DEPAC: a Corpus for Depression and Anxiety Detection from Speech

Mashrura Tasnim, Malikeh Ehghaghi, Brian Diep, Jekaterina Novikova
{mashrura, malikeh, brian, jekaterina}@winterlightlabs.com
Winterlight Labs
Toronto, Canada

Abstract

Mental distress like depression and anxiety contribute to the largest proportion of the global burden of diseases. Automated diagnosis system of such disorders, empowered by recent innovations in Artificial Intelligence, can pave the way to reduce the sufferings of the affected individuals. Development of such systems requires information-rich and balanced corpora. In this work, we introduce a novel mental distress analysis audio dataset DEPAC, labelled based on established thresholds on depression and anxiety standard screening tools. This large dataset comprises multiple speech tasks per individual, as well as relevant demographic information. Alongside, we present a feature set consisting of hand-curated acoustic and linguistic features, which were found effective in identifying signs of mental illnesses in human speech. Finally, we justify the quality and effectiveness of our proposed audio corpus and feature set in predicting depression severity by comparing the performance of baseline machine learning models built on this dataset with baseline models trained on other well-known depression corpora.

1 Introduction

Effective treatment for psychiatric diseases requires characterizing disease profiles with high accuracy. The traditional schema for diagnosis is based on clustering of non-specific physical and behavioral symptoms, which makes the diagnostic process challenging. For example, in major depressive disorder (MDD), high disease heterogeneity and lack of agreed-upon assessment standards necessitate a high degree of clinical experience and training to make an accurate diagnosis. Both clinician-administered and self-rated clinical assessments for MDD, such as the Hamilton Depression Scale (HAM-D) (Hamilton and Guy, 1976), Montgomery Asberg Depression Scale (MADRS) (Montgomery and Åsberg, 1979), Beck Depression Inventory (BDI) (Beck et al., 1988), and Patient Health Questionnaire (PHQ-9) (Löwe et al., 2004) are suboptimal in many ways. Each assess the illness through different symptom domains, have low construct validity, lack specific behavioral references, and are subjective (Berman et al., 1985; Nemeroff, 2007; Wakefield, 2013). Moreover, participants are often reluctant to fill-out the self rated assessment in regular intervals. These issues can lead to misdiagnosis, which impacts treatment timelines and can lead to poor clinical outcomes.

In contrast, language can be an effective alternative to objectively characterize psychiatric illness. For example, emotion and cognition are both affected in MDD. As a result, depressed patients demonstrate negative emotional bias in memory, attention, and event-interpretation (Mathews and MacLeod, 2005), as well as more general impairment in attention, memory, and decision-making (Cohen et al., 1982; Blanco et al., 2013). These effects are manifested in patients' language in a variety of ways, for example, slowed rate of speech, volume, prosody, as well as increased use of first-person pronouns, negatively valenced speech content, and use of absolute words (Flint et al., 1992; Fineberg et al., 2016). Therefore, automated computational analysis of speech represents an excellent data source to develop digital biomarkers for mental illness. This kind of automated assessment takes only a few minutes of audio recording, therefore is less time-consuming, and would reduce burden on the individuals. However, such model development requires access to datasets of sufficient quality and size.

The recent development of speech-based computational models for measuring depression prevalence and severity has been accelerated by the introduction of Audio-Visual Emotion Recognition Challenge (AVEC) in 2013. A subset of the audio-visual depressive language corpus (AViD-Corpus) was introduced as challenge corpus for 2013 (Valstar et al., 2013) and 2014 (Valstar et al., 2014) Depression Recognition Sub-Challenge (DSC) of the event. This dataset comprises 150 recordings in German language, divided equally into training, development and test partitions. Predicting depression severity on BDI-II scale was the challenge specified task.

Another popular dataset in this area is the Distress Analysis Interview Corpus (DAIC) (Gratch et al., 2014). It contains semi-structured clinical interviews in English language formulated to support diagnosis of psychological conditions such as anxiety, depression, and post-traumatic stress disorder. Different subsets of this dataset were used as the challenge corpus of AVEC 2016, 2017 and 2019 (Valstar et al., 2016; Ringeval et al., 2017, 2019) where challenge participants reported PHQ-8 scores predicted by their respective regression models.

However, the depression corpora used in previous research suffer from two vital limitations. Firstly, the small sample size in the existing depression datasets increases the risk of overfitting in the machine learning models. For example, the number of recordings in the AVEC challenges available for model training range from 50 to 189, which is far from sufficient. Secondly, the datasets in the previous works lack in linguistic variety, as they only contain a small number (only one or two) of samples per subject. To mitigate these challenges, in this work we introduce the **DEP**ression and **A**nxiety **C**rowdsourced corpus (DEPAC) as a novel dataset that is rich in the diversity of speech tasks and subjects and is tailored to capture the signs of anxiety and depression to make accurate prediction on subjects' psychological state. We also present a set of acoustic and linguistic features extracted from the corpus which incorporates domain knowledge of clinical and machine learning experts. Finally, we benchmark our dataset with several baseline machine learning models that use this set of features, to show that this novel dataset is well-suited for the machine learning-based methods with the goal of generating speech biomarkers for depression.

2 DEPAC Corpus

The DEPAC corpus introduced in this work was collected with the goal of gathering a large training dataset to identify candidate speech and language features that are specific to a given psychiatric disease. Data collection for the corpus was approved by the Institutional Review Board (IRB). This is a proprietary dataset, collected via crowdsourcing and consists of a variety of self-administered speech tasks. The participants completed these tasks using Amazon Mechanical Turk[1] (mTurk), a platform where individuals are paid to complete short tasks online (Paolacci et al., 2010). The speech samples were then manually transcribed and compiled along with participant demographic information into the final corpus.

2.1 Platform and Instrumentation

Once recruited for this study via mTurk, participants were able to remotely complete a range of tasks including surveys and responding to audio prompts. Participants were required to have:

1. A desktop or laptop computer
2. A working microphone
3. Chrome or Mozilla Firefox browser

Amazon facilitated payment between the experimenter and the participant.

2.2 Recruitment and Screening

Participation in the study was voluntary. Participant eligibility was configured to only permit individuals located in Canada and the United States. Amazon verified the location of participants by confirming their address and associated credit card. Locations were used to assess eligibility only.

The platform also restricted participation to individuals with an mTurk approval rating of at least 95%. This preliminary criterion attempted to ensure that participation was restricted to those who had historically consistently followed task instructions.

During the study, participants saw a short description of the task, the approximate length of the task (5 to 8 minutes, depending on the condition they were randomly placed into), and the per-minute payment for their time. Participants were compensated at a rate of $0.16 per minute. This is well above the average payment rate for mTurk tasks and above the recommended rate of $0.10/minute (Chandler and Shapiro, 2016).

As part of our exclusion criteria, individuals with a history of chronic alcohol or drug dependency within the past 5 years, as well as participants with clinically significant vision or hearing impairment, were excluded from the study.

[1] https://www.mturk.com

2.3 Transcription and Quality Assurance

Each audio sample gathered from the mTurk platform was assigned to a trained transcriptionist to follow the protocols and annotation formats detailed in the CHAT manual (MacWhinney, 2000) that was used to transcribe TalkBank, which is the largest open repository of spoken data (MacWhinney, 2007). The transcriptionists annotated via an internally developed tool where they had access to the recording and a platform for transcribing the content of the audio file, separating the audio file into utterances, and performing quality assurance. Samples with minor audio issues not impacting the transcriptionist's ability to produce an accurate transcript were processed as normal. Samples that could not be properly transcribed due to excessive background noise, poor audio quality, or other external issues such as the presence of multiple speakers in the file were tagged as unusable and were omitted from the corpus. In total, 91 samples out of 2765 collected samples were tagged as such and omitted.

2.4 Demographic Data Collection

Upon consenting, participants were asked to indicate whether they are native English speakers (i.e., whether they learned the English language before the age of 5 years old). They were also asked to indicate their age, gender, and education level.

2.5 Speech tasks

During each recording session, the subjects completed the following standard tasks, selected to elicit speech patterns that can be analyzed for acoustic and linguistic features that correlate to psychiatric state:

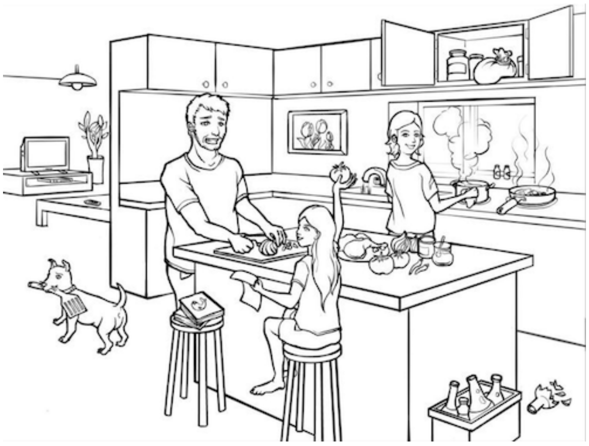
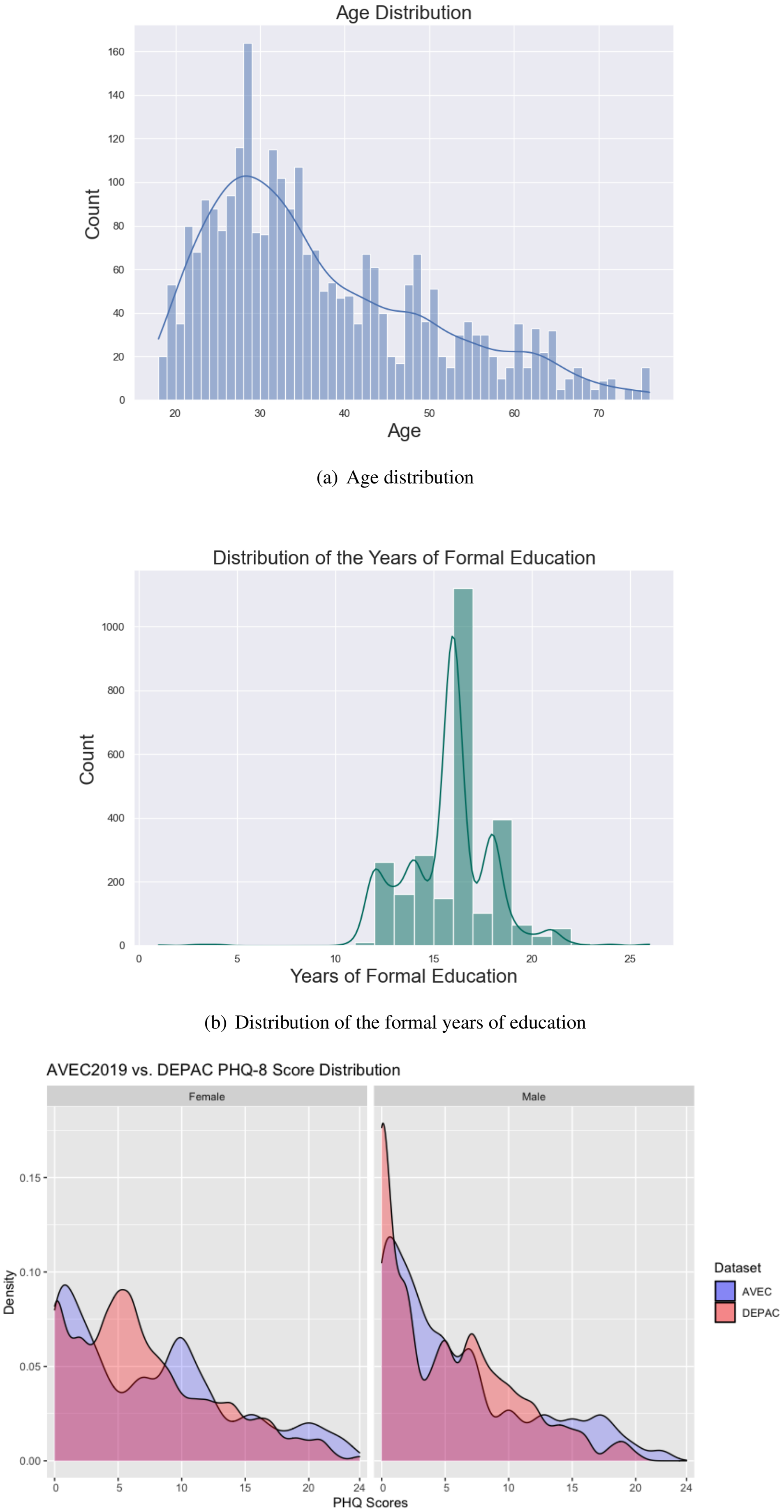


Figure 1: 'Family in the kitchen' image used in the picture description task.

Criteria	AVEC2013, 2014	DAIC-WoZ	DEPAC (our)
Language	German	English	English
# of speech tasks	2	1	5
# of samples total / per subj.	150 / 2	189 / 1	2674 / 5
Depression scale	BDI-II	PHQ-8	PHQ-9
Anxiety scale	-	-	GAD-7
Avg. depression score	15.34(± 12.13)	6.65 (± 6.11)	6.56 (± 5.56)
Depression score range in the corpus	0-45	0-23	0-27

Table 1: Description of our DEPAC dataset and its comparison to existing depression/anxiety corpora.

- **Phoneme Task:** Participants were asked to sustain a phoneme sound (e.g., /ā/) for as long as they could, up to one minute. They could cease making the sound whenever they choose. Due to difficulty in finding voiced parts in continuous speech, sustaining vowels would be optimal for measuring source and respiration features (e.g., shimmer) (Low et al., 2020).
- **Phonemic fluency:** Phonemic verbal fluency was evaluated using the FAS ('F', 'A', 'S') (Borkowski et al., 1967) task (letter “F"). This assessment has been used widely in a variety of populations, including individuals with Alzheimer's Disease (AD). The average duration of this speech task was 22.13 seconds in DEPAC dataset.
- **Picture description**: A static image depicting an event was presented to the subject, and they were asked to describe what is happening in their own words. The average length of picture-based narratives was 46.60 seconds. Tasks of this type have been shown to be good proxies for spontaneous discourse (Giles et al., 1996). Picture description was found to be an effective speech task in evoking situations that required more cognitive effort and caused noticeable changes in speech for detecting depression (Jiang et al., 2017). In this study, a proprietary image 'Family in the kitchen' (Figure 1) was used, which was designed to match the 'Cookie theft' picture (Goodglass et al., 2001) in style and content units. The picture was a line drawing of an everyday scene, containing three to four characters, two salient action items (e.g., broken bottle, or steaming pot), and a similar number of object units (20-25), action items (9-10) and locations (2) (Forbes-McKay and Venneri, 2005). Our core design guidelines to develop this picture are



(a) Age distribution

(b) Distribution of the formal years of education

(c) Distribution of PHQ-8 scores in AVEC 2019 and DEPAC dataset by gender

Figure 2: Distribution of the participants' demographics in mTurk Study.

listed in A.1.

- **Semantic fluency:** Participants were asked to list as many positive future experiences as they can within one minute. They were given time parameters to guide them, such as future events predicted to happen within three weeks, within one month, within one year, and so on. They were allowed to describe as little or as much as they choose. Performance on verbal fluency tasks are found to correlate with executive deficits caused by depression (Fossati et al., 2003). The length of speech in this task was 43.76 seconds on average.
- **Prompted narrative:** Participants were asked to describe an event, interest, or hobby based on a single prompt, e.g., "Describe your day", "Describe a travel experience" and "Describe a hobby you have". Participants were allowed to describe as much or as little as they choose. Narrative speech provides an opportunity to elicit speech containing the linguistic structures and acoustic information that is known to contain indicators of depression (Trifu et al., 2017). The average duration of the prompted speech in the collected dataset was 45.34 seconds.

2.6 Clinical Assessments

The following two mental health assessment questionnaires were completed by the participants after the recording session:

Patient Health Questionnaire (PHQ-9): The PHQ-9 is a well established 3-point self-rated measure for **depressive symptoms** that has been validated against clinician rated measures (Kroenke et al., 2001). It contains 9 questions which correspond to the core criteria of the Diagnostic And Statistical Manual of Mental Disorders (DSM) for depression. Scores on this scale range from 0 to 27 with diagnostic cut-off thresholds for depression severity. Scores less than 5 represent the individuals with no depression; individuals with a mild or subthreshold depressive disorder are reflected by scores from 5 to 9; scores between 10 and 14 indicate moderate severity level of depression, and scores 15 or higher signify major depressive disorder in the participants (Kroenke et al., 2001).

Generalized Anxiety Disorder - 7 (GAD-7): The GAD-7 is a popular self-rated measure of general **anxiety symptoms** that is scored from 0 to 21 (Spitzer et al., 2006). It has been validated against clinical diagnosis and has been shown to be robust as a screening tool and a continuous measure of symptom severity. Scores of 10 or above indicate a reasonable threshold for detecting individuals with generalized anxiety disorder. Similar to the levels of depressive disorder in PHQ-9, 5, 10, and 15 are the cut points on the GAD-7 scale to classify anxiety severity level into minimal, mild, moderate and severe groups (Spitzer et al., 2006).

2.7 Corpus Composition

The dataset consists of 2,674 audio samples collected from 571 subjects (Table 1). 54.67% of the study subjects are female and 45.33% are male. The age of the subjects ranges between 18 and 76, and they received 1 to 26 years of formal education.

Figure 2 illustrates the demographic distribution of the mTurk study. The age distribution is shifted toward the left around its average value, which is equal with 36.85, indicating that most of the dataset is made up of young or middle-aged adults (Figure 2(a)). Moreover, it is witnessed in the education level distribution plot that the most of the participants received higher education, with on average around 15 years of formal education (Figure 2(b)).

Figure 3 (Appendix A.2) demonstrates that the distribution of both GAD-7 and PHQ-9 scores are skewed-right, representing that the majority of the dataset is composed of either no or subthreshold level of the disorders. In addition, the number of samples with moderate to severe level of both disorders are higher among women compared with men.

3 Feature Sets

In this section, we introduce a set of hand-crafted features extracted from the DEPAC audio records and the associated transcripts. The set of features comprises various linguistic and acoustic features that have been found in previous psychiatric literature to be effective in detection of depression and anxiety from spoken language (Low et al., 2020; Smirnova et al., 2018).

3.1 Acoustic Features:

We extracted 220 acoustic features from each audio sample. The feature set includes:

Generic Linguistic Features

Feature Category	**Description**
Discourse mapping (18)	**Utterance distances** and **speech-graph** (Mota et al., 2012) features extracted from the graph representation of the transcripts.
Local coherence (15)	Average, maximum, and minimum similarity between Word2Vec (Mikolov et al., 2013) representations of the successive utterances.
Lexical complexity and richness (103)	**Vocabulary richness**: Such as Brunet's index (Brunet et al., 1978) and Honore's statistic (Honoré et al., 1979). **Psycholinguistics norms**: Average norms across all words, nouns only and verbs only for imageability, age of acquisition, familiarity (Stadthagen-Gonzalez and Davis, 2006) and frequency (commonness) (Brysbaert and New, 2009). **Grammatical constituents**: The constituents comprising the parse tree in a set of Context-Free Grammar (CFG) features.
Syntactic complexity (143)	**Constituency-parsing based features**: Scores based on the parse tree (Chae and Nenkova, 2009) (e.g., the height of the tree, the statistical functions of Yngve depth (a measure of embeddedness) (Yngve, 1960), and the frequencies of various production rules(Chae and Nenkova, 2009)). **Lu's syntactic complexity features**: Metrics of syntactic complexity suggested by (Lu, 2010) such as the length of sentences, T-units, and clauses, etc. **Utterance length**: Average, maximum and minimum utterance length.
Utterance cohesion (1)	Number of switches in verb tense across utterances divided by total number of utterances.
Sentiment (9)	Variables such as valence, arousal, and dominance scores (Warriner et al., 2013) for all words and word types describing the sentiment of the words used.
Word finding difficulty (11)	**Pauses and fillers**: Variables like speech rate, hesitation, duration of words and number of filled (e.g., um, uh) and unfilled pauses as signs of word finding difficulty, which result in less fluid or fluent speech (Pope et al., 1970). **Invalid words**: Not in Dictionary (NID) indicating proportion of words not in the English dictionary.

Task-Specific Linguistic Features

Speech Task	**Description**
Phonemic Fluency (2)	Includes the raw number of words starting with the correct letter with/without explicit filtering out of proper nouns by their Part of Speech (POS) tags.
Picture Description (25)	**Global coherence**: Average, minimum and maximum cosine distance between GloVe (Pennington et al., 2014) word vector representation of each utterance and its closest content unit centroid utterances. **Information units**: The number of objects, subjects, locations and actions used to measure the number of items correctly named in the picture description task.
Semantic Fluency (1)	Includes the raw number of words of the correct category.

Table 2: List of linguistic features in our conventional feature set. The number of features in each subtype is shown in the parentheses.

- **Spectral features:** Intensity (auditory model based), MFCC 0-12, Zero-Crossing Rate (ZCR)
- **Voicing-related features:** Fundamental frequency (F_0), Harmonic-to-Noise Ratio (HNR), shimmer and jitter, durational features, pauses and fillers, phonation rate

Statistical functionals including minimum, maximum, average, and variance were computed on the low-level descriptors. Additionally, skewness and kurtosis were calculated on MFCCs, first and second order derivatives of MFCCs, and Zero Crossing Rate (ZCR) (Low et al., 2020) (Table 7 in appendix elaborates on detailed descriptions of these features as well as previous literature motivating their selection as the indicators of psychiatric conditions).

A Python implementation of Praat phonetic analysis toolkit (Boersma and Van Heuven, 2001) has been used to extract the majority of these features. The MFCC features and their functionals were computed using `python_speech_features`[2] library.

3.2 Linguistic Features:

We also applied standard natural language processing libraries (e.g., spaCy[3] and Stanford Parser[4]) to extract 300 generic and 28 task-specific linguistic features from the associated transcripts of the audio files (Table 2). For simplification, we classified the generic features into the categories including discourse mapping, local coherence, lexical complexity and richness, syntactic complexity, utterance cohesion, sentiment, and word finding difficulty (the selection motivations of our linguistic features are explained in Appendix A.3, Table 6).

4 Intended Usage

The study aimed to collect a high quality training dataset with the intention of developing a speech-based digital biomarker for the psychiatric diseases of depression and anxiety. The dataset is well-suited for exploratory analysis involving statistical and machine learning methods to generate potential speech biomarkers and test their validity. In Section 5, we present the baseline models to predict depression severity using this dataset, that can be used as benchmarks for the future research.

Range of scores	AVEC PHQ-9	DEPAC PHQ-9	DEPAC GAD-7
[0 - 5)	77	240	261
[5 - 10)	36	178	152
[10 - 15)	26	84	87
[15 - 20)	17	40	45
[20 - 27]	7	10	7

Table 3: Counts for the PHQ-8/GAD-7 scores in AVEC and DEPAC datasets

5 Baseline Models for Depression Analysis

5.1 Data Preprocessing

Standardization: Once the acoustic and linguistic features were extracted from the data records, we standardized them using z-scores, i.e., subtracting the mean and dividing by standard deviation. The standard score of a sample x of feature f_i is calculated as:

$$y = \frac{x - \mu}{\sigma} \tag{1}$$

here μ and σ are the mean and standard deviation of the values of f_i in all training samples.

5.2 Model Training

To compare the efficacy of different modalities in predicting depression, we trained a combination of linear and non-linear Machine Learning (ML) models: Support Vector Regressors (SVR), Linear Regression (LR), and Random Forest Regressor (RF) separately on the following feature categories:

1. Demographic features (i.e., age, gender, and education)
2. Acoustic features
3. Linguistic features

We further investigated the effectiveness of each speech task for predicting depression severity on the PHQ-8 scale. The main reason for excluding the last question in PHQ-9 questionnaire was to make the results comparable to the performances with AVEC 2016 (Valstar et al., 2016) and AVEC 2019 (Ringeval et al., 2019) baselines, which are reported on PHQ-8 scale. The audio samples in AVEC challenges are subsets of Distress Analysis Interview Corpus (DAIC-WoZ) (Gratch et al.,

[2] https://pypi.org/project/python_speech_features/

[3] https://spacy.io/

[4] https://nlp.stanford.edu/software/lex-parser.html

2014), which includes interviews of the participants conducted by a virtual agent. The length of the speech samples of the DAIC-WoZ dataset range from 5 to 25 minutes, including both participants' and interviewer's speech.

Figure 2(c) compares how the PHQ-8 scores are distributed in male versus female participants in AVEC 2019 and DEPAC datasets. Higher PHQ scores indicates the higher depression severity in the subjects. The distributions are skewed-right both for the male and female participants, representing that the majority of both datasets is composed of either no or mild level of depression. The number of samples in each level of depression in each of the two datasets is summarized in Table 3.

To validate the comparison of our models' performance with the ones trained on the AVEC datasets, we performed independent t-test on the PHQ-8 score distribution of the DEPAC dataset and AVEC 2019 corpus. The outcome of the test showed that the two datasets do not exhibit significant differences ($t = 0.65, p > 0.05$)and as such, these two datasets are similar enough to compare the performance of the baseline ML models.

Compared with previous datasets, our dataset is enriched with a greater variety of speech tasks. Thus, in addition to an analysis using data from all the included tasks, we evaluate models trained on task-subsets of the corpus and report their performance in predicting depressive disorder. Each model is evaluated with regard to the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) scales, following the baseline set by AVEC challenge (Valstar et al., 2016), (Ringeval et al., 2017). The performance metrics are described in Appendix A.4.

We trained an SVR model on the combination of acoustic and linguistic features extracted from all five speech tasks (See Section 2.5), and also separately on each of the speech (See Table 5).

For all the experiments, all model hyperparameters were set to their default values as on the Scikit-learn implementation (Pedregosa et al., 2011). Models were trained using grouped 10-fold cross validation, where samples from the same individual do not appear in both the training folds and test fold. All results are reported as the mean MAE/RMSE scores across the 10 folds.

Features	Algorithm	RMSE	MAE
Demographic	LR	6.94	5.18
	RF	6.34	4.93
	SVR	**5.20**	**4.06**
Acoustic	LR	7.51	5.86
	RF	5.41	4.41
	SVR	5.48	4.40
	AVEC 2016 baseline (Valstar et al., 2016)	7.78	5.72
	AVEC 2019 baseline (Ringeval et al., 2019)	8.19	-
Linguistic	LR	5.72	4.60
	RF	5.40	4.37
	SVR	5.37	4.24

Table 4: Regression results of the models predicting PHQ-8 score on different categories of features. Bold indicates the best performance.

5.3 Baseline Model Result and Discussion

We present and discuss the results of baseline model training across different modalities of input features, i.e. demographic, acoustic and linguistic, as well as across five different speech tasks, using DEPAC speech data.

Model Performance across Modalities: Among the three modalities, SVR model trained on demographic features performs the best, achieving the lowest MAE and RMSE, followed by the SVR model trained on linguistic features. Both acoustic and linguistic baseline models attain less than 20% MAE in the range of scores (0 to 24). Marginal deviation of both MAE and RMSE between acoustic and linguistic models suggests that these two modalities are effective for the task of recognizing signs of depression from speech. It is noteworthy that, the audio files did not undergo any preprocessing or enhancement before extracting the acoustic features. Yet, models trained on acoustic features exhibit competitive performance with the linguistic model, indicating that the quality of the recordings is sufficient and is a valuable foundation for future research.

In terms of predicting PHQ-8 scores, our baseline models perform substantially better than the baseline models specified by challenge organizers

Speech task	RMSE	MAE
Phoneme Task	5.49	4.32
Phonemic fluency	5.44	4.31
Picture description	5.36	4.25
Positive fluency	**5.19**	**4.11**
Prompted narrative	5.30	4.20
All tasks	5.38	4.27

Table 5: Regression results of SVR models predicting PHQ-8 score on different speech tasks. Bold indicates the best performance.

of AVEC 2016 (Valstar et al., 2016) and AVEC 2019 (Ringeval et al., 2019) (Table 4), despite the shorter length of samples than the AVEC corpus, which justify the robustness of the hand-curated acoustic features introduced in this work, as well as the quality of the dataset.

Surprisingly, the SVR model using only demographic features outperforms both acoustic and linguistic models (Table 4). This demographic information was previously found to be highly correlated to one's level of depression in literature (Akhtar-Danesh and Landeen, 2007). However, in real-world application, the demographic model may not be completely reliable due to ambiguity of these features.

Model Performance across Speech Tasks: In our task-specific analysis, comparatively lower RMSE and MAE are scored by models trained on picture description, positive fluency and prompted narrative than the phoneme task, phonemic fluency and all tasks combined. The possible reason behind this observation is that the picture description, positive fluency and prompted narrative tasks produce longer audio samples, resulting in more informative acoustic and linguistic features, leading to more accurate models. This observation shows that long recordings of narrative tasks can be rich sources of markers to predict depressive disorder from speech.

6 Conclusion

In this work, we introduce DEPAC, a rich audio dataset for mental health research which is labelled with scores on standard scales of two highly prevalent mental disorders: PHQ-9 scores for depression and GAD-7 scores for anxiety assessment. The dataset offers a remarkably larger sample size in comparison to other publicly available corpora. One other source of novelty of the presented corpus is its richness in the diversity of speech tasks and participants with various degrees of education, genders, and age groups. We also introduce a hand-curated set of acoustic and linguistic features incorporating domain knowledge of clinical and ML experts, which are used as the predictors of models for quantifying depression severity. We present the performance of baseline models in prediction of depression severity level, that can be applied by future researchers as a benchmark. Our baseline models achieve competitive performance when compared to the AVEC 2016 and AVEC 2019 baseline models and demonstrate the quality of the DEPAC dataset and effectiveness of our proposed feature set in measuring depression severity.

References

Noori Akhtar-Danesh and Janet Landeen. 2007. Relation between depression and sociodemographic factors. *International journal of mental health systems*, 1(1):1–9.

Murray Alpert, Enrique R Pouget, and Raul R Silva. 2001. Reflections of depression in acoustic measures of the patient's speech. *Journal of affective disorders*, 66(1):59–69.

RG Bachu, S Kopparthi, B Adapa, and BD Barkana. 2008. Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In *American Society for Engineering Education (ASEE) zone conference proceedings*, pages 1–7. American Society for Engineering Education.

Megan S Barker, Breanne Young, and Gail A Robinson. 2017. Cohesive and coherent connected speech deficits in mild stroke. *Brain and language*, 168:23–36.

Aaron T Beck, Robert A Steer, and Margery G Carbin. 1988. Psychometric properties of the beck depression inventory: Twenty-five years of evaluation. *Clinical psychology review*, 8(1):77–100.

Jeffrey S Berman, R Christopher Miller, and Paul J Massman. 1985. Cognitive therapy versus systematic desensitization: Is one treatment superior? *Psychological bulletin*, 97(3):451.

Nathaniel J Blanco, A Ross Otto, W Todd Maddox, Christopher G Beevers, and Bradley C Love. 2013. The influence of depression symptoms on exploratory decision-making. *Cognition*, 129(3):563–568.

Paul Boersma and Vincent Van Heuven. 2001. Speak and unspeak with praat. *Glot International*, 5(9/10):341–347.

John G Borkowski, Arthur L Benton, and Otfried Spreen. 1967. Word fluency and brain damage. *Neuropsychologia*, 5(2):135–140.

Étienne Brunet et al. 1978. *Le vocabulaire de Jean Giraudoux structure et évolution*. Slatkine.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Jieun Chae and Ani Nenkova. 2009. Predicting the fluency of text with shallow structural features: Case studies of machine tanslation and human-written text.

Jesse Chandler and Danielle Shapiro. 2016. Conducting clinical research using crowdsourced convenience samples. *Annual review of clinical psychology*, 12:53–81.

Robert M Cohen, Herbert Weingartner, Sheila A Smallberg, David Pickar, and Dennis L Murphy. 1982. Effort and cognition in depression. *Archives of general psychiatry*, 39(5):593–597.

Sarah Kathryn Fineberg, J Leavitt, Sasha Deutsch-Link, Samson Dealy, Christopher D Landry, Kevin Pirruccio, Samantha Shea, Savannah Trent, Guillermo Cecchi, and Philip R Corlett. 2016. Self-reference in psychosis and depression: a language marker of illness. *Psychological medicine*, 46(12):2605–2615.

Alastair J Flint, Sandra E Black, Irene Campbell-Taylor, Gillian F Gailey, and Carey Levinton. 1992. Acoustic analysis in the differentiation of parkinson's disease and major depression. *Journal of Psycholinguistic Research*, 21(5):383–399.

Katrina E Forbes-McKay and Annalena Venneri. 2005. Detecting subtle spontaneous language decline in early alzheimer's disease with a picture description task. *Neurological sciences*, 26(4):243–254.

Philippe Fossati, Anne-Marie Ergis, Jean-François Allilaire, et al. 2003. Qualitative analysis of verbal fluency in depression. *Psychiatry research*, 117(1):17–24.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2016. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Lior Galili, Ofer Amir, and Eva Gilboa-Schechtman. 2013. Acoustic properties of dominance and request utterances in social anxiety. *Journal of social and clinical psychology*, 32(6):651–673.

Eva Gilboa-Schechtman, Lior Galili, Yair Sahar, and Ofer Amir. 2014. Being "in" or "out" of the game: subjective and acoustic reactions to exclusion and popularity in social anxiety. *Frontiers in human neuroscience*, 8:147.

Elaine Giles, Karalyn Patterson, and John R Hodges. 1996. Performance on the boston cookie theft picture description task in patients with early dementia of the alzheimer's type: missing information. *Aphasiology*, 10(4):395–408.

Harold Goodglass, Edith Kaplan, and Sandra Weintraub. 2001. *BDAE: The Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins Philadelphia, PA.

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)*, pages 3123–3128.

M Hamilton and W Guy. 1976. Hamilton depression scale. *Group*, 1:4.

Antony Honoré et al. 1979. Some simple measures of richness of vocabulary. *Association for literary and linguistic computing bulletin*, 7(2):172–177.

Christine Howes, Matthew Purver, and Rose McCabe. 2014. Linguistic indicators of severity and progress in online text-based therapy for depression. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*. Association for Computational Linguistics.

Haihua Jiang, Bin Hu, Zhenyu Liu, Lihua Yan, Tianyang Wang, Fei Liu, Huanyu Kang, and Xiaoyu Li. 2017. Investigation of different speech types and emotions for detecting depression using different classifiers. *Speech Communication*, 90:39–46.

Emil Kraepelin. 1921. Manic depressive insanity and paranoia. *The Journal of Nervous and Mental Disease*, 53(4):350.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Kwok-Keung Leung, Tatia MC Lee, Paul Yip, Leonard SW Li, and Michael MC Wong. 2009. Selective attention biases of people with depression: Positive and negative priming of depression-related information. *Psychiatry research*, 165(3):241–251.

Daniel M Low, Kate H Bentley, and Satrajit S Ghosh. 2020. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1):96–116.

Bernd Löwe, Jürgen Unützer, Christopher M Callahan, Anthony J Perkins, and Kurt Kroenke. 2004. Monitoring depression treatment outcomes with the patient health questionnaire-9. *Medical care*, pages 1194–1201.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Brian MacWhinney. 2000. The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database.

Brian MacWhinney. 2007. The talkbank project. In *Creating and digitizing language corpora*, pages 163–180. Springer.

Andrew Mathews and Colin MacLeod. 2005. Cognitive vulnerability to emotional disorders. *Annu. Rev. Clin. Psychol.*, 1:167–195.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Stuart A Montgomery and MARIE Åsberg. 1979. A new depression scale designed to be sensitive to change. *The British journal of psychiatry*, 134(4):382–389.

Natalia B Mota, Nivaldo AP Vasconcelos, Nathalia Lemos, Ana C Pieretti, Osame Kinouchi, Guillermo A Cecchi, Mauro Copelli, and Sidarta Ribeiro. 2012. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PloS one*, 7(4):e34928.

James C Mundt, Peter J Snyder, Michael S Cannizzaro, Kara Chappie, and Dayna S Geralts. 2007. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. *Journal of neurolinguistics*, 20(1):50–64.

Charles B Nemeroff. 2007. Prevalence and management of treatment-resistant depression. *Journal of Clinical Psychiatry*, 68(8):17.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419.

Rupal Patel and Kathryn Connaghan. 2014. Park play: A picture description task for assessing childhood motor speech disorders. *International Journal of Speech-Language Pathology*, 16(4):337–343.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Benjamin Pope, Thomas Blass, Aron W Siegman, and Jack Raher. 1970. Anxiety and depression in speech. *Journal of Consulting and Clinical Psychology*, 35(1p1):128.

Thomas F Quatieri and Nicolas Malyska. 2012. Vocal-source biomarkers for depression: A link to psychomotor activity. In *Thirteenth annual conference of the international speech communication association*.

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*, pages 81–88.

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*, 71:103107.

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*, pages 3–12.

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*, pages 3–9.

Daun Shin, Won Ik Cho, C Hyung Keun Park, Sang Jin Rhee, Min Ji Kim, Hyunju Lee, Nam Soo Kim, and Yong Min Ahn. 2021. Detection of minor and major depression through voice as a biomarker using machine learning. *Journal of clinical medicine*, 10(14):3046.

Daria Smirnova, Paul Cumming, Elena Sloeva, Natalia Kuvshinova, Dmitry Romanov, and Gennadii Nosachev. 2018. Language patterns discriminate mild depression from normal sadness and euthymic state. *Frontiers in psychiatry*, 9:105.

Hannah R Snyder, Roselinde H Kaiser, Mark A Whisman, Amy EJ Turner, Ryan M Guild, and Yuko Munakata. 2014. Opposite effects of anxiety and depressive symptoms on executive function: The case of selecting among competing options. *Cognition & emotion*, 28(5):893–902.

Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097.

Hans Stadthagen-Gonzalez and Colin J Davis. 2006. The bristol norms for age of acquisition, imageability, and familiarity. *Behavior research methods*, 38(4):598–605.

Raluca Nicoleta Trifu, Bogdan NEMEŞ, Carolina Bodea-Haţegan, and Doina Cozman. 2017. Linguistic indicators of language in major depressive disorder (mdd). an evidence based research. *Journal of Evidence-Based Psychotherapies*, 17(1).

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*, pages 3–10.

Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*, pages 3–10.

Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10.

Jerome C Wakefield. 2013. The dsm-5 debate over the bereavement exclusion: Psychiatric diagnosis and the future of empirically supported treatment. *Clinical psychology review*, 33(7):825–845.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Victor H Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466.

Jörg Zinken, Katarzyna Zinken, J Clare Wilson, Lisa Butler, and Timothy Skinner. 2010. Analysis of syntax and word use to predict successful participation in guided self-help for anxiety and depression. *Psychiatry research*, 179(2):181–186.

Chiara Zucco, Barbara Calabrese, and Mario Cannataro. 2017. Sentiment analysis and affective computing for depression monitoring. In *2017 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 1988–1995. IEEE.

A Appendix

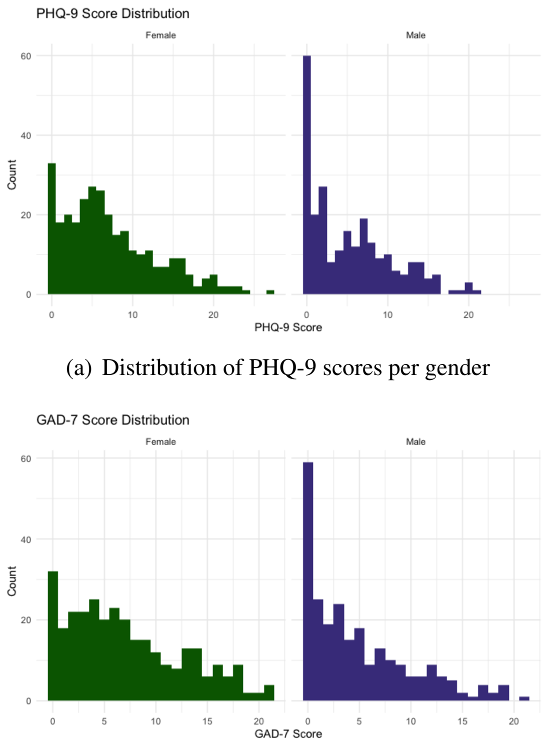
A.1 Picture Design Guidelines

To develop the ’Family in the kitchen’ image (Figure 1) for our picture description task, we used the core design principles (Patel and Connaghan, 2014) described below:

1. Image content breakdown should contain:
   (a) **2 scenes/locations** (e.g., kitchen, or living room)
   (b) **20 to 25 objects** (e.g., knife, pan, or cupboard)
   (c) **9 to 10 actions** (e.g., chop, cook, steam, or fall)
   (d) **3 to 4 people/subjects** (e.g., dad, dog, mom, or daughter)
   (e) **2 “dangerous” elements** (e.g., broken bottle, or steaming pot)
2. Images should display **relationships between components** in a scene.
3. Images should depict **familiar themes**, but they must be accessible to adults with diverse cultural backgrounds, sexual orientations, and various socioeconomic strata.
4. Images should be designed appropriately for **older adults** with varied levels of visual impairment.
5. Images should provoke spontaneous discourse useful in **diagnosis and assessment** of mental health conditions. It should:
   (a) Elicit tokens whose labels **span the phonetic range** useful in diagnosing motor speech difficulty.
   (b) Elicit tokens whose labels **span lexical norms** (varying age of acquisition (AoA), familiarity, and imageability). Representing a varied range of lexical norms allows for using the same image to test speakers with varying degrees of cognitive and language impairment.
   (c) **Contain sub-scenarios** (Patel and Connaghan, 2014) which would be useful generally for generating longer speech samples, and specifically in assessing discourse structure (e.g., coherence, repetition, trajectory (what order are the sub-scenarios described in), content units (which sub-scenarios are mentioned and which left out), reasoning/inferences (e.g., interconnections and causation between the sub-scenarios)).

The goal of these guidelines was to keep the content generalizable across diverse cultures and to control the similarity with the ’Cookie theft’ (Goodglass et al., 2001) image in lexico-syntactic complexity and the amount of information content units.

A.2 Distribution of Assessment Scores



(a) Distribution of PHQ-9 scores per gender

(b) Distribution of GAD-7 scores per gender

Figure 3: Distribution of the participants’ PHQ-9 and GAD-7 scores in mTurk Study.

A.3 Feature Selection Motivations

The prior studies supporting the choice of our conventional feature set are described in Table 6 and 7. Table 7 displays the selection motivations of our acoustic features derived from the audio files, including spectral and energy related as well as voicing related features. In addition, Table 6 represents the motivations behind the choice of the generic and task-specific linguistic features extracted from the associated transcripts.

Generic Linguistic Features

Feature Category	**Motivations**
Discourse mapping	Techniques to formally quantify utterance similarity and disordered speech via distance metrics or graph-based representations have been used to differentiate speech from those suffering from various other mental health issues that are known to affect speech production (Mota et al., 2012; Fraser et al., 2016).
Local coherence	Coherence and cohesion in speech is associated with the ability to sustain attention and executive functions (Barker et al., 2017). Depression and anxiety are both known to impair such cognitive processes (Leung et al., 2009; Snyder et al., 2014).
Lexical complexity and richness	Language pattern changes in particular related to the irregular usage patterns of words of certain grammatical categories such as pronouns or verb tenses have been found to differentiate depression from normal fluctuations in mood from healthy individuals (Smirnova et al., 2018).
Syntactic complexity	Previous literature suggests that syntactic complexity of utterances, can be used to predict symptoms of depression (Smirnova et al., 2018), including utterances elicited in self-administered contexts (Zinken et al., 2010).
Utterance cohesion	Rates of verb tense use (in particular the past-tense) is known to be changed in individuals with depression. (Smirnova et al., 2018).
Sentiment	Emotional state and speech are connected, and sentiment scores in speech have been used to predict depression and anxiety levels in past research (Howes et al., 2014; Zucco et al., 2017).
Word finding difficulty	Previous work has found relationships between speech disturbance, filled, and unfilled speech of individuals with anxiety and depression (Pope et al., 1970).

Task-Specific Linguistic Features

Speech Task	**Motivations**
Phonemic Fluency	Measures of individual performance at the phonemic fluency task (Borkowski et al., 1967).
Picture Description	Measures of individual performance at picture description task as defined in (Giles et al., 1996; Jiang et al., 2017).
Semantic Fluency	Measures of individual performance at the semantic fluency task (Fossati et al., 2003).

Table 6: Support literature motivating the selection of the linguistic features in our conventional feature set.

Spectral and Energy Related Features

Feature	**Motivations**
Intensity (auditory model based)	Perceived loudness in dB relative to normative human auditory threshold. In 1921, Emil Kraepelin recognized lower sound intensity in the voices of depressed patients (Kraepelin, 1921).
MFCC 0-12	MFCC 0-12 and energy, their first and second order derivatives are calculated on every 16 ms window and step size of 8 ms, and then, averaged over the entire sample. MFCCs and their derivatives were included as baseline features in AVEC since 2013 (Valstar et al., 2013), (Valstar et al., 2016), (Ringeval et al., 2019) and found to be effective in predicting depression severity in the literature (Ray et al., 2019), (Rejaibi et al., 2022).
Zero-crossing rate (ZCR)	Zero crossing rate across all the voiced frames showing how intensely the voice was uttered. It was used as a speech biomarker of depression in previous studies (Bachu et al., 2008; Shin et al., 2021).

Voicing Related Features

F_0	Fundamental frequency in Hz. A drop in F_0 and F_0 range indicates monotonous speech, which is common in depression (Low et al., 2020). In addition, many studies have discovered a considerable rise in mean F_0 in people suffering from social anxiety disorder (Gilboa-Schechtman et al., 2014; Galili et al., 2013).
Harmonics-to-noise-ratio (HNR)	Degree of acoustic periodicity in dB using both auto-correlation and cross-correlation method. Decreasing HNR ratio has been found to correlate with increasing severity of depression (Quatieri and Malyska, 2012).
Jitter and shimmer	Jitter is the period perturbation quotient and shimmer is the amplitude perturbation quotient representing the variations in the fundamental frequency. In previous studies, anxious patients indicated substantially higher shimmer and jitter. In addition, rise in jitter and shimmer variability was observed in subjects with major depressive disorder (Low et al., 2020).
Durational features	Total audio and speech duration in the sample. In prior studies, depression severity increased the total duration of speech because of longer pauses resulting in lower speech to pause ratio (Alpert et al., 2001; Mundt et al., 2007).
Pauses and fillers	Number and duration of short ($< 1s$), medium ($1 - 2s$) and long ($> 2s$) pauses, mean pause duration, and pause-to-speech ratio. Depression and anxiety are known to affect the rate of pauses/speech in individuals (Pope et al., 1970).
Phonation rate	Number of voiced time windows over the total number of time windows in a sample.

Table 7: Support literature motivating the selection of the acoustic features in our conventional feature set.

A.4 Performance Metrics

Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are calculated using the formulas shown below.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \tag{2}$$

$$MAE = \frac{\sum_{i=1}^{N} |x_i - y_i|}{N} \tag{3}$$

In the above, x_i and y_i are the true and predicted scores respectively.