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.

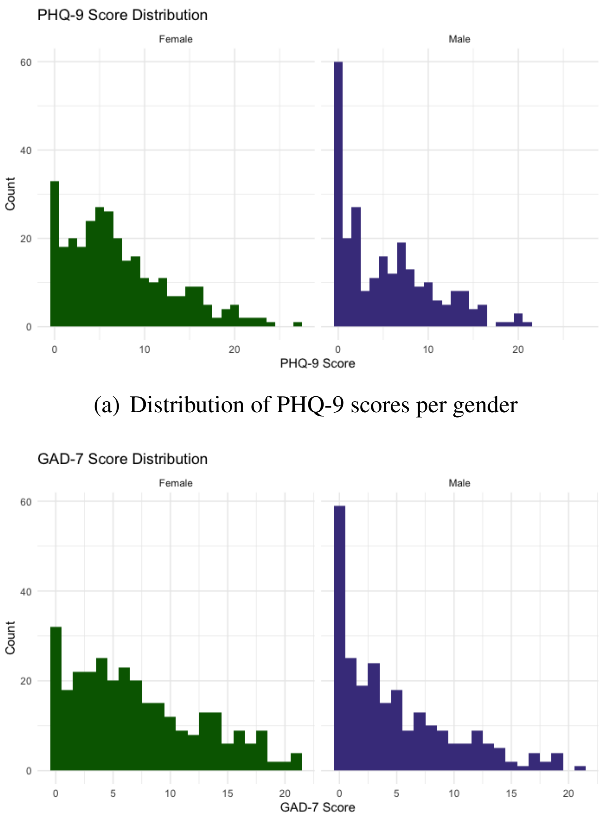
3. Images should depict **familiar themes**, but they must be accessible to adults with diverse cultural backgrounds, sexual orientations, and various socioeconomic strata.

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**
Semantic Fluency	Measures of individual performance at the phonemic fluency task (Borkowski et al., 1967).
Picture Description	Includes measures of individual performance at picture description task as defined in (Giles et al., 1996; Jiang et al., 2017).
Semantic Fluency	Measures 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.