Leveraging Pre-Trained Acoustic Feature Extractor For Affective Vocal Bursts Tasks

Bagus Tris Atmaja*†, and Akira Sasou*
* National Institute of Advanced Industrial Science and Technology, Japan
† Sepuluh Nopember Institute of Technology, Indonesia

***Abstract*—Understanding humans' emotions is a challenge for computers. Nowadays, research on speech emotion recognition has been conducted progressively. Instead of a speech, affective information may lay on short vocal bursts (i.e., cry when sad). In this study, we evaluated a recent self-supervised learning model to extract acoustic embedding for affective vocal bursts tasks. There are four tasks investigated on both regression and classification problems. Using similar architectures, we found the effectiveness of using a pre-trained model over the baseline methods. The study is further expanded to evaluate the different number of seeds, patiences, and batch sizes on the performance of the four tasks.**

***Index Terms*—affective computing, affective vocal bursts, pre-trained model, wav2vec 2.0, speech emotion recognition**

I. Introduction

Affective computing is an emerging field in technology for understanding human emotions by computers. It is a combination of interdisciplinary fields, from computer science to psychology. The goal of affective computing is to analyze and synthesize human emotions, i.e., creating a system that can interact with humans naturally. The system can be used in various fields, such as healthcare, education, and entertainment. The information for the system can be collected from various sources, such as speech, facial expressions, and body language.

One of the key components of affective computing is speech emotion recognition, which uses speech (verbal communication) as a source to analyze and synthesize human emotions. Speech emotion recognition is a challenging task since it is difficult to distinguish between emotional states. For example, the emotion of anger and fear are similar since both of them are expressed by raising the pitch of the voice. The emotion of happiness and sadness are also similar since both of them are expressed by lowering the pitch of the voice. Given these difficulties, analyzing emotions from humans' non-verbal communication may improve our understanding of human emotions.

Vocal bursts, non-verbal vocalizations like laughter and cries, constitute a potential source of information for emotion [1]. Scherer [2] proposed to model vocal communication as Brunswik's lens model from encoding (expression) to representation (perception). There is no exact number of emotion categories resulting from this study. The authors mentioned eight examples of emotion categories with ranges of importance for their design features delimitation (e.g., intensity). A study by Cowen et al. [3] has found that vocal bursts are rich in emotional information that can be conceptualized into 24 emotion categories. Nevertheless, the research on affective vocal bursts has been limited by the lack of data until recently.

The workshops, challenges, and competitions on affective vocal bursts accelerate research on affective vocalizations and, at the same time, provide a dataset to experiment with [4], [5], [6]. In [4], one of the challenges is the vocalization sub-challenge. The participants are tasked to predict one of the six classes of categorical emotions for vocal bursts provided in the test set. In [5], two out of three tasks are vocal bursts recognition from multitask and few-shot learnings. The challenge employed a large Hume-VB dataset [7] with 59201 samples. In [6], the Hume-VB dataset is again used for all four tasks of affective vocal burst recognition. Three tasks are regression problems for measuring either the intensity of emotion categories or valence and arousal; another task is for predicting the type of vocal burst. This study partially intended to participate in this challenge.

Previous works on affective vocal burst have been proposed to accept the challenges in competitions and workshops. Belanic et al. [8] evaluated two classifiers, ResNet and Conformer, for multitask vocal bursts modeling and achieved improvements over the baseline method on the validation set. Anuchitanukul and Specia [9] proposed Burst2Vec approach and achieved state-of-the-art results on ExVo multitask learning track. Purohit et al. [10] evaluated supervised and semi-supervised learning for the same ExVo multitask learning problem and showed that semi-supervised learning outperforms supervised learning in terms of the overall score. As it has been noted in [8], these results may only apply to the associated dataset/challenge and may not necessarily generalize to other datasets.

This study contributes to the affective computing research by evaluating a pre-trained acoustic feature extractor for all affective vocal tasks in The ACII Affectice Vocal Bursts Workshop & Competition [6]. Using the same architecture for all tasks, we obtained remarkable improvements over the baseline methods provided by the organizer. We also evaluated the effect of the different numbers of seeds, patiences, and batch sizes on the performance of the four tasks.

The next sections describe the methods, results and discussion, and the conclusion. The methods include the dataset and tasks, a pre-trained feature extractor, and the classifier. A pre-trained acoustic feature extractor is the key method in this work, which is compared to the baseline methods on the four tasks with the same classifier. The results and discussion

section discuss the experiment results on different conditions; the conclusion section summarizes the findings.

II. Dataset and Tasks

This study employed Hume-VB dataset [6], [7] to predict four tasks of affective vocal bursts. The dataset consists of 36 hours of recording from 1702 speakers across four countries: China, South Africa, the US, and Venezuela. Each vocal burst was labeled on an integer scale from 1 to 100 for ten expressed emotions: amusement, awe, awkwardness, distress, excitement, fear, horror, sadness, surprise, and triumph. The integer scale was scaled to [0,1] during the experiments. The original raw audio data were sampled at 48 kHz (but resampled to 16 kHz for the experiments). The data were already partitioned into training, validation, and test sets for each task by the organizer of the challenge. No label is provided for the test set; the score to obtain the test set's performance is obtained by emailing the predictions to the organizers.

There are four tasks provided in the ACII 2022 affective vocal bursts workshop and competition. The first task, called "High", is to predict the intensity of 10 aforementioned emotions. The second task, called "Two", is to predict the degree of valence and arousal for given vocal bursts in the test set. The third task, called "Culture", is to predict the intensity of 40 emotions (10 from each culture) as a multioutput regression problem. The fourth task, called "Type", is to predict the type of given vocal bursts in the test set. There are eight types of vocal bursts in the fourth task: gasp, laugh, cry, scream, grunt, groan, pant, and other. The first to third tasks are regression problems with concordance correlation coefficient (CCC) as the evaluation metric; the last task is a classification problem with unweighted average recall (UAR) as the evaluation metric.

III. Pre-Trained Acoustic Feature Extractor

Pre-trained models recently gained more attention due to their effectiveness in modeling data based on self-supervised learning, including pre-trained models for speech emotion recognition. One of the models built for this specific purpose, i.e., speech emotion recognition, is proposed by Wagner et al. by incorporating a large and robust version of wav2vec 2.0 on the affective speech dataset [11], [12]. The base model is Robust wav2vec 2.0 [13] trained on MSP-Podcast dataset [14]. One of the outputs of the model is the hidden states which can be used as acoustic features or acoustic embedding for affective-related tasks. Another output is logits, the degree of arousal, dominance, and valence in a range [0, 1]. We concatenated the hidden states (1024-dims) and logits (3-dims) as acoustic embedding for all vocal burst tasks (w2v2-r-vad, 1027-dims). The acoustic embedding was then fed to the regression or classification model, depending on the task.

IV. Classifiers

Research on deep neural networks has been developed progressively with several new architectures. Nevertheless, classical approaches such as fully-connected networks or multilayer perceptron have shown competitiveness against newer architectures like LSTM or CNN [15]. In this study, we employed a fully-connected network as the classifier. The fully-connected network is a feed-forward neural network with three hidden layers. The number of nodes for each layer is 128, 64, and 32, respectively. Each layer is connected to batch normalization [16] layer and leaky rectified linear unit (LeakyReLU) activation function. The number of nodes at the output layers depends on the task, i.e., 10 for High, 2 for Two, 10 for Culture, and 8 for Type. The output layer for regression problems is activated with a sigmoid function.

Fig. 1 and Table I depict the architecture and hyperparameters of the fully-connected network. The architecture and hyperparameters are the same for all tasks. The learning rate is set to 0.001, weight decay is set to 0.01, and the maximum number of epochs is 25. The optimizer is AdamW [17] with a weight decay of 0.01. For ablation experiments, the authors varied the number of seeds from 5 to 20, the patience from 5 to 25, and the batch size from 4 to 1024.

Three tasks, High; Two, and Culture, employed MSE loss to be minimized. The Type task employed cross-entropy loss to be minimized. The evaluation metric is the concordance correlation coefficient (CCC) for High, Two, Culture and is an unweighted average recall (UAR) for Type. The range for CCC scores is in [-1, 1], whereas the UAR score is in [0, 1]. The reported scores for evaluation on the validation set are the maximum or average of five runs.

TABLE I
Hyperparameters of the classifier

Hyper-parameter	Value
Layer	MLP
N_layers	3
Nodes	(128, 64, 32)
Normalization	BatchNorm1d
Layer activation	LeakyReLU
Output activation	Sigmoid
Optimizer	AdamW
Learning rate	0.001
Weight decay	0.01
Number of seeds	5 – 20
Batch size	4 – 1024
Patience	5 – 25
Patience delta	0.01

V. Results and Discussion

We present our experimental results in different ways. First, we evaluated the performance of the proposed acoustic features, w2v2-r-vad, on the same classifier for four different tasks. Then, we reported ablation studies on different batch sizes and numbers of patience to optimize our methods. Finally, we reported the test results on the hidden set.

A. Comparison of Different Acoustic Features

At first, we evaluated different acoustic embeddings on the same classifier explained in the previous section. As the baselines are ComParE feature set [18] with 6373-dims and eGeMAPS feature set [19] with 88-dims. The proposed acoustic embedding is w2v2-r-vad with 1027-dims. The experiments

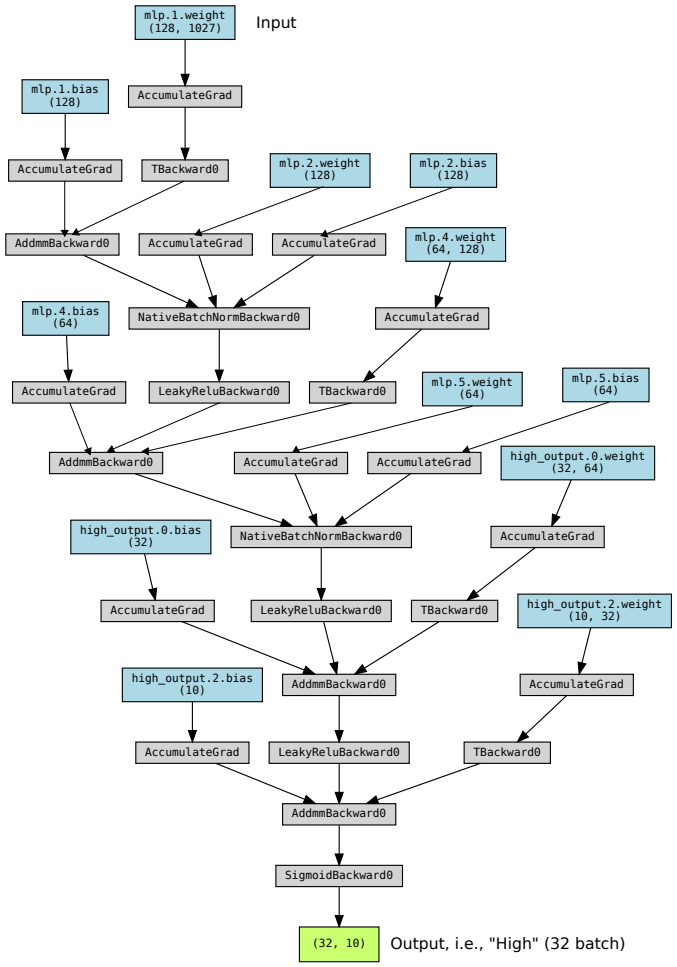

Fig. 1. Architecture of MLP networks for all tasks, the example is for regression "High" task with 32 batch size for 10 outputs.

for comparing these acoustic embeddings are set at batch size 32, the patience of 10 epochs, and a single seed of 109. The architecture and hyperparameters are optimized for w2v2-r-vad embedding.

Table II shows evaluation of different acoustic embedding on validation set. The table clearly indicates the superiority of w2v2-r-vad among other acoustic embeddings. For all four tasks, w2v2-r-vad achieved the top performance with remarkable gaps. The assumption that the self-supervised learning (SSL) model will achieve better performances than traditional acoustic features has been proven in this study. Moreover, the effectiveness of this wav2vec 2.0-based model trained on the affective speech dataset may be compared with other SSL methods in the future.

B. Effect of Seed Number

Seed number to initiate the process of deep learning methods has been crucial for finding the best performance and stability of the model [20], [21]. While the previous study only evaluated single seeds to determine the best model with associated seed [22] in repeated experiments, this study evaluated different numbers of seeds for initiations. The authors

TABLE II
MAXIMUM VALIDATION SCORES ON DIFFERENT ACOUSTIC FEATURES ON THE SAME CLASSIFIER (CCC FOR HIGH, TWO, AND TYPE; UAR FOR TYPE)

Feature	High	Two	Culture	Type
ComParE	0.4734	0.4648	0.3762	0.3694
EgeMAPS	0.1790	0.1049	0.1672	0.2767
w2v2-r-vad	**0.6427**	**0.6023**	**0.4802**	**0.4502**

evaluated 5, 10, 15, and 20 different seed numbers to observe the performance of the model. The model randomly chooses these numbers from 101 to 120. Both maximum and average scores from different seed numbers are reported (Table III and Fig. 2).

Table III shows maximum scores using different seed numbers. It is shown that the effect of different seed numbers is minimum. No remarkable difference has been found in using different seed numbers from 5 to 20 numbers. The gap between the highest and lowest scores on four tasks are 0.0008, 0.0030, 0.0013, and 0.0111 for High, Two, Culture, and Type, respectively. The largest gap is between 5 seed numbers and ten seed numbers on the Type task. Given this finding, it is reliable to evaluate the model with a minimum number of 5 seeds to report the average performance of the model. The average score of different seed numbers is still needed since the model is non-deterministic, meaning that the model will produce different results in different runs.

To evaluate the stability of the model over different seed numbers, Fig. 2 summarizes the experiments by showing average scores and their standard deviations. As in the maximum scores report, there is no remarkable difference in the average scores of the different seed numbers. The average scores and their deviation shows the stability of the evaluated model, a three-layer MLP with 128, 64, and 32 nodes, for all affective vocal burst tasks.

TABLE III
MAXIMUM SCORES AS AN EFFECT OF THE NUMBER OF SEEDS ON THE VALIDATION SET (CCC FOR HIGH, TWO, AND TYPE; UAR FOR TYPE)

N_seeds	High	Two	Culture	Type
5	0.6448	0.6136	0.4876	0.4786
10	0.6466	0.6146	0.4898	0.4638
15	0.6457	0.6097	0.4876	0.4685
20	0.6466	0.6108	0.4929	0.4709

C. Effect of Batch Sizes

The next evaluation is the different batch sizes (for each task), which is crucial in training deep learning methods. As shown in this study, using different batch sizes lead to different performances. The authors evaluated 32, 64, 128, 256, 512, and 1024 batch sizes. Both maximum and average scores are reported in Table IV and Fig. 3. In contrast to the previous evaluation of the different numbers of seeds, the gaps between the highest and lowest scores in maximum scores are 0.0383, 0.0621, 0.0392, and 0.0172 for High, Two, Culture, and Type, respectively. As shown in maximum scores in Table 3, using

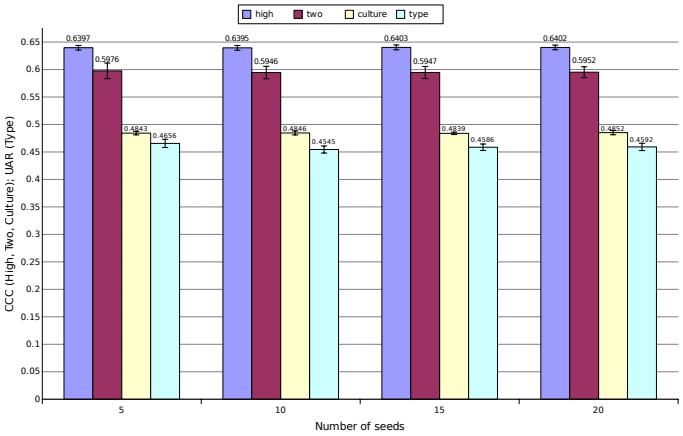

Fig. 2. Average scores (CCC and UAR) from the different number of seeds with their standard deviations

batch sizes of 16, 32, 64, or 128 achieves the same level of performance among other batch sizes.

For choosing the optimal batch size for each task, the average scores are more meaningful than the maximum scores. These scores show the stability of the model (with evaluated batch size) over different runs. The evaluation (Fig. 3) leads to the following optimal batch size for each task: High with 32 batch size, Two and Culture with 64 batch size, and Type with 128 batch size. Hence, for submitting the test evaluations, the authors used the optimal batch size for each task in addition to the same batch size for all tasks.

TABLE IV
MAXIMUM SCORES OF DIFFERENT BATCH SIZES ON THE VALIDATION SET (CCC FOR HIGH, TWO, AND TYPE; UAR FOR TYPE). FOR AVERAGE SCORES SEE FIG. 3.

Batch size	High	Two	Culture	Type
4	0.6093	0.5524	0.4587	0.4595
8	0.6343	0.5932	0.4796	0.4696
16	0.6429	0.6056	0.4902	0.4773
32	0.6426	0.6095	0.4913	0.4765
64	0.6430	0.6070	0.4911	0.4743
128	0.6419	0.6026	0.4842	0.4786
256	0.6404	0.6022	0.4837	0.4698
512	0.6370	0.6072	0.4735	0.4769
1024	0.6311	0.6048	0.4632	0.4702

D. Effect of Number of Patiences

The final evaluation is the effect of the different numbers of patience. The number of patience is needed to terminate the learning process once there is no improvement in the given number of patience. In this study, the authors evaluated five different numbers of patience: 5, 10, 15, 20, and 25. Note that using 25 of patience means no early stop criterion was used since the maximum number of epochs is also 25. The results are shown in Table V and Fig. 4 for maximum and average scores.

Both maximum scores in Table V and average scores in Table 4 indicate no remarkable difference in using different numbers of patience, as the authors found on the seeds' evaluation. The gaps between the highest and lowest scores in maximum scores are 0.0044, 0.0083, 0.0032, and 0.0226 for High, Two, Culture, and Type, respectively. Similar gaps were also observed in average scores. Given this finding, using five patiences is enough to terminate the learning process and to make predictions based on that model.

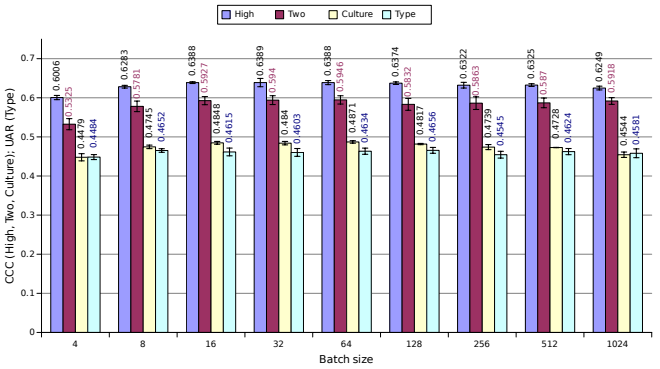

Fig. 3. Average scores (CCC and UAR) from five seeds with their standard deviations as a function of batch size

TABLE V
MAXIMUM SCORES AS THE EFFECT OF THE NUMBERS OF PATIENCE ON THE VALIDATION SET (CCC FOR HIGH, TWO, AND TYPE: UAR FOR TYPE)

N_patiences	High	Two	Culture	Type
5	0.6438	0.6084	0.4851	0.4758
10	0.6452	0.6106	0.4903	0.4786
15	0.6425	0.6122	0.4876	0.4592
20	0.6460	0.6036	0.4875	0.4672
25	0.6416	0.6088	0.4920	0.4496

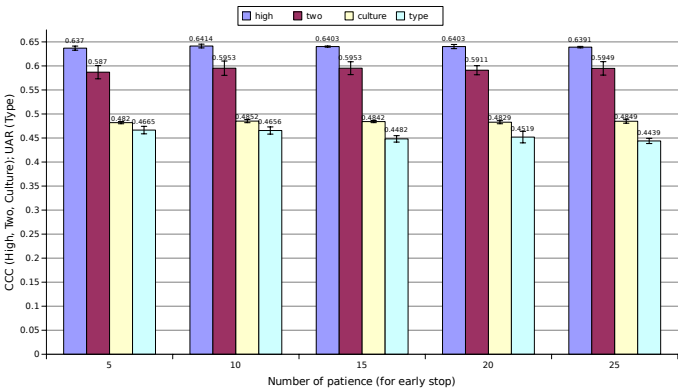

Fig. 4. Average scores (CCC and UAR) from the different numbers of patience with their standard deviations

E. Test Benchmark

The previous evaluations were conducted on the validation set since the labels of the test set are not provided; the following evaluation report scores on the test set. The scores were obtained by submitting the predictions of the test set to the organizer of The ACII 2022 Affective Vocal Bursts Workshop and Competition. The authors submitted two predictions. One

TABLE VI
PERFORMANCE SCORES ON THE TEST SET (CCC FOR HIGH, TWO, AND TYPE; UAR FOR TYPE)

Feature	High		Two		Culture		Type	
	Val.	Test	Val.	Test	Val.	Test	Val.	Test
ComParE	0.5154	0.5214	0.4942	0.4986	0.3867	0.3887	0.3913	0.3839
eGeMAPS	0.4484	0.4496	0.4114	0.4143	0.3229	0.3214	0.3608	0.3546
End2You	0.5638	0.5686	0.4988	0.5084	0.4359	0.4401	0.4166	0.4172
w2v2-r-vad #1	0.6427	0.6440	0.6023	0.5948	0.4802	0.4835	0.4502	0.4560
w2v2-r-vad #2	0.6466	**0.6478**	0.6156	**0.6142**	0.4929	**0.4962**	0.4810	**0.4791**

is with the same batch size of 32 for all tasks, and the other is with the optimal batch size for each task (32 for High, 64 for Two and Culture, and 128 for Type). The results are shown in Table VI. As the baselines, the authors quoted results from [6] with eGeMAPS [19], ComParE [18], and End2You [23] approaches.

It clearly shows that our two approaches with w2v2-r-vad gain remarkably better scores than the best method with End2You approach. On the first submission with the same 32 batch size for all tasks (w2v2-r-vad #1), we improved the End2You approach by absolute margins of about 0.0388 to 0.09 for all tasks. By employing our findings on batch size evaluation, i.e., using a specific batch size for each task, we slightly improved our scores from the first submission to the second submission (w2v2-r-vad #2) by absolute margins of about 0.0561 to 0.1094 from the best baseline. This finding indicates that the optimal batch size for each task is better than the same batch size for all tasks. It also can be noted in the comparison of validation and test scores that the gap between the two scores is not large, which means that the model is not overfitting.

VI. CONCLUSION

In this study, we evaluated leveraging the pre-trained self-supervised learning model, trained on an affective speech dataset, for affective vocal bursts tasks. The authors evaluated the same architecture for four affective vocal burst tasks and obtained improvements over the baseline methods using traditional acoustic features and an end-to-end approach. The best scores from baseline methods were obtained by End2You, while the best scores from the proposed approach were obtained by w2v2-r-vad second submission with 32, 64, 64, and 128 batches for High, Two, Culture, and Type tasks, respectively. The gains for these tasks are from 0.5686 to 0.6478, 0.5084 to 0.6142, 0.4401 to 0.4962, and 0.4172 to 0.4791. Our methods improved the affective vocal burst recognition maximally at predicting valence and arousal (the Two task) and minimally at the Culture task. Although the expression of vocal burst is general across cultures, there is a need for adjustment for different cultures in building an affective recognition model, which is shown by different results (High vs. Culture) and improvement (on the Culture task).

ACKNOWLEDGMENT

This paper is partly based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO), Japan.

REFERENCES

[1] N. Holz, P. Larrouy-Maestri, and D. Poeppel, "The paradoxical role of emotional intensity in the perception of vocal affect," *Sci. Rep.*, vol. 11, no. 1, p. 9663, 2021.

[2] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, no. 1-2, pp. 227–256, 2003.

[3] A. S. Cowen, H. A. Elfenbein, P. Laukka, and D. Keltner, "Mapping 24 emotions conveyed by brief human vocalization." *Am. Psychol.*, vol. 74, no. 6, pp. 698–712, sep 2019.

[4] B. W. Schuller, A. Batliner, S. Amiriparian, C. Bergler, M. Gerczuk, N. Holz, P. Larrouy-Maestri, S. P. Bayerl, K. Riedhammer, A. Mallol-Ragolta, M. Pateraki, H. Coppock, I. Kiskin, M. Sinka, and S. Roberts, "The ACM Multimedia 2022 Computational Paralinguistics Challenge: Vocalisations, Stuttering, Activity, & Mosquitoes," 2022.

[5] A. Baird, P. Tzirakis, G. Gidel, M. Jiralerspong, E. B. Muller, K. Mathewson, B. Schuller, E. Cambria, D. Keltner, and A. Cowen, "The ICML 2022 Expressive Vocalizations Workshop and Competition: Recognizing, Generating, and Personalizing Vocal Bursts," 2022.

[6] A. Baird, P. Tzirakis, J. A. Brooks, C. B. Gregory, B. Schuller, A. Batliner, D. Keltner, and A. Cowen, "The ACII 2022 Affective Vocal Bursts Workshop & Competition: Understanding a critically understudied modality of emotional expression," in *ACII Work. Demos*, 2022.

[7] A. Cowen, A. Bard, P. Tzirakis, M. Opara, L. Kim, J. Brooks, and J. Metrick, "The Hume Vocal Burst Competition Dataset (H-VB) — Raw Data [ExVo: updated 02.28.22] [Data set]," *Zenodo*, 2022.

[8] J. Belanich, K. Somandepalli, B. Eoff, and B. Jou, "Multitask vocal burst modeling with ResNets and pre-trained paralinguistic Conformers," in *39th Int. Conf. Mach. Learn.*, Baltimore, Maryland, 2022.

[9] A. Anuchitanukul and L. Specia, "Burst2Vec: An Adversarial Multi-Task Approach for Predicting Emotion, Age, and Origin from Vocal Bursts," in *39th Int. Conf. Mach. Learn.*, 2022. [Online]. Available: http://arxiv.org/abs/2206.12469

[10] T. Purohit, I. B. Mahmoud, B. Vlasenko, and M. M. Doss, "Comparing supervised and self-supervised embedding for ExVo Multi-Task learning track," in *39th Int. Conf. Mach. Learn.*, jun 2022. [Online]. Available: http://arxiv.org/abs/2206.11968

[11] B. W. Wagner, Johannes, Triantafyllopoulos, Andreas, Wierstorf, Hagen, Schmitt, Maximilian, Eyben, Florian, Schuller, "Model for Dimensional Speech Emotion Recognition based on Wav2vec 2.0 (1.1.0)," 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6221127

[12] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," mar 2022. [Online]. Available: http://arxiv.org/abs/2203.07378

[13] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," in *Interspeech 2021*, vol. 3. ISCA: ISCA, aug 2021, pp. 721–725.

[14] R. Lotfian and C. Busso, "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, 2019.

[15] B. T. Atmaja and M. Akagi, "Deep Multilayer Perceptrons for Dimensional Speech Emotion Recognition," in *2020 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA ASC 2020 - Proc.*, Auckland, 2020, pp. 325–331.

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 1, 2015, pp. 448–456.

[17] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *7th Int. Conf. Learn. Represent. ICLR 2019*, nov 2017. [Online]. Available: http://arxiv.org/abs/1711.05101

[18] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. August, pp. 148–152, 2013.

[19] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, apr 2016.

[20] M. Macary, M. Lebourdais, M. Tahon, Y. Estève, and A. Rousseau, "Multi-corpus Experiment on Continuous Speech Emotion Recognition: Convolution or Recurrence?" in *Int. Conf. Speech Comput.*, 2020, pp. 304–314.

[21] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion Approaches for Emotion Recognition from Speech Using Acoustic and Text-Based Features," in *ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, may 2020, pp. 6484–6488.

[22] B. T. Atmaja, Zanjabila, and A. Sasou, "Jointly Predicting Emotion, Age, and Country Using Pre-Trained Acoustic Embedding," in *10th Int. Conf. Affect. Comput. Intell. Interact. Work. Demos*, jul 2022.

[23] P. Tzirakis, S. Zafeiriou, and B. W. Schuller, "End2You-The Imperial Toolkit for Multimodal Profiling by End-to-End Learning," *arXiv Prepr. arXiv 1802.01115*, 2018.