



Self-Supervised Embeddings for Detecting Individual Symptoms of Depression

Sri Harsha Dumpala^{1,2}, Katerina Dikaivos³, Abraham Nunes¹, Frank Rudzicz^{1,2},
Rudolf Uher^{1,4}, Sageev Oore^{1,2}

{¹Dalhousie University, ²Vector Institute, ³McMaster University, ⁴Nova Scotia Health}, Canada
{sriharsha.d, nunes, frank, uher, sageev}@dal.ca, dikaiosk@mcmaster.ca

Abstract

Depression, a prevalent mental health disorder impacting millions globally, demands reliable assessment systems. Unlike previous studies that focus solely on either detecting depression or predicting its severity, our work identifies individual symptoms of depression while also predicting its severity using speech input. We leverage self-supervised learning (SSL)-based speech models to better utilize the small-sized datasets that are frequently encountered in this task. Our study demonstrates notable performance improvements by utilizing SSL embeddings compared to conventional speech features. We compare various types of SSL pretrained models to elucidate the type of speech information (semantic, speaker, or prosodic) that contributes the most in identifying different symptoms. Additionally, we evaluate the impact of combining multiple SSL embeddings on performance. Furthermore, we show the significance of multi-task learning for identifying depressive symptoms effectively.

Index Terms: Self-supervised learning, depression, depressive symptoms, multi-task learning, semantic, speaker, prosody

1. Introduction

Major depressive disorder, commonly known as depression, is the most prevalent mental health disorder, and a leading cause of global disability [1]. Yet, it remains largely under-detected and under-treated [2]. Therefore, accurately identifying and treating depression is a significant public health priority aimed at reducing its prevalence. However, scarcity of trained clinicians and other resources hinder this objective. While deep learning-based mental health assessment systems have shown promise in tasks including patient screening and monitoring, further research is necessary to develop more informative and interpretable deep learning models for widespread deployment in clinical settings.

Recent research has demonstrated the efficacy of utilizing natural speech as a potentially inexpensive and easily scalable indicator for assessing depression [3–7]. Previous studies primarily focused on either detecting depression [8–10] or predicting its severity [11–13]. Potential vocal biomarkers for depression identified in these studies encompass a spectrum of traditional acoustic features, including prosodic elements (such as pitch and speech rate), spectral characteristics (like Mel-frequency cepstral coefficients and formant frequencies), and glottal attributes (related to vocal fold behaviour) [4, 6, 9], as well as neural embeddings from pretrained models like x -vectors, wav2vec 2.0, and HuBERT [13, 14]. However, these studies often operate under the assumption that of all the measurable aspects of speech, depression severity is the most significant and influential [15]. Yet, merely identifying or estimating depression severity may prove inadequate compared to the clin-

ical setting of monitoring the core subset of depressive symptoms for effective treatment outcomes [16]. In this study, we adopt a clinically oriented approach to depression assessment – aiming to detect depressive symptoms while also predicting overall depression severity. This method also offers an implicit explanation for the decisions made by the automated models.

Only a few studies have explored the task of identifying individual symptoms of depression using speech as input [15, 17, 18]. Fara et al. [15, 17] used acoustic features such as eGeMAPS [19] and prosodic features extracted from spontaneous speech (in English) to train random forest classifiers for detecting individual symptoms of depression based on the self-reported Patient Health Questionnaire (PHQ-8) [20]. Similarly, Takano et al. [18] used spectral, prosodic, and glottal features extracted from read speech (in Japanese) to train decision tree models for detecting individual symptoms of depression based on the clinician-rated HAM-D scale [21]. These works used conventional speech features to train simple machine learning models for detecting individual symptoms of depression. In this work, we use neural embeddings extracted from deep learning models to detect individual symptoms of depression based on the Montgomery and Åsberg Depression Rating Scale (MADRS) [22], which is considered a gold standard for clinical assessment of depression severity.

One of the major hindrances in applying deep learning techniques to the medical domain, including depression assessment, is insufficient training data [23]. To address the challenge of low-resourced data, we leverage SSL-based training to exploit the availability of large amounts of unlabeled speech data. Previous studies have employed SSL-based speech models to detect depression [14, 24, 25]. We extend these approaches by leveraging SSL-based speech models to both (A) identify individual symptoms of depression and (B) predict depression severity. Specifically, we analyze several SSL-based speech models pretrained with different objective functions such as masked language modeling (MLM), contrastive learning, and speaker-invariant training [26–28]. The choice of the pre-training objective function defines the information encoded by the SSL speech models, such as semantic, speaker, and prosodic information. Using these models, we assess the significance of each information type in identifying individual symptoms of depression. Additionally, we examine the impact of combining embeddings obtained from different SSL models in identifying individual symptoms of depression.

2. SSL-models for Analysis

In this paper, we analyze several SSL-based speech models pretrained using different objective functions. The type of speech information, whether semantic, speaker, or prosodic, encoded

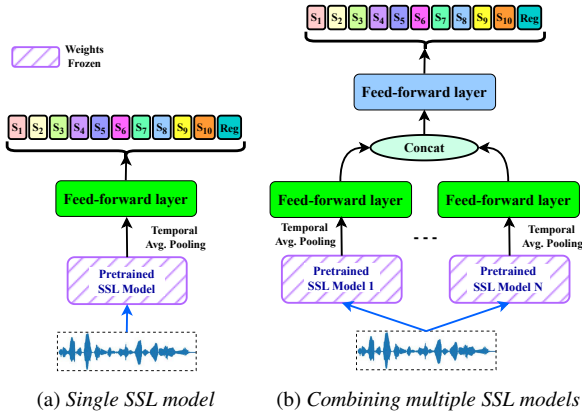


Figure 1: Schematic diagram of the symptom detection model using (a) single and (b) multiple SSL models (N models). In this work, $N = 2$ or 3 . Multi-task setting will have 10 symptom detection heads in the output layer (S_i for $1 \leq i \leq 10$) along with one regression (Reg) head. In single-task setting, there will be only one head (specific-symptom detection or regression).

by the SSL model depends on the pre-training objective function. Different pre-training objectives used to train SSL-based speech models include masked language modeling (MLM), contrastive pre-training, speaker-invariant training, and student-teacher training with cross-entropy loss [26–29]. Below, we explain the SSL models that we use in this paper.

- HuBERT [26] was pretrained using MLM which uses an offline clustering step to align target labels. This model predominantly encodes semantic information in speech.
- WavLM [30] was pretrained using MLM and denoising objectives. WavLM predominantly encodes semantics with some speaker-specific information. Here, we use the WavLM base model.
- BEATS [31] uses an iterative framework to optimize an acoustic tokenizer and an audio SSL model. The audio SSL model is optimized using MLM where the discrete labels are generated by the acoustic tokenizer. This objective allows BEATS to learn semantic-rich embeddings.
- ContentVec [28] uses a speaker-invariant training objective to suppress speaker information contained in the pretrained HuBERT model. ContentVec encodes semantic information while discarding speaker-specific information.
- RDINO [29] uses a regularized distillation with no labels (DINO) framework to pre-train models to encode speaker-specific information.
- AudioMAE [32] is an encoder-decoder model pretrained using MLM with a very high masking ratio – around 80% of frames masked. The pre-training task of reconstruction makes the model encode semantic, speaker, and prosodic information.
- BYOL-Audio [27] was pretrained by maximizing the cosine similarity between multiple augmented versions of the same speech signal. This model encodes the global information contained in speech, i.e., mainly speaker and prosodic.

Note that WavLM, BEATS, HuBERT, ContentVec, and RDINO take raw speech as input whereas BYOL-Audio and AudioMAE take spectrograms. As shown in Figure 1, we freeze all the weights of these pretrained models, and train only the feed-forward and output layers to detect the individual symp-

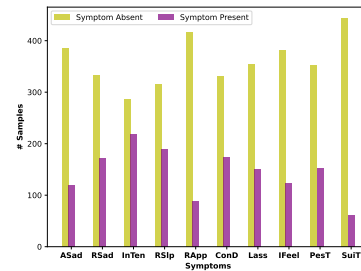


Figure 2: Distribution of samples between class-0 (symptom absent) and class-1 (symptom present) for each symptom in the MADRS. Symptom abbreviations are provided in Table 1.

toms of depression. We analyze SSL models using two different architectures: (1) single SSL model – here we use the embeddings of a single SSL model (Figure 1a), and (2) Multiple SSL models – here we combine embeddings of multiple SSL models (Figure 1b). Moreover, we evaluate each of this architecture in single-task and multi-task settings. For the multi-task setting, the output layer consists of 10 classification heads, each with two softmax units to detect each symptom, and a single linear unit for predicting the overall depression severity. For the single-task setting, we train a separate model for each symptom which consists of a single classification head as the output layer.

We also use conventional speech features such as spectrograms, COVAREP [33], and the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [19] as baselines. We extract 80-dimensional spectrograms, 88-dimensional eGeMAPS features using OpenSMILE [34], and 74-dimensional COVAREP features using COVAREP [33] to train convolutional neural network (CNN) models. The CNNs trained using conventional speech features consist of two convolutional layers, each with 100 channels. The kernels have sizes of 3 and 5 for the first and second layers, respectively. The outputs of the second layer are then passed through a fully connected layer with 100 sigmoid linear units (SiLU) [35]. Subsequently, a symptom-specific classification head with two softmax units follows the fully connected layer. Separate CNN models are trained for detecting each symptom. The output layer of the CNN model trained to predict overall depression severity consists of a linear unit.

Table 1: Individual symptoms of depression as defined in MADRS and their abbreviations, *Abbr.*

Symptom	Abbr.	Symptom	Abbr.
1) Apparent Sadness	ASad	6) Concentration Difficulties	ConD
2) Reported Sadness	RSad	7) Lassitude	Lass
3) Inner Tension	InTen	8) Inability to Feel	IFeel
4) Reduced Sleep	RSlp	9) Pessimistic Thoughts	PesT
5) Reduced Appetite	RApp	10) Suicidal Thoughts	SuiT

3. Dataset Details

We use an internal speech-based depression dataset which we call the Clinically-Labelled with INdividual symptoms of Depression (CLIND) dataset. CLIND consists of speech samples in English language collected from 505 participants (351 female and 154 male). Each participant was asked to speak about their experiences from the past few weeks with different prompts. The three prompts were designed to evoke neutral, positive, and negative contexts, respectively. The interviewer let the partici-

Table 2: **Performance of different self-supervised learning speech models in single task setting:** Comparison of the model performance (for each individual symptom of MADRS). Each cell contains F-scores in the form: macro F-score F_M (Symptom absent F_A , Symptom present F_P). RMSE shown for the regression task in predicting the overall depression severity score according to MADRS. Best performance (in terms of F_M) for each symptom is presented in bold. a-priori refers to the system that always predicts the output as the majority class i.e., symptom absent.

Symptoms	a-priori	WavLM	HuBERT	BEATS	ContentVec	RDINO	BYOL-Audio	AudioMAE
(1) ASad	43 (86, 0)	63 (84, 42)	60 (79, 41)	59 (81, 36)	57 (73, 41)	62 (81, 43)	57 (77, 37)	62 (83, 41)
(2) RSad	40 (79, 0)	72 (79, 65)	71 (74, 68)	68 (75, 61)	61 (64, 57)	69 (72, 66)	68 (74, 61)	68 (72, 64)
(3) InTen	36 (72, 0)	68 (71, 65)	68 (70, 66)	67 (69, 64)	58 (57, 59)	69 (70, 68)	64 (65, 63)	66 (64, 67)
(4) RSlp	38 (76, 0)	66 (75, 57)	67 (73, 61)	64 (73, 55)	58 (61, 55)	66 (73, 59)	62 (64, 59)	65 (73, 56)
(5) RApp	45 (90, 0)	47 (84, 10)	48 (87, 08)	52 (87, 18)	41 (75, 07)	47 (88, 06)	42 (79, 05)	53 (88, 17)
(6) ConD	40 (79, 0)	70 (80, 59)	72 (79, 65)	72 (82, 61)	65 (70, 61)	67 (74, 59)	69 (75, 63)	69 (77, 61)
(7) Lass	41 (82, 0)	68 (77, 58)	66 (78, 54)	61 (75, 47)	57 (64, 51)	64 (76, 52)	59 (73, 45)	65 (78, 52)
(8) IFeel	43 (86, 0)	68 (76, 59)	68 (77, 60)	63 (79, 46)	60 (63, 57)	73 (81, 64)	61 (77, 45)	66 (77, 55)
(9) PesT	41 (82, 0)	72 (80, 64)	72 (75, 70)	69 (73, 64)	65 (68, 62)	73 (70, 75)	74 (81, 66)	71 (73, 68)
(10) SuiT	47 (93, 0)	59 (82, 36)	60 (83, 36)	54 (79, 29)	55 (79, 31)	63 (82, 43)	60 (83, 37)	61 (82, 39)
MADRS(RMSE)	–	8.53	8.44	8.92	9.46	8.76	8.85	9.05

pant speak uninterrupted and only used a standardized prompt if the participant was silent for 30 seconds. At least 4 minutes of speech, including pauses, were recorded following each prompt, providing a total of approximately 10 minutes of speech per participant. Trained clinicians used structured interviews to score each participant’s current depression severity on MADRS, which consists of 10 individual symptoms (see Table 1), with each symptom scored in the range of 0 – 6, and a total score in the range of 0 – 60. The range of total MADRS scores in our dataset range from 0 to 47. In this study, we consider a symptom absent (Class-0) if the score for a symptom (S_s) is $0 \leq S_s \leq 1$, and a symptom to be present (Class-1) if $2 \leq S_s \leq 6$. The distribution of class-level samples for each symptom is illustrated in Figure 2. To train and test the models, we segment the speech samples into 10-second segments, resulting in a total of 25,471 segments (we segment after segregating the speech samples into train and test sets to avoid speaker overlap). We use the sample-level labels, encompassing both symptoms and overall depression severity scores, for all segments. Test performance is reported by majority voting across all segments of the test samples.

It is essential to note that the majority of publicly available depression datasets were labeled using self-reported scores like the Patient Health Questionnaire (PHQ-8) [20]. In contrast, our dataset is rated by trained clinicians using the MADRS, which is regarded as a gold standard for clinical assessment of depression severity.

Table 3: Comparing performance (in terms of F_M (F_A , F_P)) of conventional speech features to HuBERT (SSL-based model). ‘Spectro’ refers to spectrograms.

Symptoms	Spectro	COVAREP	eGeMAPS	HuBERT
(1) ASad	47(51, 43)	52(61, 43)	53(63, 43)	60(79, 41)
(2) RSad	60(63, 56)	58(64, 52)	59(65, 52)	71(74, 68)
(3) InTen	54(60, 48)	55(64, 46)	58(61, 54)	68(70, 66)
(4) RSlp	52(54, 50)	57(65, 49)	55(58, 52)	67(73, 61)
(5) RApp	42(72, 11)	41(77, 5)	42(74, 10)	48(87, 08)
(6) ConD	48(55, 41)	53(62, 44)	61(72, 50)	72(79, 65)
(7) Lass	46(53, 39)	48(59, 37)	50(66, 34)	66(78, 54)
(8) IFeel	56(66, 46)	53(70, 36)	54(72, 37)	68(77, 60)
(9) PesT	61(67, 54)	60(68, 52)	62(66, 59)	72(75, 70)
(10) SuiT	52(72, 32)	49(77, 21)	53(73, 32)	60(83, 36)
MADRS (RMSE)	10.81	10.18	9.82	8.44

4. Experiments

4.1. Model training

For each pretrained SSL model (see Figure 1), we freeze all its weights and extract embeddings. We perform average pooling on the embeddings across the temporal dimension before passing as input to a feed-forward layer comprising 100 SiLU. In the scenario of combining multiple SSL models, we pass the output embedding of each model through a 100-dimensional feed-forward layer with SiLU units before concatenating the embeddings. Subsequently, the concatenated output is passed through a 100-dimensional feed-forward layer with SiLU activation, followed by an output layer. We train both the feed-forward and output layers using the Adam optimizer with a learning rate of $1e-3$, a weight decay of $1e-5$, and a batch size of 32. Negative log-likelihood (NLL) and mean squared error (MSE) loss functions are utilized for training models on classification and regression tasks, respectively. All experiments are conducted using a single A40 GPU. Each model is trained for 5 epochs, and the results are reported using 5-fold cross-validation, where the folds are speaker independent (no overlap of speakers between folds). A randomly selected subset (10%) of the training set is allocated as the validation set for hyperparameter tuning.

We train CNNs on the conventional speech features using Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, an initial learning rate of 0.001, weight decay of $1e-5$, and a batch size of 32. Dropout rates of 0.3 and 0.4 were used for the convolutional and fully connected layers, respectively to avoid model overfitting. NLL and MSE loss functions were used to train models on classification and regression tasks, respectively.

4.2. Results

The performance of the models is evaluated in terms of F-scores. We provide three F-scores: F-score of the symptom being absent (F_A), F-score of the symptom being present (F_P), and the macro F-score (F_M), computed as $F_M = (F_A + F_P)/2$. We perform 5-fold cross validation, and provide the performance metrics as the mean across the 5-folds.

Table 2 provides the performance of different SSL models, in single-task setting, for each of the symptoms as defined in the MADRS. Models such as WavLM and HuBERT which encode semantic information along with some speaker and prosodic information achieved better performance on ASad, RSad, RSlp, ConD, and Lass symptoms. By contrast, RDINO, which en-

Table 4: **Comparing single-task (ST) and multi-task (MT) learning:** Performance of the model (in terms of macro F-score) for each individual item of MADRS. MADRS (RMSE) refers to the RMSE error in estimating total depression severity based on MADRS score.

Symptoms	WavLM		HuBERT		BEATS		ContentVec		RDINO		BYOL-Audio		AudioMAE		{CV,RD,BY}		{HB,RD,BY}	
	ST	MT	ST	MT	ST	MT	ST	MT	ST	MT	ST	MT	ST	MT	ST	MT	ST	MT
(1) ASad	63	65(↑)	60	59(↓)	59	58(↓)	57	57	62	64(↑)	57	57	62	63(↑)	62	63(↑)	65	66(↑)
(2) RSad	72	72	71	71	68	70(↑)	61	62(↑)	69	70(↑)	68	70(↑)	68	69(↑)	69	71(↑)	73	74(↑)
(3) InTen	68	68	68	67(↓)	67	66(↓)	58	56(↓)	69	71(↑)	64	65(↑)	66	64(↓)	68	67(↓)	70	70
(4) RSlp	66	66	67	67	64	65(↑)	58	59(↑)	66	66	62	62	65	66(↑)	67	67	70	69(↓)
(5) RApp	47	45(↓)	48	47(↓)	52	53(↑)	41	42(↑)	47	47	42	43(↑)	53	50(↓)	49	48(↓)	53	52(↓)
(6) ConD	70	71(↑)	72	73(↑)	72	72	65	62(↓)	67	66(↓)	69	70(↑)	69	71(↑)	72	72	74	75(↑)
(7) Lass	68	69(↑)	66	67(↑)	61	63(↑)	57	60(↑)	64	67(↑)	59	61(↑)	65	66(↑)	64	65(↑)	68	70(↑)
(8) IFeel	68	69(↑)	68	70(↑)	63	61(↓)	60	62(↑)	73	75(↑)	61	61	66	67(↑)	73	72(↓)	73	75(↑)
(9) PesT	72	74(↑)	72	75(↑)	69	71(↑)	65	63(↓)	73	75(↑)	74	75(↑)	71	69(↓)	77	77	79	79
(10) SuiT	59	62(↑)	60	63(↑)	54	59(↑)	55	57(↑)	63	65(↑)	60	64(↑)	61	63(↑)	67	68(↑)	70	71(↑)
MADRS (RMSE)	8.5	7.7(↓)	8.4	7.9(↓)	8.9	8.2(↓)	9.5	8.9(↓)	8.8	8.1(↓)	8.9	8.2(↓)	9.1	8.5(↓)	8.1	7.8(↓)	7.9	7.5(↓)

codes speaker information, achieved better performance on InTen, IFeel, and SuiT symptoms. Model BYOL-Audio, which encodes prosodic information, achieved better performance on PesT. ContentVec, which encodes semantic information and suppresses speaker information, achieved the lowest performance across all the models. These results suggest that semantic information alone, without speaker or prosodic information (as encoded by ContentVec), may not be sufficient to detect different symptoms of depression.

Table 3 compares the performance of conventional speech features (Spectrogram, COVAREP and eGeMAP) with the embeddings extracted from SSL-based speech models. SSL-based models outperformed the conventional speech features. Among the conventional features, eGeMAPS perform better than Spectrogram and COVAREP features for most symptoms. Tables 2 and 3 show that most of the SSL-based models outperform the conventional speech features across all symptoms. For the regression task, SSL models perform better than the conventional speech features with HuBERT achieving the best performance.

Table 4 compares performance between single-task and multi-task learning. The systems trained with multi-task learning perform comparably or better than the single task learning systems. Moreover, the computational load of multi-task learning is much lower than the single-task learning system. For all the models, the performance on the RApp (Reduced Appetite) symptom is very low. This could be due to the least correlation between the symptom RApp and speech [36].

Table 5 provides the performance when different SSL-based speech embeddings are combined. These model combinations are selected such that different types of information (semantic, speaker and prosodic) can be combined. The performances are provided in terms of macro F-scores (F_M), and absolute improvement in performance when compared to the best performing single model in the set is provided in parenthesis. Combining semantic (ContentVec) with speaker (RDINO) or prosodic information (BYOL-Audio), i.e., {CV,RD} or {CV,BY}, improves performance. But combining semantic with speaker and prosodic information ({CV,RD,BY} and {HB,RD,BY}) further improves performance for most of the symptoms, with {HB,RD,BY} achieving the best performance on most of the symptoms. This shows that the semantic, speaker, and prosodic information might all be essential for most of the tasks.

We employ the bootstrapping approach to compute confidence intervals for comparing SSL-based models with those trained using conventional features for each symptom [37].

Table 5: Performance (in terms of macro F-score) obtained by combining different SSL-based speech models. CV, RD, BY and HB refer to ContentVec, RDINO, BYOL-A, HuBERT, respectively. The absolute improvement in performance (in terms of macro F-score) compared to the best performing single model in the set is given in parenthesis.

Symptoms	{CV,RD}	{CV,BY}	{RD,BY}	{CV,RD,BY}	{HB,RD,BY}
(1) ASad	63 (1↑)	61 (4↑)	64 (2↑)	62	65 (3↑)
(2) RSad	70 (1↑)	68	70 (1↑)	69	73 (2↑)
(3) InTen	67 (2↓)	66 (2↑)	70 (1↑)	68 (1↓)	70 (1↑)
(4) RSlp	67 (1↑)	63 (1↑)	69 (3↑)	67 (1↑)	70 (3↑)
(5) RApp	48 (1↑)	44 (2↑)	52 (5↑)	49 (2↑)	53 (5↑)
(6) ConD	68 (1↑)	72 (3↑)	70 (1↑)	72 (3↑)	74 (2↑)
(7) Lass	63 (1↓)	61 (2↑)	64	64	68 (2↑)
(8) IFeel	72 (1↓)	63 (2↑)	72 (1↓)	73	73
(9) PesT	76 (3↑)	75 (1↑)	78 (4↑)	77 (3↑)	79 (5↑)
(10) SuiT	66 (3↑)	62 (2↑)	68 (5↑)	67 (4↑)	70 (7↑)
MADRS Score (RMSE)	8.29 (0.47↓)	8.35 (0.50↓)	8.20 (0.56↓)	8.12 (0.64↓)	7.92 (0.52↓)

The 95% confidence interval for the difference between group means do not contain zero, implying a significantly better performance for SSL-based models than conventional features. We observe this behaviour for all the symptoms except for the reduced appetite for which the performance of all the models is very low.

5. Conclusions

In this study, we explore the application of self-supervised learning (SSL)-based speech models for detecting individual symptoms of depression. We demonstrate that SSL embeddings yield huge performance improvements compared to conventional speech features such as spectrograms, eGeMAPS, and COVAREP. Upon comparing different SSL-based speech models, we find that models encoding predominant semantic information, combined with other information, demonstrate improved performance in symptoms such as apparent and reported sadness, concentration difficulties. Meanwhile, models encoding speaker and prosodic information excel in symptoms such as inability to feel, pessimistic thoughts, and suicidal tendencies. Importantly, combining multiple SSL-based embeddings that encode various speech aspects, enhances detection performance across most depressive symptoms, signifying the importance of semantic, speaker and prosodic information. Additionally, we show that multi-task learning, while being efficient, performs in-par or better than single-task learning.

6. References

- [1] J. Rehm and K. D. Shield, "Global burden of disease and the impact of mental and addictive disorders," *Current psychiatry reports*, vol. 21, no. 2, p. 10, 2019.
- [2] H. Herrman, V. Patel, C. Kieling, M. Berk, C. Buchweitz, P. Cuijpers, T. A. Furukawa *et al.*, "Time for united action on depression: a lancet–world psychiatric association commission," *The Lancet*, vol. 399, no. 10328, pp. 957–1022, 2022.
- [3] 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, 2015.
- [4] F. Ringeval, B. Schuller, M. Valstar, N. Cummins *et al.*, "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings Audio/visual Emotion Challenge and Workshop*, 2019, pp. 3–12.
- [5] M. Yamamoto, A. Takamiya, K. Sawada, M. Yoshimura, M. Kitazawa, K.-c. Liang, T. Fujita, M. Mimura, and T. Kishimoto, "Using speech recognition technology to investigate the association between timing-related speech features and depression severity," *PLoS one*, vol. 15, no. 9, 2020.
- [6] 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, 2020.
- [7] K. Dikaios, S. Rempel, S. H. Dumpala, S. Oore, M. Kieft, and R. Uher, "Applications of speech analysis in psychiatry," *Harvard Review of Psychiatry*, vol. 31, no. 1, pp. 1–13, 2023.
- [8] A. Harati, E. Shriberg, T. Rutowski, P. Chlebek, Y. Lu, and R. Oliveira, "Speech-based depression prediction using encoder-weight-only transfer learning and a large corpus," in *ICASSP*. IEEE, 2021, pp. 7273–7277.
- [9] S. P. Dubagunta, B. Vlasenko, and M. Magimai-Doss, "Learning voice source related information for depression detection," in *IEEE ICASSP*, 2019, pp. 6525–6529.
- [10] V. Ravi, J. Wang, J. Flint, and A. Alwan, "Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement," *Computer Speech & Language*, vol. 86, 2024.
- [11] 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 *ICASSP*. IEEE, 2021.
- [12] N. Seneviratne and C. Espy-Wilson, "Multimodal depression severity score prediction using articulatory coordination features and hierarchical attention text embeddings," *Interspeech*, 2022.
- [13] S. H. Dumpala, K. Dikaios, S. Rodriguez, R. Langley, S. Rempel, R. Uher, and S. Oore, "Manifestation of depression in speech overlaps with characteristics used to represent and recognize speaker identity," *Scientific Reports*, vol. 13, no. 1, p. 11155, 2023.
- [14] W. Wu, C. Zhang, and P. C. Woodland, "Self-supervised representations in speech-based depression detection," in *ICASSP*. IEEE, 2023, pp. 1–5.
- [15] S. Fara, O. Hickey, A. Georgescu, S. Gorla, E. Molimpakis, and N. Cummins, "Bayesian networks for the robust and unbiased prediction of depression and its symptoms utilizing speech and multimodal data," in *Proc. INTERSPEECH*, 2023, pp. 1728–1732.
- [16] S. Yadav, J. Chauhan, J. P. Sain, K. Thirunarayan, A. P. Sheth, and J. Schumm, "Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework," in *Proceedings of COLING*, 2020, pp. 696–709.
- [17] S. Fara, S. Gorla, E. Molimpakis, and N. Cummins, "Speech and the n-back task as a lens into depression. how combining both may allow us to isolate different core symptoms of depression," in *Interspeech*. ISCA, 2022, pp. 1911–1915.
- [18] T. Takano, D. Mizuguchi, Y. Omiya, M. Higuchi, M. Nakamura, S. Shinohara, S. Mitsuyoshi, T. Saito, A. Yoshino, H. Toda *et al.*, "Estimating depressive symptom class from voice," *International Journal of Environmental Research and Public Health*, vol. 20, no. 5, p. 3965, 2023.
- [19] F. Eyben, K. Scherer, B. Schuller *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, pp. 1–1, 01 2015.
- [20] 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.
- [21] J. B. Williams, K. A. Kobak, P. Bech, N. Engelhardt, K. Evans, J. Lipsitz, J. Olin, J. Pearson, and A. Kalali, "The grid-hamd: standardization of the hamilton depression rating scale," *International clinical psychopharmacology*, vol. 23, no. 3, pp. 120–129, 2008.
- [22] S. A. Montgomery and M. Åsberg, "A new depression scale designed to be sensitive to change," *The British Journal of Psychiatry*, 1979.
- [23] T. Rutowski, A. Harati, E. Shriberg, Y. Lu, P. Chlebek, and R. Oliveira, "Toward corpus size requirements for training and evaluating depression risk models using spoken language," in *INTERSPEECH*, 2022, pp. 3343–3347.
- [24] P. Zhang, M. Wu, H. Dinkel, and K. Yu, "Depa: Self-supervised audio embedding for depression detection," in *Proceedings of ACM international conference on multimedia*, 2021, pp. 135–143.
- [25] E. L. Campbell, J. Dineley, P. Conde, F. Matcham, K. M. White, C. Oetzmann, S. Simblett, S. Bruce, A. A. Folarin, T. Wykes *et al.*, "Classifying depression symptom severity: Assessment of speech representations in personalized and generalized machine learning models," in *INTERSPEECH*, 2023, pp. 1738–1742.
- [26] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.
- [27] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for Audio: Exploring pre-trained general-purpose audio representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, p. 137–151, 2023.
- [28] K. Qian, Y. Zhang, H. Gao, J. Ni, C.-I. Lai, D. Cox, M. Hasegawa-Johnson, and S. Chang, "Contentvec: An improved self-supervised speech representation by disentangling speakers," in *ICML*. PMLR, 2022, pp. 18 003–18 017.
- [29] Y. Chen, S. Zheng, H. Wang, L. Cheng, and Q. Chen, "Pushing the limits of self-supervised speaker verification using regularized distillation framework," in *ICASSP*. IEEE, 2023, pp. 1–5.
- [30] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [31] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," in *ICML*, 2023.
- [32] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," *NeurIPS*, vol. 35, pp. 28 708–28 720, 2022.
- [33] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep— a collaborative voice analysis repository for speech technologies," in *IEEE ICASSP*, 2014, pp. 960–964.
- [34] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. ACM conference on Multimedia*, 2010, pp. 1459–1462.
- [35] S. Elfving, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural networks*, vol. 107, pp. 3–11, 2018.
- [36] R. Uher, A. Farmer, W. Maier *et al.*, "Measuring depression: comparison and integration of three scales in the gendep study," *Psychological medicine*, vol. 38, pp. 289–300, 2008.
- [37] L. Ferrer and P. Riera, "Confidence intervals for evaluation in machine learning [computer software]," Github, 2022. [Online]. Available: <https://github.com/luferrer/ConfidenceIntervals>