sample size of each full-agreement label, we chose the positive emotions classes as '*amusement*' and '*satisfaction*', while the negative emotions classes were '*anger*', '*stress*', and '*sadness*'.

The *Unweighted Accuracies* (UAs) and *Weighted Accuracies* (WAs) [6] for these experiments are given in Table 3. Note that the gold-standard results refer to the case in which we use the emotional attributes of each test sample for recognition, which leads to emotion recognition being achieved only through the subjectively-rated attributes. The results in Table 3 indicate that the zero-shot SER approaches perform significantly better compared with the chance level; at a significance level of 0.05 using one-tailed z-test [4, 6]. Further, the regression approaches of MLP and Kernel SVR often outperform the other regressors across the different classifiers. However, there is a large gap between the gold-standard results and the accuracies of the zero-shot approaches. We speculate that this may have resulted from misalignments between different emotions, or the limited sample size in the attribute learning.

We also ran tests in which *Principal Component Analysis* (PCA) was included to investigate the influence of dimensionality reduction (Fig. 2a), only using the attribute learning data in training. Our results indicate that AC2 and AC3 perform much better than the AC1, inferring that larger κ values do not always indicate better performance in zero-shot PN applications. Moreover, the accuracy tendency of AC2 and AC3 is to reduce with decreasing PCA dimensionality. Furthermore, we added the combination of AC4 (A3 to A6), in order to assess the influence of the energy-related attributes (Table 4). It can be inferred from Table 4 that the energy-based attributes do not heavily affect the performance of a zero-shot SER for the PN case. Interestingly, we observed that the attributes of *control*, *suddenness*, and *naturalness* are also able to define the extent of valence.

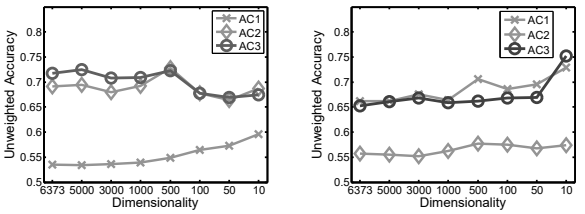
5.2. Anger-Disappointment Experiments

We also investigated the experimental performance of zero-shot recognition for *within-negative* (WN) emotions, within which we attempt to separate '*anger*' and '*disappointment*', as these invoke very different reactions in individuals across a society [31]. The UAs and WAs of these experiments are presented in Table 3. Within this experiments, we observed that the zero-shot SER significantly outperforms the gold-standard system at significance levels of 0.05 using one-tailed z-test. These results indicate that subjective emotional descriptors may include less effective information for emotion recognition when compared with objective paralinguistic features.

The UA and WA results for the WN PCA experiments are given in Fig. 2b. Differing greatly from the PN setting, we

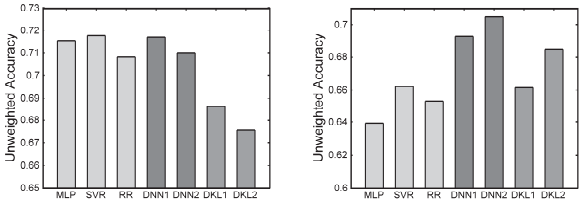
Table 3: *UAs and WAs (%; represented by 'UA / WA') using multiple regressors and classifiers for the Positive-Negative (PN) and Within-Negative (WN) cases respectively, when employing the ComParE feature set. The significant results (compared with the corresponding gold-standard results) for the WN case are marked using double underlines.*

Regressor\Classifier		kNN		NB		Linear SVM		Kernel SVM	
		PN	WN	PN	WN	PN	WN	PN	WN
Gold-Standard (Attr.)		97.6 / 97.3	55.7 / 53.8	96.9 / 97.3	60.4 / 62.3	97.5 / 97.5	56.4 / 57.8	97.8 / 97.6	59.8 / 61.6
MLP Regressors		70.3 / **77.4**	59.8 / 58.5	62.6 / **73.6**	60.9 / 63.0	**67.0 / 76.1**	63.9 / 63.5	**71.5 / 78.8**	62.4 / 63.9
SVR	Linear SVR	66.5 / 75.5	58.2 / 56.9	63.0 / 72.2	61.7 / 64.1	65.2 / 72.9	62.8 / 63.0	67.2 / 76.1	63.0 / 63.4
	Kernel SVR	**71.7** / 77.3	**60.5 / 59.0**	**63.7** / 73.3	**66.2 / 68.3**	66.0 / 75.4	**65.6 / 67.2**	71.4 / 77.3	63.7 / 64.8
RR	Linear RR	61.9 / 68.1	57.6 / 56.5	60.5 / 68.9	56.6 / 55.8	62.1 / 68.7	57.6 / 58.3	62.9 / 69.1	55.9 / 57.4
	Kernel RR	70.2 / 76.8	58.4 / 57.1	61.8 / 72.4	61.4 / 64.0	65.6 / 74.7	65.3 / 64.8	70.8 / 77.4	**64.1 / 65.9**



(a) (b)

Figure 2: *The best PN-case (a) and WN-case (b) UAs for different dimensions through PCA, when using AC1 to AC3.*



(a) (b)

Figure 3: *The best PN-case (a) and WN-case (b) UAs of all the regressors, when using the ACs of AC1 to AC3.*

Table 4: *The best UAs and WAs (%; represented by 'UA / WA') of the gold-standard setup, PN, and WN cases respectively for the ACs of AC1 to AC4, when using the ComParE feature set.*

ACs	PN		WN	
	Gold-Standard	Zero-Shot	Gold-Standard	Zero-Shot
AC1	52.4 / 65.1	53.5 / 66.2	60.4 / 62.3	**66.2 / 68.3**
AC2	97.7 / 97.5	69.1 / 76.7	50.8 / 53.6	55.7 / 55.1
AC3	97.8 / 97.6	**71.8 / 78.8**	59.8 / 61.6	65.3 / 65.3
AC4	97.9 / 97.6	70.5 / 77.8	55.4 / 58.0	63.7 / 65.0

Table 5: *The ZSL-set RMSE, PCC, and accuracies (%) corresponding to the best UAs of the PN and WN cases (represented by 'PN / WN') using different attribute-learning approaches.*

Reg.	RMSE	PCC	UA	WA
MLP	.239 / .192	.395 / .255	71.5 / 63.9	**78.8** / 63.9
SVR	.240 / .193	.386 / .278	**71.7** / 66.2	77.3 / 68.3
RR	.241 / .180	.386 / .266	70.8 / 65.3	77.4 / 65.9
DNN	.241 / .188	.376 / .242	**71.7 / 70.5**	77.9 / **70.2**
DKL	.237 / .186	.391 / .246	69.3 / 68.5	77.8 / 68.1

achieved the highest WN accuracies (75.2%) with a PCA dimensionality of 10. Considering the WN results in Table 3 and Table 4, we conclude that energy-based information may play a critical role in classifying *anger* and *disappointment*.

5.3. Analysis of Attribute Learning

Having proved the applicability of the proposed zero-shot SER, we now investigate the influence from the attribute-learning phase. Thus, we introduced the deep methods of *Deep Neural Networks* (DNN) and *Deep Kernel Learning* (DKL) [32, 33], which include the hidden-layer structures of $(800, 400, 50, 2)$ and $(800, 800, 400, 50, 2)$ noted as (DNN1, DKL1) and (DNN2, DKL2) respectively, considering the scale of samples. We also considered using either a $tanh$ and *Rectified Linear Unit* (ReLU) activation function in the two methods within 30 epochs in training, employing the ACs of AC1 to AC3. Fig. 3 presents the best UAs for the regressors of MLP, SVR, RR, DNN, and DKL for the PN and WN cases. One can observe that the two methods are capable of achieving better performance in zero-shot SER, through using deep structures.

Further, we present the measures of *Root-Mean-Square Error* (RMSE) and *Pearson Correlation Coefficient* (PCC) (averaging from A1 to A6) on the ZSL set, for each regressor corresponding to its best UA for the PN and WN cases (Table 5). It can be seen in these results that the best RMSE and PCC do not always reflect the best accuracies. This may be due to the interference from redundant attributes, and the possible gap between attribute learning and label learning.

6. Conclusions

Within this paper, we proposed and developed a novel autonomous zero-shot *Speech Emotion Recognition* (SER). This approach utilised the emotional dimensions as the attributes connecting paralinguistic features and emotional states. The experimental results gained on CINEMO corpus indicate that this zero-shot approach is effective at recognising unseen emotional states in speech for the positive-negative and within-negative recognition tasks. Furthermore, the results also indicate that the utilised regression and classification approaches, the attribute selection process, and the paralinguistic feature representation all affect the performance of our zero-shot system. Our future work will focus on two aspects: designing better zero-shot SER systems, and exploring other emotional descriptors.

7. Acknowledgements

This work was supported by the Natural Science Foundation of China under Grants No. 61673108 and No. 61801241, the Natural Science Foundation for Jiangsu Higher Education Institutions under Grants 18KJB510029 and 16KJB510031, the Natural Science Foundation of Jiangsu under Grant BK20180746, and the NUPTSF under Grant NY217149, and the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE).

8. References

[1] 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.

[2] 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, 2018.

[3] A. Metallinou, M. Wöllmer, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, 2012.

[4] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 115–126, 2015.

[5] 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.

[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] 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.

[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] 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.

[10] 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.

[11] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning - The good, the bad and the ugly," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, 2017, pp. 3077–3086.

[12] 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.

[13] N. Madapana and J. P. Wachs, "A semantical & analytical approach for zero shot gesture learning," in *Proc. IEEE International Conference on Automatic Face Gesture Recognition (FG)*. Washington, DC: IEEE, 2017, pp. 796–801.

[14] 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.

[15] 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.

[16] D. Keltner, J. Tracy, D. A. Sauter, D. C. Cordaro, and G. McNeil, "Expression of emotion," *Handbook of Emotions*, pp. 467–482, 2016.

[17] A. S. Cowen, H. A. Elfenbein, P. Laukka, and D. Keltner, "Mapping 24 emotions conveyed by brief human vocalization," *American Psychologist*, pp. 1–15, 2018.

[18] N. Rollet, A. Delaborde, and L. Devillers, "Protocol CINEMO: The use of fiction for collecting emotional data in naturalistic controlled oriented context," in *Proc. International Conference on Affective Computing and Intelligent Interaction and Workshops*. Amsterdam, Netherlands: CTIT, 2009, pp. 1–6.

[19] B. Schuller, R. Zaccarelli, N. Rollet, and L. Devillers, "CINEMO - a French spoken language resource for complex emotions: Facts and baselines," in *Proc. International Conference on Language Resources and Evaluation (LREC)*. Valletta, Malta: ELRA, 2010, pp. 1643–1647.

[20] B. Schuller and L. Devillers, "Incremental acoustic valence recognition: An inter-corpus perspective on features, matching, and performance in a gating paradigm," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Makuhari, Chiba, Japan, 2010, pp. 801–804.

[21] 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.

[22] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Stockholm, Sweden: ISCA, 2017, pp. 1103–1107.

[23] A. Chasaide, I. Yanushevskaya, and C. Gobl, "Voice-to-affect mapping: Inferences on language voice baseline settings," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Stockholm, Sweden: ISCA, 2017, pp. 1258–1262.

[24] A. Kazemzadeh, S. Lee, and S. Narayanan, "Fuzzy logic models for the meaning of emotion words," *IEEE Computational Intelligence Magazine*, vol. 8, no. 2, pp. 34–49, 2013.

[25] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K. Scherer, "On the acoustics of emotion in audio: What speech, music and sound have in common," *Frontiers in Psychology, Emotion Science, Special Issue on Expression of emotion in music and vocal communication*, vol. 4, no. Article ID 292, pp. 1–12, May 2013.

[26] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. International Conference on Multimedia*. Barcelona, Catalunya, Spain: ACM, 2013, pp. 835–838.

[27] J. Rynkiewicz, "General bound of overfitting for mlp regression models," *Neurocomputing*, vol. 90, no. 8, pp. 106–110, 2012.

[28] W. Han, "Considering relative order of emotional degree in dimensional speech emotion recognition," *Signal Processing*, vol. 27, no. 11, pp. 1658–1663, 2011.

[29] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity and native language," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*. San Francisco, CA: ISCA, 2016, pp. 2001–2005.

[30] S. H. Kim and S. Hamann, "Neural correlates of positive and negative emotion regulation," *Journal of Cognitive Neuroscience*, vol. 19, no. 5, pp. 776–798, 2007.

[31] 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, 2014.

[32] A. G. Wilson, Z. Hu, R. R. Salakhutdinov, and E. P. Xing, "Deep kernel learning," in *Proc. International Conference on Artificial Intelligence and Statistics*. Cadiz, Spain: PMLR, 2016, pp. 370–378.

[33] A. G. Wilson, Z. Hu, R. R. Salakhutdinov, and E. P. Xing, "Stochastic variational deep kernel learning," in *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016, pp. 2586–2594.