### Research Plan: A Multimodal Fusion of Affective, Linguistic, and Semantic Cues for Dementia Detection

#### 1. Rationale and Hypothesis

This research will investigate a novel multimodal approach for dementia detection by fusing three distinct but complementary feature sets: affective acoustic cues that reflect a speaker's emotional state, content-based linguistic cues that reflect cognitive function and grammatical structure, and semantic cues from large language models that capture contextual meaning.

The central hypothesis is that a model combining these three modalities will achieve significantly higher classification accuracy than models trained on any single modality or a two-modality combination. Furthermore, this approach will allow for a more nuanced, interpretable analysis of which specific emotional, cognitive, and semantic markers are most predictive of dementia, both within a single language and across different languages.

#### 2. Detailed Methodology

This plan now involves three parallel feature extraction streams that will be combined before classification.

##### Stream A: Affective Acoustic Feature Extraction

This stream remains focused on the emotional tone of the voice.

* **Tool**: Use the openSMILE toolkit.
* **Feature Set**: Extract the eGeMAPS (extended Geneva Minimalistic Acoustic Parameter Set). This will produce a vector of **88 acoustic features** for each audio file, capturing information about pitch, intensity, jitter, and other vocal characteristics related to emotion.

**Stream B: Content-Based Linguistic Feature Extraction**

This stream focuses on analyzing the structure and content of the transcripts.

* **Preprocessing**: Ensure transcripts are clean and contain only the participant's speech.
* **Feature Engineering**: From each transcript, you will calculate a set of powerful, interpretable linguistic features.  
  + **Key Element Ratio (Content Density)**: Calculate the ratio of unique key elements mentioned from a predefined list to the total number of words. This measures how much relevant information the person is conveying.
  + **Lexical Diversity (Vocabulary Richness)**: Calculate the Type-Token Ratio (TTR), which is the ratio of unique words to the total number of words. A lower TTR can indicate repetitive speech.
  + **Pronoun-to-Noun Ratio**: Calculate the ratio of pronouns to nouns. Patients with dementia may overuse pronouns when they struggle to recall specific nouns.
  + **Syntactic Complexity**: Analyze the grammatical structure to measure cognitive processing.
    - **Sentence Length**: Calculate the average number of words per sentence.
    - **Parse Tree Depth**: Use a syntactic parser to measure the average depth of sentence parse trees. A shallower depth can indicate simpler grammatical   
      constructions.\

This stream will produce a vector of **5-7 linguistic features** for each transcript.

**Action:** From each clean transcript, we will calculate the following 8 features:

1. **Cue Count:** The total number of times words from the cue\_list appear.
2. **Cue Proportion (New):** The Cue Count divided by the total number of words in the transcript. This normalizes the feature.
3. **Key Element Ratio (Content Density):** The number of *unique* cues mentioned from the list, divided by the total number of words.
4. **Lexical Diversity (TTR):** The ratio of unique words to total words.
5. **Pronoun Count:** The total count of pronouns.
6. **Noun Count:** The total count of nouns.
7. **Pronoun-to-Noun Ratio:** The ratio of Pronoun Count to Noun Count.
8. **Syntactic Complexity:**
   * Average Sentence Length.
   * Average Parse Tree Depth.

##### Stream C: Semantic Feature Extraction using BERT

This new stream focuses on capturing the deep contextual meaning of what is said, going beyond simple word counts or ratios.

* **Tool**: Use a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model.
* **Feature Engineering**:
  + For each transcript, generate a sentence-level or document-level embedding. This is a dense vector representation (typically 768 dimensions) that encodes the contextual meaning of the entire transcript.
  + This single vector serves as a powerful semantic "fingerprint" of the participant's speech.

This stream will produce a **semantic feature vector** for each transcript.

#### 3. Model Fusion and Classification

This is the critical step where you combine the three streams.

* **Fusion Method (Early Fusion)**: The most straightforward and effective method is feature concatenation. For each participant, you will combine the **88 eGeMAPS features** from Stream A, the **5-7 linguistic features** from Stream B, and the **semantic feature vector** from Stream C into a single, unified feature vector.
* **Classification Models**: Train the following machine learning models on this new, combined feature vector.  
  1. Support Vector Machine (SVM)
  2. Random Forest
  3. XGBoost
  4. **Feed-Forward Neural Network (FNN)**: A multi-layer perceptron that can capture complex, non-linear relationships between the combined features and the diagnosis.

#### 4. Experiment, Analysis, and Interpretation

Your analysis will now be much richer and will form the core of your paper.

1. **Baseline Models**: First, train and test models using only the features from each stream individually (A, B, and C) to get baseline accuracy scores.
2. **Multimodal Model**: Train and test the models using the combined (concatenated) feature vector from all three streams.
3. **Comparative Analysis**: Compare the performance (accuracy, F1-score, AUC) of the full multimodal model against the baseline models. The key finding will likely be that the three-stream multimodal model outperforms all others.
4. **Interpretability**: Use the feature importance scores from your trained XGBoost or Random Forest model. This will allow you to rank all features (acoustic, linguistic, and now semantic dimensions if using feature ablation) and answer critical questions.  
   * Overall, is the model relying more on affective cues, linguistic structure, or semantic content?
   * What are the top 5 most predictive features? Is it a mix of all three types?

#### 5. Cross-Lingual Validation on Taukodial

Apply the entire multimodal pipeline to the Taukodial dataset.

* **Linguistic Feature Adaptation**: Features like TTR, pronoun-to-noun ratio, and syntactic complexity are language-independent in concept and can be calculated for both English and Chinese transcripts.
* **Semantic Feature Adaptation**: To handle the Chinese transcripts in Stream C, switch from a standard BERT model to a pre-trained **Multilingual BERT (mBERT)**. This allows you to generate comparable semantic embeddings across different languages without needing a model trained specifically on Chinese medical text.
* **Cross-Lingual Comparison**: You can now perform a much deeper comparison. Are the most important affective, linguistic, and semantic features consistent across languages? This will allow you to contribute novel findings on the universality of these markers for dementia.