On The Optimal Classifier For Affective Vocal Bursts And Stuttering Predictions Based On Pre-Trained Acoustic Embedding

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

***Abstract*—Speech emotion recognition currently gained more interest from researchers due to its potential applications in the market. Instead of a speech, vocal bursts are understudied and may contain richer affective information than speech for recognizing emotion (e.g., laugh for happiness and cry for sadness). On the other side, acoustic features used for the affective vocalization may also be helpful for the stuttering evaluation task. Instead of handcrafted acoustic features, a pre-trained model feature extractor is now attaining more attention due to its competitiveness in modeling universal speech embedding. However, the previous speech embedding evaluations are not well-suited for emotion recognition. In this study, the researchers evaluated acoustic embedding extracted from a model fine-tuned on an affective speech dataset for affective vocalization and stuttering predictions using different classifiers. The methods were evaluated on a baseline classifier from the previous study and five new different classifiers, including an ensemble classifier. The results show improvements over the baseline methods; the ensemble classifier consistently resulted in the optimal performance on new validation sets with balanced and unnormalized data for both affective vocal bursts and stuttering predictions.**

I. Introduction

Recognizing mental and physical health via speech is an emerging field of research. Several pieces of information regarding specific mental and physical health conditions can be collected from speech. Emotion detection and cough diagnosis are examples of cases where speech-based technology could be implemented. These kinds of assessments now have potential applications for bypassing the previous complex diagnosis and therapy in healthcare industries.

Speech emotion recognition is a useful tool for customer satisfaction and mental healthcare purposes. In [1], the authors propose an ensemble method for recognizing emotion in call centers. By evaluating natural and acted speeches, the advantages and disadvantages of the two methods could be studied. A simple voting scheme among different classifiers also improves the accuracy modestly. In [2], the authors also employed an ensemble method from acoustic and linguistic modalities for predicting the emotion of older people in hospitals.

Instead of a speech, bursts of vocalization are another way to find the affective state of humans from their voices. Unlike speech emotion recognition (SER) which is well established in more than 20-year studies, the automatic recognition of affect burst is a new study that is supported by the availability of datasets and challenges [3], [4], [5]. The current methods for predicting affective vocal burst (AVB) adopted the workflow from SER. In general, acoustic features are extracted from the dataset. The features are then fed into a classifier or regressor. The output is either prediction of the emotion class (classification) or the degree/score of each class (regression).

Stuttering is a type of speech disorder, and therefore the diagnosis and treatment could be traced back to speech. The disorder can be identified by prolongation of the sounds, words, or syllables; repetition; and blocks [6]. Speech processing could be used to assess prolongation in stuttering; the decrease in speaking rate, the worse naturalness of speech [7]. There is a correlation between unnatural speech and stuttering [8] which can be used as a tool to monitor stuttering.

Instead of conducting affect bursts and stuttering predictions separately, this study contributes to the previous works by evaluating the same acoustic embedding – acoustic features extracted from a pre-trained model – for both affective vocal burst and stuttering predictions on different classifiers. The pre-trained model is built from wav2vec 2.0 large and robust model and is fine-tuned on an affective speech dataset. These acoustic embeddings may contain richer affect information than traditional acoustic features (e.g., MFCC and Melspectrogram). These acoustic embeddings are then fed into different classifiers to find the optimal one. The researchers evaluated five classical machine learning classifiers (Linear SVM, SVM, MLP, KNN, and XGB) and ensemble learning to study the effect of such parameters on a specific task. Both tasks were evaluated independently in a single-task learning manner.

II. Methods

A. Datasets

The researchers employed datasets from The ACM Multimedia 2022 Computational Paralinguistic Challenge (ComParE). Two tasks (challenges) with two datasets are evaluated with the same acoustic embedding and similar methods. The first dataset comes from the Vocalization Sub-Challenge (VOC-C). The second dataset comes from the Stuttering Sub-challenges (Kassel State of Fluency Corpus, KSF-C [6], [9]). There are 1361 samples in the original VOC-C dataset and 4601 in the original KSF-C dataset.

978-616-590-477-3 ©2022 APSIPA

The original VOC-C dataset has split the training set (6 speakers, 625 samples) and dev(elopment) set (5 speakers, 460 samples) with their labels. All speakers in both train and dev sets are female. The male vocalizations (2 speakers, 276 samples) are for the test without labels provided. The task is modeling a 6-class emotion prediction with achievement, anger, fear, pain, pleasure, and surprise. The distribution of the original vocalization dataset is near balance, ranging from 89 samples to 114 samples for each class.

The original KSF-C dataset has split the training set (23 speakers), dev set (6 speakers), and test set (8 speakers). There is no splitting of sex between training, dev, and test sets. All sets include male and female participants. The stuttering classes are block, prolongation, sound repetition, word/phrase repetition, modified speech technique, interjection, and no disfluency. The distribution of samples for the original stuttering dataset is unbalanced (ranging from 52 samples to 830 samples for each class).

To enable experiments with unseen data for evaluating several methods, the researchers arrange new vocalization and stuttering datasets. A number of 276 samples from development were moved to training data, while the rest of 184 is for (val)idation data in the new vocalization dataset. For the stuttering dataset, we reserved 282 samples from dev to the validation set and moved the remainder of 700 samples to the training set. Table I shows the comparison of the original and new distributions for each partition for both datasets.

TABLE I
COMPARISON OF THE ORIGINAL DATASET AND NEW DATASET FOR EACH PARTITION; THE NEW DATASET ONLY CONTAINS TRAINING AND VAL.

Task	Training	Dev	Test	Total
		Original		
VOC-C	625	460	276	1361
KSF-C	2489	982	1130	4601
		New		
VOC-C	901	184		1085
KSF-C	3189	282		3471

B. Pre-trained acoustic embedding (w2v2-R-er)

The feature is one of the important factors in machine learning. Many authors proposed specific features for specific goals, e.g., eGeMAPS and adieu features for extracting affective information from speech [10], [11]. Recent studies showed that acoustic features extracted from pre-trained self-supervised learning (SSL) models perform well across several tasks [12]. In line with these recent findings, we use a pre-trained model to extract acoustic embedding as the input of machine learning. The pre-trained model [13], [14] is based on wav2vec 2.0 large and robust model [15] fine-tuned on MSP-Podcasts dataset [16]. The acoustic embedding used in this study (called w2v2-R-er, wav2vec 2.0 Robust emotion recognition) is the hidden states of the pre-trained model applied to the audio data. For each sample, the generated acoustic embedding has 1024 dimensions. Since the pre-trained model is fine-tuned on an affective dataset, the researchers hypothesize an improvement in the performance evaluation compared to the baseline models for the vocalization task. The improvement may also apply to the stuttering task.

C. Data Balancing

Two data balancing methods were evaluated: synthetic minority over-sampling technique (SMOTE) [17] and random over-sampling (ROS). The choice of these balancing methods is based on the ablation studies in the early stage of this research. Only one balancing method is reported in this study for each task/challenge (which one is better), SMOTE for the vocalization task and ROS for the stuttering task.

D. Classifiers

The researchers evaluated five individual classifiers and an ensemble classifier. The first classifier is Linear SVC (SVC with a linear kernel), which obtained prominent results from the previous studies [5], [18], [19]. SVC is SVM for classification tasks. The second classifier is SVC with an RBF kernel. The third classifier, MLP, showed better performances than LSTM in the previous research on dimensional emotion recognition [20]. For comparisons, we evaluated KNN and XGB [21] as the fourth and fifth classifiers. Finally, an ensemble classifier was added due to its effectiveness in the previous related studies and challenges [22], [1], [2].

E. Evaluation Metric

The researchers measured the evaluated methods with a single metric, namely unweighted accuracy (UA). This metric is also known as balanced accuracy or unweighted average recall (UAR). It calculates the mean of recall for all classes.

F. Open Repositories

The methods to obtain the results are hosted in the open repositories. For the vocalization task, the repository is https://github.com/bagustris/vocc_w2v2. For the stuttering task, the repository is https://github.com/bagustris/ksfc_w2v2.

III. RESULTS AND DISCUSSION

We present our results in various tasks and conditions. For clarity, we split the results for different tasks and test results. In each task, we report unweighted accuracies or unweighted average recalls for both unbalanced and balanced data. For each unbalanced and balanced data, we report scores for unnormalized (original) and normalized audio signals. A discussion on a mismatch between validation and test sets is also provided to highlight the findings.

A. Vocalization

In the first result, we present UA scores of the original unbalanced data for the stuttering task (VOC-C). Five individual classifiers are compared along with an ensemble classifier. All classifiers are solid in the training phase; however, the scores in the validation phase differ remarkably. The different scores between the training and validation phases are observed for submitting predictions for both vocalization and stuttering challenges.

In the early stages of experiments, we only experimented with individual classifiers. We found that SVC performed best at this stage, and we followed by building an ensemble classifier based on this SVC. Four different SVC are combined with different regularization C values (10, 8, 6, 4) with a maximum of 100 iterations. The kernel is RBF, which is the default kernel in scikit-learn SVC implementation. The voting method is a hard voting type that uses majority voting of four SVC models. The ensemble method shows its superiority among individual classifiers, supporting the previous finding on the use of ensemble methods for various tasks [23], [22], [1].

No remarkable improvement has been found between UA scores with and without waveform normalization. For MLP, XGB, and Ensemble, normalized waveform improves the UA scores slightly. In contrast, for SVC and KNN, the normalized waveform decreases the UA scores slightly. As in [24], normalizing waveform may not affect the accuracy (UA) significantly. The best score in this unbalanced vocalization is achieved by Ensemble with normalized waveforms, a UA score of 59.34% on the new validation set. Since the results are deterministic for every trial, it is not possible to calculate the significance of differences between normalized and unnormalized waveforms from a statistics point of view.

TABLE II
UNWEIGHTED ACCURACY (UA, %) ON NEW VALIDATION SET FOR **VOCALIZATION UNBALANCED TRAINING DATA**; NORMALIZED/UNNORMALIZED APPLIES FOR AUDIO SIGNALS.

Classifier	Unnormalized		Normalized	
	Training	Val	Training	Val
Linear SVC	100	41.58	100	41.74
SVC	83.11	55.43	83.10	54.99
MLP	99.40	49.26	99.41	49.48
KNN	100	45.21	100	43.68
XGB	100	47.24	100	51.57
Ensemble	100	57.71	100	**59.34**

The next evaluation is vocalization with balanced data. It has been found that data balancing is an essential factor for machine learning, including SVC, MLP, KNN, and XGB. Unbalance data usually suffers from a low score of UA, in which the least data will perform worst while the most data will perform best. To tackle the limitation of unbalanced data, the researchers performed data balancing using SMOTE on imbalanced-learn toolkit [25]. The results for this balanced evaluation of the vocalization challenge are shown in Table III.

As shown in Table III, there are improvements in balancing the training data. For example, the UA improves from 55.43% to 59.86 % from unnormalized-unbalanced to unnormalized-balanced data. The highest UA score for this vocalization data is obtained with unnormalized-balanced data with 61.33 % of UA. This highest score is obtained by an ensemble classifier, as in unbalanced data (with the same configurations). The order of classifiers from the best to the worst is consistent for both unbalanced and balanced data, i.e., Ensemble, SVC, MLP, XGB, and KNN.

TABLE III
UNWEIGHTED ACCURACY (UA, %) ON NEW VALIDATION SET FOR **VOCALIZATION BALANCED TRAINING DATA**; NORMALIZED/UNNORMALIZED APPLIES FOR AUDIO SIGNALS.

Classifier	Unnormalized		Normalized	
	Training	Val	Training	Val
Linear SVC	100	43.44	100	45.09
SVC	100	59.86	100	56.19
MLP	99.32	56.19	99.50	53.81
KNN	100	42.34	100	42.29
XGB	100	46.02	100	47.88
Ensemble	100	**61.33**	100	57.71

Fig. 1 shows the confusion matrix of the best prediction for the vocalization task, i.e., prediction from an ensemble classifier with balanced-unnormalized data. The figure depicts the usability of the ensemble method by gaining the best recall for each class (the darker colors in a diagonal direction). The best recall is on pleasure detection with 80% of UA. The worst recall is on achievement detection with 44% of UA. Nevertheless, the average UA of 61.33% on the vocalization's validation score remains room for improvement.

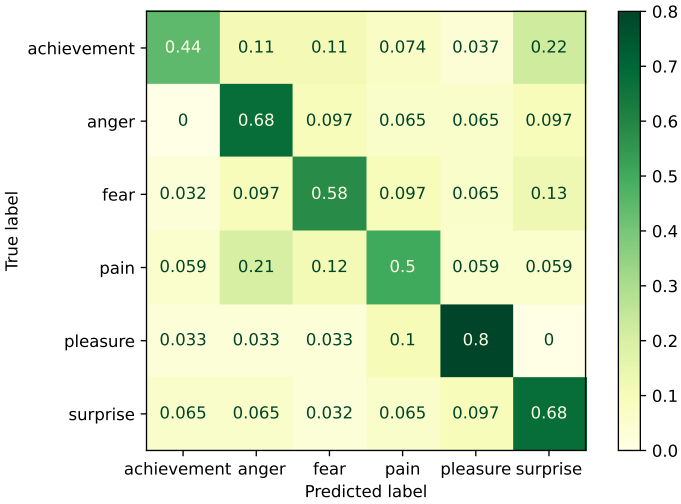

Fig. 1. Confusion matrix on (new) validation set using the best prediction (ensemble balanced-unnormalized) for Vocalization

B. Stuttering

The trend for stuttering is similar to that of vocalization. Both differ only on the least performed classifiers. In the stuttering task, the worst classifier is XGB for unbalanced and balanced data evaluations. In the previous vocalization, XGB is only better than KNN. The scores for training and validation are also similar in both cases; most classifiers perform well on training (UA around 90-100%) but not on validation (UA around 40-55%).

The highest UA score for the stuttering task was obtained by the ensemble classifier with balanced-unnormalized data, the same as the vocalization task. In this case, the UA is 55.08%. For an individual class in the stuttering task, the ensemble classifier fails to obtain the highest recall for classes

TABLE IV
UNWEIGHTED ACCURACY (UA, %) ON NEW VALIDATION SET FOR **STUTTERING UNBALANCED TRAINING DATA**; NORMALIZED/UNNORMALIZED APPLIES FOR AUDIO SIGNALS.

Classifier	Unnormalized		Normalized	
	Training	Val	Training	Val
Linear SVC	98.50	46.54	98.45	47.98
SVC	93.27	53.63	93.07	54.53
MLP	100	53.37	100	50.87
KNN	100	46.28	100	46.28
XGB	100	40.10	100	38.02
Ensemble	93.79	53.63	93.91	**54.79**

TABLE V
UNWEIGHTED ACCURACY (UA, %) ON NEW VALIDATION SET FOR **STUTTERING BALANCED TRAINING DATA**; NORMALIZED/UNNORMALIZED APPLIES FOR AUDIO SIGNALS.

Classifier	Unnormalized		Normalized	
	Training	Val	Training	Val
Linear SVC	99.30	47.70	99.36	48.54
SVC	93.51	54.40	93.30	54.22
MLP	100	50.96	100	50.86
KNN	100	46.28	100	46.28
XGB	100	42.71	100	38.65
Ensemble	93.14	**55.08**	93.14	53.38

SoundRepetition and WordRepetition (Fig. 2). Interestingly, both are repetitions of the same sound or word. It is presumed that *the model (the ensemble classifier) cannot distinct repeated events*, indicated by low recall scores in both cases (15% and 14%). Future studies may accommodate this issue to recognize repeated events in the stuttering task.

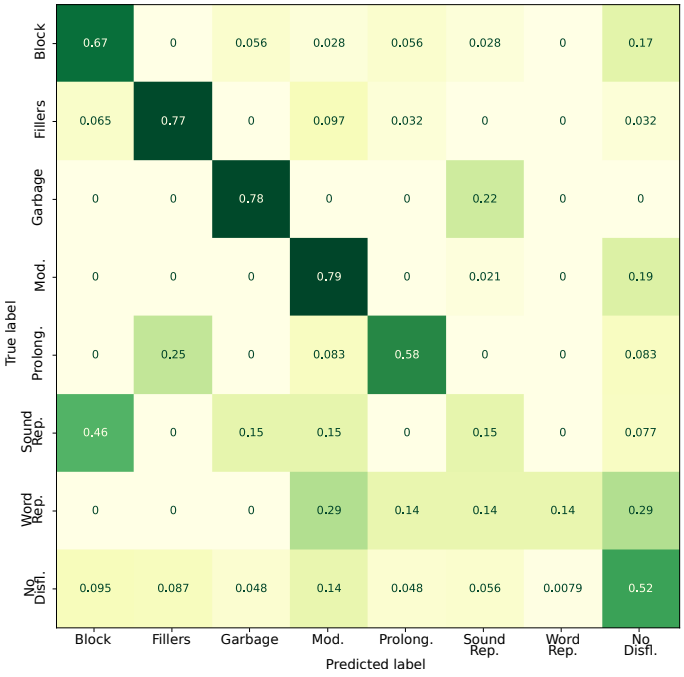

Fig. 2. Confusion matrix on (new) validation set using the best prediction (ensemble balanced-unnormalized) for Stuttering; the optimal classifier fails on predicting repetition events.

C. Test Results

Not all results, particularly the best results for vocalization and stuttering, were submitted to obtain the test results. Both datasets contain audio files for the test set, but their labels are not available. The prediction of each audio file in the test set along with the predicted label (e.g., 'test_0001.wav anger' for vocalization) in a CSV file is sent to the organizer of ComParE 2022. The organizer then sent the UA score for the corresponding submitted file. Given the limited time and number of trials, it is difficult to experiment with the test results. The researchers sent three predictions for the vocalization task and two predictions for the stuttering task.

Table VI shows the results of submitted predictions for vocalization and stuttering tasks using various methods. In both tasks, the researchers surpassed the UA baseline scores with 1% for vocalization and 7% for stuttering on the test set. For the dev set, which has no official baseline scores, but the scores were provided in [5], two submitted predictions surpass the highest baseline dev score. Two submitted results for stuttering also surpass the highest baseline score. The original dev set and test set (via submissions) are evaluated for obtaining the scores reported in Table VI. Hence, the comparison of the evaluated methods on both dev and test scores is fair since they use the same data as the baseline.

The best test results for both vocalization and stuttering tasks were obtained by the SVC classifier. In the vocalization task, the best result is obtained by the SVC classifier with the default configuration of the scikit-learn toolkit (C=1, max_iter = -1). In this task, using an ensemble did not improve the test scores in contrast to the validation scores. Regarding the stuttering task, the SVC classifier was modified with C=10 and max_iter=100. Changing the regularization parameter (C) and hard limit on iteration within solver (max_iter) improved the test performance remarkably, highlighting the importance of these parameters on SVC for the stuttering task. From these results, it can be seen that using pre-trained acoustic embedding improved the performance of vocalization and stuttering predictions.

TABLE VI
DEV & TEST (ORIGINAL PARTITIONS) SCORES (% UA) ON VOCALIZATION AND STUTTERING CHALLENGES

Method	Dev	Test
Vocalization		
Best baseline: BoAWs + LinearSVC [5]	39.60	37.4
Our w2v2-R-er + SVC	38.44	**38.4**
Our w2v2-R-er + XGB	36.21	36.5
Our w2v2-R-er + Ensemble	40.46	35.8
Stuttering		
Best baseline: DeepSpectrum + LinearSVC [5]	28.10	40.4
Our w2v2-R-er + SVC (C = 1, max_iter=-1)	33.25	39.2
Our w2v2-R-er + SVC (C = 10, max_iter=100)	36.44	**47.4**

D. Discussion Regarding New Validation and Test Scores

The researchers have observed a mismatch between validation and test scores in this study. The validation set is designed to evaluate the performance of different methods for submitting

the predictions for the hidden test set. The mismatch was observed for vocalization, in which the best method in the validation set with ensemble performed worse than SVC. A similar mismatch between development/validation and test sets was also observed in the previous study with a small-size dataset [18]. For the stuttering dataset with a larger number of samples (stuttering: 4601 samples, vocalization: 1361 samples, both include test set), the researchers did not evaluate the best validation method (ensemble) for the prediction of the test set due to time limitation of the challenge.

For choosing the more reliable results, the researchers chose the results of the validation set presented in this paper. These results are validated across different conditions (unbalanced vs. balanced, unnormalized vs. normalized) and show consistency. The best result was obtained by the ensemble method with the same condition (balanced-unnormalized). The trend was also similar on both tasks, with a small change in the order of the worst classifier.

Another issue in the validation and test results is the bias due to different sex for validation and test data on the vocalization task. The validation data employed female speakers, while the test data employed male speakers. Although the pre-trained model to extract the acoustic embedding is trained on diverse speakers (both male and female speakers), the model might not be able to fine-tune the embeddings for male speakers. Since the test set with male speakers is hidden, it is not possible to evaluate the bias due to sex differences in this study. Future studies could address this issue if male and female speakers are available for the validation set.

IV. Conclusions

In this paper, we evaluated the use of pre-trained acoustic embedding for predicting affective vocalization and stuttering tasks. Both tasks are run independently (in a single-task learning manner, not multitask learning). The acoustic embeddings are extracted using a pre-trained model of wav2vec 2.0 large and robust version fine-tuned on affective speech dataset. The researchers evaluated five different classifiers using these acoustic embeddings. The results show the superiority of an ensemble classifier trained on balanced-unnormalized data among other classifiers. The ensemble classifier consists of four SVC models with different regularization parameters with hard voting.

For the stuttering task, the ensemble model fails to detect the repeated events, i.e., SoundRepetition and WordRepetition classes. Future research could be directed to focus on this repetition issue. Since the model is quite well on both tasks, combining these two tasks into multitask learning may also beneficial for future research.

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] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Commun.*, vol. 49, no. 2, pp. 98–112, 2007.

[2] G. Soğancıoğlu, O. Verkholyak, H. Kaya, D. Fedotov, T. Cadée, A. A. Salah, and A. Karpov, "Is Everything Fine, Grandma? Acoustic and Linguistic Modeling for Robust Elderly Speech Emotion Recognition," in *Proc. Interspeech 2020*. ISCA: ISCA, oct 2020, pp. 2097–2101.

[3] 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.

[4] 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.

[5] 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.

[6] S. P. B. B and H. Florian, *Text, Speech, and Dialogue*, ser. Lecture Notes in Computer Science, P. Sojka, I. Kopeček, K. Pala, and A. Horák, Eds. Cham: Springer International Publishing, 2020, vol. 12284. [Online]. Available: https://link.springer.com/10.1007/978-3-030-58323-1

[7] C. Jessen, "Acoustics correlates of speech naturalness in Post-Treatment Adults who Stutter: Role of Speaking Rate," Honor Theses, Western Michigan University, 2016.

[8] R. R. Martin, S. K. Haroldson, and K. A. Triden, "Stuttering and speech naturalness," *J. Speech Hear. Disord.*, vol. 49, no. 1, pp. 53–58, feb 1984.

[9] S. P. Bayerl, A. W. von Gudenberg, F. Hönig, E. Nöth, and K. Riedhammer, "KSoF: The Kassel State of Fluency Dataset – A Therapy Centered Dataset of Stuttering," 2022. [Online]. Available: http://arxiv.org/abs/2203.05383

[10] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2016-May, 2016, pp. 5200–5204.

[11] 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.

[12] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech 2021*. ISCA: ISCA, aug 2021, pp. 1194–1198.

[13] 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

[14] 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

[15] 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.

[16] 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.

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, no. 2, pp. 321–357, jun 2002.

[18] B. T. Atmaja, Y. Hamada, and M. Akagi, "Predicting Valence and Arousal by Aggregating Acoustic Features for Acoustic-Linguistic Information Fusion," in *2020 IEEE Reg. 10 Conf.* IEEE, nov 2020, pp. 1081–1085.

[19] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore Sound Classification Using Image-Based Deep Spectrum Features," in *Interspeech 2017*. ISCA: ISCA, aug 2017, pp. 3512–3516.

[20] 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.

[21] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," mar 2016. [Online]. Available: http://arxiv.org/abs/1603.02754 http://dx.doi.org/10.1145/2939672.2939785

[22] E. A. Mohammed, M. Keyhani, A. Sanati-Nezhad, S. H. Hejazi, and B. H. Far, "An ensemble learning approach to digital corona virus preliminary screening from cough sounds," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, 2021.

[23] N. K. Chowdhury, M. A. Kabir, M. M. Rahman, and S. M. S. Islam, "Machine learning for detecting COVID-19 from cough sounds: An ensemble-based MCDM method," *Comput. Biol. Med.*, vol. 145, no. March, p. 105405, 2022.

[24] 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.

[25] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, sep 2016.