ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096552

EVALUATING VARIANTS OF WAV2VEC 2.0 ON AFFECTIVE VOCAL BURST TASKS

Bagus Tris Atmaja and Akira Sasou

National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan

ABSTRACT

The search for emotional biomarkers within the human voice is a challenging research area. Previous studies focused on predicting affective state from speech; this study explores various tasks on affective vocal bursts. Borrowing the success of self-supervised learning in automatic speech recognition, we extracted acoustic embedding using variants of wav2vec 2.0 for four affective vocal bursts tasks: High, Two, Culture, and Type. Using a similar architecture for all tasks, the evaluation of acoustic embeddings reveals the potential use of wav2vec 2.0 variants over the conventional acoustic features in affective vocal bursts tasks. We evaluated both conventional acoustic features and these acoustic embeddings on the different number of twenty seeds evaluation and reported the maximum and average scores with their standard deviations in the validation set. Three high scores from these validations for all tasks assist the generation of predictions for the test set. We compared the test scores with previous studies and obtained remarkable improvements.

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

1. INTRODUCTION

Vocal bursts may have richer affective information than speech. However, speech emotion recognition, rather than vocal bursts, is currently gaining more attention from researchers due to its potential implementation and the availability of the datasets. Instead of a speech, affective information may also lay on short vocal bursts (i.e., crying when sad). In contrast to speech emotion recognition which may have difficulties in distinguishing between emotions, different vocal bursts may reflect different affective states more distinctly. For instance, the emotion of sadness and fear in speech are similar since both of them are expressed by higher pitch [1]. A specific pattern of crying may indicate sadness, whereas laughter produces happiness. Given these benefits, analyzing emotions from humans' vocal bursts may improve our understanding of human emotions.

Humans communicate through verbal and vocal communication, including communicating emotions [2]. Verbal communication includes chosen words in the speech that has semantic meaning. Vocal communication includes prosody: intonation, intensity, and rhythm. A study in cognitive brain research suggests that brain activity in emotional prosody detection is higher than in verbal detection [3]. Further studies by Tian et al. [4, 5, 6] suggest that adding non-verbal vocalization information to acoustic features improved the recognition rate of speech emotion recognition in IEMOCAP [7] and AVEC2012 [8] datasets.

Vocal bursts – a non-verbal communication – constitute a potential source of information for emotion [9]. A study by Cowen et al. [10] has found that vocal bursts are rich in emotional information that can be conceptualized into 24 emotion categories. A previous study by Scherer [11] has proposed a model of vocal communication as Brunswik's lens model from expression (encoding) to perception (representation). There is no exact number of emotion categories emerging from this study. The authors mentioned eight examples of emotion categories with ranges of importance for their design features delimitation (e.g., intensity). However, the research on affective vocal bursts since then has been limited by the lack of available datasets.

One way to speed up research on affective vocal bursts is to hold workshops and competitions in that area. In [12, 13, 14], the organizers provided datasets and baseline methods to challenge participants to explore the dataset and surpass the baseline scores. This study, in particular, is presented to report the evaluation of wav2vec 2.0 variants for the ACII 2022 Affective Vocal Bursts Workshop and Competition [14]. There are four tasks in the competition: three regression problems and a classification problem. The regression problems are intended for measuring either the intensity of ten emotion categories or valence (positive-negative of emotion) and arousal (low-high of emotion). The classification problem is for predicting the type of vocal bursts (e.g., laughter). We approached all four tasks using a similar method while observing the effect of varying the acoustic embeddings.

There are two research questions to be solved in this research. First, we evaluated the effectiveness of seven wav2vec 2.0 variants for four affective vocal bursts tasks, including variants of wav2vec 2.0 pre-trained on an affective speech dataset. Second, we combined wav2vec 2.0 embeddings with valence, arousal, and dominance (vad) predictions in a variant to evaluate their benefits. The architecture of deep learning

E-mail: b-atmaja@aist.go.jp; funded by NEDO Japan (project JPNP20006)

models was similar for all tasks except in the output layer and the loss function (which depends on the task).

Those contributions can be compared further with previous speech emotion recognition studies. Although a direct benchmark cannot be made (due to differences in datasets), the performance scores could be observed to gain insights into the two worlds (speech vs. vocalization). For instance, in this study, we reported average concordance correlation coefficients (CCC) of valence and arousal predictions of 0.629 from task Two of Hume A-VB vocal burst data. The previous studies have shown CCC scores for valence and arousal of 0.691 [15] and 0.431 [16] for MSP-Podcast data, 0.569 [15] and 0.566 [16] for IEMOCAP data [7].

We built our method based on the previous studies [14, 17, 18] with distinctive modifications. These modifications include an evaluation of different acoustic embeddings and a loss function to minimize concordance correlation coefficient loss for regression problems. Both modifications are intended to improve the previous performance by using specialized finetuned models, including the one finetuned on the affective speech dataset [19], and matching the loss function to the evaluation metric.

2. DATASET

This study employed a single Hume-VB dataset [20] for all four affective vocal bursts tasks. The dataset consists of 59201 samples in more than 36 hours of recordings. These samples were recorded from 1702 speakers across four countries (China, South Africa, the US, and Venezuela). Each sample was labeled with an integer scale from 1 to 100 (scaled to [0,1] during the experiments) for ten expressed emotions, scores of valence and arousal (in [0, 1]), and the type of vocal bursts. The ten emotion categories are amusement, awe, awkwardness, distress, excitement, fear, horror, sadness, surprise, and triumph. The types of vocal bursts are Cry, Gasp, Groan, Grunt, Laugh, Pant, Scream, and Other. The original raw audio data were sampled at 48 kHz and resampled to 16 kHz in the experiments of acoustic embedding extraction. The organizer partitioned the data into training, validation, and test sets for each task. Training and validation sets are open with their labels, while the test set is closed without labels. The participants need to send the test set predictions to obtain the performance scores on the test set.

3. TASKS

Four tasks provided in The ACII 2022 Affective Vocal Bursts Workshop & Competition [14] are summarized as follows:

1. The "High" task is to predict the intensity of 10 aforementioned emotions,
2. The "Two" task is to predict the degrees of valence and arousal for the given vocal bursts,
3. The "Culture" task is to predict the intensity of 40 culture-specific emotions (10 aforementioned emotions from each culture) as a multioutput regression problem,
4. The "Type" task is to predict the type of given vocal bursts. Eight types of vocal bursts in the fourth task are gasp, cry, laugh, scream, groan, grunt, pant, and other.

The first three tasks are regression problems evaluated with concordance correlation coefficients (CCC); the fourth task is a classification problem evaluated with unweighted average recall (UAR), also known as unweighted accuracy (UA). These evaluation metrics are obtained by submitting the predictions for the hidden test set to the organizers.

4. BASELINE FEATURES

As the baseline features, we experimented with ComParE [21] and eGeMAPS [22] feature sets. These acoustic features are utterance aggregations by calculating high-level statistical functions from the low-level descriptors (extracted per frame). We re-trained these acoustic features on the modified classifiers (see Section 6) to obtain the validation scores. The classifiers are modified from the previous studies [14, 18].

5. WAV2VEC 2.0 AND ITS VARIANTS

The wav2vec 2.0 model is the improved version of wav2vec (version 1.0), unsupervised pre-training for speech recognition [23]. Instead of unsupervised learning, wav2vec 2.0 uses a self-supervised approach for generating speech representations [24]. Given the success of wav2vec 2.0 in speech emotion recognition [25], we evaluated seven variants of wav2vec 2.0 for four affective vocal bursts tasks. These variants are

1. wav2vec 2.0 base model (w2v2-base) [24],
2. wav2vec 2.0 large model (w2v2-large) [24],
3. wav2vec 2.0 cross-lingual model (w2v2-xlsr) [26],
4. wav2vec 2.0 large and robust model (w2v2-lr) [27],
5. wav2vec 2.0 large and robust model fine-tuned on 300h Switchboard dataset (w2v2-lr-300) [27],
6. wav2vec 2.0 large and robust model fine-tuned on 960h Librispeech dataset (w2v2-lr-960) [27], and
7. wav2vec 2.0 large and robust model fine-tuned on MSP-Podcast dataset (w2v2-r-er and w2v2-r-vad) [15, 28].

For the last model [28] built from the MSP-Podcast dataset [19], we extracted two representations of each given vocal burst audio file. The first is 'w2v2-r-er' with 1024-dims of the hidden states. The second is 'w2v2-r-vad' with a concatenation of the hidden states and logits (1027-dims). These acoustic embeddings were then fed to the regression model for the first three tasks and to the classification model for the last task four ("Type"). The model IDs (in https://huggingface.co) are given in Table 1.

Table 1. Name of wav2vec 2.0 variants and their Hugging Face models (from facebook and audeering); the last two rows (audeering) used the same model but different sizes.

Name	Hugging Face Model ID
w2v2-base	wav2vec2-base-960h
w2v2-large	wav2vec2-large-960h
w2v2-xlsr	wav2vec2-large-xlsr-53
w2v2-lr	wav2vec2-large-robust
w2v2-lr-300	wav2vec2-large-robust-ft-swbd-300h
w2v2-lr-960	wav2vec2-large-robust-ft-libri-960h
w2v2-r-er	wav2vec2-large-robust-12-ft-emotion-msp-dim
w2v2-r-vad	wav2vec2-large-robust-12-ft-emotion-msp-dim

6. CLASSIFIERS

This study employed a three-layer fully-connected network as the classifier. The last fully-connected (FC) layer is then coupled with an output layer. The number of nodes for each FC layer is 128, 64, and 32, respectively. Each layer is connected to layer normalization [29] and leaky rectified linear unit (LeakyReLU) activation function. The number of nodes at 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. These classifiers are modifications of the previous studies [14, 18].

Fig. 1 depicts the architecture of the fully-connected network. The architecture and hyperparameters are the same for all tasks. The learning rate is set to 0.0005, weight decay is set to 0.01, the optimizer is AdamW [30], the batch size is 8, and the maximum number of epochs is 100. For the Type task, an early stopping criterion was set to a patience of 10 epochs with delta patience of 0.01. Other tasks are trained without an early stopping criterion of patience.

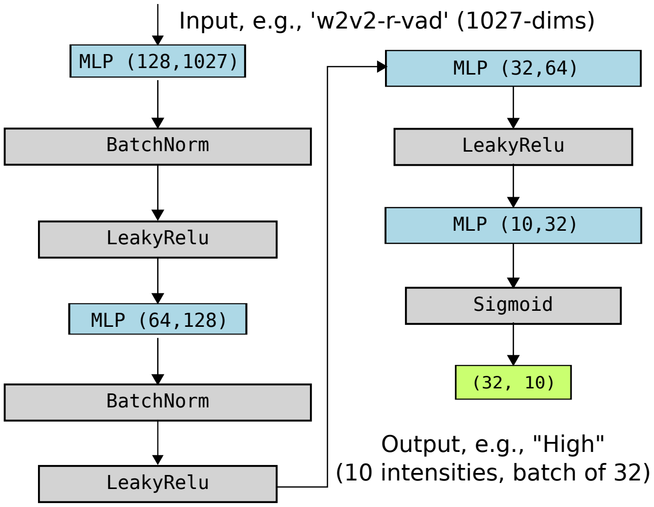

Fig. 1. Architecture of MLP networks for all tasks, the example is for ”High” classification task with 32 batch size for 10 outputs. Brackets show (output, input) nodes.

The Type task minimized the cross-entropy loss function. The UAR score for the Type task is the normalized score in a range of [0, 1]. Three tasks: High, Two, and Culture, minimized CCC loss (since the metric is CCC in a range of [-1, 1]). This loss function is the main difference between this study from the previous studies [14, 18] in terms of the model's architecture. CCC loss (CCCL) is formulated

$$CCC = \frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{1}$$

$$CCCL = 1 - CCC, \tag{2}$$

where μ_x and μ_y are the means of the predicted and ground truth values, respectively. σ_x and σ_y are the standard deviations of the predicted and ground truth values, respectively, and ρ_{xy} is the Pearson correlation between the predicted and ground truth values. CCC loss is arguably more effective than other error-based loss functions, especially when the metric is CCC [31]. CCC is more effective than other correlation functions since it not only accounts for the relation of the two variables but also for the exact difference in values.

7. RESULTS AND DISCUSSION

7.1. Validation benchmark

Since the labels of the test set are hidden by the organizer of the A-VB 2022, the authors at first only experimented with the validation set to measure the performance of the evaluated methods (Table 2). Two baseline feature sets are compared with seven acoustic embeddings. Table 2 shows the maximum scores from 20 seed numbers evaluations. We validate 20 seeds to enable statistical analysis. The scores are CCC for High, Two, and Culture; and UAR for Type. Bolds indicate the highest score for each task with 'w2v2-r-er' for Two and Culture and 'w2v2-r-vad' for Culture and Type. Both embeddings obtained the same score on the Culture task.

It is clearly shown that SSL-based acoustic embeddings outperformed conventional acoustic features on the same classifiers. Most acoustic embeddings obtained better scores with smaller dimensions than ComParE dimensions (1024 vs. 6373). Finetuning wav2vec 2.0 on the affective speech dataset improved scores for the tasks; however, comparable scores were also obtained by the large and robust model without (w2v2-lr) and with finetuning on the non-affective speech dataset. The latter (w2v2-lr-300) obtained the third position after 'w2v2-r-er' and 'w2v2-r-vad' for the High, Two, and Culture tasks. It is interesting here to see that finetuning on a smaller dataset led to higher performance (w2v2-lr-300 vs. w2v2-lr-960), highlighting the importance of choosing the right data for finetuning. In this case, the 300h Switchboard dataset [32] may contain more affective information than the 960h Librispeech dataset [33].

On the average performance scores from 20 validations (Fig. 2), both the large-robust model (w2v2-lr) and large-

Table 2. Validation best scores from 20 seed numbers (CCC for High, Two, and Culture; UAR for Type)

Feature	Dims.	High	Two	Culture	Type
eGeMAPS	88	0.4896	0.4850	0.3880	0.3784
ComParE	6373	0.5336	0.5122	0.4254	0.3909
w2v2-base	768	0.4892	0.4446	0.3742	0.3323
w2v2-large	1024	0.4783	0.4406	0.3694	0.3414
w2v2-xlsr	1024	0.5718	0.5651	0.4341	0.4394
w2v2-lr	1024	0.6292	0.6100	0.4885	0.4734
w2v2-lr-300	1024	0.6317	0.6285	0.5119	0.4559
w2v2-lr-960	1024	0.5984	0.6138	0.4961	0.4656
w2v2-r-er	1024	0.6521	**0.6312**	**0.5138**	0.4822
w2v2-r-vad	1027	**0.6523**	0.6296	**0.5138**	**0.4829**

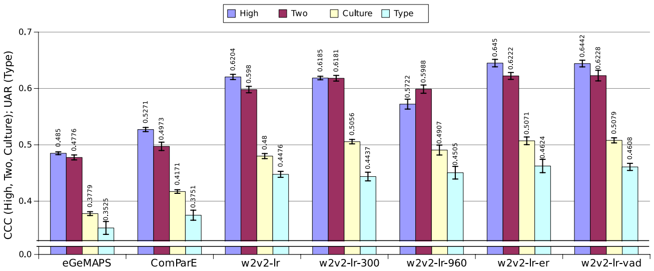

Fig. 2. Average scores from 20 seeds on the validation set for each embedding with their standard deviation bars

robust model finetuned on 300h Switchboard dataset (w2v3-lr-300) also show competitive results, among others. The 'w2v2-lr' obtained a high score for the High task while 'w2v2-lr-300' scored competitively on the Two task. Finetuning wav2vec 2.0 large-robust on affective dataset showed superiority on both average and maximum scores evaluations.

The authors also performed two sample tests (paired samples) to observe if there is any significant difference between the two means of w2v2-r-er and w2v2-r-vad. Since the latter only concatenates the former with values of arousal, dominance, and valence, the difference might not be significant (1027-dims vs. 1024-dims). The statistic test shows that the $p - values$ are large for all tasks, indicating that there is no significant difference between the two embeddings ($p - values$ = 0.50, 0.66, 0.51, and 0.80 for High, Two, Culture, and Type tasks, respectively). The degrees of valence, arousal, and dominance are important for obtaining categorical emotions; future research may find different ways of employing these values in the acoustic embeddings instead of a simple concatenation performed in this study.

7.2. Test benchmark

Table 3 shows the test scores of three submitted predictions (last three rows) along with the baseline test results. These three predictions are based on the highest scores on the validation set (Table 2). These test scores obtained by three acoustic embeddings are similar to that of validation scores; the only remarkable gap between test and validation is obtained by w2v2-lr-300 (Others) for the Culture task. The overall best score was obtained by w2v2-r-er; the fusion of w2v2-r-er with the logits (valence, arousal, dominance) did not improve the scores except for the High task. The w2v2-lr-300 gained comparable high scores to these scores on the High and Two tasks, although it was finetuned on the non-affective speech dataset.

Table 3. Performance scores on the test set (CCC for High, Two, and Culture; UAR for Type)

Feature	High	Two	Culture	Type
ComParE [21, 14]	0.5214	0.4986	0.3887	0.3839
eGeMAPS [22, 14]	0.4496	0.4143	0.3214	0.3546
End2You [34, 14]	0.5686	0.5084	0.4401	0.4172
w2v2-r-vad [18]	0.6440	0.5948	0.4835	0.4560
w2v2-r-vad [18]	0.6478	0.6142	0.4962	0.4791
w2v2-r-er	0.6545	**0.6290**	**0.5199**	**0.4902**
w2v2-r-vad	**0.6554**	0.6244	0.5178	0.4834
Others*	0.6552	0.6224	0.4206	0.4713

*Others: High, Two, Culture: w2v2-lr-300; Type: w2v2-lr

Compared to [14] and [18], we improved the performance scores by remarkable margins. In [18], the authors evaluated different acoustic embeddings and modified the classifier of the baseline methods [14]. In this study, we evaluated different characteristics of acoustic embedding with CCC loss. Both modifications improved the performance scores. The ablation study showed no improvement had been made if the loss was MSE (relative to [18]); the acoustic embedding instead showed improvements with the existing architecture proposed in [14] using either MSE or CCC loss functions. This finding suggests that acoustic embedding (input data) is more important than the classifier architecture for improving the performance scores of affective vocal bursts tasks.

8. CONCLUSION

This study evaluated four affective vocal burst tasks with two conventional acoustic features (eGeMAPS and ComParE) and seven variants of wav2vec 2.0. The results showed that the acoustic embeddings from wav2vec 2.0 outperformed the conventional acoustic features in all tasks. The best overall performance was obtained by the wav2vec 2.0 model finetuned on MSP-Podcast affective speech dataset. Fusing the hidden states with logits on this model did not improve the performance significantly from the hidden states only. The wav2vec 2.0 finetuned on Switchboard attained comparable test scores to that of MSP-Podcast on High and Two tasks. We achieved comparable results to previous studies on nonvocalization speech emotion recognition. Further exploration is needed to improve the performance since vocalization is believed to have richer affective information than speech.

9. REFERENCES

[1] Disa A. Sauter, Frank Eisner, Andrew J. Calder, and Sophie K. Scott, “Perceptual cues in nonverbal vocal expressions of emotion,” *Q. J. Exp. Psychol.*, vol. 63, no. 11, pp. 2251–2272, 2010.

[2] Celine Berckmoes and Guy Vingerhoets, “Neural Foundations of Emotional Speech Processing,” *Curr. Dir. Psychol. Sci.*, vol. 13, no. 5, pp. 182–185, oct 2004.

[3] Tony W. Buchanan et al., “Recognition of emotional prosody and verbal components of spoken language: An fMRI study,” *Cogn. Brain Res.*, vol. 9, no. 3, pp. 227–238, 2000.

[4] Leimin Tian, Johanna Moore, and Catherine Lai, “Recognizing emotions in dialogues with disfluencies and non-verbal vocalisations,” in *Proc. ofthe 4th Interdiscip. Work. Laugh. Other Non-verbal Vocalisations Speech*, 2015, pp. 39–41.

[5] Leimin Tian, Johanna Moore, and Catherine Lai, “Emotion recognition in spontaneous and acted dialogues,” in *2015 Int. Conf. Affect. Comput. Intell. Interact. ACII 2015*, 2015, pp. 698–704.

[6] Leimin Tian, Johanna Moore, and Catherine Lai, “Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features,” in *2016 IEEE Spok. Lang. Technol. Work.* dec 2016, pp. 565–572, IEEE.

[7] Carlos Busso et al., “IEMOCAP: Interactive emotional dyadic motion capture database,” *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[8] Björn Schuller et al., “AVEC 2012 - The continuous audio/visual emotion challenge,” in *ICMI’12 - Proc. ACM Int. Conf. Multimodal Interact.*, 2012, pp. 449–456.

[9] 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, pp. 9663, 2021.

[10] Alan S. Cowen et al., “Mapping 24 emotions conveyed by brief human vocalization.,” *Am. Psychol.*, vol. 74, no. 6, pp. 698–712, sep 2019.

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

[12] Alice Baird et al., “The ICML 2022 Expressive Vocalizations Workshop and Competition: Recognizing, Generating, and Personalizing Vocal Bursts,” in *Proc. ICML Expressive Vocalizations Work. Compet.*, 2022.

[13] Björn W. Schuller et al., “The ACM Multimedia 2022 Computational Paralinguistics Challenge: Vocalisations, Stuttering, Activity, & Mosquitoes,” 2022.

[14] Alice Baird et al., “The ACII 2022 Affective Vocal Bursts Workshop & Competition: Understanding a critically understudied modality of emotional expression,” in *10th Int. Conf. Affect. Comput. Intell. Interact. Work. Demos (ACIIW)*, 2022.

[15] Johannes Wagner et al., “Dawn of the transformer era in speech emotion recognition: closing the valence gap,” mar 2022.

[16] Bagus Tris Atmaja and Masato Akagi, “Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using SVM,” *Speech Commun.*, vol. 126, pp. 9–21, feb 2021.

[17] Bagus Tris Atmaja et al., “Jointly Predicting Emotion, Age, and Country Using Pre-Trained Acoustic Embedding,” in *10th Int. Conf. Affect. Comput. Intell. Interact. Work. Demos*, 2022.

[18] Bagus Tris Atmaja, Zanjabila, and Akira Sasou, “On The Optimal Classifier For Affective Vocal Bursts And Stuttering Predictions Based On Pre-Trained Acoustic Embedding,” in *APSIPA Annu. Summit Conf.*, 2022.

[19] Reza Lotfian and Carlos 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.

[20] Alan Cowen et al., “The Hume Vocal Burst Competition Dataset (H-VB) — Raw Data [ExVo: updated 02.28.22] [Data set],” *Zenodo*, 2022.

[21] Björn Schuller et al., “The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism,” in *Interspeech 2013*, ISCA, 2013, pp. 148–152, ISCA.

[22] Florian Eyben et al., “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.

[23] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, “wav2vec: Unsupervised Pre-Training for Speech Recognition,” in *Interspeech 2019*, ISCA, sep 2019, pp. 3465–3469, ISCA.

[24] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Adv. Neural Inf. Process. Syst.*, 2020.

[25] Leonardo Pepino, Pablo Riera, and Luciana Ferrer, “Emotion Recognition from Speech Using wav2vec 2.0 Embeddings,” in *Interspeech 2021*, ISCA, 2021, pp. 3400–3404, ISCA.

[26] Zhaohang Zhang, Xiaohui Zhang, Min Guo, Wei-qiang Zhang, Ke Li, and Yukai Huang, “A Multilingual Framework Based on Pre-training Model for Speech Emotion Recognition,” in *APSIPA Annu. Summit Conf.*, 2021, pp. 750–755.

[27] Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli, “Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training,” in *Interspeech 2021*. aug 2021, vol. 3, pp. 721–725, ISCA.

[28] Johannes Wagner et al., “Model for Dimensional Speech Emotion Recognition based on Wav2vec 2.0 (1.1.0),” 2022.

[29] Jimmy Lei Ba, Jamie R. Kiros, and Geoffrey E. Hinton, “Layer Normalization,” *arXiv:1607.06450v1*, 2015.

[30] Ilya Loshchilov and Frank Hutter, “Decoupled Weight Decay Regularization,” *7th Int. Conf. Learn. Represent. ICLR*, 2019.

[31] Bagus Tris Atmaja and Masato Akagi, “Evaluation of error- and correlation-based loss functions for multitask learning dimensional speech emotion recognition,” *J. Phys. Conf. Ser.*, vol. 1896, no. 1, pp. 012004, 2021.

[32] John Godfrey and Edward Holliman, “Switchboard-1 Release 2 LDC97S62,” *Linguist. Data Consort.*, 1993.

[33] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, 2015, pp. 5206–5210.

[34] Panagiotis Tzirakis, Stefanos Zafeiriou, and Björn W. Schuller, “End2You-The Imperial Toolkit for Multimodal Profiling by End-to-End Learning,” *arXiv Prepr. arXiv 1802.01115*, 2018.