



# Test-Time Training for Speech-based Depression Detection

Sri Harsha Dumpala, Chandramouli S. Sastry, Rudolf Uher, Sageev Oore

Dalhousie University, Canada

{sriharsha.d, cssastry, uher, sageev}@dal.ca

## Abstract

Previous works on speech-based depression detection typically use datasets collected in similar environments for both training and testing the models. However, in practice, the training and testing distributions often differ. Distributional shifts in speech can result from various factors, such as differences in recording environments (e.g., background noise) and demographic attributes (e.g., gender, age). These shifts can significantly degrade the performance of depression detection models. In this paper, we analyze the application of test-time training (TTT) to improve the robustness of depression detection models against such shifts. Our results demonstrate that TTT can substantially enhance model performance under various distributional shifts, including those caused by (a) background noise, (b) gender bias, and (c) differences in data collection and curation procedures, where training and testing samples originate from different datasets.

**Index Terms:** distributional shifts, test-time training, depression, masked autoencoders, self-supervised models

## 1. Introduction

Depression is one of the most common mental health disorders and a leading cause of disability worldwide [1]. According to the World Health Organization [2], depressive disorders are highly prevalent globally, yet remain largely under-detected and under-treated [3]. Extensive screening and early diagnosis are crucial to controlling the prevalence of depression. Inadequate accessibility to clinical services, further exacerbated by associated stigma impedes early detection. Automated assessment systems can aid in early detection, enabling individuals to seek timely professional help. To enable the widespread adoption of automated depression assessment systems, it is important to develop reliable and robust systems, which is the focus of this work.

Recent research highlights natural speech as a cost-effective and scalable indicator for depression assessment [4, 5, 6]. Following standard machine learning practices, most prior studies have focused on audio models trained and tested under similar conditions i.e., under matching training and testing conditions. However, in practice, training and testing distributions often differ, leading to distributional shifts between the training and testing data. Such shifts in speech can arise due to (a) inter-speaker variations, such as speaking style, gender, and age; and (b) recording environment, which can introduce various background noises, such as babble, living room sounds, or traffic. These shifts can severely degrade even state-of-the-art deep learning models performance [7, 8, 9, 10], especially given the limited availability of large-scale depression datasets. Additionally, large-scale depression datasets that cover multi-

ple distributional shifts are not only expensive and challenging to acquire but also cannot guard against shifts not represented in the training data. To address this, we aim to enhance the robustness of depression detection models using test-time training (TTT).

TTT [11, 12, 13] has recently been studied in applications such as image classification and has shown improved robustness against various (unseen) distribution shifts. In TTT, a portion of the model parameters is updated based on a self-supervised loss objective (i.e., without requiring any labels) using only the test sample. These updates are specific to each test sample and are reset after the prediction is made.

The effectiveness of TTT depends on the self-supervised learning (SSL) task, and research [12], [13] demonstrates that masked auto-encoding is a suitable SSL objective that offers reliable improvements against distribution shifts. Motivated by the success of transformer-based masked autoencoders (MAE) for speech [14], we extend a TTT approach based on MAE [13, 15] to depression detection in this work.

To the best of our knowledge, this is the first study to apply test-time adaptation to depression detection. We show that TTT-MAE achieves significant improvements under various distributional shifts in depression detection. Specifically, we experiment with the following types of distributional shifts in this work: 1) **Background noise:** Models trained on clean speech but tested on speech corrupted with background noise; 2) **Gender-bias:** Models trained on speech samples from female speakers and evaluated on speech samples from male speakers, and vice versa; 3) **Dataset:** Models where the training and testing data are obtained from different datasets. The distribution shift between these samples is not easily defined and is rooted in the data collection and curation processes.

## 2. Related Work

**Depression detection under distributional shifts:** Most previous studies on depression detection have used speech recordings collected in controlled laboratory settings. Deploying depression screening systems in natural environments, such as through smartphones, remains a challenge and is relatively under-explored. A few studies have shown that depression detection systems experience significant performance degradation when there is a distributional shift between the training and testing sets [16, 17, 18, 19, 20, 21]. For instance, SVMs trained on data from one country exhibited substantial degradation when tested on data from another, despite using the same language (English) [18]. Similarly, SVMs trained on data from one set of smartphones and tested on another set showed significant performance drops, with further degradation observed due to variations in gender and speech elicitation methods [19]. Addition-

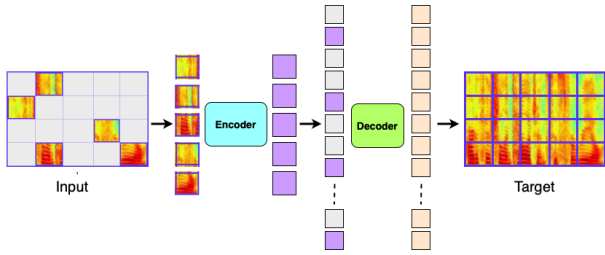


Figure 1: Schematic outline of AudioMAE pre-training.

ally, models trained on speech data from the general population (ages 18 to 65) performed poorly when evaluated on an older population (ages 45 to 75), even when using a large dataset and a pre-trained CNN-based acoustic model [20]. Variations in treatment type also resulted in performance degradation, as seen when dilated CNN models trained on datasets from participants with and without pharmacotherapy or psychotherapy treatment were tested [21].

These studies demonstrate the vulnerability of simple depression detection systems, such as SVMs, CNNs, and dilated CNNs trained on conventional acoustic features, to distributional shifts. In this work, we evaluate SSL-based models for their robustness to distributional shifts in the context of depression detection. While self-supervised training on large unlabelled datasets makes these models more robust than those using conventional acoustic features, significant performance degradation can still occur under distributional shifts.

Robustness to distributional shifts is an underexplored topic in depression detection. In [22], domain adaptation techniques were applied to improve performance in cross-corpus testing. Domain adaptation involves training the model to generalize across various distributional shifts. However, anticipating every possible shift during training is impractical, especially in real-world applications. These models remain fixed during inference, even when the test distribution changes. In this work, we extend TTT techniques for depression detection to address distributional shifts.

**Pre-trained models for depression detection:** Recent studies have explored pre-trained models, particularly SSL models, for depression detection [23, 24, 25, 26, 27]. However, all of these studies evaluated SSL-based depression detection under matched train-test conditions. None have assessed these models under mismatched conditions. In this paper, we aim to close this gap by analyzing SSL speech models such as Wav2Vec 2.0 [28], HuBERT [29], WavLM [30], and AudioMAE [14] when evaluated with distributionally shifted test instances.

**Test-time training (TTT):** The basic paradigm in TTT [11] involves using a test-time task, typically a self-supervised learning task, alongside the main task during training. The pre-trained model is then updated using test data with the self-supervised test-time objective before making the final prediction. In vision tasks, various self-supervised tasks for TTT include rotation prediction [11], contrastive loss [12], and masked autoencoding [13]. TTT-MAE was later extended to speech-based tasks such as speaker recognition, emotion classification, and short word detection [15]. In this work, we extend the TTT-MAE framework to depression detection.

### 3. Method

**Pre-training MAE:** In this paper, we use the audio masked autoencoder (AudioMAE) [14], which is pre-trained to recon-

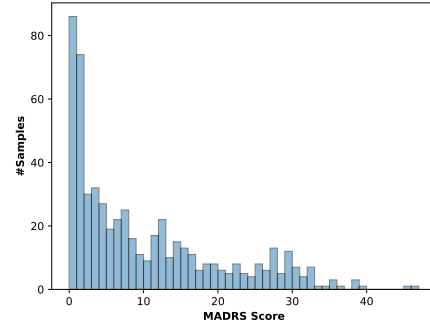


Figure 2: Distribution of VMD samples with respect to depression severity as measured by MADRS.

struct masked patches of speech Mel-spectrograms using an asymmetrical encoder-decoder architecture. Below is a brief overview of the MAE pre-training process (see Figure 1).

The input speech waveform is first transformed into Mel-spectrograms, which are then divided into non-overlapping grid patches. These patches are flattened and embedded through a linear projection layer. Fixed sinusoidal positional embeddings are added to provide positional information. Next, 80% of the patches are randomly masked while retaining the positional indices of all patches to enable the decoder to reconstruct the spectrogram. The encoder processes only the unmasked patches to generate latent representations. The decoder then attempts to reconstruct the original spectrogram using the encoder’s latent representations and the masked patches, organized in their original order. The training objective is to minimize the mean squared error (MSE) between the reconstructed and input spectrograms, averaged over the masked patches. For all experiments in this paper, we use the pre-trained AudioMAE model released by [14].

#### 3.1. TTT using Pre-trained AudioMAE

**Architecture:** Similar to [11, 13], we employ a Y-shaped architecture: a shared encoder network  $e$  followed by two heads, a self-supervised head  $g$  and a depression detection head  $d$ . In this work,  $e$  and  $g$  are the encoder and decoder networks of the pre-trained AudioMAE, respectively.  $d$  comprises of two feed-forward layers with SiLU activation, followed by a softmax layer for depression detection.

TTT involves two levels of training:

**1) Train-time training:** In train-time training, we train the model using labeled data for the downstream task—in this case, depression detection. During this phase, only the encoder  $e$  is used, and  $g$  is discarded. The encoder’s weights are frozen and used as a feature extractor, while only  $d$  is trained. Here, the latent representations generated by  $e$  serve as input to  $d$ .

**2) Test-time training (TTT):** At test time, we start with the trained depression detection head and the pre-trained AudioMAE encoder  $e_0$  and decoder  $g_0$ . For each test input, while keeping the decoder’s weights frozen, we optimize only the encoder using the self-supervised loss function used for pre-training AudioMAE—minimizing the MSE between the reconstructed and input masked spectrograms. During test-time, all parameters of the shared encoder are updated for  $t$ -steps, from  $e_0$  to  $e_t$ , to minimize the MSE across various augmentations of a single test sample. After this process, we use  $e_t$  and  $d$  to make a decision on the test sample.

Table 1: Comparison of AudioMAE-TTT with non-TTT methods (AudioMAE, AudioMAE-FT, other SSL methods such as Wav2Vec, HuBERT, WavLM, and conventional speech features like COVAREP and eGeMAPS) under various shifts caused by background noise (noises added at 5 dB). AudioMAE-TTT significantly outperforms all non-TTT approaches across these distributional shifts. Each cell contains F-scores in the form:  $F_M (F_H, F_D)$ .

Model	Clean	AWGN	AC	Babble	Living Room	Park	Reverberation	Traffic	Average
COVAREP	59.2 (71.8, 46.6)	35.8 (68.2, 3.6)	38.7 (56.1, 20.8)	33.2 (66.4, 0.0)	40.4 (62.3, 18.1)	41.9 (68.6, 15.1)	40.7 (54.5, 26.8)	49.6 (68.6, 30.7)	42.4 (64.6, 20.2)
eGeMAPS	63.5 (72.4, 54.7)	35.1 (61.4, 8.8)	40.6 (53.6, 27.4)	33.2 (66.4, 0.0)	42.1 (61.5, 22.6)	43.5 (69.4, 17.6)	44.6 (57.9, 31.3)	51.4 (66.7, 36.1)	44.3 (63.7, 24.8)
Wav2Vec 2.0	67.7 (68.5, 66.8)	37.2 (56.8, 17.5)	52.4 (67.3, 36.9)	35.8 (65.5, 4.1)	50.7 (68.9, 32.5)	51.5 (68.1, 34.7)	48.2 (59.2, 37.2)	54.3 (70.4, 38.2)	49.7 (65.7, 33.5)
HuBERT	69.8 (70.3, 69.1)	40.6 (52.3, 28.6)	54.1 (65.7, 42.5)	36.2 (65.1, 7.3)	52.0 (54.3, 49.6)	53.0 (60.3, 46.2)	49.1 (58.5, 39.7)	55.9 (63.4, 48.4)	51.4 (61.3, 41.4)
WavLM	70.9 (73.2, 68.5)	38.6 (51.1, 26.2)	57.4 (67.5, 47.1)	38.3 (63.7, 12.6)	53.2 (68.4, 38.1)	55.8 (68.9, 42.6)	50.7 (63.4, 38.1)	57.2 (67.3, 47.1)	52.7 (65.3, 40.1)
AudioMAE	69.4 (70.7, 68.2)	39.7 (55.6, 24.4)	55.2 (63.2, 46.9)	35.9 (63.3, 8.4)	51.8 (61.3, 42.4)	53.4 (62.2, 44.6)	48.8 (60.5, 37.2)	56.3 (66.4, 46.1)	51.3 (62.8, 39.7)
AudioMAE-FT	71.1 (73.6, 68.6)	39.2 (57.1, 21.2)	53.6 (62.8, 44.4)	35.5 (64.5, 6.4)	51.1 (60.7, 41.5)	52.3 (65.1, 39.7)	47.5 (59.4, 35.6)	55.8 (67.1, 44.5)	50.8 (63.7, 37.8)
AudioMAE-TTT	<b>71.4 (72.5, 70.3)</b>	<b>52.5 (62.8, 41.7)</b>	<b>62.1 (60.4, 63.7)</b>	<b>47.3 (68.5, 26.3)</b>	<b>58.6 (64.2, 53.1)</b>	<b>60.3 (63.7, 57.2)</b>	<b>59.4 (67.6, 51.3)</b>	<b>63.4 (68.5, 58.3)</b>	<b>59.5 (66.2, 52.7)</b>

## 4. Dataset Details

In this work, we use two different depression datasets:

1) Vocal Mind dataset (VMD): VMD consists of speech samples collected from 559 (401 female and 158 male) participants. Each participant was prompted with three different prompts to speak about their experiences from the past few weeks. The three prompts were designed to evoke neutral, positive, and negative context, respectively. Each participant spoke uninterruptedly for at least 3 minutes per prompt, resulting in approximately ten minutes of speech per participant. Depression severity of each speech sample was scored on the Montgomery and Asberg Depression Rating Scale (MADRS) [31], which is in the range of 0 – 60. The range of total MADRS scores in our dataset range from 0 to 47. Participants with MADRS  $\geq 10$  are classified as depressed, and the remaining are classified as non-depressed (healthy). The distribution of the samples is shown in Figure 2. The dataset is divided into train and test set with 417 (307 female and 110 male) and 142 (94 female and 48 male) recordings, respectively.

2) DAIC-WoZ corpus [32]: We use the DAIC-WoZ dataset for cross-data evaluation, which includes 219 clinical interviews labeled with PHQ-8 scores (0-24) to indicate depression severity. Participants with PHQ-8 scores  $\geq 10$  were classified as depressed, and those with PHQ-8 scores  $< 10$  as healthy. We use the train, validation and test splits of DAIC-WOZ as defined in [33, 32]

Manual transcripts with timestamps of the DAIC-WOZ and VMD datasets were used to discard the interviewer speech and retain only the participant’s speech. To train and test the models, we segment each speech recording into 7-second segments. The depression label of each segment is same as the label of the entire speech sample. Performance is reported by using majority voting across all segments of test samples.

## 5. Experiments

**Models for comparison:** In this work along with AudioMAE, we will evaluate the performance of following SSL-based speech models for depression detection under different distributional shifts: (1) Wav2Vec 2.0 [28], (2) HuBERT [29] and (3) WavLM [30]. We freeze the weights of the pre-trained models and train a one-layer fully connected neural network (with 100 sigmoid linear units (SiLU) [34]) with an output softmax layer on top of these models for depression detection.

We also evaluate CNN models trained using conventional speech features which include COVAREP [35] and eGeMAPS [36] as baselines. These CNN models comprise two convolutional layers, each with 100 channels. The kernels have sizes of 4 and 5 for the first and second layers, respectively. Outputs from the second convolutional layer are flattened before passing through a fully-connected layer with 100 units and

an output layer. We extract 88-dimensional eGeMAPS and 74-dimensional COVAREP using OpenSMILE [37] and COVAREP toolkits, respectively.

**Model training:** To train models (both TTT and non-TTT) on the depression detection task, we freeze all the weights of the pre-trained model and only train the feed-forward neural network and the output layer. We train 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. We use Negative log-likelihood as the training objective. Each model is trained for 5 epochs.

**TTT at inference for AudioMAE-TTT:** Following [13, 15], for each test sample, we train only the encoder (freezing the decoder weights) for 20 steps with a batch-size of 128 using SGD optimizer with a fixed learning rate of  $2.5e-3$ , momentum of 0.9 and weight decay of 0.2. During TTT, we follow similar procedure as pre-training: mask 80% of the input patches and provide the unmasked patches as input to the encoder whereas all the patches are provided as input to the decoder. The encoder weights are then updated to optimize the reconstruction loss (MSE) over the masked patches. We note that we do not use any augmentation beyond random masking for TTT. We performed experiments using 4 Nvidia A40 48GB GPUs.

We train CNNs on the conventional speech features using Adam optimizer with a 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. A randomly selected subset (10%) of the training set is allocated as validation set for selecting the model hyperparameters.

**Results:** We report the performance of the models in terms of macro F-score ( $F_M$ ) which is computed as  $F_M = (F_1(H) + F_1(D))/2$ , where  $F_1(H)$  and  $F_1(D)$  are the F-scores of the healthy class and the depressed class, respectively. Unless specified otherwise, we report results for AudioMAE-TTT after 20 TTT steps.

Table 1 compares the performance of TTT (AudioMAE-TTT) with no-TTT (Wav2Vec 2.0, HuBERT, WavLM) under different distributional shifts caused due to background noises. To evaluate models’ performance under different distributional shifts, we introduce diverse background noises sourced from Microsoft’s Scalable Noisy Speech Dataset (MS-SNSD) [38]. These noises are exclusively added during the testing phase and are not used in pre-training or train-time training of the self-supervised models. When trained and tested with clean speech, non-TTT methods such as SSL-based models (Wav2Vec 2.0, HuBERT, WavLM and AudioMAE) and models trained using conventional speech features (COVAREP and eGeMAPS), show significant degradation in performance with SSL-based models performing better than conventional speech models. Whereas, AudioMAE with TTT (AudioMAE-TTT) achieves

Table 2: *Distributional shift due to gender variations. Performance (in  $F_M$ ) when models are trained on one gender and tested on the other. AudioMAE-TTT significantly outperforms all other methods.*

Train set	Female		Male	
Test set	Female	Male	Male	Female
Wav2Vec 2.0	70.6	51.8	71.6	53.3
HuBERT	71.8	52.3	72.3	54.2
WavLM	<b>73.6</b>	47.3	<b>73.1</b>	52.9
AudioMAE	72.2	49.6	71.8	51.8
AudioMAE-TTT	73.4	<b>63.1</b>	71.8	<b>62.2</b>

best performance under all distribution shifts, with lower degradation in performance compared to testing with clean speech. It is interesting to see that Audio-FT (where we finetune the encoder of AudioMAE along with the depression detection head for depression detection task) performs inferior to AudioMAE (where we freeze the weights of the AudioMAE encoder). This is in agreement with the findings in [39].

Tables 2 and 3 compare the performance of TTT with non-TTT techniques under (a) gender-based and (b) dataset-based distributional shifts, respectively. For gender-based experiments (see Table 2), we train the models with all speech samples from the same gender (Female or Male) and test the models on samples other gender (Male or Female). While non-TTT techniques show significant performance degradation for cross-gender testing compared to same gender testing, AudioMAE-TTT shows relatively low degradation in performance.

For the case of cross-dataset testing (see Table 3), we train the models with one dataset (VMD or DAIC) and test the datasets with another dataset (DAIC or VMD), where VMD is spontaneous speech whereas DAIC is interview-based speech. While non-TTT techniques show significant performance degradation, AudioMAE-TTT outperforms all the other methods with relatively very low degradation in performance.

Figure 3 shows how the performance of AudioMAE-TTT varies with the number of TTT steps at inference. We show performance curves for 20 TTT steps, when tested with distributional shifts due to background noises. We can see that the performance of TTT improves as we increase the number of TTT steps. But just a few TTT steps are sufficient to achieve significant improvements in performance. In this paper, we report most of the results after 20 TTT steps.

We evaluate the statistical significance of the improvements obtained by TTT by computing the confidence intervals be-

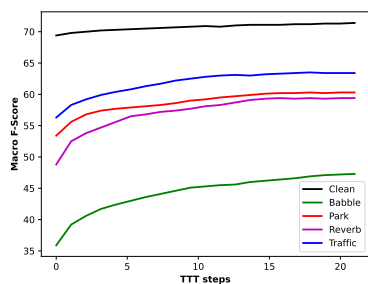


Figure 3: *Performance (in terms of  $F_M$ ) of AudioMAE-TTT across TTT steps. Performance of TTT improves with the number of steps. In this paper, we report results after 20 TTT steps.*

Table 3: *Distributional shift due to dataset variations. Performance (in  $F_M$ ) when models are trained on one dataset and tested on another dataset (cross-dataset evaluation). AudioMAE-TTT significantly outperforms all other methods.*

Train Dataset	VMD		DAIC	
Test Dataset	VMD	DAIC	DAIC	VMD
Wav2Vec 2.0	67.7	37.9	60.4	41.2
HuBERT	69.8	40.6	64.2	43.7
WavLM	70.9	41.5	<b>66.7</b>	44.8
AudioMAE	69.4	39.4	63.1	44.2
AudioMAE-TTT	<b>71.4</b>	<b>48.7</b>	65.9	<b>56.8</b>

tween TTT and non-TTT approaches [40]. For the 95% confidence intervals, we observed no significant overlap between TTT and non-TTT approaches. Furthermore, the confidence interval for the difference in metrics does not contain 0, implying a statistically significant difference in performance between the two types of models.

Table 4 highlights the impact of the masking ratio selection during TTT. Higher masking ratios can lead to significant performance degradation. In this study, we set the masking ratio to 75%, as it provides a good balance between performance and computational efficiency.

A key limitation of TTT is the additional computational overhead incurred during inference. This overhead can potentially be alleviated by adopting higher masking ratios, reducing the number of TTT steps, and possibly by utilizing parameter-efficient fine-tuning techniques—need further exploration in future work.

Table 4: *Effect of masking ratio during TTT on performance. Steps provides the performance after 10 steps of TTT. Reverb. refers to reverberation noise.*

Mask ratio	AC	Babble	Park	Reverb.	Traffic
50%	61.4	<b>45.7</b>	<b>59.5</b>	<b>58.6</b>	62.6
75%	<b>62.1</b>	45.3	59.2	58.1	<b>62.8</b>
90%	55.6	39.2	52.3	51.4	57.2

## 6. Conclusions

Robust and reliable depression detection in the presence of distribution shifts is a challenging problem for both conventional deep neural networks and self-supervised models pre-trained over large datasets. We discuss and evaluate test-time training technique as a solution to achieve robust depression detection even on distributionally shifted instances. In summary, we consider the following categories of distribution shifts: (a) background noise, (b) gender-biased training data, (c) cross-corpus generalization. While the factors of distribution shift are different in each of these cases, we consistently observe that TTT enables robust identification across all evaluations over both in-distribution testing samples and distributionally-shifted instances.

## 7. 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] W. H. O. WHO, "World mental health report: transforming mental health for all," Website, 2022. [Online]. Available: <https://iris.who.int/bitstream/handle/10665/356119/9789240049338-eng.pdf?sequence=1>
- [3] H. Herrman, V. Patel, C. Kieling, M. Berk *et al.*, "Time for united action on depression: a lancet–world psychiatric association commission," *The Lancet*, vol. 399, no. 10328, pp. 957–1022, 2022.
- [4] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder *et al.*, "A review of depression and suicide risk assessment using speech analysis," *Speech communication*, vol. 71, pp. 10–49, 2015.
- [5] 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.
- [6] K. Dikaos, 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.
- [7] T. Likhomanenko, Q. Xu, V. Pratap *et al.*, "Rethinking evaluation in ASR: are our models robust enough?" in *Interspeech*, 2021.
- [8] D. Garcia-Romero, D. Snyder, S. Watanabe, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition benchmark using the chime-5 corpus." in *Interspeech*, 2019, pp. 1506–1510.
- [9] J. Parry, D. Palaz, G. Clarke, P. Lecomte *et al.*, "Analysis of deep learning architectures for cross-corpus speech emotion recognition." in *Interspeech*, 2019, pp. 1656–1660.
- [10] C. Botelho, T. Schultz, A. Abad, and I. Trancoso, "Challenges of using longitudinal and cross-domain corpora on studies of pathological speech." in *INTERSPEECH*, 2022, pp. 1921–1925.
- [11] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *ICML*, 2020, pp. 9229–9248.
- [12] Y. Liu, P. Kothari, B. Van Delft, B. Bellot-Gurlet, T. Mordan, and A. Alahi, "Ttt++: When does self-supervised test-time training fail or thrive?" *NeurIPS*, vol. 34, pp. 21 808–21 820, 2021.
- [13] Y. Gandelsman, Y. Sun, X. Chen, and A. A. Efros, "Test-time training with masked autoencoders," in *NeurIPS*, 2022.
- [14] 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.
- [15] S. H. Dumpala, C. S. Sastry, and S. Oore, "Test-time training for speech," in *ICML workshop on Efficient Systems for Foundation Models*, 2023.
- [16] V. Mitra, E. Shriberg, D. Vergyri, B. Knoth, and R. M. Salomon, "Cross-corpus depression prediction from speech," in *ICASSP*. IEEE, 2015, pp. 4769–4773.
- [17] M. Gerczuk, S. Amiriparian, A. Kathan, J. Bauer, M. Berking, and B. W. Schuller, "Noise robust recognition of depression status and treatment response from speech via unsupervised feature aggregation," in *EMBC*. IEEE, 2023, pp. 1–4.
- [18] S. Alghowinem, R. Goecke, J. Epps, M. Wagner, and J. F. Cohn, "Cross-cultural depression recognition from vocal biomarkers." in *Interspeech*, 2016, pp. 1943–1947.
- [19] Z. Huang, J. Epps, D. Joachim, and M. Chen, "Depression detection from short utterances via diverse smartphones in natural environmental conditions." in *INTERSPEECH*, 2018, pp. 3393–3397.
- [20] 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.
- [21] N. Seneviratne and C. Espy-Wilson, "Generalized dilated cnn models for depression detection using inverted vocal tract variables," *arXiv preprint arXiv:2011.06739*, 2021.
- [22] Z. Huang, J. Epps, D. Joachim, B. Stasak, J. R. Williamson, and T. F. Quatieri, "Domain adaptation for enhancing speech-based depression detection in natural environmental conditions using dilated cnns." in *INTERSPEECH*, 2020, pp. 4561–4565.
- [23] 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.
- [24] 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.
- [25] S. H. Dumpala, C. S. Sastry, R. Uher, and S. Oore, "On combining global and localized self-supervised models of speech," in *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022.
- [26] S. H. Dumpala, K. Dikaos, 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.
- [27] W. Wu, C. Zhang, and P. C. Woodland, "Self-supervised representations in speech-based depression detection," in *ICASSP*. IEEE, 2023, pp. 1–5.
- [28] 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. 34, 2020.
- [29] 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.
- [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. A. Montgomery and M. Åsberg, "A new depression scale designed to be sensitive to change." *The British Journal of Psychiatry*, 1979.
- [32] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews," in *Proc. International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 3123–3128.
- [33] 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*, 2016, pp. 3–10.
- [34] 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.
- [35] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *ICASSP*. IEEE, 2014, pp. 960–964.
- [36] 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.
- [37] 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.
- [38] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A scalable noisy speech dataset and online subjective test framework," *arXiv preprint arXiv:1909.08050*, 2019.
- [39] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," in *ICLR*, 2022.
- [40] L. Ferrer and P. Riera, "Confidence intervals for evaluation in machine learning [computer software]," Github, 2022. [Online]. Available: <https://github.com/luferrer/ConfidenceIntervals>