EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY

Open Access

Estimating depression and anxiety scores from conversational speech in females with and without comorbidity

Aslı Beşirli[1*], Tuğrahan Karakadıoğlu[2†] and Cenk Demiroğlu[2†]

Abstract

Major depressive disorder and generalized anxiety disorder together account for a substantial share of global disability. They occur more frequently in women and their frequent co-occurrence complicates both diagnosis and treatment. To investigate how comorbidity affects speech-based assessment, we collected conversational speech data covering both negative and positive topics from 130 Turkish women at a psychiatric clinic. Each patient was diagnosed with depression, anxiety, or comorbid depression and anxiety by a psychiatrist. From each recording, we extracted classic acoustic descriptors (MFCCs and the AVEC-2013 feature set) alongside learned embeddings from three pretrained self-supervised models—HuBERT, WavLM, and wav2vec2—each fine-tuned with a lightweight regression head. The LoRA fine-tuned wav2vec2 achieved the best accuracy (depression: RMSE 8.44, MAE 6.84; anxiety: RMSE 12.49, MAE 10.71), surpassing WavLM, HuBERT, and all baseline feature sets while using significantly fewer trainable parameters. To evaluate architectural variants, we experimented with five alternative heads on the wav2vec2 backbone. Recordings in response to negative question consistently performed better. Comorbid depression and anxiety cases were assessed more accurately than depression-only cases (RMSE 7.34 vs. 8.03), underscoring the value of modeling clinical heterogeneity. Together, our research shows that co-occurring anxiety enhances the acoustic salience of depressive symptoms.

Keywords Depression assessment, Comorbid anxiety and depression, Acoustic features, Self-supervised speech models

1 Introduction

Depression has become a major global public health concern. The World Health Organization reports that around 280 million people across the world are affected by depression [1]. Similarly, anxiety is another leading mental health disorder with 301 million people living with an anxiety disorder. Moreover, according to the same report 75% of people with mental health conditions receive no treatment in low- and middle-income countries. Thus, automatic and low cost assessment of depression and anxiety and their co-occurrence emerged as an active research area.

Extracting speech features for depression assessment has been studied extensively in the literature. In addition to the popular MFCC features [2], there are hand-crafted features such as GeMAPS [3], COMPARE [4]. Prosodic features such as pitch, tone, rhythm, speaking rate, and pausing are important for understanding emotional states and mental health conditions [5]. Similarly, vocal source features such as jitter and shimmer were also shown to be depression indicators [5].

In depressive disorders, neuromotor outputs slow down, which causes changes in the coordination of

†Tuğrahan Karakadıoğlu and Cenk Demiroğlu contributed equally to this work.

*Correspondence:
Aslı Beşirli
besirliasli@gmail.com
[1] Department of Psychiatry, University of Health Sciences, Şişli Hamidiye Etfal Training and Research Hospital, Istanbul, Turkey
[2] Department of Engineering, Ozyegin University, Istanbul, Turkey

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

articulatory features during speech. Articulatory Coordination Features (ACFs) detects changes in the timing of speech gestures [6] for depression assessment.

Most of the features in the literature are focused on using amplitude spectrogram. However, phase also contains information about the state of the speaker. In [7], authors use a Complex Squeeze-and-Excitation Network where complex spectrogram is used as input features.

Feature extraction using pretrained deep learning models has been getting increasingly more attention in the literature. In [8], a pretrained convolutional neural network (CNN), trained for image classification, was used as feature extractor for depression detection. In [9], an x-vector extraction network, which was originally developed for speaker verification, is used for depression detection. In [10], self-supervised contrastive learning was used for detecting depression. The pretrained CNN10 network was trained for emotion detection, which is then used as a guide for contrastive learning to select positive and negative samples.

For modeling depressed speech, deep learning techniques were commonly used. Convolutional neural networks are one of the widely used architectures [11], typically operating on spectrogram representations of speech signals, or on cepstrograms when MFCC features are employed. Dilated convolution, which increases the receptive field size, was used in [12]. A multi-task training approach was used in [13] where the CNNs were used to estimate the depression category, depression level, emotion and sentiment. Similarly, LSTM networks were used successfully [14, 15]. To address the vanishing gradients problem with long sequences, attention networks were used in combination with LSTMs [16, 17].

CNN and LSTM networks were also used together for depression assessment [18]. In those hybrid networks, CNNs were used to extract spatial features, while LSTMs were used to model temporal dependencies. In [19], CNNs were used to extract features from the spectrograms, and LSTMs then processed these features to model the temporal dynamics of speech. In [15], segment-level depression severity estimation was performed using dilated CNNs, and LSTMs were used.

There are other hybrid architectures that use CNNs, LSTMs, and the attention mechanism together. In [20], CNNs are used together with a self-attention mechanism. In [21], LSTM and CNN networks are used together and their outputs are weighted based on importance using self-attention. Similarly, in [22], time-frequency attention was used to focus on the most relevant parts of the input data.

One of the problems with depression assessment is the lack of sufficient data to train complex networks. To overcome those limitations, models that are pretrained with huge amounts of speech data are further fine-tuned for assessment of depression. In [23], three models that are trained using self-supervised approach, wav2vec [24], HuBERT [25], and WavLM [26] models, were successfully fine-tuned for depression detection. In [27], unsupervised instance discriminative learning was used with an embedding-generating network where the goal is to get the embeddings of the same instance clustered together and embeddings of different instances away from each other. The network is then used in a downstream depression detection task. These approaches, validated on public datasets such as AVEC2013, AVEC2014, and E-DAIC [28–30], have achieved strong performance in predicting standardized clinical scores such as Beck Depression Index (BDI) [31], typically evaluated using root mean square error (RMSE) or mean absolute error (MAE). To further refine these features, architectures such as multi-head self-attention (MHSA) and State Space Models (SSMs) are used to capture long-range dependencies and focus on the most relevant temporal segments within the speech signal [32, 33].

Beyond model architecture, research has also focused on how speech data is processed and contextualized to improve prediction accuracy. While segmenting long audio recordings into shorter, manageable windows is a common practice, studies argue that this can disrupt critical temporal interconnections necessary for accurate assessment. This has motivated the development of methods specifically designed for modeling long speech sequences to better reflect real-world conversational patterns [33]. Another key insight is the importance of emotional context; research demonstrates that emotionally rich speech segments, identified by their arousal and valence scores, contain more salient cues for depression than emotionally neutral segments [34].

Similar to depression, anxiety can also affect through muscle tension and shortness of breath [35]. Moreover, varied speech rates, hesitations and other speech disturbances may also be indicators of anxiety [35]. Anxiety-level assessment was performed using sustained vowels in [35]. In [36], short clips of speakers from their self-recorded speech journals were used for detecting anxiety using hand-crafted acoustic features, transcript-based embeddings, wav2vec audio embeddings, and a multimodal approach. In [35], high and low levels of anxiety were detected from sustained vowels using openSMILE [37] and DeepSpectrum features [38].

Comorbid anxiety and depressive disorders are common, with 45–67% of patients with major depressive disorder (MDD) having at least one comorbid anxiety disorder, and 30–63% of anxiety disorder patients

meeting the criteria for concurrent MDD [39]. Moreover, misdiagnosis risks are heightened due to diagnostic complexities and the lack of distinct categorizations in diagnostic manuals such as DSM-5, which uses a specifier rather than a separate diagnosis for anxious depression [39].

Despite the prevalence and importance of the problem, comorbid anxious depression has not yet been investigated much by the machine learning community partly because of the lack of data. A longitudinal bilingual dataset of speech samples labeled with depression, anxiety, and stress scores based on the DASS-21 questionnaire in [11]. Moreover, CNN-based models were used for estimating anxiety and depression scores in [11].

Another typically ignored factor in depression assessment is the gender effect. For instance, how gender biases in datasets can affect the performance of machine learning models in detecting depression is investigated in [40]. Results in [40] underscore the necessity of addressing gender-specific factors to enhance the accuracy of depression detection models. In addition, higher rates of depression incidence, 50% more [1], and morbidity among females compared to males [41] also increase the importance of working on gender-specific datasets.

To investigate speech-based depression and anxiety score estimation with and without comorbidity in female speakers, we collected Turkish conversational speech data at the psychiatry department of a local hospital. Along with the speech data, we collected Beck depression and anxiety scores and a psychiatrist's diagnosis. Fine-tuning of large pretrained speech models for depression and anxiety score prediction was investigated. Both disease-specific models and joint models were used and compared. Results indicate that co-occurrence of both disorders should be taken into account during evaluation for more accurate depression score prediction.

2 Dataset description

The dataset comprises recordings from 130 Turkish-speaking female participants. The recordings were collected in an interview format, where the interviewer engaged each patient in a structured question–answer dialogue. All participants were diagnosed by an experienced psychiatrist using a structured clinical interview consistent with DSM-5 criteria.

Recordings were conducted during patients' hospital visits with their informed consent. The following exclusion criteria were applied: individuals who were illiterate; had intellectual disability; met DSM-5 diagnostic criteria for delirium, dementia, amnestic and other cognitive disorders, schizophrenia or other psychotic disorders, mood disorders (including bipolar disorder), alcohol use disorder, or other substance use disorders; used tobacco; or presented with speech impairments. speech impairments included those arising from acute conditions (e.g., the common cold) as well as chronic tobacco-related speech changes (dysphonia), which have been described in the literature as hoarseness, roughness, and lowered fundamental frequency associated with long-term smoking [42–44].

There are two recordings for each patient. In one of the recordings, questions that are on positive topics, such as "did you have any recent event that made you happy?" were used. In the second recording, questions that are on negative topics, such "can you talk about a past event that upset you?" were used. The interview was in a quiet room and the patient used a headset with a close-talking microphone for high-quality recordings at 48 kHz. The dataset spans a total duration of 20 h and 11 min. The average duration of each recording was 4 min and 29 s.

The absence of speech problems was determined through both participant self-assessment and the physician's clinical evaluation during the hospital visit. All recordings for each participant were conducted on the same day to ensure consistency. For all participants, positive questions were asked first, followed by negative questions, with the same fixed order applied across subjects.

Each subject was diagnosed with anxiety, depression, or comorbid depression and anxiety by a psychiatrist. In addition to doctor's diagnosis, for each subject, Beck depression [31] and Beck anxiety [45] scales were used to assess the levels of the two diseases regardless of the diagnosis. The distribution of participants across diagnostic groups, along with their demographic and clinical characteristics, is summarized in Table 1.

2.1 Ethical considerations

The study protocol was approved by the Ethics Committee of Şişli Hamidiye Etfal Training and Research Hospital (Date: 16.05.2023; Decision No: 2302). Written informed consent was obtained from all participants after they were informed about the purpose and procedures of the study. All procedures involving human participants

Table 1 Cohort characteristics of the study groups

Characteristic	Depression	Comorbid	Anxiety
Number of patients	78	42	10
Age range (years)	19–71	19–73	24–57
Age (mean ± SD, years)	43.46 ± 13.00	42.10 ± 13.78	40.0 ± 13.42
Clinical assessment score (mean ± SD)	21.71 ± 9.23	21.05 ± 8.29	17.7 ± 8.2
Mean speech-active duration (minutes, post-VAD)	2.64	3.01	3.20

were conducted in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

3 Methods

Self-supervised models such as wav2vec2 [46] generate powerful, context-dependent speech representations that were successfully used in various downstream tasks including depression detection [23]. Here, we also leverage those models to extract speech representations that are then used by a regression head to assess depression and anxiety levels. Classification heads of the base models were replaced with regression heads for depression/anxiety score estimation. Models were then fine-tuned using Low Rank Adaptation (LoRA) [47] because of the relatively small amounts of data available.

Three pretrained self-supervised models: HuBERT, WavLM, and wav2vec2 were used. Each of those models are described below. A comparison of those models are shown in Table 2. LoRA technique is explained in Sect. 3.2 and the regression heads are described in Sect. 3.3.

3.1 Description of the base models

3.1.1 HuBERT

Hidden Unit BERT (HuBERT) [25] is a self-supervised learning framework designed to extract meaningful representations from speech data. HuBERT implements a fundamental approach involving offline clustering to generate pseudo-labels, which are used in a BERT-style masked prediction task. Initial clustering is performed on MFCCs (HuBERT Stage 1), while later iterations use learned feature representations from previous model checkpoints.

HuBERT compels the model to infer acoustic and linguistic representations from continuous inputs. By iteratively refining the clustering and prediction processes, the model improves its ability to learn high-quality speech representations. The training objective is based on a masked prediction task, where the cross-entropy loss is minimized over the clusters by the model. This loss function is expressed as

$$\mathcal{L} = -\left[\alpha \sum_{t\in M} \log p_f(z_t \mid \tilde{X}, t) + (1-\alpha) \sum_{t\notin M} \log p_f(z_t \mid \tilde{X}, t)\right]. \tag{1}$$

The first term in (1) focuses on learning representations by predicting the targets for masked frames ($\mathcal{L}_m$), while the second term optionally includes the loss for unmasked frames ($\mathcal{L}_u$). To prioritize masked predictions and allow the model to learn robust representations of the underlying sequential structure of a speech signal, $\alpha = 1$ is commonly used.

In this study, we employed the HuBERT base (90 M parameters) as the underlying pretrained model to take advantage of its robust speech representation capabilities.

3.1.2 WavLM

WavLM [26] is a self-supervised learning framework designed for comprehensive speech representation learning across various speech tasks. It reuses the same masked prediction objective as HuBERT, where the model minimizes the cross-entropy loss over pseudo-labels derived from k-means clustering—initially on MFCCs in early training stages, and later on intermediate hidden representations.

Unlike HuBERT, WavLM introduces simulated noisy and overlapped speech during pretraining to improve robustness. This forces the model to learn to extract speaker identity and linguistic content from acoustically complex environments. Its architecture is based on a transformer backbone, enhanced with a gated relative position bias that enables adaptive encoding of position information based on speech content.

The masked prediction loss, which is represented as follows, is minimized by the training objective:

$$\mathcal{L} = -\sum_{l\in K} \sum_{t\in M} \log p(z_t \mid h_t^L), \tag{2}$$

Table 2 Comparison of self-supervised speech models

Feature	HuBERT	WavLM	Wav2vec2
Architecture	CNN + Transformer	CNN + Transformer with gated relative position bias	CNN + Transformer
Pretraining objective	Masked prediction of clustered targets	Masked prediction of clustered targets	Contrastive loss over quantized targets
Pseudo-label source	K-means on MFCCs or features	K-means (same as HuBERT)	Learned quantizer with Gumbel softmax
Masking strategy	Predict masked tokens (BERT-style)	Predict masked tokens with noise augmentation	Contrastive task on masked tokens
Noise robustness	No	Yes (noise + overlap simulation)	Limited (no augmentation)
Quantization	No	No	Yes (via codebooks)
Diversity loss	No	No	Yes

where M represents masked time steps, and z_t are pseudo-labels derived from clustering latent features.

We leverage the robust and ubiquitous speech representation capabilities of WavLM Base (94.70 M parameters).

3.1.3 Wav2vec2

Wav2vec2 [24] is a self-supervised learning framework designed to learn speech representations directly from raw audio. It operates end-to-end by combining feature extraction, quantization, and contextual representation learning within a unified architecture, offering significant advancements over its predecessors. Unlike HuBERT and WavLM, wav2vec2 includes a learned quantization module trained jointly with the rest of the model using the Gumbel-softmax trick.

The framework first extracts latent speech representations using a convolutional feature encoder, followed by a Transformer network that builds contextualized representations. During training, portions of the latent representations are masked, and the model performs a contrastive task: it distinguishes the true quantized target from a set of distractors. The contrastive loss is defined as:

$$\mathcal{L}_m = -\log \frac{\exp(\text{sim}(c_t, q_t)/\kappa)}{\sum_{\tilde{q}\in Q_t} \exp(\text{sim}(c_t, \tilde{q})/\kappa)}, \tag{3}$$

where c_t denotes the contextualized representation at time t, q_t is the true quantized latent vector, and Q_t is the set of distractor vectors.

Using this contrastive objective, the model learns to identify the correct quantized representation among negative samples drawn from the same batch or temporal window. Additionally, a diversity loss is employed to encourage uniform usage of the quantized codebook entries, promoting richer and more informative representations.

In this study, we used the wav2vec2 Base model (94.39 M parameters) to extract contextualized speech features effectively.

3.2 Parameter-efficient fine-tuning approach: LoRA

As pretrained models grow in scale, fine-tuning all parameters for task-specific applications becomes increasingly computationally expensive and inefficient. Full fine-tuning of large models, such as Llama 3.1 with 405 billion parameters, requires enormous storage and memory, making deployment of separate task-specific models impractical. Low-Rank Adaptation (LoRA) [47] offers a parameter-efficient method of fine-tuning models in order to overcome this difficulty.

LoRA hypothesizes that during fine-tuning, the updates to the model weights can be well approximated by a low-rank decomposition. Specifically, for a pretrained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, the fine-tuned weight matrix W is expressed as

$$W = W_0 + \Delta W, \quad \Delta W = BA, \tag{4}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are low-rank matrices with rank $r \ll \min(d, k)$.

During fine-tuning, the pretrained weights W_0 are frozen, and only the low-rank matrices B and A are optimized. The resulting forward pass becomes:

$$h = W_0 x + BAx, \tag{5}$$

where x represents the input to the weight matrix.

In this study, we investigate the effects of different LoRA ranks, specifically 4, 16, 32, and 48 on the performance of a regression task.

3.3 Depression and anxiety score prediction model

Five regression models were developed to predict depression and anxiety scores on top of a self-supervised speech representation model. The pretrained base model consists of a CNN-based feature encoder followed by transformer blocks. To maintain computational efficiency and prevent overfitting, all convolutional layers were frozen, and only the transformer block parameters were fine-tuned using Low-Rank Adaptation (LoRA). LoRA introduces trainable low-rank matrices into specific projections—namely, the query (**q**) and value (**v**) projections—within each transformer block, enabling lightweight yet effective adaptation of the model.

On top of these speech representation models, five distinct regression heads were designed and compared to explore the impact of different modeling strategies on depression-anxiety score prediction:

- **Head-1: average pooling.** A simple but effective approach for estimating depression-anxiety score where the time frames are averaged in hidden representations, followed by a fully connected linear layer.
- **Head-2: attention + linear.** This architecture employs an attention mechanism [48] to compute a weighted sum of the temporal embeddings, allowing the model to focus on the most relevant parts of the time frames. The attention output is passed through a linear layer for regression.
- **Head-3: BiLSTM + attention.** A bidirectional LSTM [49] is first applied to the transformer outputs to model sequential dependencies. The LSTM outputs are then processed with an attention mechanism, and

the resulting representation is passed through a linear regression layer.

- **Head-4: BiLSTM final hidden state.** Similar to Architecture 3, a bidirectional LSTM is used, but instead of applying attention, only the final hidden states from both directions are concatenated and passed through a linear layer for depression-anxiety score prediction.
- **Head-5: BiLSTM + mean pooling.** This head includes a bidirectional LSTM followed by layer normalization and mean pooling across the time dimension. The pooled representation is then passed through a two-layer fully connected regression head. ReLU activations and dropout regularization are applied to enhance non-linearity and reduce overfitting.

These five architectures allow for an in-depth comparison of different temporal aggregation strategies—ranging from simple pooling to attention and recurrent modeling—in the context of depression and anxiety score prediction from speech.

4 Experimental setup

4.1 Data preprocessing

To prepare the dataset for model training, we first performed silence detection and removal using the Pyannote Voice Activity Detector [50, 51]. This ensured that all segments contained active speech, eliminating non-informative silent intervals.

Recordings were then manually segmented into 20-s clips to maintain temporal consistency across samples. The 20-s duration was selected based on empirical results, as shown in Table 3, which indicated it offered the best trade-off between performance and granularity.

To prevent the model from overfitting to individual speakers rather than learning generalized depression markers, we limited the number of segments per patient. Although each speaker had on average 13–14 usable segments, we randomly selected five segments per speaker for training. This sampling strategy reduces the risk of speaker memorization, especially since the 20-s segments are long enough for the model to infer the patient's label. Without this constraint, we observed that including more segments per speaker led to reduced generalization, despite lower training loss—suggesting the model was learning speaker-specific patterns rather than task-relevant features.

During inference, all 20-s clips belonging to the target speaker were individually scored by the model. The final prediction was obtained by taking the median of these scores, providing a robust decision for each speaker.

Table 3 Performance metrics of the wav2vec2 model with Head 1 across varying audio clip durations

Duration	RMSE	MAE
2 s	8.96	7.35
5 s	8.63	7.05
10 s	8.54	6.87
20 s	8.45	6.84
30 s	8.46	6.84
50 s	8.48	6.87

The objective is to investigate how different input lengths, ranging from 2 to 50 s, influence the model's regression performance in terms of RMSE and MAE. The model was trained and tested with negative clips from all patients

Table 4 Training parameters used in the regression

Parameter	Value
Batch size	32
Epochs	10
Learning rate (model)	1×10^{-3}
Learning rate (LoRA modules)	1×10^{-5}
Learning rate scheduler	Linear

Table 5 Results with MFCC features and three regression models

Topic	Model	Depression score		Anxiety score	
		RMSE	MAE	RMSE	MAE
Negative	XGRegressor	9.65	7.71	14.77	12.51
	SVR	**8.92**	**7.32**	**12.91**	**11.01**
	Random Forest	9.30	7.53	13.67	11.63
Positive	XGRegressor	10.47	8.56	14.77	12.60
	SVR	**9.41**	**7.63**	**13.11**	**11.25**
	Random Forest	9.79	7.99	13.92	11.96
All	XGRegressor	10.11	8.19	14.30	12.08
	SVR	**9.34**	**7.56**	**13.09**	**11.22**
	Random Forest	9.66	7.84	13.68	11.73

Experiments were conducted using negative, positive, and combined clips, and the results from these conditions were compared. 10-fold cross validation was used to calculate RMSE and MAE metrics

Bold values indicate the best (lowest) RMSE/MAE within each topic group

4.2 Hyperparameter tuning

The model is trained using the parameters summarized in Table 4, which provides a comprehensive overview of the hyperparameter settings adopted during training. The batch size was set to 32, ensuring efficient utilization of computational resources while maintaining stability in gradient updates. The training process was conducted over 10 epochs, allowing the model sufficient iterations to learn complex patterns from the data.

The learning rate for the regression heads was initialized at 1×10^{-3}, while a more conservative learning rate of 1×10^{-5} was employed for fine-tuning the LoRA (Low-Rank Adaptation) modules of the pretrained

Table 6 Results with AVEC 2013 features and three regression models

Topic	Model	Depression score		Anxiety score	
		RMSE	MAE	RMSE	MAE
Negative	XGRegressor	9.50	7.56	13.89	11.76
	SVR	**8.91**	**7.20**	**12.93**	**11.12**
	Random Forest	9.21	7.39	13.17	11.25
Positive	XGRegressor	9.95	8.03	14.62	12.11
	SVR	**9.40**	**7.59**	**13.11**	**11.25**
	Random Forest	**9.40**	7.60	13.55	11.65
All	XGRegressor	10.09	8.04	13.99	11.77
	SVR	**9.32**	**7.51**	**13.13**	**11.24**
	Random Forest	9.44	7.57	13.26	11.38

Experiments were conducted using negative, positive, and combined clips, and the results from these conditions were compared. 10-fold cross validation was used to calculate RMSE and MAE metrics

Bold values indicate the best (lowest) RMSE/MAE within each topic group

model. This dual learning rate approach facilitates effective adaptation of pretrained parameters while enabling robust learning in the task-specific layers. The learning rate schedule followed a linear decay strategy, gradually reducing the learning rate over time to promote convergence and prevent overfitting. This careful tuning of hyperparameters ensured an optimal balance between learning efficiency and model generalization.

5 Results and discussion

Depression and anxiety models were first tested with the 42-dimensional MFCC features and the 2268-dimensional AVEC 2013 feature set [28]. Root mean square error (RMSE) and mean absolute error (MAE) results are presented in Tables 5 and 6. Experiments were conducted separately for recordings with positive topics, negative topics, and a combination of both. Performance was consistently better with the negative topics. Although the AVEC 2013 features generally outperformed the MFCC features, the differences were not substantial.

Three different regression algorithms were employed to estimate depression and anxiety scores from acoustic features. The first, Extreme Gradient Boosting Regressor (XGRegressor) [52], is based on gradient-boosted decision trees and incrementally improves predictive accuracy by correcting residual errors of prior trees. The second, support vector regression (SVR) [53], utilizes kernel-based learning to capture potentially non-linear relationships. We adopted the radial basis function (RBF) kernel, as it provided better generalization than linear kernels in our experiments. The third, Random Forest Regressor [54], aggregates predictions from an ensemble of decision trees built on bootstrapped samples, reducing variance and improving robustness to noise.

These three models were deliberately chosen to represent complementary regression paradigms—boosting, kernel-based methods, and bagging ensembles. This design enabled a more comprehensive assessment of how low-dimensional MFCCs and high-dimensional AVEC 2013 features interact with different learning mechanisms. Across all conditions, SVR consistently achieved the lowest RMSE and MAE, suggesting that the ability to model non-linear feature–target mappings is particularly advantageous for depression–anxiety estimation from speech. All experiments were conducted using 10-fold cross validation to compute RMSE and MAE metrics.

The performance of the proposed system for depression assessment is presented in Table 7. The system was fine-tuned using LoRA, and performance was compared across different adaptation ranks. Full fine-tuning results are also shown in Table 7. Similar to baseline systems, negative clips

Table 7 Depression results are shown for fine-tuning HuBERT, wav2vec2, and WavLM using LoRA with ranks 4, 16, and 32

Topic	Model	$r=4$		$r=16$		$r=32$		Without LoRA	
		RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Negative	HuBERT	14.40	12.24	12.79	10.65	10.07	8.02	14.85	12.71
	Wav2vec2	**8.86**	**7.13**	**8.44**	**6.84**	**8.47**	**6.89**	**9.42**	**7.50**
	WavLM	10.45	8.39	10.42	8.37	10.44	8.39	10.42	8.37
Positive	HuBERT	14.84	12.60	13.04	10.80	10.23	8.15	15.20	13.11
	Wav2vec2	**9.28**	**7.41**	**8.76**	**7.06**	**8.71**	**7.07**	**9.91**	**7.82**
	WavLM	10.79	8.61	10.81	8.62	10.80	8.62	10.82	8.63
All	HuBERT	14.84	12.59	13.15	10.91	10.49	8.38	15.28	13.05
	Wav2vec2	**9.31**	**7.44**	**8.77**	**7.07**	**8.73**	**7.09**	**9.90**	**7.83**
	WavLM	10.92	8.72	10.91	8.71	10.91	8.71	10.92	8.72

Results without LoRA are also shown. Experiments were conducted using negative, positive, and combined clips, and the results from these conditions were compared. Head-1 was used. 80% of the data was used for training and 20% was used for testing

Bold values indicate the best (lowest) RMSE/MAE for each LoRA setting (r=4, 16, 32, and Without LoRA) within each topic group

Table 8 Anxiety results are shown for fine-tuning HuBERT, wav2vec2, and WavLM using LoRA with ranks 4, 16, and 32

		r=4		*r*=16		*r*=32		*r*=48		Without LoRA	
Topic	**Model**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
Negative	HuBERT	19.32	15.77	17.67	14.22	14.73	11.78	19.87	16.13	19.87	16.32
	Wav2vec2	**13.42**	**11.01**	**12.65**	**10.68**	**12.49**	**10.71**	**14.18**	**11.47**	**14.14**	**11.44**
	WavLM	15.20	12.11	15.22	12.13	15.19	12.10	15.23	12.14	15.20	12.12
Positive	HuBERT	19.63	16.07	17.86	14.43	14.75	11.91	20.10	16.53	20.11	16.53
	Wav2vec2	**13.79**	**11.38**	**12.92**	**10.96**	**12.69**	**10.90**	**14.52**	**11.82**	**14.50**	**11.82**
	WavLM	15.50	12.49	15.49	12.49	15.50	12.49	15.49	12.49	15.50	12.50
All	HuBERT	19.83	16.27	18.09	14.63	14.99	12.01	20.29	16.73	20.27	16.70
	Wav2vec2	**13.75**	**11.32**	**12.87**	**10.88**	**12.65**	**10.86**	**14.57**	**11.83**	**14.59**	**11.83**
	WavLM	15.61	12.55	15.67	12.58	15.69	12.60	15.67	12.58	15.69	12.60

Results without LoRA are also shown. Experiments were conducted using negative, positive, and combined clips, and the results from these conditions were compared. Head-1 was used. 80% of the data was used for training and 20% was used for testing

Bold values indicate the best (lowest) RMSE/MAE for each LoRA setting (*r*=4, 16, 32, 48, and Without LoRA) within each topic group

Table 9 RMSE and MAE values together with confidence intervals are compared for depression assessment with different regression heads

Model/metric	Avg RMSE	Avg MAE	Confidence interval RMSE	Confidence interval MAE
Head-1	8.45	6.84	(7.43, 9.52)	(5.82, 7.74)
Head-2	8.59	7.00	(7.55, 9.56)	(6.09, 7.89)
Head-3	10.86	8.77	(9.71, 12.15)	(7.72, 9.77)
Head-4	8.48	6.87	(7.37, 9.53)	(5.98, 7.75)
Head-5	8.50	6.92	(7.47, 9.43)	(5.96, 7.81)

10-fold cross-validation was used to estimate confidence intervals. Base model was wav2vec2 and only the negative samples were used

Table 10 RMSE and MAE values together with confidence intervals are compared for anxiety assessment with different regression heads

Model/Metric	Avg RMSE	Avg MAE	Confidence interval RMSE	Confidence interval MAE
Head-1	12.66	10.70	(10.85, 14.50)	(8.86, 12.50)
Head-2	12.62	10.99	(11.14, 14.15)	(9.47, 12.43)
Head-3	15.99	12.80	(13.74, 18.22)	(10.77, 14.83)
Head-4	12.61	10.70	(10.61, 14.65)	(8.88, 12.48)
Head-5	13.01	11.37	(11.30, 14.76)	(9.84, 12.98)

10-fold cross-validation was used to estimate confidence intervals. Base model was wav2vec2 and only the negative samples were used

performed the best. Moreover, setting the LoRA rank to 16 and using the negative clips, wav2vec2 model outperformed the baseline systems in Tables 5 and 6.

Anxiety score estimation performance with the proposed system is shown in Table 8. Similar to depression, wav2vec2 outperformed HuBERT and WavLM systems as well as the baseline systems when negative clips are used with a rank of 32.

For best performing systems, wav2vec2 with negative clips and LoRA (rank=16 for depression and rank=32 for anxiety), regression heads described in Sect. 3.3 were compared for depression and anxiety score estimation. Results are shown in Tables 9 and 10. Head-1 and Head-4 consistently achieved the lowest RMSE and MAE across both tasks, while Head-3 significantly underperformed compared to others.

Depression and anxiety co-occur in patients and their entanglement can affect the speech markers. Thus, we designed a joint model that estimates both depression and anxiety using wav2vec2 with LoRA (rank=16).

Table 11 RMSE and MAE values with confidence intervals are shown for depression scores, with the model trained to jointly predict depression and anxiety

Model/Metric	Avg RMSE	Avg MAE	Confidence interval RMSE	Confidence interval MAE
Head-1	8.44	6.83	(7.34, 9.47)	(5.89, 7.68)
Head-2	8.60	7.00	(7.64, 9.55)	(6.08, 7.94)
Head-3	11.27	9.19	(10.09, 12.53)	(8.31, 10.21)
Head-4	8.55	6.93	(7.39, 9.59)	(6.05, 7.83)
Head-5	8.48	6.89	(7.51, 9.44)	(5.91, 7.75)

10-fold cross validation was used to calculate confidence intervals. Base model is wav2vec2 and only the negative samples were used

Results are shown in Tables 11 and 12 for depression and anxiety respectively. Even though RMSE and MAE improved with the joint estimators, improvement was not significant.

Table 12 RMSE and MAE values with confidence intervals are shown for anxiety scores, with the model trained to jointly predict depression and anxiety

Model/Metric	Avg RMSE	Avg MAE	Confidence interval RMSE	Confidence interval MAE
Head-1	12.61	10.61	(10.47, 14.58)	(8.78, 12.35)
Head-2	12.56	10.91	(10.86, 14.01)	(9.32, 12.56)
Head-3	15.96	12.63	(13.72, 18.17)	(10.40, 14.60)
Head-4	12.71	10.68	(10.64, 14.73)	(8.86, 12.71)
Head-5	12.51	10.83	(10.97, 14.28)	(9.30, 12.32)

10-fold cross validation was used to calculate confidence intervals. Base model is wav2vec2 and only the negative samples were used

To analyze the role of comorbidity in depression and anxiety score estimation, we conducted a dedicated set of leave-one-patient-out experiments. The design of these experiments is summarized below for clarity (see Table 14 for dataset sizes and performance results):

5.1 Comorbidity experiments (Experiments 1–5)

To specifically analyze the role of comorbidity in depression and anxiety score estimation, we conducted a dedicated set of leave-one-patient-out experiments. The design of these experiments is summarized in Table 13, while the corresponding performance results are shown in Table 14 and the statistical significance analysis in Table 15. We considered $p < 0.05$ statistically significant.

The comparison between Experiment 2 (train on all patients, test on comorbid cases) and Experiment 3 (train on all patients, test on depression-only cases) revealed a statistically significant difference ($p = 0.0396$). This result indicates that the model achieved lower error rates when depression and anxiety co-occurred in the test subjects, even though depression-only patients constituted the larger portion of the training set.

A second significant difference was observed between Experiment 4 (train and test on depression-only cases) and Experiment 5 (train and test on comorbid cases), with $p = 0.0447$. Consistent with the earlier finding, models trained and tested on comorbid cases yielded more accurate predictions than those trained and tested on depression-only cases. While both models demonstrated overall performance improvements under matched training and testing conditions, the magnitude of the performance gap (in terms of RMSE) remained relatively small.

Taken together, the results in Table 15 suggest that comorbid depression and anxiety manifest more distinctly in speech. Consequently, comorbidity should be considered an important confounding factor in depression score estimation.

Table 14 Leave-one-patient-out experiment results for testing how comorbid cases affect depression score estimation performance

Experiment	Train	Test	Train size	Test size	RMSE
1	All	All	130	130	7.78
2	All	Depr+Anx	130	42	7.77
3	All	Depr	130	78	8.64
4	Depr	Depr	78	78	8.03
5	Depr+Anx	Depr+Anx	42	42	7.34

Subjects are grouped into depression-only (depr), comorbid depression and anxiety (depr+anx) and all

Table 15 Statistical significance test results

Experiment	*p*-value	*U*-value
2–3	0.0396	1738.00
4–5	0.0447	1705.00

Mann–Whitney *U* tests were conducted between experiment groups from Table 14. Reported are the comparisons with significant differences in performance

6 Conclusion and future work

We investigated speech-based automatic score estimation for depression, anxiety, and their co-occurring presentations in female speech. Evaluations were performed using both classical acoustic features (MFCC, AVEC 2013) and modern LoRA fine-tuned self-supervised speech models equipped with regression heads. Wav2vec2 consistently and significantly outperformed both other self-supervised models (WavLM and HuBERT) and classical acoustic feature sets. Our

Table 13 Definition of comorbidity experiments (Experiments 1–5)

Experiment	Description
1	Train on all patients, test on all patients
2	Train on all patients, test only on comorbid depression+anxiety cases
3	Train on all patients, test only on depression-only cases
4	Train and test exclusively on depression-only cases
5	Train and test exclusively on comorbid depression+anxiety cases

results showed that LoRA fine-tuning provides an additional boost in accuracy. Moreover, across all experimental settings, negative topics consistently yielded the lowest estimation errors for depression and anxiety, suggesting that emotionally negative content makes the acoustic markers more salient.

In this study, we focused on the comorbidity of depression and anxiety, in contrast to most existing literature, which mostly addresses these conditions in isolation. We first evaluated models trained to estimate depression and anxiety scores separately. Subsequently, we assessed joint models designed for simultaneous estimation, but found that they did not significantly outperform the models trained on individual conditions. To further explore the impact of comorbidity, we then developed models to assess depression scores for individuals with depression alone and for those with comorbid depression and anxiety. Our findings indicate that depression manifests more clearly in speech when it co-occurs with anxiety, suggesting that comorbidity may improve the depression score estimation performance from speech.

Overall, this investigation affirms the feasibility and clinical significance of LoRA fine-tuned self-supervised speech models for depression and anxiety score estimation in females, while highlighting the importance of negative emotional content and comorbidity.

Since the current dataset includes only Turkish-speaking female participants, the findings may not generalize across genders, languages, or cultural backgrounds. In future work, we will extend this framework to larger, more diverse, and multilingual cohorts, and incorporate additional data modalities (e.g., speech, facial, and textual cues) as an essential next step toward reliable, real-world deployment.

Acknowledgements
Not applicable.

Authors' contributions
AB collected, analyzed, and interpreted the patient data regarding anxiety and depression. TK and CD implemented algorithms for estimation models. AB interpreted the results. All authors read and approved the final manuscript.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability
The authors do not have permission to share data in accordance with the local ethics committee approval.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 21 May 2025 Accepted: 13 November 2025
Published online: 13 December 2025

References

1. World Health Organization: Depression. (2024). https://www.who.int/health-topics/depression. Accessed 14 Apr 2025
2. S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. **28**(4), 357–366 (1980)
3. F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, E. André, C. Busso, L.Y. Devillers, J. Epps, P. Laukka, S.S. Narayanan et al., The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE Trans. Affect. Comput. **7**(2), 190–202 (2015)
4. B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J.K. Burgoon, A. Baird et al., in *Proceedings of Interspeech 2016*. The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language (ISCA, San Francisco, 2016), pp. 2001–2005
5. N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, T.F. Quatieri, A review of depression and suicide risk assessment using speech analysis. Speech Commun. **71**, 10–49 (2015)
6. Seneviratne. N, Espy-Wilson. C, Multimodal depression severity score prediction using articulatory coordination features and hierarchical attention based text embeddings. In Proc. INTERSPEECH, (ISCA, Baixas, France, 2022) pp. 3353–3357.
7. Fan, C. Lv, Z. Pei, S. Niu, M. CSENet: Complex squeeze-and-excitation network for speech depression level prediction. In Proc. ICASSP, (IEEE, Piscataway, NJ, USA, 2022) pp. 546–550.
8. Ringeval. F, Valstar. M, Schuller. B, Cowie. D, Cowie. R, Pantic. M, et al. AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition. In Proceedings of the 9th International Audio/Visual Emotion Challenge and Workshop (ACM, New York, NY, USA, AVEC 2019) pp. 3–12.
9. Egas-López. J.V, Kiss. G, Sztahó. D, Gosztolya. G, Automatic assessment of the degree of clinical depression from speech using x-vectors. In Proc. ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE, Piscataway, NJ, USA 2022) pp. 8502–8506.
10. Zhang. P, Wu. M, Yu. K, ReCLR: Reference-enhanced contrastive learning of audio representation for depression detection. In Proc. INTERSPEECH (ISCA, Baixas, France 2023) pp. 2998–3002.
11. Tasnim. M, Ramos. R.D, Stroulia. E, Trejo. L.A. A machine-learning model for detecting depression, anxiety, and stress from speech. In Proc. ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE, Piscataway, NJ, USA 2024) pp. 7085–7089.
12. Seneviratne. N, Espy-Wilson. C. Generalized dilated CNN models for depression detection using inverted vocal tract variables. In Proc. INTERSPEECH (ISCA, Baixas, France 2021) pp. 4513–4517.
13. Dumpala. S.H, Rempel. S, Dikaios. K, Sajjadian. M, Uher. R, Oore. S. Estimating severity of depression from acoustic features and embeddings of natural speech. In Proc. ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE, Piscataway, NJ, USA 2021) pp. 7278–7282.
14. Shen. Y, Yang. H, Lin, L. Automatic depression detection: an emotional audio-textual corpus and a GRU/BiLSTM-based model. In Proc. ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE, Piscataway, NJ, USA 2022) pp. 6247–6251.
15. Seneviratne. N, Espy-Wilson, C. Speech based depression severity level classification using a multi-stage dilated CNN-LSTM model. In Proc. INTERSPEECH (ISCA, Baixas, France 2021) pp. 2526–2530.
16. Zhao. Z, Bao. Z, Zhang. Z, Cummins. N, Wang. H, Schuller, B. Hierarchical attention transfer networks for depression assessment from speech. In Proc. ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE, Piscataway, NJ, USA 2020) pp. 7159–7163.
17. Tao. F, Ge. X, Ma. W, Esposito. A, Vinciarelli, A. Multi-local attention for speech-based depression detection. In Proc. ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE, Piscataway, NJ, USA 2023) pp. 1–5.
18. Campbell. E.L, Dineley. J, Conde. P, Matcham. F, White. K.M, Oetzmann, C. et al. Classifying depression symptom severity: assessment of speech

representations in personalized and generalized machine learning models. In Proc. INTERSPEECH (ISCA, Baixas, France 2023) pp. 1738–1742.
19. Ravi. V, Wang. J, Flint. J, Alwan. A. Fraug: A frame rate based data augmentation method for depression detection from speech signals. In Proc. ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE, Piscataway, NJ, USA, 2022) pp. 6267–6271.
20. Zhao. Z, Li. Q, Cummins. N, Liu. B, Wang. H, Tao. J. Schuller, B.W. Hybrid network feature extraction for depression assessment from speech. In Proc. INTERSPEECH (ISCA, Baixas, France 2020) pp. 4956–4960.
21. Li. Y, Niu. M, Zhao. Z, Tao, J. Automatic depression level assessment from speech by long-term global information embedding. In Proc. ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE, Piscataway, NJ, USA 2022) pp. 8507–8511.
22. Li. Q, Wang. D, Ren. Y, Gao. Y, Li. Y. FTA-net: A frequency and time attention network for speech depression detection. In Proc. INTERSPEECH (ISCA, Baixas, France 2023) pp. 1723–1727.
23. Wu. W, Zhang. C, Woodland. P.C. Self-supervised representations in speech-based depression detection. In Proc. ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE, Piscataway, NJ, USA 2023) pp. 1–5.
24. Schneider. S, Baevski. A, Collobert. R, Auli. M. wav2vec: Unsupervised pre-training for speech recognition. Proc. Interspeech (Graz, Austria 2019) pp. 3465–3469.
25. Hsu. W. N, Bolte. B, Tsai. Y. H. H, Lakhotia. K, Salakhutdinov. R, Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM transactions on audio, speech, and language processing. **29**, 3451-3460 (2021)
26. S. Chen, C. Wang, Y. Wu, S. Liu, Z. Chen, Z. Chen et al., Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Booktitle Sel. Top. Signal Process. **16**(6), 1505–1518 (2022)
27. Wang. J, Ravi. V, Flint. J, Alwan. A. Unsupervised instance discriminative learning for depression detection from speech signals. In Proc. INTERSPEECH (ISCA, Baixas, France 2022) pp. 2018–2022.
28. Valstar. M, Schuller. B, Smith. K, Eyben. F, Jiang. B, Bilakhia. S, et al. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '13), Barcelona, Spain, (ACM, New York, NY, USA 2013) pp. 3–10.
29. Valstar. M, Schuller. B, Smith. K, Almaev. T, Eyben. F, Krajewski. J, Cowie. R, Pantic, M. AVEC 2014: 3D dimensional affect and depression recognition challenge. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC '14), Orlando, FL, USA, (ACM, New York, NY, USA 2014) pp. 3–10.
30. Gratch. J, Artstein. R, Lucas. G.M, Stratou. G, Scherer. S, Nazarian. A, Wood. R, Boberg. J, DeVault. D, Marsella. S, et al. The distress analysis interview corpus of human and computer interviews. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), (ELRA, Reykjavik, Iceland, 2014) pp. 3123–3128.
31. A.T. Beck, C.H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An inventory for measuring depression. Arch. Gen. Psychiatr. **4**, 561–571 (1961)
32. Liu, S. Multi-head self-attention network for depression level estimation from speech. In Proc. ICSP 2024 – 2024 9th International Conference on Intelligent Computing and Signal Processing (IEEE, Piscataway, NJ, USA, 2024) pp. 1088–1091.
33. Li. S, Xie. Z, Naqvi. S.M. Efficient long speech sequence modelling for time-domain depression level estimation. In Proc. ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing; Hyderabad, India, (IEEE, Piscataway, NJ, USA 2025) pp. 1–5.
34. Yu. J, Kaya. H. Using emotionally rich speech segments for depression prediction. In Proc. ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing; Hyderabad, India, (IEEE, Piscataway, NJ, USA 2025) pp. 1–5.
35. Baird. A, Cummins. N, Schnieder. S, Krajewski. J. Schuller, B.W. An evaluation of the effect of anxiety on speech—computational prediction of anxiety from sustained vowels. In Proc. INTERSPEECH (ISCA, Baixas, France 2020) p. 4951–4955.
36. Agarwal, P., Jindal, A., & Singh, S. Detecting anxiety from short clips of free-form speech. (2023). arXiv preprint arXiv:2312.15272.
37. Eyben. F, Wöllmer. M, Schuller. B. openSMILE: The Munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia (MM '10) (ACM, New York, NY, USA 2010) pp. 1459–1462.
38. S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy et al., in *Interspeech 2017*. Snore sound classification using image-based deep spectrum features (ISCA, 2017), pp. 3512–3516
39. Choi, K.W. Kim, Y.K. Jeon, H.J. Comorbid anxiety and depression: Clinical and conceptual consideration and transdiagnostic treatment. In Anxiety Disorders: Rethinking Understanding of Recent Discoveries. Adv Exp Med Biol **1191**, 219–235 (2020)
40. Bailey, A. Plumbley, M.D. Gender bias in depression detection using audio features. In Proc. EUSIPCO 2021 – 29th European Signal Processing Conference (IEEE, Piscataway, NJ, USA 2021) pp. 596–600.
41. M. Piccinelli, G. Wilkinson, Gender differences in depression: Critical review. Br. J. Psychiatry **177**(6), 486–492 (2000)
42. V. Uloza, V. Saferis, I. Uloziene, Smoking and voice: Acoustic analysis in men and women. Medicina (Kaunas) **41**(9), 757–765 (2005). Demonstrates acoustic changes in smokers' voices compared to non-smokers
43. K. Nemr, A. Amar, M. Abrahão, G. Leite, H. Kohler, A. Santos, Smoking and voice: Correlation between acoustic analysis and self-assessment. Braz. J. Otorhinolaryngol. **79**(4), 461–466 (2013). https://doi.org/10.5935/1808-8694.20130083. Explores both objective acoustic parameters and subjective self-assessment in smokers
44. P.H.R. Amaral, L.W. Lopes, A.S. Ferreira, The impact of smoking on vocal quality: A systematic review. J. Voice **33**(5), 765–2376534 (2019). https://doi.org/10.1016/j.jvoice.2018.02.006. Summarizes acoustic and perceptual changes in smokers' voices
45. A.T. Beck, N. Epstein, G. Brown, R.A. Steer, An inventory for measuring clinical anxiety: Psychometric properties. J. Consult. Clin. Psychol. **56**, 893–897 (1988)
46. A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. **33**, 12449–12460 (2020)
47. E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen et al., Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
48. Vaswani. A, Shazeer. N, Parmar. N, Uszkoreit. J, Jones. L, Gomez. A.N, Kaiser. Ł, Polosukhin, I. Attention is all you need. Adv Neural Inf Process Syst. **30**, 5998–6008 (2017)
49. A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)
50. Bredin, H. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In Proc. INTERSPEECH 2023, (ISCA, Baixas, France 2023)
51. Plaquet, A. Bredin, H. Powerset multi-class cross entropy loss for neural speaker diarization. In Proc. INTERSPEECH 2023, (ISCA, Baixas, France 2023)
52. Chen, T. Guestrin, C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, (ACM, New York, NY, USA, 2016) pp. 785–794.
53. A.J. Smola, B. Schölkopf, A tutorial on support vector regression. Stat. Comput. **14**(3), 199–222 (2004)
54. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.