INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Classifying depression symptom severity: Assessment of speech representations in personalized and generalized machine learning models.

Edward L. Campbell[1,2], *Judith Dineley*[2], *Pauline Conde*[2], *Faith Matcham*[2,3], *Katie M. White*[2], *Carolin Oetzmann*[2], *Sara Simblett*[2], *Stuart Bruce*[4], *Amos A. Folarin*[2,5,6], *Til Wykes*[2,5], *Srinivasan Vairavan*[7], *Richard J.B. Dobson*[2,6], *Laura Docío-Fernández*[1], *Carmen García-Mateo*[1], *Vaibhav A. Narayan*[8], *Matthew Hotopf*[2,5], *Nicholas Cummins*[2], *The RADAR-CNS Consortium*[9]

[1] GTM research group, AtlanTTic Research Center, University of Vigo, Spain
[2] Institute of Psychiatry, Psychology and Neuroscience, King's College London, UK
[3] School of Psychology, University of Sussex, Falmer, UK
[4] RADAR-CNS Patient Advisory Board, King's College London, UK
[5] NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, UK
[6] Institute of Health Informatics, University College London, UK
[7] Janssen Research and Development LLC, Titusville, NJ, United States
[8] Davos Alzheimer's Collaborative
[9] www.radar-cns.org

ecampbell@gts.uvigo.es, nick.cummins@kcl.ac.uk

Abstract

There is an urgent need for new methods that improve the management and treatment of Major Depressive Disorder (MDD). Speech has long been regarded as a promising digital marker in this regard, with many works highlighting that speech changes associated with MDD can be captured through machine learning models. Typically, findings are based on cross-sectional data, with little work exploring the advantages of personalization in building more robust and reliable models. This work assesses the strengths of different combinations of speech representations and machine learning models, in personalized and generalized settings in a two-class depression severity classification paradigm. Key results on a longitudinal dataset highlight the benefits of personalization. Our strongest performing model set-up utilized self-supervised learning features and convolutional neural network (CNN) and long short-term memory (LSTM) back-end.

Index Terms: Major depressive disorder, personalization, self-supervised learning, remote monitoring technologies

1. Introduction

Due to the prevalence and high socioeconomic costs associated with Major Depressive Disorder, several digital health initiatives have started to explore new ways to improve the management and treatment of MDD [1, 2, 3]. Speech is uniquely placed as a health signal in such projects due to its pyramidal structure of information [4, 5]. This structure runs from acoustic information at the lowest level, then onto prosodic, phonetic and finally conversational at the highest level [4, 5]. The acoustic, phonetic and prosodic levels have been of particular interest in speech-based depression detection, with a rich set of supporting literature strengthening the case for speech to be considered a valuable marker of depression [6, 7].

The majority of machine learning works in this field have focused on developing generalizable machine learning models to detect the presence or absence of depression in speech samples from cross-sectional datasets [8, 9]. However, the complexity of speech and the natural variety of human voices make robust extraction of speech patterns associated with depression a highly non-trivial task. Adding this is the ordinal nature of depression scores, meaning we cannot assume a continuous and well-behaved relationship between changes in speech features and assessment scores [10]. Given these difficulties, there is a strong case for exploring personalization to improve the performance of speech-based systems [11].

Herein, we compare the performance of different speech representations and machine learning models in generalized and personalized settings. The main focus of the work is the 2-class automatic speech-based classification of MDD severity. We use a experimental corpus of *scripted* and *free-response* speech samples collected longitudinally over a period of 18 months from 271 individuals with a history of recurrent MDD [1]. Our goal is to demonstrate the advantages of personalized models as opposed to generalizable methods. Additionally, we showcase the discriminative ability of features obtained through self-supervised models compared to conventional acoustic features. We also create systems that are not reliant on specific tasks and can be applied to real-world situations.

2. Experimental Corpus

Due to a lack of large, clinical longitudinal datasets [6, 7], we could not use publicly available corpora in our analysis. Our experimental data was collected as part of a large observational cohort study of individuals with a history of recurrent MDD [1]. Briefly, core eligibility criteria for inclusion in the study were meeting the DSM-5 diagnostic criteria [12] for non-psychotic MDD within the two years prior to enrolment and having recurrent MDD (lifetime history of at least two episodes). Exclusion criteria included having a history of bipolar disorder, schizophrenia, MDD with psychotic features, or schizoaffective disorder; having dementia; and moderate or severe drug or al-

10.21437/Interspeech.2023-1721

Table 1: *Sociodemographic characteristics, audio files distribution and depression score distribution of the RADRA-MDD used in this paper.*

Task	Scripted	Unscripted
Participants	271	258
Gender M/F	59/212	56/202
Mean Age	45(±16)	
No. Files	4,504	3,500
Size (hours)	17.62	18.92
Files per participant	12(±9)	11(±8)
Mean file length	14.5s	19.0s
Median PHQ-8 Score (IQR)	8 (4–13)	

cohol use in the six months prior to enrolment. All participants were aged over 18, and were able to give informed consent. The full eligibility and exclusion criteria are published in [1]. Ethical approval was obtained from the Camberwell St. Giles Research Ethics Committee (17/LO/1154) in London.

2.1. Speech Collection and Preparation

English speech data was collected in the study for a period of 18 months, from 2019 to 2021. During this time study participants were asked to complete two speech-recording tasks every two weeks. First, a purpose-built smartphone application [13] produced notifications each time speech recordings were scheduled. Before starting each recording task, participants were reminded, via on-screen instructions, to find a quiet place and to complete two recordings in their normal voice.

The first recording was a *scripted task*, in which the participants read aloud an extract from Aesop's fable, The North Wind and the Sun [14]. The other task was *Free Response*, participants were asked to speak about *something they were looking forward to in the next seven days*. Once recorded, the speech data were encrypted and sent to a secure server. When on the server the collected data were separated into the respective tasks and decrypted into 16 kHz, 16-bit mono *Waveform Audio File Format* (WAV) files. All files under five seconds in length were not considered in our analysis. The total number of participants and information on the distribution of the audio files used in our analysis are presented in Table 1.

2.2. Depression Severity Scores

We assigned a level of depression severity to each file using concurrently collected 8-item Patient Health Questionnaire (PHQ-8) scores [15]. The PHQ-8 is a standardized and validated self-report questionnaire for MDD detection [15]. We divided the speech files into two classes: (i) a *low* class (PHQ-8 < 10); and (ii) a *high* class (PHQ-8 ≥ 10). Visualizations of the PHQ-8 distribution are given in Figure 1, with the matching median PHQ scores and Interquartile Range given in Table 1.

2.3. Patient Involvement

The experimental protocol was co-developed with a patient advisory board who shared their opinions on several user-facing aspects of the study, including the choice and frequency of survey measures, the usability of the study app, participant-facing documents, selection of optimal participation incentives, selection, and deployment of wearable device as well as the data analysis plan. The speech task and subsequent analysis have been discussed specifically with a Patient Advisory Board.

Table 2: *Total number of parameters in, and average training time of the neural networks used in Section 4.3.*

System	Parameters	Train Time
ComParE (MLP)	9,177	19.46 sec
TRILLsson (CNN-LSTM)	856,705	9.6 min
WAV2VEC2 (CNN-LSTM)	955,009	51.65 min

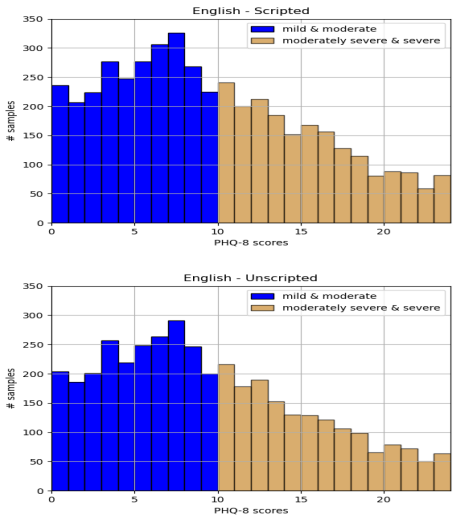

Figure 1: *PHQ-8 scores distributions collected concurrently with the speech files*

2.4. Data Availability

Due to the confidential nature of speech data, we are unable to make our data publicly available. Access to the data can be made through reasonable requests to the RADAR-CNS consortium and will be subject to local ethics clearances. Please email the senior author for details.

3. Methodology

3.1. Speech Representations and Models

The first system we used combines *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) functionals [16] with a support vector machine (SVM) classifier. We include this model as it is a widely used 'baseline' system in paralinguistic analysis. The second system utilizes *Computational Paralinguistics Challenge* (ComParE) functions [17] with a multilayer perceptron (MLP) classifier, again this set-up has performed well over a range of different paralinguistic task; e. g. [18]. Our third and fourth systems are based on more contemporary representations, in particular, self-supervised models. We assess the efficacy of transfer learning from *wav2vec 2.0* base architecture [19] and TRILLsson [20] models.

A context window and hop size of 5 seconds was applied to extract sequences of TRILLsson and wav2vec 2.0 features from each audio file. Records longer than 5 seconds, we split it into entire sequences of 5 seconds. Both systems utilize a combined Convolutional Neural Networks (CNN) and Long Short-Term Memory(LSTM) back end to perform classification and inference. We compute the average severity MDD score of the sequences, which represents the final classification score of the audio recording.

Table 3: *Comparison of speaker-independent (generalized) and speaker-dependent (personalized) 2-class depression severity detection. Results, Accuracy (ACC), precision (PR) and recall (RCL), are the average after 10-Fold cross-validation.*

System	**Generalized**						**Personalized**					
	Scripted			*Free Response*			*Scripted*			*Free Response*		
	ACC	PR	RCL	ACC	PR	RCL	ACC	PR	RCL	ACC	PR	RCL
eGeMAPS (SVM)	59.56	56.11	47.82	62.86	62.53	52.47	**73.30**	71.34	64.50	**72.01**	68.78	58.63
ComParE (MLP)	63.57	56.09	43.04	67.10	68.97	62.32	70.78	68.37	61.39	70.41	66.78	55.96
WAV2VEC2 (CNN-LSTM)	**72.19**	73.77	58.59	**69.31**	74.48	58.50	70.87	68.26	62.11	70.53	68.01	56.22
TRILLsson (CNN-LSTM)	66.77	58.61	45.03	64.63	68.35	55.65	70.77	68.08	62.74	70.73	66.34	58.85

3.2. Models configuration

The SVM model was configured using scikit-learn [21] as a non-linear classifier with a radial basis function kernel. We tested *cost* values between 0.1 and 20, optimal model performance was achieved at 5. The neural network models were initialized by a uniform distribution and optimized by the ADAM algorithm [22]. One cycle learning rate policy [23] was applied to decrease the training convergence time, setting the maximum learning rate to 0.1 and 1e-4 for the MLP and CNN-LSTM models, respectively. Early stopping criteria was applied to all networks' training. Training was terminated once a model achieved a binary-cross entropy loss of 0.1.

The MLP contains a hidden layer with 100 neurons. In initial experimentation, we analyzed 3 other set-ups that included increasing the number of hidden neurons and layers. However, the performance did not improve, and as a result, we decided to use an architecture that was less complex but more effective. The CNN-LSTM architecture was: a one-dimensional CNN and MaxPooling layers (applied across frequencies) with a kernel size of 3, two-bidirectional LSTM layers (128 hidden units each) and two dropout layers with a rate of 30%, one before the CNN block and another before the LSTM blocks. Finally, it is worth noting that we evaluated a sequence-to-sequence system with local/global attention mechanisms. Its performance was quite similar to the CNN-LSTM but its execution time was about 10 times slower. As a result, we made the decision to exclude this system from our experimental framework.

All systems were implemented on Python 3.8.16. Extraction of eGeMAPS and ComParE was done by openSMILE software [24]. Wav2vec 2.0 were extracted by the torchaudio.pipelines module packages and the pre-trained TRILLsonn model was downloaded from [1] and run with TensorFlow 2.4. Torch 1.13.1 was used for the development of the MLP and CNN-LSTM classifiers. The number of model parameters and average training time of our networks are given in Table 2.

4. Results and Discussion

We report all model performances in terms of accuracy, precision and recall.

4.1. Personalised and speaker-independent cross-validation experiments.

The aim of the work in this section is to compare the performance of various speech representations and ML classifiers and in *speaker-independent* (generalized) and *speaker-dependent* (personalized) modeling paradigms for the 2-class classification of MDD symptom severity. We did this testing separately for the scripted and free-response prompts, with all results verified using 10-fold cross-validation (Table 3). To help ensure that the personalized models learned patterns related to MDD severity symptoms and not speaker identity, we included samples from both low and high MDD classes per speaker in training.

We observed that the personalized training led to higher performance in most systems, as evidenced by higher accuracy, recall, and precision values. Moreover, results across tasks (i. e. scripted / free-response) were more steady. The CNN-LSTM model featuring self-supervised features, especially the wav2vec 2.0 feature, had comparable effectiveness across the personalized and speaker-independent approaches. This highlights the *robustness* of self-supervised features. The SVM-based system achieved the highest performance scores among the classifiers evaluated. This could be due to the amount of available training data being better suited to the lower complexity of the SVM.

4.2. Predictive modeling per collection year

In a second set of experiments, we explore the effects of generalization and personalization in two different prediction modeling tests. In these tests, samples are arranged and classified according to collection year, which also allows us to observe the effect of training set size on our models. To limit uncertainty due to acoustic variability in these tests, only recordings from the scripted tasks are analyzed.

Experiment 1 - Generalized: Models are first trained and evaluated with samples collected in 2019 and 2020 respectively. As a second step, the training set size is increased by adding samples collected from 2020 and using samples from 2021 as an evaluation set. The goal is to analyze the influence of the corpus size on the MDD severity detection rate. Additionally, speakers who are already present in the training set are deleted from the testing set. This experiment also allows for assessing the speaker-independent ability of the trained models.

Experiment 2 - Personalized: The same procedures as Experiment 1 are followed, except that the testing samples from speakers who are already present in the training set are retained. This means models can retain MDD severity symptoms specific to the examined patient. We make sure the model learns patterns related to MDD symptoms and not speaker identity, by having samples from both classes (low/high MDD) per speaker in training.

Table 4 shows the results of such experiments. The MDD severity detection rate of every system increased when more training samples were added for both experiments. However, systems show more stable results in the personalized framework. The recall values in particular in our generalized testing do not appear stable. We suspect that this is related to the smaller number of test instances used to ensure our models were truly speaker independent (Table 5).

[1] https://tfhub.dev/google/nonsemantic-speech-benchmark/trill/2

Table 4: *Personalized and generalization prediction modeling experiments. Results are reported for Accuracy (ACC), precision (PR) and recall (RCL).*

	eGeMAPS (SVM)					
	Generalized			**Personalized**		
Test year	ACC	PR	RCL	ACC	PR	RCL
2020	57.19	54.89	54.27	60.39	60.51	60.73
2021	66.67	63.33	61.74	67.55	69.24	68.76
	ComParE (MLP)					
	Generalized			**Personalized**		
Test year	ACC	PR	RCL	ACC	PR	RCL
2020	57.84	57.84	50.00	62.38	61.91	58.82
2021	63.49	31.75	50.00	67.24	66.83	66.49
	WAV2VEC2 (CNN-LSTM)					
	Generalized			**Personalized**		
Test year	ACC	PR	RCL	ACC	PR	RCL
2020	59.09	59.93	89.72	63.36	61.03	40.16
2021	63.49	63.49	100.00	67.45	68.71	52.22
	TRILLsson (CNN-LSTM)					
	Generalized			**Personalized**		
Test year	ACC	PR	RCL	ACC	PR	RCL
2020	61.66	61.43	85.65	61.02	57.47	35.87
2021	63.49	63.49	100.00	67.05	61.42	70.76

4.3. Personalised and task-independent training

In our final set of experiments, we assess the independent-task ability (i. e. we combine the scripted and free-speech data) of the proposed models in a personalized framework. In this analysis, we add the Area Under the Curve (AUC) metric to highlight the discriminative capacity of our models. Our results show that all models achieve an increase in performance with an increased amount of training data (Table 6). In particular, the TRILLsson–CNN-LSTM system which achieves our strongest accuracy (78.42%) and AUC (84.56%). These results highlight the benefit of including both scripted and free-response data to maximize the amount of available training data.

5. Conclusions

This paper presents a new experimental database for evaluating MDD severity symptoms using speech. This dataset contains longitudinal speech samples from a clinical population of individuals with a history of recurrent MDD. We used this dataset to highlight the benefits of personalization when predicting MDD symptom severity from speech. In particular, our prediction models demonstrated consistently higher performance in the personalized setting. We observed that the best-performing system depends on the amount of available training data, with our eGeMAPS and SVM set-up performing well with fewer instances, while a combination of TRILLsson features and CNN-LSTM-based classifier performed best when we maximized training data by combining our scripted and free-response samples.

A limitation of our work would be the binary nature of our classification task and the smaller number of systems tested. These were deliberate choices as this work represents the initial machine learning analysis on this data. In future work, we will expand to regression analysis. To help achieve this, we will explore augmentation strategies to increase the amount of available data. Additionally, given the robustness of the self-supervised frameworks, we will also explore the benefits of fine-tuning these features toward more MDD-specific latent spaces. Such as approach could enhance the MDD severity classification system's performance while also providing a more precise representation of MDD speech patterns.

Table 5: *Number of instances per class for prediction modeling experiments 1 and 2. Low represents mild and moderate depression severity (PHQ-8 $<$ 10). High represents moderately severe and severe depression (PHQ-8 $\geq$ 10)*

test-fold	Experiment 1		Experiment 2	
	Low	High	Low	High
2020	129	177	1,809	1,357
2021	23	40	356	288

Table 6: *Evaluation of the top-performing models in the fusion corpus (Scripted & Free-response). Personalised 10-Fold cross-validation experiment. Accuracy (ACC), precision (PR) and recall (RCL) are shown in percent*

System	ACC	PR	RCL	AUC
eGeMAPS (SVM)	74.04	70.46	66.17	79.14
ComParE (MLP)	73.20	70.53	63.94	78.32
WAV2VEC2 (CNN-LSTM)	74.34	73.87	73.27	79.55
TRILLsson (CNN-LSTM)	**78.42**	**75.42**	**72.8**	**84.56**

6. Acknowledgements

Funding The RADAR-CNS project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115902. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA (www.imi.europa.eu). This communication reflects the views of the RADAR-CNS consortium and neither IMI nor the European Union and EFPIA are liable for any use that may be made of the information contained herein. The funding body has not been involved in the design of the study, the collection or analysis of data, or the interpretation of data. We thank our colleagues both within the RADAR-CNS consortium and across all involved institutions for their contribution to the development of this protocol. We thank all the members of the RADAR-CNS patient advisory board for their contribution to the device selection procedures, and their invaluable advice throughout the study protocol design. This paper also represents independent research part funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. C.O. is supported by the UK Medical Research Council (MR/N013700/1) and King's College London member of the MRC Doctoral Training Partnership in Biomedical Sciences This work has also received financial support from Axudas propias para a mobilidade de Persoal Investigador da Unversidade de Vigo 2021, the Xunta de Galicia (Centro singular de investigación de Galicia accreditation 2019-2022), Consellería de Cultura (Educación e Ordenación Universitaria; axudas para a consolidación e estruturación de unidades de investigación competitivas do Sistema Universitario de Galicia -ED431B 2021/24), and the European Union (European Regional Development Fund - ERDF).

7. References

[1] F. Matcham, C. Barattieri di San Pietro, V. Bulgari, G. De Girolamo, R. Dobson, H. Eriksson, A. Folarin, J. M. Haro, M. Kerz, F. Lamers *et al.*, "Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol," *BMC psychiatry*, vol. 19, no. 1, pp. 1–11, 2019.

[2] C. G. Fairburn and V. Patel, "The impact of digital technology on psychological treatments and their dissemination," *Behaviour Research and Therapy*, vol. 88, pp. 19–25, 2017.

[3] N. Topooco, H. Riper, R. Araya, M. Berking, M. Brunn, K. Chevreul, R. Cieslak, D. D. Ebert, E. Etchmendy, R. Herrero *et al.*, "Attitudes towards digital treatment for depression: a european stakeholder survey," *Internet interventions*, vol. 8, pp. 1–9, 2017.

[4] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu *et al.*, "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. Hong Kong, China: IEEE, 2003, pp. 784–787.

[5] D. Reynolds, J. Campbell, B. Campbell, B. Dunn, T. Gleason, D. Jones, T. Quatieri, C. Quillen, D. Sturim, and P. Torres-Carrasquillo, "Beyond cepstra: exploiting high-level information in speaker recognition," in *Workshop on Multimodal User Authentication*, Santa Barbara, CA, 2003, pp. 223–229.

[6] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech communication*, vol. 71, pp. 10–49, 2015.

[7] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[8] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. Amsterdam, The Netherlands: ACM, 2016, pp. 3–10.

[9] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition," in *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*. Nice, France: ACM, 2019, pp. 3–12.

[10] N. Cummins, V. Sethu, J. Epps, J. R. Williamson, T. F. Quatieri, and J. Krajewski, "Generalized two-stage rank regression framework for depression score prediction from speech," *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 272–283, 2020.

[11] R. V. Shah, G. Grennan, M. Zafar-Khan, F. Alim, S. Dey, D. Ramanathan, and J. Mishra, "Personalized machine learning of depressed mood using wearables," *Translational psychiatry*, vol. 11, no. 1, pp. 1–18, 2021.

[12] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association, 2013, vol. 5.

[13] Y. Ranjan, Z. Rashid, C. Stewart, P. Conde, M. Begale, D. Verbeeck, S. Boettcher, R. Dobson, A. Folarin, R.-C. Consortium *et al.*, "Radar-base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices," *JMIR mHealth and uHealth*, vol. 7, no. 8, p. e11734, 2019.

[14] International Phonetic Association, *Handbook of the International Phonetic Association*. Cambridge University Press, 1999.

[15] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[16] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[17] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The Interspeech 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *Proc. Interspeech 2016*. San Francisco, CA, USA: ISCA, 2016, pp. 2001–2005.

[18] S. H. Dumpala, S. Rempel, K. Dikaios, M. Sajjadian, R. Uher, and S. Oore, "Estimating severity of depression from acoustic features and embeddings of natural speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. Toronto, Canada: IEEE, 2021, pp. 7278–7282.

[19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[20] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 5036–5040. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-3015

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[23] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and machine learning for multi-domain operations applications*, vol. 11006. SPIE, 2019, pp. 369–386.

[24] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.