Phoneme-Based Multi-Task Assessment of Affective Vocal Bursts

Tobias Hallmen[1], Silvan Mertes[1], Dominik Schiller[1], Florian Lingenfelser[1], and Elisabeth André[1]

Chair for Human-Centered Artificial Intelligence, University of Augsburg, Germany
https://hcai.eu
{tobias.hallmen, silvan.mertes, dominik.schiller, florian.lingenfelser, elisabeth.andre}@uni-a.de

Abstract. Affective speech analysis is an ongoing topic of research. A relatively new problem in this field is the analysis of affective vocal bursts, which are non-verbal vocalisations such as laughs or sighs. The current state of the art in the analysis of affective vocal bursts is predominantly based on wav2vec2 or HuBERT features. In this paper, we investigate the application of the wav2vec2 successor data2vec and the extension wav2vec2phoneme in combination with a multi-task learning pipeline to tackle different analysis problems at once, e.g., type of burst, country of origin, and conveyed emotion. Finally, we present an ablation study to validate our approach. We discovered that data2vec appears to be the best option if time and lightweightness are critical factors. On the other hand, wav2vec2phoneme is the most appropriate choice if overall performance is the primary criterion.

Keywords: data2vec · wav2vec2 · wav2vec2phoneme · vocal bursts · affective vocal bursts

1 Introduction

The human voice is a fundamental means of communication. While it can be used to produce spoken language, it can also carry an enormous amount of information on its own. Especially in the field of affect, non-verbal patterns are often even more important than linguistic content [21]. This becomes particularly apparent when listening to *vocal bursts*, which are short and intense vocalizations, often expressing strong emotions. The fact that vocal bursts can effectively communicate affective information without using verbal language makes them an interesting object of research. However, computational analysis of affective vocal bursts still remains a challenging topic [23, 24, 8, 10]. As such, it is surprising that the current state-of-the-art approaches for affective vocal burst analysis rely on wav2vec2 [5] or HuBERT [14] models that were trained on speech data, which has a substantially different structure than non-verbal vocal bursts. Therefore, in this paper, we examine the use of a successor and an extension of wav2vec2:

- First, we study if *data2vec* [4], a more generic version of wav2vec2, can be used to effectively infer various characteristics from vocal bursts.
- Second, we conduct various experiments using *wav2vec2phoneme* [28], which, instead of being trained on raw audio data, makes use of a phoneme vocabulary. As vocal bursts can, similar to speech, also be seen as a series of phonemes, we examine if using that intermediate representation additionally can improve automatic affect analysis pipelines.

We evaluate both architectures in a multi-task setting using HUME-VB dataset [9], a dataset of vocal bursts annotated regarding 5 different tasks. To further our understanding of how to build a successful system for analysing affective voice breaks, we subject our approach to an ablation study. Therefore, we investigate how different aspects of our training pipeline contribute to the performance of the analysis pipeline.

2 Related Work

Multi-task learning for vocal bursts recently became a popular research topic, partly because it was addressed in multiple conference challenges. E.g., in the *ExVo2022* challenge, participants were asked to predict the expression of 10 emotions along with the age and native country of the speaker at the same time [7]. [25] approached the task by experimenting with various encoder frontends as well as handcrafted features. They found that using the HuBERT model [14], which is closely related to the wav2vec architecture and training approach, as a backbone yielded the best performance. Purohit et al. [22] compared various embeddings that have been either trained using self-supervision or directly in a task-dependent manner. They found that overall, the self-supervised embeddings are outperforming the task-dependent ones, which, supports the choice of data2vec for our experiments. Anuchitanukul and Specia [1] also rely on wav2vec and HuBERT backbones to extract embeddings for their multi-task training system. They further utilise an adversarial training approach to disentangle the input representations into shared, task-specific ones. Their experiments showed that the wav2vec-based model performs best, but using ensemble techniques to combine multiple variations of their wav2vec and HuBERT models can achieve even higher performance.

Another challenge that addressed similar tasks was the *ACII-VB* challenge [6]. Here, participants of the challenge had to assess the type, valence/arousal, intensity of the emotion type, and the emotional type specific to certain countries. Again, the majority of contributions made use of either HuBERT or wav2vec2 models [27, 19, 2, 26, 15, 3, 11].

All those works indicate that self-supervision in general and wav2vec specifically are building a good foundation for the task at hand and confirming our choice of data2vec and wav2vec2phoneme as the successor of wav2vec.

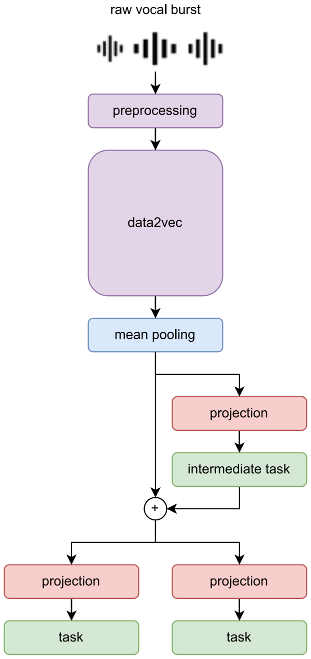

Fig. 1. Overview of the data2vec architecture.

3 Dataset

For our experiments, we utilised the HUME-VB dataset [9], which consists of emotional non-linguistic vocal bursts. Overall, there are roughly 37 hours of recorded audio clips at a 16 kHz sampling rate spread over 59,201 files. The data has been recorded in 4 countries (China, South Africa, U.S. and Venezuela) representing different cultures, totalling in 1702 speakers with ages from 20 to 39 years. Each vocal burst was rated on average by 85 raters from the same country as the vocal burst's origin. For our experiments, we use the *Train* and *Validation* splits provided by the authors of dataset. The corpus provides multiple annotations for each sample that we use to train and evaluate our multi-task learning architecture:

- *High* refers to the intensity of 10 different emotions: *Awe, Excitement, Amusement, Awkwardness, Fear, Horror, Distress, Triumph, Sadness, Surprise.*
- *Country* labels inform about the origin of the person a vocal burst was recorded from.
- *Culture* labels provide the country-specific annotations of the 10 different emotions. As such, for each country, a 10 different emotion gold standard is given that was derived from annotators of the same country, resulting in $4 \cdot 10 = 40$ dimensions.
- *Two* refers to the two-dimensional continuous emotion representation of the samples, i.e., valence/arousal labels.

- *Type* annotations are given to divide the samples into 8 different expression classes, i.e., *Gasp, Laugh, Cry, Scream, Grunt, Groan, Pant* and *Other*.

High, *Culture*, and *Two* are multi-label regressions with each label ranging from 0 to 1. *Type* and *Country* are classifications, having 8 respectively 4 classes.

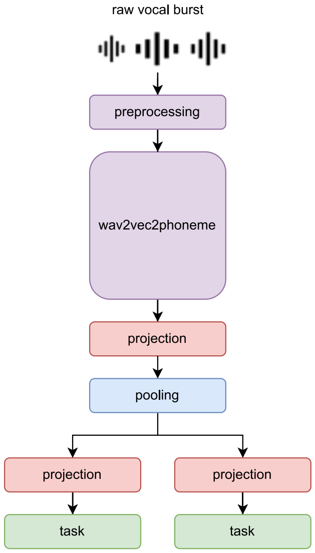

Fig. 2. Overview of the wav2vec2phoneme architecture.

4 Methodology

data2vec Since vocal bursts are not "speech made out of words", but rather "speech made out of vocalised emotions", the not fine-tuned version of data2vec is used for our first experiments. Instead of handcrafted features which are often specifically engineered for spoken language, data2vec uses dataset-specific representations learnt in an unsupervised way.

To follow data2vec's modality-agnostic approach (i.e., it can be applied to either audio, video, or textual data) no or as few assumptions as possible are posed for the downstream supervised fine-tuning. To use all the provided labels of task $i = 1, \ldots, n$, multi-task learning is applied in a self-learning way [16], which approximates optimal task weights by the learning task uncertainty σ_i:

$$L = \sum_{i}^{n} \frac{1}{\sigma_i^2} L_i + \sum_{i}^{n} \log \sigma_i \tag{1}$$

In figure 1, the network architecture is depicted. The raw audio is fed into the pretrained data2vec model, including its preprocessor. Variable sequence lengths

Table 1. Results for the experiment sets on *Intermediate Tasks* and *Task Loss*.

	High	Culture	Two	Type	Country	All
Method	CCC	CCC	CCC	UAR	UAR	HMean
Baseline	.564	.436	**.499**	.412	–	.472
2/3	.639	.625	.252	.562	.631	.476
1/4	.628	.614	.245	.542	.603	.463
0/5	.650	.636	.254	.564	.633	.481
MSE	.624	.598	.244	.553	.591	.460
MAE	.640	.573	.251	.575	.655	.474
-High	–	.608	.244	.543	.600	.432
-Culture	.625	–	.255	.539	.599	.442
-Two	.651	.637	–	.574	.645	**.625**
-Type	.649	.634	.265	–	**.665**	.476
-Country	**.659**	**.642**	.266	**.577**	–	.467

are zero-padded to the longest seen sequence. Because of the attention mask, the extracted features vary in length. Therefore, these are mean-pooled and fed into downstream projection layers. To investigate the question of whether knowing certain tasks before predicting other tasks is helpful, the intermediate tasks are separated and their prediction is fed along with the extracted features to the remaining tasks.

The projection layers reduce the output dimension of data2vec (768 base, 1024 large) to 256, apply GELU [12], and further reduce the dimension down to the five task's required dimensions (high: 10, culture: 40, two: 2, type: 8, country: 4). The tasks' layers then apply softmax for classification and compute the loss via cross-entropy using inversely proportional class weights. For regression, we apply sigmoid and compute the loss using the concordance correlation coefficient. The losses are then linearly combined through the learnt optimal uncertainty-based weights. This architecture serves as a starting point for the ablation study, where the design is iteratively improved.

wav2vec2phoneme Using the insights drawn from our experiments with data2vec (see section 5), the architecture was slightly altered for wav2vec2phoneme, as seen in figure 2 - the intermediate tasks were removed. With the projection of wav2vec2's 1024-dimensional feature vector down to a 392-dimensional phoneme vocabulary, the resulting sequence had to be aggregated to a fixed length for downstream tasks.

Training The whole architecture has a size of 360MB for the base and 1.2GB for the large version of data2vec, and 1.4GB for wav2vec2phoneme. They are trained in a two-stage matter. First, we freeze the net and train only the tasks with their projection layers. Second, we unfreeze the net and fine-tune the whole architecture. Both stages are trained for a maximum of 30 epochs with early stopping using a patience of 2 on the validation split to avoid overfitting on the training data. We use a batch size of 32 for the base version of data2vec, 24 for the

Table 2. Results for the experiment sets on *Weighting*, *Fine-Tuning Splits* and *Network Size*.

	High	Culture	Type	Country	All
Method	CCC	CCC	UAR	UAR	HMean
-CW	.645	.631	.276	.632	.480
+SW	.643	.629	.319	.644	.511
-SM	.654	.639	**.584**	**.657**	**.632**
B10m	.656	**.642**	.577	.655	.631
B100h	.639	.624	.552	.634	.610
B960h	**.658**	**.642**	.567	.647	.626
L	.635	.622	.554	.603	.602
L10m	.641	.624	.555	.607	.605
L100h	.497	.601	.540	.578	.551
L960h	.613	.617	.540	.539	.575

large version, and 16 for wav2vec2phoneme. As optimiser, we used AdamW [18]. For the first stage, the optimiser was initialised with default parameters and a learning rate of $1 \cdot 10^{-3}$. For the second stage, the learning rate is set to $4 \cdot 10^{-5}$ and follows a cosine schedule with a warmup of 1 epoch.

The training runs for our experiments were done on a single Nvidia A40 GPU and took 5 to 6.5 hours for the base version of data2vec and 5.5 to 8.5 hours for the large version. For wav2vec2phoneme, training took between 10 to 13 hours. Replacing the optimiser with ASAM [17], which we did for some experiments as described in Section 5, increased the training times by a factor of 2.5 to 3.5.

5 Experiments & Results

data2vec In our experiments, we varied certain details of the data2vec network architecture as listed below. The results for the validation set are shown in tables 1 and 2, additionally to the official baseline results that were published by the authors of the dataset [6]. Note that the single experiments were conducted iteratively, carrying the best configuration of the previous set of experiments over to the next set of experiments. Since we use a multi-task approach, comparing different approaches can be difficult if not all tasks perform better or worse than in the experiment being compared to. Therefore, we calculate the harmonic mean of the tasks' metrics to provide an overall comparison. The results of the first set of experiments are listed in table 1.

Intermediate Tasks Following the dataset's motivation that vocal bursts depend on the country of origin to assess the conveyed emotions like a rater of the same origin would, we tried three different variants: (i) *Country* and *Type* as intermediate tasks before predicting the remaining three (*2/3*). (ii) Only *Country* as an intermediate task (*1/4*), assuming the type depends less on the country. (iii) No intermediate tasks, i.e., predicting all tasks simultaneously (*0/5*). (iii) performed best.

Table 3. Results for the experiment sets on *Aggregation*, *Loss adjustments*, *Features*, *Combination*, *Optimiser*, and *Loss Revision*.

	High	Culture	Type	Country	All
Method	CCC	CCC	UAR	UAR	HMean
LSTM-1	.549	.558	.468	.253	.412
Mean-1	.665	.652	.574	.660	.635
Count	.053	.055	.125	.250	.082
Regularisation	.041	.039	.125	.250	.064
Weighting	.267	.160	.166	.250	.200
LSTM-2	.673	.656	.580	.562	.614
Mean-2	.327	.642	.581	**.703**	.516
Separate	.670	.638	.595	.688	.646
Concat-1	.675	.650	.596	.689	**.650**
Mean-3	.664	.649	.583	.696	.645
ASAM-1	**.683**	**.667**	**.602**	.624	.642
ASAM-2	.673	.645	.588	.699	.649
DRUW-1	.570	.643	.584	.692	.619
DRUW-2	.680	**.667**	.593	.561	.621
DRUW-3	.666	.650	.562	.675	.635

Task Loss To investigate why task *Two* performed so poorly, we experimented with replacing the CCC loss by (i) mean squared error (*MSE*) and (ii) mean absolute error (*MAE*). Further, to investigate if there is an issue with a single task, we removed each task from training once (iii) - (vii) (e.g. *-Task*, *-High*). (v), i.e., removing task *Two*, turned out to be the best option. As removing that task improves the harmonic mean for the other tasks by large margins, we decided to drop it for the following experiments. Results of the second set of experiments are reported in table 2.

Weighting Since the cross-entropy losses in tasks *Country* and *Type* have inverse proportional class weights while the other tasks do not, we experimented with (i) removing these class weights from the training (*-CW*) so that every sample's tasks are unweighted, (ii) adding sample weights inversely proportional to *Country* (*+SW*), thereby weighting all tasks, and (iii) keeping the class weights but removing the sigmoid activation from the last layer (*-SM*) and clamping the linear output to $[0, 1]$ in the regression tasks *High* and *Culture* instead. We could marginally boost the network's performance by applying (iii).

Fine-Tuning Splits To investigate if fine-tuning the self-supervised learnt audio representations using labeled speech improves the pipeline's performance, we experimented with different versions of the base network. Each version was fine-tuned on word labels after pretraining using connectionist temporal classification (CTC) loss on a different amount of LibriSpeech [20] data: (i) 10 minutes (*B10m*), (ii) 100 hours (*B100h*), (iii) 960 hours (*B960h*). Version (i) performed best overall.

Network Size At last, we replaced the base network with (i) the large version (*L*) and (ii) the respective fine-tuning splits (*L10m*, *L100h*, *L960h*). None of the large versions could outperform the former base version experiments. For the tasks *High* and *Culture*, experiment *B960h* worked best on the validation set. For the tasks *Type* and *Country*, experiment -*SM* performed best on the validation set.

wav2vec2phoneme The insights on data2vec so far are:

- No intermediate tasks are needed.
- Removing task *Two* greatly benefits the other tasks.
- Removing the sigmoid activation in the last layer improves the performance.
- Using larger versions of the model does not improve performance.
- No fine-tuning on word labels is needed.

With those findings as a starting point, we investigated if using another vocabulary than words made out of letters to describe vocal bursts improves the performance. For this, we use wav2vec2phoneme, which transcribes audio using a phoneme vocabulary. Intuitively, it is conceivable that this allows for a better vocal burst description. The results of the following experiments are reported in table 3.

Aggregation The next set of experiments targets the aggregation of the varying-length sequence of the transcribed 392-dimensional phoneme vocabulary. We did this by (i) using a bi-directional two-layer LSTM [13] with a feature size of 768 (*LSTM-1*), the same feature size as data2vec's base version. Further, we (ii) applied mean pooling (*Mean-1*), and (iii) simply counted the occurrences of each phoneme (*Count*). (ii) performed best. The experiments also showed that experiment (iii), by aiming to learn constant predictions, artificially induced the loss to decrease – the net converges towards chance level.

Loss adjustments (i) To avoid the regularisational term from becoming negative, a lower bound was introduced to the uncertainties in equation 1 (*Regularisation*):

$$L = \sum_{i}^{n} \frac{1}{\sigma_i^2} L_i + \sum_{i}^{n} \log(1 + \sigma_i) \tag{2}$$

(ii) additionally to the above adjustment, the loss is weighted by the sequence length (*Weighting*). Both experiments did not improve the results. As such, the former best configuration remains the starting point for the next set of experiments.

Features To investigate if the transcription to human-readable 392 different phonemes or the preceding 1024-dimensional layer should be passed on, the latter is (i) fed to a bi-directional two-layer LSTM with a hidden size of 1024 (*LSTM-2*), and (ii) mean pooled (*Mean-2*). While for both the harmonic mean decreased compared to the former best, the LSTM improved the regression, while mean pooling improves the classification tasks.

Table 4. Transfer of best setup *Concat-1* to other wav2vec2-architectures. Upper half ex- and lower half includes task *Two* while applying the same configuration.

	High	Culture	Two	Type	Country	All
Method	CCC	CCC	CCC	UAR	UAR	HMean
w2v2-B	.642	.630	-	.563	.661	.622
w2v2-B960h	.293	.285	-	.125	.250	.211
w2v2-L	.651	.642	-	.579	**.711**	.642
w2v2-L960h	.366	.636	-	.584	.636	.527
Concat-1	**.675**	**.650**	-	**.596**	.689	**.650**
w2v2-B	.652	.636	.137	.581	.669	.367
w2v2-B960h	.198	.273	.232	.125	.250	.200
w2v2-L	.665	**.650**	**.263**	.588	.706	.502
w2v2-L960h	.138	.163	.137	.493	.592	.205
Concat-2	.669	.563	.255	.593	.695	.485

Combination In these experiments, we tried to find a combination of LSTM and mean pooling to improve the harmonic mean, thereby overall performance, partly at the cost of a higher dimension. In order to do so, we (i) separately used the LSTM features for regression and the mean-pooled features for classification (*Separate)*. In (ii), both were concatenated (*Concat-1*), and last, in (iii), they were averaged to keep the dimension the same (*Mean-3*). (ii) performed best.

Optimiser To investigate if the loss issues are caused by the optimiser, it is extended by applying adaptive sharpness-aware minimisation (ASAM) [17]. For parameter ρ, we experimented with (i) $\rho = .5$ (*ASAM-1*) and (ii) $\rho = .05$ (*ASAM-2*). Both slightly reduced the harmonic mean – they decrease some tasks' performance and increase the others'. As such, no overall improvement could be observed.

Loss revision By applying dynamic restrained uncertainty weighting (DRUW) [15], we tried to tackle the loss issues through further adjustments to equation 2:

$$\begin{aligned} L = &\sum_{i}^{n} (\frac{1}{\sigma_i^2} + \lambda_i) L_i + \sum_{i}^{n} \log{(1 + \log \sigma_i^2)} + \\ &|\phi - \sum_{i}^{n} |\log \sigma_i|| \end{aligned} \tag{3}$$

with the dynamic weights λ_i being:

$$\lambda_i = n \frac{\exp{(\tau \frac{L_{i,t-1}}{L_{i,t-2}})}}{\sum_i^n \exp{(\tau \frac{L_{i,t-1}}{L_{i,t-2}})}} \tag{4}$$

Using the configuration parameters proposed by [15], i.e., $\tau = 1$ for temperature and $\phi = 1$ as regularisation constant, DRUW is applied to *Concat-1* and both *ASAM-1/2* experiments, resulting in experiments (i) - (iii) (*DRUW-1/2/3*). No overall performance improvement was observed. However, (ii) managed to maintain the performance for task *Culture*.

Transfer to wav2vec2 To investigate if phonemes are really more suited than words, the experiments that worked best were transferred to the base and large versions of wav2vec2 (*w2v2-B*, *w2v2-L*). The same modifications were applied to the respective fine-tuned models (*w2v2-B960h*, *w2v2-L960h*). Additionally, all of those four experiments were run with the inclusion of task *Two* in order to validate if the aforementioned negative interferences of that task still occur for the wav2vec2 model. Results are shown in table 4. *Concat-1* remains the overall best choice.

6 Discussion

Recapitulating the conducted experiments and considering the results, the following insights can be drawn:

Intermediate Tasks Determining the country of origin before assessing the conveyed emotions like a rater of that country would may be beneficial to a human rater in order to detect and adjust emotional biases. However, it is disadvantageous for our pipeline – the extracted features already encompass these biases and need not be handcrafted into.

Task Loss Revising the task losses, adding further regularisational terms and applying a sharpness-aware optimiser did not improve the poor performance on task *Two*. Since the baseline shows double the performance here, neither of the self-supervised learnt audio or word-/phoneme-based representations are suited for estimating valence and arousal in these architectures. Therefore, the model learns to predict a rather constant output for *Two*. As such, the uncertainty in this task is artificially reduced, minimising the penalising uncertainty term, but also maximising the task weight in the computation of the MTL loss. The weight can become so large that it substantially degrades the other tasks' performance. Our experiments have shown that excluding task *Two* greatly benefits the assessment of the remaining tasks.

Weighting These experiments investigated different weighting techniques to counter imbalance in the training data. Removing the inversely proportional class weights in the calculation of the cross-entropy losses greatly reduces the performance in *Type*. Extending the cross-entropy weights over the whole sample to inversely proportional intra-batch weights depending on *Country* alleviates this only slightly, despite having the annotations made from raters of the same country as the vocal burst's utterer. Inversely proportional class weights both in *Type* and *Country* to counter class imbalance combined with the removal of sigmoid activation, is the best approach, as it increases the net's sensitivity to samples close to the boundaries of the value ranges.

Fine-Tuning Splits Here, we examined the assumption that vocal bursts are not a word-based "language". As such, we observed a decline in performance when fine-tuning the pretrained model using CTC on (English) word-based labels. Consequently, this decline is only slight when using 10 minutes of fine-tuning data, but more so for 100 hours. More than 50% performance degradations are visible when using all 960 hours of data for fine-tuning. Therefore, sticking with the pretrained network, or using only sensible fine-tuning, i.e. phonemes, seems to be the best option.

Feature summarisation When processing a sequence of phonemes, i.e. listening, one intuitively expects a recurrent neural network (RNN) to be best suited, since those type of architectures consider chronology. However, *LSTM-1* and *Mean-1* showed that simply computing the distribution of phonemes, regardless of the time of occurence, outperforms a RNN. Since *Count* performed poorly even with loss adjustments, it does not matter how often (or how long) different phonemes were uttered - only the proportion is relevant. Interestingly, when dropping the human-readable phoneme transcription and directly using wav2vec2phoneme's features, *LSTM-2* and *Mean-2* show an equal overall performance. However, each of them were better in either both regression or both classification tasks. Therefore, providing both summarisations by concatenating them (*Concat-1*) leads to the best overall performance.

Feature representations In these last experiments, we evaluated if our initial assumption, e.g., a phoneme-based feature representation, indeed has the ability to outperform the more traditional wav2vec2 approach. Therefore, we used the configuration that worked best in the preceding experiments and applied them to different versions of wav2vec2. The observation that the performance of neither data2vec nor wav2vec2phoneme could be matched supports our claim that using a phoneme-based feature representation can be a valid choice for the task at hand.

7 Conclusion

In this work, we have shown that a single network for multi-task affective vocal burst assessment is a valid choice. Per-task ensembling and large structures are not necessarily needed. Task weighting can be done automatically via task uncertainty approximation. Although being pretrained on English-only speech in a self-supervised manner, data2vec is able to assess vocal bursts originating out of different (non-English) countries after fine-tuning it for fiveish hours. Furthermore, by substituting the data2vec architecture with wav2vec2phoneme, a larger and phoneme-based net, we could further boost the pipeline's performance, while only doubling the required time for training. If time and lightweightness are of essence, data2vec seems to be the better choice. If overall performance is the most important criterion, wav2vec2phoneme fits best. Applying a sharpness-aware optimiser can yield even better results for specific subtasks, but comes with the cost of a decreased overall performance. By comparing our best configuration to a word-based wav2vec2, we can conclude that phonemes are better suited for the assessment of affective vocal bursts than words.

Acknowledgements This work was partially funded by the KodiLL project (FBM2020, Stiftung Innovation in der Hochschullehre), project TherapAI (DFG, German Research Foundation, grant number 493169211) and project Panorama (DFG, German Research Foundation, grant number 442607480).

References

1. Anuchitanukul, A., Specia, L.: Burst2vec: An adversarial multi-task approach for predicting emotion, age, and origin from vocal bursts. arXiv preprint arXiv:2206.12469 (2022)
2. Atmaja, B.T., Sasou, A.: Predicting affective vocal bursts with finetuned wav2vec 2.0. arXiv preprint arXiv:2209.13146 (2022)
3. Atmaja, B.T., Sasou, A., et al.: Jointly predicting emotion, age, and country using pre-trained acoustic embedding. arXiv preprint arXiv:2207.10333 (2022)
4. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: Data2vec: A general framework for self-supervised learning in speech, vision and language. arXiv preprint arXiv:2202.03555 (2022)
5. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems **33**, 12449–12460 (2020)
6. Baird, A., Tzirakis, P., Batliner, A., Schuller, B., Keltner, D., Cowen, A.: The acii 2022 affective vocal bursts workshop and competition: Understanding a critically understudied modality of emotional expression. arXiv preprint arXiv:2207.03572v1 (2022). https://doi.org/10.48550/arXiv.2207.03572
7. Baird, A., Tzirakis, P., Gidel, G., Jiralerspong, M., Muller, E.B., Mathewson, K., Schuller, B., Cambria, E., Keltner, D., Cowen, A.: The icml 2022 expressive vocalizations workshop and competition: Recognizing, generating, and personalizing vocal bursts. arXiv preprint arXiv:2205.01780v3 (2022). https://doi.org/10.48550/ARXIV.2205.01780
8. Cordaro, D.T., Keltner, D., Tshering, S., Wangchuk, D., Flynn, L.M.: The voice conveys emotion in ten globalized cultures and one remote village in bhutan. Emotion **16**(1), 117 (2016)
9. Cowen, A., Baird, A., Tzirakis, P., Opara, M., Kim, L., Brooks, J., Metrick, J.: The hume vocal burst competition dataset (H-VB) — raw data [exvo: updated 02.28.22] [data set]. Zenodo (2022). https://doi.org/https://doi.org/10.5281/zenodo.6308780
10. Cowen, A.S., Elfenbein, H.A., Laukka, P., Keltner, D.: Mapping 24 emotions conveyed by brief human vocalization. American Psychologist **74**(6), 698 (2019)
11. Hallmen, T., Mertes, S., Schiller, D., André, E.: An efficient multitask learning architecture for affective vocal burst analysis (2022)
12. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

14. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3451–3460 (2021)
15. Karas, V., Triantafyllopoulos, A., Song, M., Schuller, B.W.: Self-supervised attention networks and uncertainty loss weighting for multi-task emotion recognition on vocal bursts. arXiv preprint arXiv:2209.07384 (2022)
16. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7482–7491 (2018)
17. Kwon, J., Kim, J., Park, H., Choi, I.K.: Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In: International Conference on Machine Learning. pp. 5905–5914. PMLR (2021)
18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
19. Nguyen, D.K., Pant, S., Ho, N.H., Lee, G.S., Kim, S.H., Yang, H.J.: Fine-tuning wav2vec for vocal-burst emotion recognition. arXiv preprint arXiv:2210.00263 (2022)
20. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: LibriSpeech: an asr corpus based on public domain audio books. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 5206–5210. IEEE (2015)
21. Phutela, D.: The importance of non-verbal communication. IUP Journal of Soft Skills **9**(4), 43 (2015)
22. Purohit, T., Mahmoud, I.B., Vlasenko, B., Doss, M.M.: Comparing supervised and self-supervised embedding for exvo multi-task learning track. arXiv preprint arXiv:2206.11968 (2022)
23. Scherer, K.R.: Expression of emotion in voice and music. Journal of voice **9**(3), 235–248 (1995)
24. Schröder, M.: Experimental study of affect bursts. Speech communication **40**(1-2), 99–116 (2003)
25. Sharma, R., Vuong, T., Lindsey, M., Dhamyal, H., Singh, R., Raj, B.: Self-supervision and learnable strfs for age, emotion, and country prediction. arXiv preprint arXiv:2206.12568 (2022)
26. Syed, M.S.S., Syed, Z.S., Syed, A.: Classification of vocal bursts for acii 2022 a-vb-type competition using convolutional network networks and deep acoustic embeddings. arXiv preprint arXiv:2209.14842 (2022)
27. Trinh, D.L., Vo, M.C., Kim, S.H., Yang, H.J., Lee, G.S.: Self-relation attention and temporal awareness for emotion recognition via vocal burst. Sensors **23**(1), 200 (2022)
28. Xu, Q., Baevski, A., Auli, M.: Simple and effective zero-shot cross-lingual phoneme recognition. arXiv preprint arXiv:2109.11680 (2021)