Towards using Breathing Features for Multimodal Estimation of Depression Severity

Francisca Pessanha
Utrecht University
Utrecht, the Netherlands
f.pessanha@uu.nl

Heysem Kaya
Utrecht University
Utrecht, the Netherlands
h.kaya@uu.nl

Alkim Almila Akdağ Salah
Utrecht University
Utrecht, the Netherlands
a.a.akdag@uu.nl

Albert Ali Salah
Utrecht University
Utrecht, Netherlands
Boğaziçi University
Istanbul, Turkey
a.a.salah@uu.nl

ABSTRACT

Breathing patterns are shown to have strong correlations with emotional states, and hence have promise for automatic mood order prediction and analysis. An essential challenge here is the lack of ground truth for breathing sounds, especially for medical and archival datasets. In this study, we provide a cross-dataset approach for breathing pattern prediction and analyse the contribution of predicted breath signals for the detection of depressive states, using the DAIC-WOZ corpus. We use interpretable features in our models to provide actionable insights. Our experimental evaluation shows that in participants with higher depression scores (as indicated by the eight-item Patient Health Questionnaire, PHQ-8), breathing events tend to be shallow or slow. We furthermore tested linear and non-linear regression models with breathing, linguistic sentiment and conversational features, and show that these simple models outperform the AVEC17 Real-life Depression Recognition Sub-challenge baseline.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning approaches*; *Feature selection.*

KEYWORDS

Affective Computing, Paralinguistics, Depression Severity Prediction, Breathing Analysis, Interpretability, DAIC-WOZ Corpus

ACM Reference Format:

Francisca Pessanha, Heysem Kaya, Alkim Almila Akdağ Salah, and Albert Ali Salah. 2022. Towards using Breathing Features for Multimodal Estimation of Depression Severity. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3536221.3556606

ICMI '22, November 7–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-9390-4/22/11…$15.00
https://doi.org/10.1145/3536221.3556606

1 INTRODUCTION

Automatic emotional assessment from voice is a tool with great potential for mood disorder analysis and prediction. However, studies in the literature focus primarily on non-verbal paralinguistic qualities for this purpose [18, 26]. This paper investigates breathing patterns for such applications, in general, and depression, in particular. Breathing is known to correlate with emotional states, and it was previously shown that for instance, in the emotional moments during the recounting of traumatic events, there are significant changes in the use of breathing and silences [1]. Here, we propose a novel approach with interpretable features to process breathing during speech for depression prediction.

For the analysis of breathing, the lack of ground truth is a recurrent challenge. Although the ground truth is measurable with breathing belts, performing such measurements is not the standard for most naturally recorded data. Cross-dataset learning provides a potential solution for this problem: the models can be learned on external data for which ground truth is available. However, since the breathing ground truth often comes in the form of changes in diameter of the thorax/abdomen (and not necessarily as measurements related to a sound signal), it is not intuitive to evaluate its accuracy by comparing the audio and signal plots manually. Here, we propose an approach that will enable assessing the accuracy of the breathing predictions based only on the audio recordings. With this goal in mind, we use the INTERSPEECH 2020 Computational Paralinguistics Challenge (ComParE) Breathing Sub-challenge dataset [26] to observe correlations between audio and breathing signals and then extrapolate this knowledge to other settings for breathing signal prediction.

In our proposed multimodal approach, we combine linguistic and breathing features to assess psychopathology. Our ultimate goal is to understand how breathing during speech relates to emotionality in both verbal and voice features, such as pitch and voice quality. We empirically observe variations in the breathing patterns during emotional moments, such as faster and shallow breathing, holding the breath and taking deep breaths, which can serve as indicators.

This paper is structured as follows. In Section 2, we briefly summarise the related work on depression analysis from speech, as well as discuss breathing analysis in affective computing. In Section 3,

we detail the datasets we make use of, i.e. the INTERSPEECH ComParE 2020 Breathing Subchallenge corpus for cross-dataset learning and the DAIC-WOZ corpus for experimental evaluation of depression analysis. In Section 4, we detail our approach for cross-dataset breathing prediction, as well as feature extraction and depression regression. We report our experimental results and observations in Section 5, and provide future directions and a discussion in Section 6.

2 RELATED WORK

2.1 Breathing, speech and language

Breathing is a constant in a human's life and happens naturally and effortlessly. This process is adjusted continuously to the individual's needs. For example, we coordinate breathing with eating or speaking. Breathing occurs via the inspiratory pump muscles contracting to draw air into the lungs. Expiration occurs mainly passively, with the recoil of the chest wall and lungs during quiet breathing. If needed, expiratory muscles can produce active expiration. During the respiratory cycle, the volume of the lungs will vary.

Vocalisation will impose aeroacoustic constraints and so require adaptations of breathing control. However, breathing adaptations during speech go beyond that; respiration needs to be adjusted to different linguistic and communicative levels. For example, syntactic boundaries, sentence length, prosody and listener-speaker behaviour can influence this adaptation mechanism [8]. From a physiological point of view, speech breathing will involve more variable and deeper inhalations, depending on the breathing capacity needed for the spoken sentence, followed by a long exhalation. The respiratory volume during speech was studied by Winkworth et al. [32] with the help of respiratory belts (on chest and abdomen), who found that the majority of inspirations (i.e. inhalations) occurred at structural boundaries during reading and "grammatical junctures" during spontaneous speech. As predicted, the latter show a higher number of grammatically inappropriate inspirations. They also noted that the initiation lung volume (inspiration) is correlated with the breath group length. Consequently, we expect a higher inspiration volume before a long utterance, particularly during spontaneous speech, where the subject has the opportunity to make adjustments to the utterance on the go.

Emotions happen with physiological changes within the entire body, including changes in breathing. The respiratory motor system commands the contraction of the respiratory muscles following complex neural networks in our brain and primarily adapts in response to metabolic demands. However, this system's output can also be influenced by internal and external environmental changes, resulting in behavioural breathing. An example is the relationship between anxiety and breathing: studies show an increase in the respiratory rate with anticipation anxiety, which is not related to a higher demand for oxygen. Unpleasant respiratory sensations, such as an uncomfortable urge to breath, depend on the affective state of the subject and can be elicited by anxiety and distress [11].

Observations focusing on negative emotions and breathing patterns indicate that the arousal dimension is essential for the analysis. For example, although low valence and high arousal emotions such as anger or stress increase the respiratory rate and breathing depth, this phenomenon is not found for all negative emotional conditions. Emotions such as sadness or being depressed are associated with decreased respiratory rate, as well as slow and shallow breathing.

It should be noted that defining a general breathing variation pattern for clinically depressed patients is challenging. These patients will often have an anxiety disorder responsible for an increased breathing rate, which might point to a voluntarily induced slowing of the respiratory rate to cope with the stimulus [4]. Furthermore, a higher respiratory pattern variability is correlated with depression, presenting more variation in pause duration and respiratory frequency [37].

If the emotion-related changes in breathing patterns can be distinguished from other factors influencing breathing, they can serve as useful and interpretable features for the analysis of mood disorders. Efforts to provide interpretability to accurate but complex models for mood disorder recognition have received increased attention over the past years [2, 3, 19]. However these efforts still amount to only a small fraction in the computational health research domain. In the context of depression, most studies do not assess the potential of breathing features explicitly. Our main premise in this paper is that such features can provide explainable indicators, and help for diagnostics.

2.2 Depression analysis

Looking at the relevant depression analysis literature, we observe that a large number of features are potentially useful for depression detection, including speech behaviour, speech prosody, eye movements, and head pose. Neuro-physiological changes associated with depression influence motor coordination and the effects can be detected in acoustic features, such as jitter and shimmer [23, 24]. Such analysis can be used for automatically screening subjects and to facilitate diagnosis [31]. Recent studies also find speech behaviour features (e.g., pauses) to be very distinctive for diagnosis [2].

Banerjee et al. [3] described a single model for predicting three mood disorders, depression, anxiety, and anhedonia, respectively, as three binary prediction tasks. First, unimodal convolutional neural network (CNN) models are trained on audio, video, and text modalities, and the features are transferred to a multimodal model. Then, encodings are concatenated and processed further by an attention mechanism and a fully connected layer. Some features are not very informative by themselves (such as "Contrast Spectrogram 10"), while others are more interpretable, such as "Word Valence" or "Number of Characters". The fact that the top ten most important features contain many linguistic features, as well as that the linguistic model was found to be the highest performing unimodal model, indicates that linguistic features are quite important for these tasks.

Depression analysis from conversational data allowed the investigation of a range of features. The Audio/Visual Emotion Challenge (AVEC) has been instrumental in the development of new approaches, and depression analysis was specifically addressed in these challenges. During the AVEC'16, AVEC'17 and AVEC'19 Challenges on depression analysis [21, 22, 30], multiple solutions were presented for depression assessment on the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) dataset, which is part of the larger Distress Analysis Interview Corpus (DAIC) [10].

The database contained human-agent interactions, and the challenge participants were required to classify whether the human was depressed or not, where the binary ground-truth was based on the severity of self-reported depression as indicated by the Patient Health Questionnaire (PHQ-8) [13] score for each human-agent interaction. Hybrid solutions combining video, audio, and text features were shown to be the most successful, taking advantage of the transcriptions provided for participant turn segmentation and topic modelling [9, 27, 35, 36].

The modelling of the conversation topic is useful, because different topics elicit different emotions in subjects, and this leads to different amounts of information for depression analysis. In the challenge data, several questions were answered by the participants, and these are used to steer the topic. In [27], the questions posed were divided into classes inspired by the PHQ-8 domains: lack of interest, depressed feelings, sleep quality, tiredness, appetite, failure, concentration, and psycho-physiological events such as moving and speaking very slowly. Similar conclusions were reached in [33] showcasing the different predictive values of each question for post traumatic stress disorder (PTSD), on an extended set of DAIC interviews with both war veterans and non-veterans.

The lack of annotated data for the study of depression screening motivated researchers to focus solely on audio features for depression detection, achieving motivating results in the challenge development sets [18, 34]. Further work has shown the potential of speech-related features for depression assessment using deep learning approaches across various depression scored datasets [17], with the most recent work proposing a Mel Frequency Cepstrum Coefficient (MFCC)-based Recurrent Neural Network model, focusing on the effect of depression in vowel pronunciation [20].

Apart from recurrent neural networks, attention-based models have recently been employed for depression analysis. The winners of the AVEC'19 Detecting Depression with AI Sub-challenge (DDS) have developed an attention-based model, with multiple stages of attention layers using three modalities (audio, video, text) to predict the PHQ-8 scores [19]. The audio and video features were independently processed through a Bi-LSTM attention network and the text features through an ordinary Bi-LSTM layer. Then, the output from these three modalities were merged by means of an extra attention layer. In this way, the attention weights gave an indication of the importance of each modality. Indicated by the weights, the text modality was found to be very important (0.57) with respect to the visual and audio modalities (both 0.21). The authors also trained unimodal models, with a text-only model resulting in the best score. Both results indicate that the text modality, processed in this way, also provides valuable cues towards prediction of depression.

3 CORPORA

In this work, we use several corpora for breathing analysis and depression analysis. We describe these resources in this section.

3.1 Speech Breath Corpus (SBC)

The SBC database is a subset of the UCL Speech Breath Monitoring (UCL-SBM) corpus and is introduced for the breathing Sub-Challenge of the INTERSPEECH 2020 Computational Paralinguistics Challenge [26]. The dataset includes spontaneous speech about the participant's daily experiences, such as visiting a city. It consists of 49 audio interviews, each of four minutes, with the corresponding breath signal measured with a piezoelectric respiratory belt in the thorax area.

For the present work, we have further annotated audible breath events in 10 recordings from the SBC database, corresponding to 40 minutes of spontaneous speech, according to the type of event ("Inhale" or "Exhale"), as well as the location of the event in the speech signal ("Middle of the speech", or a "pause"), segmenting a total of 433 breath events.

3.2 Distress Analysis Interview Corpus-Wizard-Of-Oz (DAIC-WOZ)

The Distress Analysis Interview Corpus-Wizard-of-Oz dataset (DAIC-WOZ) consists of semi-structured clinical interviews designed to support the diagnosis of psychological distress conditions, particularly depression and post-traumatic stress disorder (PTSD). The interviews were conducted by an AI based virtual agent, under a wizard-of-oz framework, meaning that human agents controlled the agent's non-verbal behaviours and verbal utterances [6, 10]. This corpus motivated the Depression, Mood and Emotion Challenge in the Annual Workshop on Audio/Visual Emotion Challenge (AVEC) in 2016 [30], and the Real-life Depression Challenge in AVEC 2017 [22]. In the remainder of our work, we will refer to the 2017 challenge and related publications as a baseline.

For each session of the Depression corpus used in AVEC'17, audio recordings, transcriptions, and baseline audio and video features are available. The dataset includes 107, 35, and 47 subjects for training, development, and test sets, respectively. The average depression severity on the training and development set is M = 6.67 (SD = 5.75) out of a maximum score of 24.

In addition, for depression prediction, self-assessed PHQ-8 scores are provided. We use the breathing annotations presented in [12] for breath signal cross-dataset prediction evaluation. A total of 1478 breath event are annotated across 16 recordings. These annotations were performed on the extended version of the dataset proposed in the 2019 edition of the AVEC [21].

4 METHODOLOGY

Figure 1 illustrates the proposed pipeline for depression severity assessment. One of the contributions of the present work is the breath signal prediction module and the extracted breathing features. In this section, we discuss the cross-dataset prediction method, contextual segmentation, and correlation analysis for each feature proposed. Lastly, we propose a depression regression model.

4.1 Cross-dataset prediction for breath signal

A continuous breathing signal provides extensive information about the respiratory patterns, allowing the measurement of depth, respiratory speed, and pattern variability. When looking into emotions in a depressed subject, the depth of the respiratory events and overall pattern variability are essential factors, motivating us to focus on continuous breath signal prediction instead of simple audible breath event segmentation. To tackle the lack of a breath signal ground truth in the DAIC-WOZ dataset (and most other real-world datasets), we propose a cross-dataset prediction approach, based on the 1D

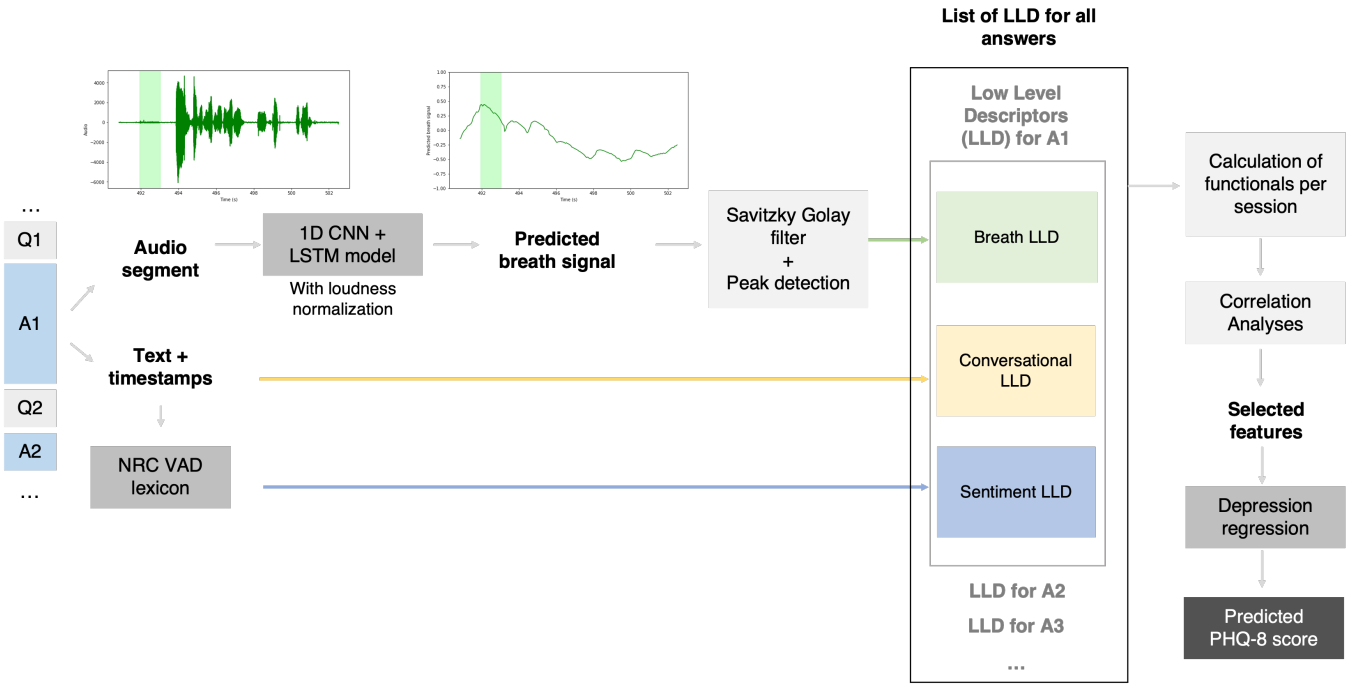


Figure 1: Proposed method for depression severity assessment. On the right, a deep breath event is highlighted on both audio and breath signals for visualisation purposes. Q1 and A1 refer to the first question and answer, with the remaining questions and answers following the same notation.

CNN + LSTM architecture proposed by the ComPaRE2020 challenge winner team, Markitantov et al. [14], on the Speech Breath Corpus (SBC). This study suggests a prediction window of 16 seconds, suitable for the continuous speech characteristics in the dataset. However, our target dataset (DAIC-WOZ) consists of dialogues between participants and an AI agent taking turns to speak. Therefore, the participant's continuous speech sections are smaller, and predicting a continuous breath signal across the session becomes highly challenging due to the AI speech component.

To adapt the 1D CNN + LSTM architecture proposed in [14] to the DAIC-WOZ, we focus solely on predicting the breath signals during the participants' reactions. The first challenge is to determine a suitable window size for the analysis of the breathing signals. Subsequently, we first analyse the DAIC-WOZ training set to find a suitable input window considering the length of the participants' reactions following a question. The resulting model performance is evaluated on the SBC, using the same cross-validation scheme proposed in the original paper.

We first segment the interview sections where the interviewee is speaking. These sections have annotations for the Participant (P), Filler (F), Breath (B), Laughter (L), and others that have a minimum duration equal to the selected window size. This division intends to mimic the participant's action segmentation implemented in the DAIC-WOZ. A small window size increases the number of participant reactions from which we can obtain breathing signal predictions. However, if the size is too small, it becomes difficult to catch breathing events. We evaluated the selected parameters on the annotated breath events introduced by Kaya et al. [12].

Lastly, we compared simple functionals of the predicted breath signal between the annotated breath events of the SBC dataset and the target dataset. If the cross-dataset breath signal prediction is successful, we expect these events to have similar characteristics.

4.2 Question segmentation and correlation analysis

The analysis of the responses of a patient can be improved by taking into account the context of the signals. The literature on DAIC-WOZ corpus presents correlations between depression recognition and certain question types, especially the ones related to PHQ-8 domains [9, 27]. Following the literature, we also segment the dataset according to question-answer pairs, and target interactions that lead to more prolonged reactions from the participants. This approach has certain challenges. Above all, the dataset is based on a semi-open interview approach, meaning that not all participants are asked the same questions. With the breath signal prediction constraints described in Section 4.1, we further impose a reduction in the number of processed answers, eliminating some additional questions. Since the interviews are not fully structured, the questions derive from a limited question pool, and hence this approach is still possible.

We use the labels provided with the train set transcriptions that annotate the AI agent's actions, for example, "dream_job" refers to the question "what is your dream job?". We extend these annotations to the development and test sets by comparing each action of the AI Agent with the "tag" - "action speech" pairs defined in the train

set. In addition, we manually annotate the actions that were not in the initial action set. We further categorise the different questions according to their polarity into five groups as "Positive" ("what do you enjoy about traveling?", "who's someone that's been a positive influence in your life?"), "Negative" ("What are some things you wish you could change about yourself?", "Tell me about a time when someone made you feel really badly about yourself?"), "Neutral" ("Why did you move to LA?", "How long ago were you diagnosed?"), "Mixed", meaning emotional questions with no implicit polarity ("Do you find it easy to be a parent?", "Tell me about your kids?") and "General", follow up questions ("Can you give me an example of that?", "Tell me more about that").

We extract the corresponding answer for each question, defined as the participant's speech between the inquiry and AI Agent's following action. Considering the window size of our breathing prediction model, we solely consider the participant reactions with a duration equal to or longer than the breath prediction window defined in the cross-dataset breath signal prediction experiment. We then explore the correlation between the continuous diagnosis score, PHQ-8 Score, and question-specific features under three main categories: conversational features, linguistic sentiment features, and breath signal features as presented in Table 1. These features are chosen due to their interpretability.

We calculate valence, arousal, and dominance (VAD) features based on the NRC VAD lexicon [15]. Each word is associated with a reliable human-rated value for VAD. Then, we calculate the answer's sentiment by applying functionals to the list of values of all the words in the response.

We pre-process the predicted breath signal for breathing feature extraction to remove noise in the prediction and produce interpretable functionals based on the literature in the field. Hence, we apply a Savitzky-Golay filter [25] with a polynomial of 2nd degree and a window of 13 samples, corresponding to 0.52 seconds, in line with the average breath event duration observed in the SBC. We extract simple functionals from the resulting smooth signal and the respective first derivative. Further, we perform peak detection over the signal to identify the local maxima and minima, expected to be associated with inhale/exhales. To filter out the smaller peaks detected, likely related to small breath events during the speech, we define the minimum prominence as 0.13, corresponding to the median prominence in the annotated breath events on the SBC, calculated on the predicted, smooth signal. Additionally, we define the minimum distance between maximum peaks as 2 seconds, approximately half of a typical breath cycle for young, healthy individuals [28]. Furthermore, to evaluate the breath signal during silences, we extracted the breathing signal for the reaction time and silences longer than 0.3 seconds. When it is not possible to calculate a feature, for instance, in the absence of two peaks in the case of peak-to-peak distance, we set its value to zero.

For each session, we extract features from 1) each answer, 2) the combined set of answers for each question type, 3) the entire set of answers, and 4) all participant's reactions. Then, we group the resulting feature vectors across the train set according to the answer selection criteria, i.e. 1) question, 2) question type, 3) all participants' answers, and 4) all participants' actions. Finally, we evaluate the Pearson Correlation Coefficient (PCC) between the individual features and the depression severity label for each subset, excluding all PCC with a p-value > 0.05.

4.3 Depression severity prediction via regression

As the last step of our processing pipeline, we want to evaluate the predictive power of the feature set and contextual segmentation proposed in the previous section for depression assessment. Since depression severity scores are continuous, we tackle this as a regression problem.

The requirement of interpretability poses some challenges in regression modelling. Non-linear models are more flexible compared to linear models, but can be less interpretable. When feature extraction approaches are used, the original feature space can be transformed into new features that are more parsimonious, but not readily interpretable. Furthermore, multiple linear regression assumes a low correlation between the independent variables, which is not always efficiently dealt with in cases when transforming the feature space is undesirable, for example, due to the loss in explainability. However, there will be some degree of collinearity in all real-world data. Previous studies report extensively on the informative value of collinearity and solutions to overcome performance loss due to redundant variables [7, 16].

Based on the correlations observed in the exploratory study, we train simple linear models (i.e. Linear Regression), and non-linear models (i.e. Random Forest), to evaluate the predictive value of the features defined and their generalisation power across different sets.

Random forests have shown to produce good results in the AVEC'17 Challenge [9, 22, 27]. The parameters of the Random Forest model were optimised using ten-fold cross-validation and experimenting with different numbers of estimators (1, 10, 30, 40, 50, 100, and 200, respectively). After selecting the best parameters using a 10-fold cross validation in the training set, a model was trained across the entire training set and evaluated on the development set. We combined the train and development sets for test set predictions and followed a similar approach. Additionally, the features were standardised according to their distribution in the train set, and the diagnosis labels were min-max normalised.

5 EXPERIMENTAL EVALUATION

5.1 Cross dataset prediction for breath signals

In this section, we discuss the adapted 1D CNN + LSTM model used and validate its performance on the annotated breath events from the SBC and the Extended DAIC-WOZ datasets.

To define a set of desirable input window sizes for the 1D CNN + LSTM model, we analysed the duration of the answers to all the questions present in this set. From the 92 questions presented in the train set, the average response time per question is 10.8 seconds, with a standard deviation of 6.1 seconds. For this reason, we explored breath signals with a window size of 4, 6, 8, and 10 seconds.

Initially, we explored the effects of EBUR128 loudness normalisation [29] in breathing signal prediction in the SBC. When applying loudness normalisation to the input for both training and evaluation phases, with the original window size, we observed a sharp

Table 1: Summary of the features used for PHQ-8 correlation analysis. The functionals will be calculated across the selected reactions, for instance, "all questions" or "positive" questions only.

	Feature (functionals)	*Description*
Conversational	Reaction time	Time between the end of AI Agent's question and the beginning of the answer.
	Word rate	Number of words in the answer divided by the speech duration.
Sentiment	Valence (max, mean)	Positive-negative dimension of the answer [15].
	Arousal (max, mean)	Active-passive dimension of the answer [15].
	Dominance (max, mean)	Powerful-weak dimension of the answer [15].
Breath	Breath signal and 1st derivative (mean, std)	Predicted chest volume and respective variation rate across time.
	Breath signal during silences and 1st derivative (mean, std)	Predicted chest volume and respective variation rate across time during silences, comparable to individual breath events.
	Inhalation slope (mean, std, max)	Slope of the line defined between the inhalation onset and the following exhalation onset.
	Peak-to-Peak distance (mean, std, max)	Duration of a breathing cycle.
	Volume (mean, std, max)	Volume of the chest post-inhalation.

decrease in the model performance on the cross-validation set, with a performance below the baseline. The 1D CNNs seem to work better with non-loudness-normalised data. Therefore, we trained the breath prediction model used in this study in the non-normalised dataset. We present the results from the cross-validation of the breathing signal prediction model, using different windows, in Table 2. We see a slight decrease in performance when using a window of six seconds; however, the prediction performance is still above the baseline defined for the ComParE2020 Challenge [26]. Thus, we consider the trade-off between window size minimisation and performance loss satisfactory.

Table 2: Pearson Correlation Coefficient (PCC) of the breathing signal prediction using a 1D CNN + LSTM model with different window sizes. Baseline corresponds to the performance of the baseline of the ComParE2020 Challenge [26] in the development set.

	Baseline	*1D CNN + LSTM*			
window size (s)	-	16 [14]	10	8	6	4
PCC	0.507	0.607	0.582	0.574	**0.583**	0.367

We evaluated the selected models under three different preprocessing conditions: 1) applying the model proposed in [14] to the complete audio signal, with no pre-cropping of the AI-speech parts; under this approach, we have a continuous signal for the entire session, 2) cropped successive non-AI instances and predicted the breath signal using the adapted model with a window of six seconds, which results in continuous predictions per answer, but not for the full interview, and lastly, 3) similar to the second condition, but applying EBUR128 normalisation to the resulting audio chunks before breath prediction.

We compared the similarity between annotated breath events for the target dataset and the SBC. For this purpose, we extracted simple functionals (mean and std of the breath events points) from the breath signal and the first derivative (see the respective functionals in Table 3). First, we evaluated how the prediction models affected the breath event characteristics in the SBC. The original and the adapted prediction models lead to a high increase in the average first derivative value and signal standard deviation compared with the respective ground truth. As anticipated, breath prediction along the entire signal leads to predictions that deviate from the expected values. The designed model produces a breath prediction using a sliding window, so if we do not remove the AI component of the speech, this part will contribute to the breath prediction of the respective window. There is no significant variation in model performance when applying the model to the recording's non-AI instances. Overall, the predicted breath signals for breath events for both sets have similar functional values with the training/development dataset predictions.

Finally, we evaluate the effect of the window size constraint and the number of individual answers extracted per session in the DAIC-WOZ. After discarding five samples with answers shorter than six seconds, the training, development, and test set have 105, 33, and 46 samples, respectively. Among these samples, 21, 7, and 14 participants have depression, according to the binary labels (PHQ-8 > 10), highlighting the importance of using continuous PHQ-8 scores to analyse the validity of the proposed feature set. The number of excluded questions due to the time restraint imposed by the breath signal prediction model in each session has a similar distribution across the three sets.

5.2 Correlation analysis between psychopathology and features

In this section, we evaluate if the proposed breath features are informative for a depressive state, we probe if a particular type of question is more predictive of depression symptoms, and we assess the advantages of answer selection versus processing all participant instances.

Table 3: Comparison of basic statistics on annotated audible breath events.

		signal value		*1st derivative*	
Dataset	***Model and input variations***	*mean*	*std*	*mean*	*std*
SBC	Ground truth	0.030	0.081	0.257	0.469
	Original model [14]	-0.074	0.239	1.016	0.846
	Adapted model (6 second input)	-0.122	0.251	1.108	0.860
Extended DAIC-WOZ [12]	Full recording (original model)	0.197	0.076	0.347	0.352
	Cropped non-AI instances (adapted model)	0.043	0.135	0.614	0.532
	Cropped normalised non-AI instances (adapted model)	0.049	0.179	0.807	0.653

Table 4: Linear regression models to predict depression severity scores. The significance of each model was defined based on the p-value of each regression F-score. The F-score values with a p-value ≤ 0.05 are highlighted. We present the root mean square error (RMSE) on the development set for the significant models. No significant correlations were found for the question types "Negative", "Neutral" and "Positive". * - p-value ≤ 0.05; ** - p-value ≤ 0.01; * - p-value ≤ 0.001.**

		All		***Conversational***		***Sentiment***		***Conv + sent***		***Breath***	
Summary	*n*	*F-score*	*RMSE*	*F-score*	*RMSE*	*F-score*	*RMSE*	*F-score*	*RMSE*	*F-score*	*RMSE*
Mixed questions	103	**2.06****	6.13	**6.85*****	6.12	**2.80****	6.66	**3.03****	6.17	**2.00****	6.35
General questions	64	**2.26****	5.93	**3.31***	5.82	1.57	-	1.92	-	**2.10***	6.10
All questions	105	**1.77***	5.10	2.52	-	**3.31****	6.40	**2.49***	6.23	**1.89***	5.98
All reactions	105	1.48	-	**3.26***	6.29	1.70	-	**2.07***	5.80	**1.73***	5.84

We present the performance of the linear regressions model for depression prediction in Table 4. The Bonferroni-corrected p-value threshold for the five tests conducted per question is 0.01. Considering the dependency between tests and the exploratory nature of this study, we report all instances with a p-value ≤ 0.05. The number of cases per question type is not the same, since some samples have no representation in the session.

The question types that show more predictive values are mixed and general questions, corresponding to questions without clear polarity, allowing for a more diverse set of answers and follow-up questions that go deeper into the participant's previous questions. Unfortunately, the differences in sample sizes do not allow us to compare the predictability performance between general and mixed questions. However, based on these preliminary results, we hypothesise that ambiguous and open questions have relevant predictive values. Furthermore, answers to negative questions do not show a clear correlation with depression in the current dataset, despite what we initially expected. There may be several reasons for this. Interviews are not standardised, making it difficult to directly assess the correlation of a specific question subgroup, since not all the participants will have the same number of negative questions asked to them, and the depression severity distribution is not consistent between question sets. Furthermore, the answer duration may have a strong impact on the performance of conversational and breath features. Finally, perceived valence may vary depending on the participant. The current question type categorisation was designed with a focus on emotion elicitation, assuming that "negative" questions will more likely elicit "negative" emotional states. In the future, we would like to extend the present question categorisation method by segmenting the questions based on key topics.

The number of questions per session is a limiting factor for robust feature summarisation; for this reason, as a preliminary study, we explore the feature/diagnosis correlation across all participant reactions. The summarisation across all responses does not lead to significant models for the sentiment features. The conversational and breath feature models present a lower p-value, ≤ 0.05, although still not significant enough after the Bonferroni correction. When evaluating only the answers, we observe a substantial performance increase for the model trained using sentiment features, suggesting that the key emotional content of the session is in the answering components of the interview. The model trained with breathing features shows an F-score with a p-value of 0.03, motivating us to hypothesise that breath features are more meaningful when applied to the answer component of the interview.

When comparing the p-values of the different models trained across all questions, breath features were the second most predictive feature set, only surpassed by the combination of all features. This is a good finding to motivate the relevance of these features for interpretable depression assessment, but a more robust distinction between answers and reactions would allow a more meaningful feature analysis. Currently, the participant speech chunks are split based on AI Agent's interactions, independent of their length. The main advantage of this strategy is that we guarantee that the AI Agent's speech does not affect the continuous breath signal prediction. However, future work on the viability of semi-continuous signals for feature analysis would be relevant to advancing the field, since it would allow us to assess breathing characteristics' variation during interviews over a more extended period.

To better understand the particular relation between each separate feature and depression, we analysed the Pearson Correlation Coefficient between each feature set and the respective PHQ-8 score in the train set (Table 5).

We observed a longer reaction time per question in depressed participants. Further, valence values are negatively correlated with depression, suggesting that participants with depression express more negative emotions across the session. Combined with the negative correlation with maximum arousal and dominance, we can infer that participants with higher levels of depression have a higher expression of emotions related to sadness [5].

For breathing, we observe lower first derivative values during silences for participants with higher PHQ-8 scores, suggesting either shallow or slow breathing events, or breath-hold events. In addition, we find that the inhalation slopes in participants with higher PHQ-8 values have a lower maximum value across the session, which supports the hypothesis that depressed participants generally have shallow and/or slow breathing episodes. Consequently, the standard deviation of the same function is smaller.

The PHQ-8 scores provided are based on a simple self-assessment questionnaire and are not comparable to a clinical diagnosis. Hence we avoid a direct comparison with breathing literature on major depressive disorders. Nevertheless, the breathing characteristics highlighted are related to depression and affect literature (particularly concerning sadness) and are consistent with the low arousal and valence values observed. Furthermore, we see a significantly higher mean breath signal value for participants with higher PHQ-8. Due to the speech variations in the breath signal, this feature is not easily interpretable. However, less interpretable features were added to the feature list to account for losses in information due to errors in peak detection, and consequent errors in volume, peak to peak distance, and inhale slope definition.

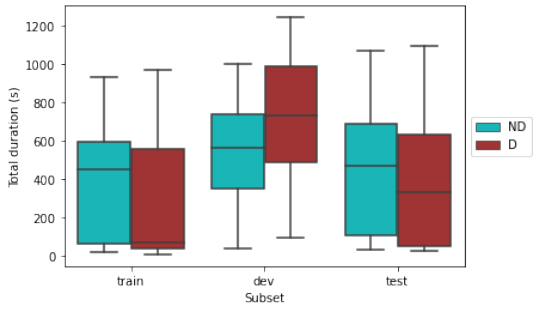
5.3 Depression score regression

We give the comparative performances of the proposed linear and non-linear models in Table 6. In addition, we present the challenge baseline and the top challenge submissions for comparison purposes.

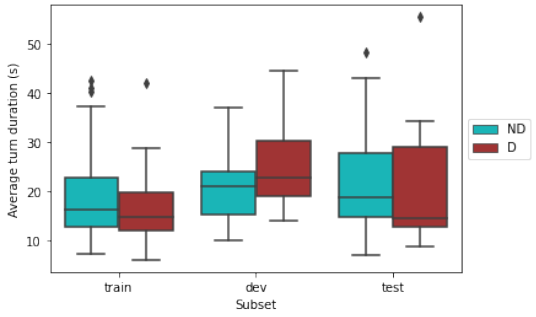
The best performance was achieved in the development set using a Linear Regression model with conversational, linguistic sentiment, and breath features. The resulting model leads to an RMSE decrease of 23% compared with the best performing baseline for the same set. However, as observed in the challenge baseline, audio-based models have significantly reduced performance on the test set, as opposed to the development set. Looking at the linear regression results for the model trained with only breath features, we observed a similar tendency, with a significant decline in performance between development and test sets. Nonetheless, the performance for this estimator outperforms the more complex baseline audio-based model, encouraging further exploration of the impact of breath features on depression assessment.

The best performing model on the test set omits breath features, pointing to a different breath feature value distribution between the combined train and development sets used to train the final model and the test set. This model surpasses the challenge baseline and is one of the top submissions in the test set. Moreover, we observed that Random Forest regressors performed more consistently between the development and test set; these models do not assume linearity between the feature and prediction, allowing more flexibility in the feature importance between sets.

To further explore the differences in acoustic information across different sets, we looked at the total duration of all participant answers (including reaction time), and the average length of a turn after a question per session. We present the respective distributions in Figure 2. Since we want to focus on depression analysis, we compared the distributions of depressed and non-depressed participants.



(a) Total turn duration



(b) Average turn duration per answer

Figure 2: Distribution of a) total answer duration and b) mean answer duration per session in each set. The division between "Non-Depressed" (ND) "Depressed" (D) was based on the binary classification provided.

There is a significant difference in the number of questions selected and total answer time between the development and test sets. This pattern is observed for both depressed and non-depressed participants. We expect the variation in the number and duration of answers to have a high impact on turn level feature extraction. Particularly, breathing signal prediction will be more significant when evaluated across a longer audio sample, since the model uses a sliding window approach. Hence, we suggest that audio features will be less robust on the test set due to the comparative lack of relevant audio information. The performances of the top challenge submissions are consistent with this conclusion, with a text-based model achieving the best performance on the test set, and the lowest variation in performance between both sets. Hybrid models

Table 5: Pearson correlation coefficients between the average of each feature across all the answers or all participant actions and the continuous diagnosis for depression. The values presented correspond to the performance on the train set (* - p-value ≤ 0.05; ** - p-value ≤ 0.01. The functionals regarding "Peak to Peak distance" and "Breath Volume" did not achieve significant correlation values in the train set.

Type	*Feature*	*Functional*	*All answers*	*All actions*
Conversational	Reaction time	-	0.214*	0.222*
Sentiment	Valence	mean	-0.293**	-0.274**
		max	-0.278**	-
	Arousal	mean	-	-
		max	-0.283**	-
	Dominance	mean	-	-
		max	-0.301**	-
Breath	Breath signal	mean	0.305**	0.235*
		std	-	-
	1st derivative	mean	-	0.273**
		std	-	-
	Breath signal during silences	mean	0.239**	0.247**
		std	-	-
	1st derivative during silences	mean	-0.221*	-
		std	-	-
	Inhale slope	mean	-	-
		max	-0.207*	-
		std	-0.195*	-

Table 6: Root mean square error (RMSE) results for the depression assessment task on development and test sets. For comparison, we provide the challenge baseline and top submissions. The best RMSE performance and corresponding MAE per subset is highlighted for each model group. LR - Linear Regression model; RF - Random Forest model; DL - Deep Learning approach; SGD-LR - Stochastic Gradient Descent Linear Regressor.

		Model	*Dev*	*Test*
Challenge baseline	audio	RF	6.74	7.78
	video	RF	7.13	**6.97**
	audio + video	RF	**6.62**	7.05
Proposed methods	conv + sent + breath	RF	6.63	5.85
		LR	**5.10**	6.80
	conv + sent	RF	6.33	5.83
		LR	6.23	**5.62**
	breath	RF	6.98	6.40
		LR	5.98	7.65
	selected features	RF	5.96	6.37
		LR	6.53	5.67
Yang et al [36]	audio + video + text	DL	**3.09**	5.40
Yang et al [35]	audio + video + text	DL	4.65	5.97
Sun et al [27]	selected-text	RF	4.97	**4.98**
Gong et al [9]	audio+video+text	SGD-LR	3.54	4.99

see a significant but less steep increase in performance between sets. Further analysis of the referred works on the contribution of each modality for the test set predictions would be useful for understanding the limitations of the dataset.

Although further work is required to confirm the potential of breathing features for depression detection, the results presented show the value of this new set of interpretable features. Additional engineering of the feature set, such as feature selection, and tackling the effects of collinearity, are the following steps to extend our understanding of the proposed approach.

6 CONCLUSIONS

In this paper, we explored the potential of simple breath features for depression assessment, based on an imperfect measurement of breath signals. We used the well-documented DAIC-WOZ dataset for the depression analysis task. However, since this dataset did not have a breathing ground truth annotation, we used a cross-dataset prediction approach, which we validated on a subset of DAIC-WOZ annotated with breath events. The interview setting of the DAIC-WOZ dataset further allowed us to test the performance of the proposed cross-dataset breathing prediction under mismatch situations that more closely resemble interactions observed during therapy sessions.

One of our premises was that during an interview, different questions would provoke different emotional tones in subjects, and questions could be grouped accordingly. Our results suggest that session summarisation based on "General" and "Mixed" questions leads to good linear models, implying that open questions will produce more meaningful reactions for mood interpretation. The

results obtained are motivating, with informative breathing features across these question types. However, negative questions did not provide any clear correlation with breathing features. One reason of this might be the shortness of the answers, which is one of the limitations we found for breathing features. Compared to conversational features, they are affected more from the answer duration since short answers do not provide an opportunity to extract useful features. Our study furthermore revealed that dataset characteristics, particularly turn duration, can impose limitations on the prediction accuracy of the breath signal.

Another limitation of our study lies in the use of PHQ-8 scores for depression severity estimation, which relies on self-assessment and cannot be directly compared with most of the literature on major depression disorder. Also, the co-morbidity between depression and anxiety disorder makes analysis more difficult, as anxiety is frequently observed in patients with PTSD, and might affect breathing features, hence limiting the assessment potential of them for depression analysis specifically.

Overall, the present study evaluated the correlation between breathing-related features and conversational, linguistic sentiment, and depression severity level, focusing on interpretable features to compare the correlations found within the literature. Our comparisons showed that features such as duration and deviations of breathing episodes provide intelligible features with the advantage of carrying non-identifiable information and hence being more privacy-preserving than the audio signal. When evaluating the individual Pearson correlation coefficients between features and the PHQ-8 scores, we observed a negative correlation between depression and arousal, valence and dominance, pointing to states such as sad and depressed. Moreover, correlations found for breathing features suggested slow and shallow breaths to be indicative of high depression scores, consistent with the detected mood.

The suggested approach could have applications in assessmentby clinicians with an interpretable automated prediction as well as emotion detection in conversational speech analysis. Our results suggest that there is room for further exploration on using breathing features for interpretable depression detection. More robust breathing rate assessment could improve the contribution of this feature even further, and frequency-domain features, such as continuous wavelet transforms, present a potential future direction.

REFERENCES

[1] Almila Akdag Salah, Albert Ali Salah, Heysem Kaya, Metehan Doyran, and Evrim Kavcar. 2021. The sound of silence: Breathing analysis for finding traces of trauma and depression in oral history archives. *Digital Scholarship in the Humanities* 36, Supplement_2 (2021), ii2–ii8.

[2] Sharifa Mohammed Alghowinem, Tom Gedeon, Roland Goecke, Jeffrey Cohn, and Gordon Parker. 2020. Interpretation of depression detection models via feature selection methods. *IEEE Transactions on Affective Computing* (2020).

[3] Tathagata Banerjee, Matthew Kollada, Pablo Gersberg, Oscar Rodriguez, Jane Tiller, Andrew E Jaffe, and John Reynders. 2021. Predicting Mood Disorder Symptoms with Remotely Collected Videos Using an Interpretable Multimodal Dynamic Attention Fusion Network. *arXiv preprint arXiv:2109.03029* (2021).

[4] Frans A Boiten, Nico H Frijda, and Cornelis JE Wientjes. 1994. Emotions and respiratory patterns: review and critical analysis. *International journal of psychophysiology* 17, 2 (1994), 103–128.

[5] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.

[6] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 1061–1068.

[7] Carsten F Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J Leitão, et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 1 (2013), 27–46.

[8] Susanne Fuchs and Amélie Rochet-Capellan. 2021. The respiratory foundations of spoken language. *Annual Review of Linguistics* 7 (2021), 13–30.

[9] Yuan Gong and Christian Poellabauer. 2017. Topic modeling based multi-modal depression detection. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 69–76.

[10] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. 3123–3128.

[11] Ikuo Homma and Yuri Masaoka. 2008. Breathing rhythms and emotions. *Experimental Physiology* 93 (9 2008), 1011–1021. Issue 9. https://doi.org/10.1113/EXPPHYSIOL.2008.042424

[12] Heysem Kaya, Dmitrii Fedotov, Denis Dresvyanskiy, Metehan Doyran, Danila Mamontov, Maxim Markitantov, Alkim Almila Akdag Salah, Evrim Kavcar, Alexey Karpov, and Albert Ali Salah. 2019. Predicting Depression and Emotions in the Cross-Roads of Cultures, Para-Linguistics, and Non-Linguistics. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop* (Nice, France) *(AVEC '19)*. Association for Computing Machinery, New York, NY, USA, 27–35. https://doi.org/10.1145/3347320.3357691

[13] Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *Journal of affective disorders* 114, 1-3 (2009), 163–173.

[14] Maxim Markitantov, Denis Dresvyanskiy, Danila Mamontov, Heysem Kaya, Wolfgang Minker, and Alexey Karpov. 2020. Ensembling end-to-end deep models for computational paralinguistics tasks: ComParE 2020 mask and breathing sub-challenges. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (2020), 2072–2076. https://doi.org/10.21437/Interspeech.2020-2666

[15] Saif M. Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*. Melbourne, Australia.

[16] Michael B Morrissey and Graeme D Ruxton. 2018. Multiple regression is not multiple regressions: the meaning of multiple regression and the non-problem of collinearity. *Philosophy, Theory, and Practice in Biology* 10, 3 (2018).

[17] Muhammad Muzammel, Hanan Salam, Yann Hoffmann, Mohamed Chetouani, and Alice Othmani. 2020. AudVowelConsNet: A phoneme-level based deep CNN architecture for clinical depression diagnosis. *Machine Learning with Applications* 2 (2020), 100005.

[18] Syed Arbaaz Qureshi, Mohammed Hasanuzzaman, Sriparna Saha, and Gaël Dias. 2019. The Verbal and Non Verbal Signals of Depression–Combining Acoustics, Text and Visuals for Estimating Depression Level. *arXiv preprint arXiv:1904.07656* (2019).

[19] Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. 2019. Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 81–88.

[20] Emna Rejaibi, Ali Komaty, Fabrice Meriaudeau, Said Agrebi, and Alice Othmani. 2022. MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control* 71 (2022), 103107.

[21] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. 2019. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*. 3–12.

[22] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 3–9.

[23] Saurabh Sahu and Carol Espy-Wilson. 2014. Effects of depression on speech. *The Journal of the Acoustical Society of America* 136, 4 (2014), 2312–2312.

[24] Saurabh Sahu and Carol Espy-Wilson. 2016. Speech Features for Depression Detection. In *Proc. INTERSPEECH*. 1928–1932.

[25] Abraham Savitzky and Marcel JE Golay. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry* 36, 8 (1964), 1627–1639.

[26] Björn W. Schuller, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, Georgios Rizos, Maximilian Schmitt, Lukas Stappen, Harald Baumeister, Alexis Deighton MacIntyre, and Simone Hantke. 2020. The INTERSPEECH 2020 Computational Paralinguistics Challenge:

Elderly Emotion, Breathing & Masks. In *Proc. Interspeech 2020*. 2042–2046. https://doi.org/10.21437/Interspeech.2020-32

[27] Bo Sun, Yinghui Zhang, Jun He, Lejun Yu, Qihua Xu, Dongliang Li, and Zhaoying Wang. 2017. A random forest regression method with selected-text feature for depression assessment. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 61–68.

[28] Martin J Tobin, Tejvir S Chadha, Gilbert Jenouri, Stephen J Birch, Hacik B Gazeroglu, and Marvin A Sackner. 1983. Breathing patterns: 1. Normal subjects. *Chest* 84, 2 (1983), 202–205.

[29] European Broadcasting Union. 2020. Loudness normalisation and permitted maximum level of audio signals. https://tech.ebu.ch/docs/r/r128.pdf

[30] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*. 3–10.

[31] Jinfang Wang, Ke Lv, Chang Liu, Xinli Nie, Dhananjaya Gowda, and Shuxin Luan. 2020. Automatic Assessment for Severe Self-Reported Depressive Symptoms Using Speech Cues. *IEEE Transactions on Cognitive and Developmental Systems* 13, 4 (2020), 875–884.

[32] A. L. Winkworth, P. J. Davis, R. D. Adams, and E. Ellis. 1995. Breathing Patterns During Spontaneous Speech. *Journal of Speech and Hearing Research* 38 (1995), 124–144. Issue 1. https://doi.org/10.1044/JSHR.3801.124

[33] Torsten Wörtwein and Stefan Scherer. 2017. What really matters—an information gain analysis of questions and reactions in automated PTSD screenings. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 15–20.

[34] Le Yang, Dongmei Jiang, and Hichem Sahli. 2020. Feature augmenting networks for improving depression severity estimation from speech signals. *IEEE Access* 8 (2020), 24033–24045.

[35] Le Yang, Dongmei Jiang, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. 2017. Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 53–59.

[36] Le Yang, Hichem Sahli, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Dongmei Jiang. 2017. Hybrid Depression Classification and Estimation from Audio Video and Text Information. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge* (Mountain View, California, USA) *(AVEC '17)*. Association for Computing Machinery, New York, NY, USA, 45–51. https://doi.org/10.1145/3133944.3133950

[37] Vera Eva Zamoscik, Stephanie Nicole Lyn Schmidt, Martin Fungisai Gerchen, Christos Samsouris, Christina Timm, Christine Kuehner, and Peter Kirsch. 2018. Respiration pattern variability and related default mode network connectivity are altered in remitted depression. *Psychological Medicine* 48, 14 (2018), 2364–2374.