A HIERARCHICAL REGRESSION CHAIN FRAMEWORK FOR AFFECTIVE VOCAL BURST RECOGNITION

Jinchao Li[1], *Xixin Wu*[1*], *Kaitao Song*[2], *Dongsheng Li*[2], *Xunying Liu*[1], *Helen Meng*[1]

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Microsoft Research Asia, Shanghai, China
[1]{jcli,wuxx,xyliu,hmmeng}@se.cuhk.edu.hk, [2]{kaitaosong,Dongsheng.Li}@microsoft.com

ABSTRACT

As a common way of emotion signaling via non-linguistic vocalizations, vocal burst (VB) plays an important role in daily social interaction. Understanding and modeling human vocal bursts are indispensable for developing robust and general artificial intelligence. Exploring computational approaches for understanding vocal bursts is attracting increasing research attention. In this work, we propose a hierarchical framework, based on chain regression models, for affective recognition from VBs, that explicitly considers multiple relationships: (i) between emotional states and diverse cultures; (ii) between low-dimensional (arousal & valence) and high-dimensional (10 emotion classes) emotion spaces; and (iii) between various emotion classes within the high-dimensional space. To address the challenge of data sparsity, we also use self-supervised learning (SSL) representations with layer-wise and temporal aggregation modules. The proposed systems participated in the ACII Affective Vocal Burst (A-VB) Challenge 2022 and ranked first in the "TWO" and "CULTURE" tasks. Experimental results based on the ACII Challenge 2022 dataset demonstrate the superior performance of the proposed system and the effectiveness of considering multiple relationships using hierarchical regression chain models.

***Index Terms*—** affective computing, vocal bursts, emotional expression, multi-label, multi-culture, multi-task learning

1. INTRODUCTION

Recognition of emotions conveyed by non-linguistic vocalizations, e.g., affective bursts, has attracted increasing research attention, as vocalizations can reliably express certain emotions and the meanings of vocal bursts are generally preserved across diverse cultures [1]. This lays the theoretical foundation for using affective vocalization information to more robustly and holistically understand emotional reactions. Despite the fact that affect vocalizations and speech-embedded prosody both utilize the same expressive (vocal) apparatus, it is also found that the accuracy of emotion decoding for non-linguistic affect vocalizations is higher than the accuracy for speech-embedded vocal prosody [2]. Much research has been conducted in speech emotion recognition (SER) with verbal speech recently, such as feature exploration [3, 4] and multilingual generalization [5]. On the other hand, nonverbal vocalizations have received less attention.

Due to the scarcity of vocal burst data and lack of understanding about mechanisms of emotion signaling via vocal bursts, developing computational models for such emotion signaling remains a challenging task. Therefore, the recent ICML Expressive Vocalisations Workshop & Competition 2022 (ExVo) and the ACII Affective Vocal Bursts Workshop & Challenge 2022 (A-VB) introduce the large-scale Hume Vocal Bursts Competition Dataset (HUME-VB) for exploring various computational approaches [6, 7]. The corpus contains about 37 hours of self-recorded data by speakers in 4 countries spanning 3 native languages, which can support investigation of affective vocal bursts from diverse perspectives. Multi-task approaches have been demonstrated to be effective in previous works, e.g., by integrating various losses [8], jointly modeling auxiliary prediction tasks of culture and age [9]. However, it is desirable to explicitly model the relationship between emotion classes in vocalization signaling and the relationship between the different related tasks. To this end, we propose a hierarchical framework based on chain regression models, which generate predictions for one task that is conditioned on the prediction from the other related tasks.

With recent advancements in self-supervised learning (SSL), the adopted speech representations for emotion recognition are shifting from hand-crafted features, e.g., acoustic pitch and energy, to high-level embeddings extracted by pretrained models, such as Wav2vec 2.0 [10]. The large, Transformer-based SSL models trained on large-scale data can learn representations for various downstream tasks, including automatic speech recognition (ASR) [10] and SER [11]. As affective vocal burst (AVB) data is generally lacking, it is important to borrow data from other speech domains to improve the AVB modeling. Purohit *et al.* [12] compared supervised and self-supervised embeddings for the affective vocal burst recognition (AVBR), and showed that SSL-based representations typically yield better performance than supervised embeddings learned by pretrained task-dependent neural networks. To further leverage these high-level features, various network architectures have been explored in the latest SER research, such as layer-wise aggregation [13], temporal attention [14] and dynamic convolution [15]. Following [13], we leverage representations from different layers of pretrained models with trainable weights.

In this work, we investigate AVBR based on a hierarchical framework using chain regression models and pretrained representations. The relationships between emotional states and diverse cultures, between low-dimension and high-dimension emotion spaces, and between various emotion classes within the high-dimension space, are explicitly modeled. Our system participated in the ACII A-VB challenge and ranked first in the task of the "TWO" and "CULTURE" tasks, and second in the "HIGH" task. The effectiveness of the regression models and the weighted aggregation of pretrained representations is also demonstrated by further experiments we conduct on the challenge dataset.

[1] The work was done during the author's internship with Microsoft.
[*] Corresponding author.

2. METHODOLOGY

The proposed hierarchical multitask learning framework is illustrated in Fig. 1, mainly consisting of a high-level feature extractor (see left side Fig. 1), and a structured output layer with a bidirectional regression chain (see right side Fig. 1). In the following, we will describe our proposed framework from the bottom levels of feature extraction, to the representation aggregation across different pre-trained model layers, and to the top structured output layer.

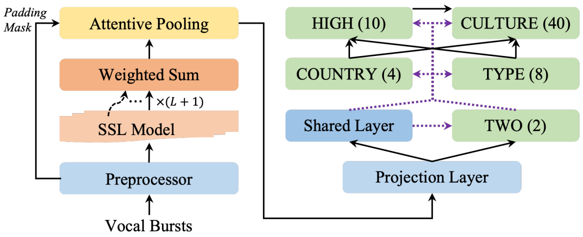

Fig. 1. Overview of hierarchical multitask learning framework for the A-VB 2022 competition. "TWO", "TYPE", "HIGH" and "CULTURE" denote the classifiers or regression models for corresponding tasks, and "Country" is the classifier of countries. "$(*)$" means the output size of the models.

2.1. Preprocessing

We preprocess the vocal burst data with peak normalization in the time domain for internal consistency of the data. In addition, we use data augmentation to enrich the data and improve the robustness by slightly changing the acoustic characteristics with minor distortions. Specifically, we applied pitch-shifting and speed-perturbation for each input waveform during the training stage (corpus size not changed) [16, 17]. The shifted ranges of pitch and speed are [-100, 100] semitones and [-0.05, +0.05] rates, respectively.

2.2. High-level Feature Extractor

The recent success of large pre-trained models motivates this work to adopt hidden embeddings from SSL models [18, 19]. We use the Wav2vec 2.0-Large XLSR ("w2v2-lg-xlsr") [10] to extract cross-lingual contextualised speech representations [20]. The Wav2vec 2.0 Large XLSR is trained on the CommonVoice corpus [21] by solving a contrastive task over masked latent speech representations and jointly learning a quantization of the latent representations shared across various languages.

The "w2v2-lg-xlsr" model contains one convolutional feature encoding layer and 24 stacked Transformer layers. The convolutional layer contains temporal convolutions with kernel widths (10,3,3,3,3,2,2) and strides (5,2,2,2,2,2,2), which yield a receptive field of about 320 samples. Through this convolutional layer, we can obtain a feature map with a shape of 49×1024 (dimensions of time and the embedding, respectively) for each one-second segment with a 16kHz sampling rate from input vocal burst signals. To avoid information loss caused by only using the last Transformer layer of the Wav2vec model, we leverage both the Transformer layers and the convolutional layer. We use learnable weights to sum up all the hidden states of the 24 stacked Transformer layers and the output feature map from the convolutional layer.

An attentive time pooling layer [22] follows the weighted summed features and is used to compress the feature sequence with variable lengths into a fixed-length vector. The attention mechanism also enables flexible focus on important frames for target prediction, by allocating more weights to the corresponding frames in the summation. Then, we project the features into a lower-dimensional vector to reduce redundancy, while retaining the intra-class variability. A batch normalization layer [23] is applied to standardize the high-level features before the features are fed to the subsequent classifiers and regressors.

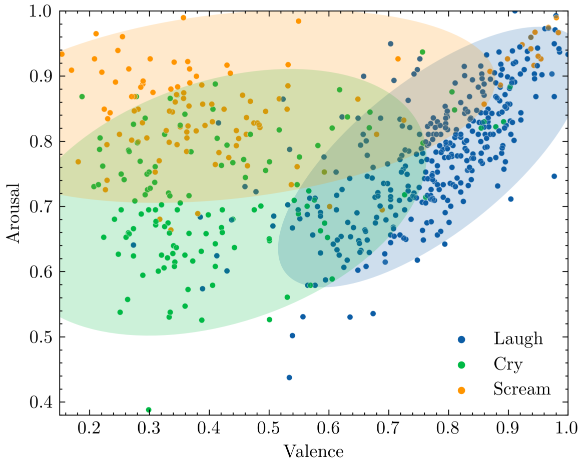

Fig. 2. Distribution of *laugh*, *cry* and *scream* of the TYPE task in the arousal-valence space of the TWO task. It can be found that different VB types have different distributions that could be modeled.

2.3. Hierarchical Multi-task Learning

We propose an elaborate hierarchical framework to explicitly model the relationships between the tasks. There are five tasks investigated in our framework [7]:

- **TWO** This task aims to predict the emotion of AB in a space with two dimensions, i.e., arousal and valence, based on the circumplex model of affect [24].
- **HIGH** The HIGH task is to predict the emotion intensity in a higher-dimensional space of 10 emotion classes, including *surprise*, *sadness*, *excitement*, *fear*, etc.
- **COUNTRY** We design this task to consider the relationship between VB and habitation locations. There are 4 countries considered in this task, i.e., U.S., China, Venezuela and South Africa.
- **CULTURE** This is a cross-cultural emotion task to predict the intensity of the 10 emotions associated with the above 4 countries.
- **TYPE** This task focuses on the prediction of 8 VB types, i.e., *cry*, *gasp*, *groan*, *grunt*, *laugh*, *pant*, *scream*, and *other*.

Following [25], we used different layers to disentangle task-agnostic and task-specific information. The shared feature extractor is trained to extract features that are generally useful for the different prediction tasks, while each task-specific feature extractor captures information that is more related to the corresponding task.

In terms of the relationship between emotion dimensions, the arousal and valence values in the TWO task can imply the emotion classes in the high-dimensional emotion space in the HIGH or

TYPE tasks [26]. As shown in Fig. 2, distributions of the VB types, *laugh*, *cry* and *scream*, are different in the arousal-valence space. The labels in the CULTURE task are combinations of emotions and countries. Therefore, the predictions of the HIGH, TYPE, CULTURE and COUNTRY tasks are conditioned on the predicted results of the TWO task, i.e., the predicted arousal and valence values. Since CULTURE task targets are combinations of emotion classes and countries, the system is designed to generate the CULTURE outputs based on the predictions from the HIGH and the COUNTRY tasks.

2.4. Bi-directional Regression Chain

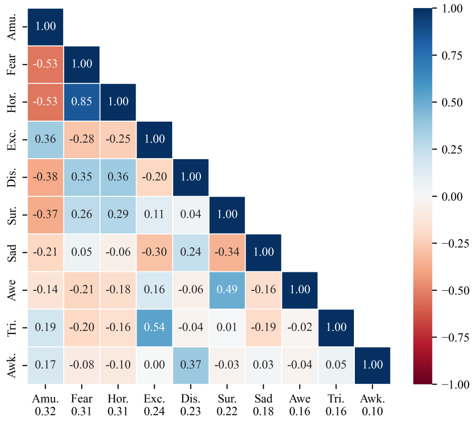

Fig. 3. Pearson correlation coefficients between emotion classes in the HIGH task based on training data.

It is noteworthy that the emotion classes are not independent, for example, a higher score in *amusement* implies higher score in *excitement* and lower scores in *fear* and *horror*. We visualize the Pearson correlation coefficients between the emotion classes on the training subset in Fig. 3. It is clearly shown that some pairs demonstrate significant correlation, which needs to be explicitly considered.

To model such relationships between emotion classes, we used a bi-directional regression chain to explicitly model the label dependency for the HIGH and CULTURE tasks. In a regression chain, with the predictor of the i-th emotion and the extracted feature denoted as f_i and z, respectively, the emotion score is calculated by: $\hat{y}_i = \sigma(f_i(z \bigoplus \hat{y}_{<i}))$, where σ is the sigmoid function, $\bigoplus$ is the vector concatenation operator, and $\hat{y}_{<i}$ is the previous predicted emotion scores before the i-th emotion prediction.

To mitigate the effect from the emotion order of the chain, we accumulate absolute coefficients of each emotion, and arrange the order from higher accumulated values to lower ones, as shown in the x-axis in Fig. 3. We then modify the chain regression layer to be bi-directional by adding another chain in the reverse direction.

2.5. Loss Functions

For the countries and labels in TYPE task, categorical cross entropy (CE) is used as the main loss function. For the labels in the TWO, HIGH and CULTURE tasks, averaged biased concordance correlation coefficient (CCC) is adopted as the main loss function [27]. The biased CCC is defined in Eq. 1.

$$CCC(x_i, y_i) = \frac{1}{N} \sum \frac{2 * cov(x_i, y_i)}{\sigma_{x_i}^2 + \sigma_{y_i}^2 + (\mu_{x_i} + \mu_{y_i})^2}, \qquad (1)$$

where N is the number of labels, and μ_{x_i}, μ_{y_i} is the mean of i-th prediction and the corresponding ground truth values, respectively. The biased covariance is defined as $cov(x_i, y_i) = \sum(x_i - \mu_{x_i})(y_i - \mu_{y_i})$.

The total loss function is a weighted combination of the losses in the main task and the auxiliary tasks:

$$\mathcal{L}_{Target} = \lambda \mathcal{L}_{Target} + (1 - \lambda) * \sum(\mathcal{L}_{Auxiliary}), \qquad (2)$$

where $\mathcal{L}_{Target}$ and $\mathcal{L}_{Auxiliary}$ are the loss functions for the target and the auxiliary tasks, respectively, and λ is a hyperparameter of the loss weights.

3. EXPERIMENTS

3.1. The A-VB Data

We use the HUME-VB dataset of emotional non-linguistic vocalizations (vocal bursts) [28] that is used in the ACII A-VB Competition 2022 [29]. The competition aims to promote research on modeling emotion in vocalizations, and proposes four tasks utilizing the HUME-VB data: the Two-Dimensional (TWO), High-Dimensional (HIGH), Cross-Cultural High-Dimensional (CULTURE) regression tasks, and the Expressive Burst-Type (TYPE) classification task. The HUME-VB data contains about 37 hours of audio data from 1702 speakers from China, South Africa, the U.S., and Venezuela. Each vocal burst is labeled from an average of 85.2 raters with intensities in [1:100] of ten different expressed emotions, *amusement*, *awe*, *awkwardness*, *distress*, *excitement*, *fear*, *horror*, *sadness*, *surprise*, and *triumph*. The data is subsequently partitioned into training, validation, and test splits, with consideration of speaker independence and balances across countries and vocalization types.

In this work, our target tasks are the TWO, HIGH and CULTURE tasks, while the COUNTRY and TYPE are used as auxiliary tasks. The TWO task aims to predict values of arousal and valence (based on 1=unpleasant/subdued, 5=neutral, 9=pleasant/stimulated), while The HIGH task aims to predict a higher dimension, i.e., the intensity of the aforementioned 10 emotions. The CULTURE task is a 10-dimensional, 4-country culture-specific emotion intensity regression task, i.e., it aims to predict the 40 intensity values of emotion (10 from each culture).

3.2. Experimental Setup

In this work, we set the dimensions of projection and shared layers to 128 and 64, respectively. The task-specific Bi-directional chains consist of two linear layers with sigmoid activation that are concatenated and averaged. The λ in Eq. 2 is set to 0.9. We use AdamW [30] as our optimizer with a learning rate of $1e - 5$ for the Wav2vec 2.0 model finetuning and $1e - 3$ for the downstream module training. To obtain a stabler CCC loss and alleviate the variance from the large pretrained model, we train the system with a large batch size of 1024 and a weight decay of $1e - 3$. A 0.25 dropout is added between every two modules. We also apply early stopping (patience of 10, maximum of 25 epochs) to avoid overfitting the model. The systems are evaluated on the validation and test datasets with the averaged biased CCC metric for the target tasks.

3.3. Baselines

The baseline systems in this challenge include feature-based and end-to-end methods [29]. The feature-based approach extracts 6,373-dimensional ComParE [31], 88-dimensional eGeMAPS [32] acoustic feature sets, and models the features with three fully-connected layers with layer normalization. While the end-to-end approach uses Emo-18 [33] convolutional neural networks followed by a 2-layer Long-short term memory (LSTM) network.

3.4. Experimental Results

Approach	TWO		HIGH		CULTURE	
	Val.	Test	Val.	Test	Val.	Test
ComParE	.4942	.4986	.5154	.5214	.3867	.3887
eGeMAPS	.4114	.4143	.4484	.4496	.3229	.3214
END2YOU	.4988	.5084	.5638	.5686	.4359	.4401
Ours	**.6966**	**.6854**	**.7351**	**.7237**	**.6464**	**.6017**

Table 1. Experimental results on the TWO, HIGH, and CULTURE tasks of the ACII A-VB challenge 2022. the mean concordance correlation coefficient (CCC) is reported.

We compare our system with the baselines on the TWO, HIGH and CULTURE tasks in Table 1. It can be found that the proposed system outperforms the baselines on all three tasks by a significant margin. This demonstrates the effectiveness of the proposed hierarchical framework.

Approach	Averaged CCC
ComParE	.5154
eGeMAPS	.4484
END2YOU	.5638
Ours	.7351
- Finetune	.6103
- Regression Chain	.6513
- Finetune & Regression Chain	.5540

Table 2. Ablation study for the HIGH task on the validation data. “Ours” means the hierarchical framework with chain regression and finetuned Wav2vec 2.0, “-” means removing corresponding module from “Ours”.

We also conducted experiments to verify the effectiveness of the integrated pre-trained representations and the regression chains on the HIGH task. As shown in Table 2, when the SSL representations are directly used without further fine-tuning on the HUME-VB dataset, the performance drops from 0.7351 to 0.6103, but still outperforms the baseline systems. If the regression chains are removed, the performance also decreases significantly, which demonstrates the effectiveness of the regression chains for the HIGH task. These results also suggest that the combination of fine-tuned SSL representations that implicitly borrow from external data, and the regression chains that model interactions between emotion classes, are both beneficial for performance.

Approach	Valence	Arousal	Average
Ours	.7622	.6309	.6966

Table 3. Performance (CCC) of proposed system for the TWO task on all validation data.

We further analyze the performance of the arousal and valence prediction in the TWO task. The breakdown of performance is shown in Table 3. It can be observed that the CCC of predicted valence values is much higher than that of predicted arousal values. This matches well with the characteristics of the HUME-VB dataset – that the distribution of human valence annotation is more diffuse than the arousal distribution [7].

Approach	Awe	Excite.	Amuse.	Awkward.	Fear
Ours	.8169	.6962	.7928	.6085	.7742
Approach	Horror	Distress	Triumph	Sadness	Surprise
Ours	.7528	.7010	.6914	.7110	.8063

Table 4. Performance (CCC) of the proposed method for the HIGH task on the validation data.

For the HIGH task, the performances of different emotion classes are shown in Table 4. It can be seen that all 10 classes have satisfactory performance. In particular, the *awkward* class is relatively more difficult with a slightly lower performance, which is also observed in [34].

Countries	Average	Train	Val.
China	.6149	79	76
U.S.	.7302	206	206
South Africa	.6520	244	244
Venezuela	.5885	42	42

Table 5. Performance (CCC) of the proposed method for the CULTURE task on the validation dataset. Distribution of recording numbers for the four countries on the training and validation sets is also shown.

In the CULTURE task, it can be found that the performance for the data from Venezuela is significantly worse than the other locations. This is probably caused by the unbalanced distribution in the dataset. This is shown in Table 5, where the training and validation data for Venezuela is much less compared to the data for U.S. and South Africa. Similarly, the performance for China is also inferior to the those for U.S. and South Africa.

4. CONCLUSION

In this paper, we investigate affective vocal burst recognition (AVBR) by proposing a hierarchical framework with bi-directional regression chains to explicitly consider multiple relationships, (i) between emotional states and diverse cultures, (ii) between low-dimensional and high-dimensional emotion spaces, and (ii) between various emotion classes within the high-dimensional space. To address the data sparsity problem in AVBR, we also integrate SSL representations via a trainable aggregation method. The proposed framework achieves significantly better performance than baseline systems on the HUME-VB dataset. Data analysis on the dataset and the experimental results also supports the necessity of modeling the inherent relationships. In the future, we will investigate imbalanced learning w.r.t. cultures and labels in the AVBR task. We will also try to interpret the affective cues from the high-level embeddings for VBs.

5. REFERENCES

[1] A. S. Cowen, H. A. Elfenbein, P. Laukka, and D. Keltner, "Mapping 24 emotions conveyed by brief human vocalization.," *American Psychologist*, vol. 74, no. 6, pp. 698, 2019.

[2] S. T. Hawk, G. A. Van Kleef, A. H. Fischer, and J. Van Der Schalk, ""worth a thousand words": absolute and relative decoding of nonlinguistic affect vocalizations.," *Emotion*, vol. 9, no. 3, pp. 293, 2009.

[3] P. P. Liang, Y. Lyu, X. Fan, et al., "Multibench: Multiscale benchmarks for multimodal representation learning," *arXiv preprint arXiv:2107.07502*, 2021.

[4] J. Li, S. Wang, Y. Chao, et al., "Context-aware multimodal fusion for emotion recognition," *INTERSPEECH*, pp. 2013–2017, 2022.

[5] Y. B. Singh and S. Goel, "A systematic literature review of speech emotion recognition approaches," *Neurocomputing*, 2022.

[6] A. Baird, P. Tzirakis, G. Gidel, et al., "The icml 2022 expressive vocalizations workshop and competition: Recognizing, generating, and personalizing vocal bursts," *arXiv preprint arXiv:2205.01780*, 2022.

[7] A. Baird, P. Tzirakis, J. A. Brooks, et al., "The acii 2022 affective vocal bursts workshop & competition: Understanding a critically understudied modality of emotional expression," *arXiv preprint arXiv:2207.03572*, 2022.

[8] X. Jing, M. Song, A. Triantafyllopoulos, et al., "Redundancy reduction twins network: A training framework for multi-output emotion regression," *arXiv preprint arXiv:2206.09142*, 2022.

[9] M. Song, Z. Yang, A. Triantafyllopoulos, et al., "Dynamic restrained uncertainty weighting loss for multitask learning of vocal expression," *arXiv preprint arXiv:2206.11049*, 2022.

[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[11] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *ICASSP*. IEEE, 2022, pp. 6907–6911.

[12] T. Purohit, I. B. Mahmoud, B. Vlasenko, and M. M. Doss, "Comparing supervised and self-supervised embedding for exvo multi-task learning track," *arXiv preprint arXiv:2206.11968*, 2022.

[13] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *INTERSPEECH*, pp. 3400–3404, 2021.

[14] J. Liu, Z. Liu, L. Wang, et al., "Temporal attention convolutional network for speech emotion recognition with latent representation.," in *INTERSPEECH*, 2020, pp. 2337–2341.

[15] H. Wen, S. You, and Y. Fu, "Cross-modal dynamic convolution for multi-modal emotion recognition," *Journal of Visual Communication and Image Representation*, vol. 78, pp. 103178, 2021.

[16] P. A. Cariani and B. Delgutte, "Neural correlates of the pitch of complex tones. ii. pitch shift, pitch ambiguity, phase invariance, pitch circularity, rate pitch, and the dominance region for pitch," *Journal of neurophysiology*, vol. 76, no. 3, pp. 1717–1734, 1996.

[17] J. A. Colosi and M. G. Brown, "Efficient numerical simulation of stochastic internal-wave-induced sound-speed perturbation fields," *The Journal of the Acoustical Society of America*, vol. 103, no. 4, pp. 2232–2235, 1998.

[18] A. F. Adoma, N.-M. Henry, and W. Chen, "Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition," in *Proc. ICCWAMTIP*. IEEE, 2020, pp. 117–121.

[19] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[20] A. Conneau, A. Baevski, R. Collobert, et al., "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[21] R. Ardila, M. Branson, K. Davis, et al., "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[22] C. d. Santos, M. Tan, B. Xiang, and B. Zhou, "Attentive pooling networks," *arXiv preprint arXiv:1602.03609*, 2016.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[24] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161, 1980.

[25] A. Anuchitanukul and L. Specia, "Burst2vec: An adversarial multi-task approach for predicting emotion, age, and origin from vocal bursts," *arXiv preprint arXiv:2206.12469*, 2022.

[26] E. Schubert, "Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space," *Australian Journal of Psychology*, vol. 51, no. 3, pp. 154–165, 1999.

[27] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[28] A. Cowen, A. Bard, P. Tzirakis, et al., "The hume vocal burst competition dataset (H-VB)," *Zenodo*, 2022.

[29] A. Baird, P. Tzirakis, A. Batliner, et al., "The ACII 2022 affective vocal bursts workshop and competition: Understanding a critically understudied modality of emotional expression," 2022.

[30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[31] B. Schuller, S. Steidl, A. Batliner, et al., "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *INTERSPEECH*, 2013.

[32] F. Eyben, K. R. Scherer, B. W. Schuller, et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[33] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *ICASSP*. IEEE, 2018, pp. 5089–5093.

[34] D. Xin, S. Takamichi, and H. Saruwatari, "Exploring the effectiveness of self-supervised learning and classifier chains in emotion recognition of nonverbal vocalizations," *arXiv preprint arXiv:2206.10695*, 2022.