Appendix I - Data Dictionary

Original Metadata

- **path**:
  The file path to the corresponding audio recording (.mp3 format). This column serves as a reference for linking the extracted features back to the original audio files.
- **sentence**:
  The transcribed text of the sentence spoken in the audio recording.
- **sentence_domain**:
  An optional descriptor indicating the thematic domain of the spoken sentence (e.g., legal, medical, financial). This information is available for a subset of the recordings and remains missing for the majority.
- **age**:
  The age of the speaker as an age range (e.g. thirties, forties). Some entries are missing.
- **gender**:
  The self-identified gender of the speaker, filtered to only male and female.
- **accents**:
  A descriptor of the speaker's accent or regional variation of English. This information is available for a majority, but not all, of the samples.

Extracted Acoustic Features

The acoustic properties of each audio file were extracted using standard signal processing techniques implemented with the `librosa` library. The extracted features aim to capture diverse aspects of the speech signal, including its energy distribution, tonal structure, and spectral characteristics. All features were computed over the full duration of each recording, and both mean and standard deviation values were included where relevant.

Mel-Frequency Cepstral Coefficients (MFCCs)

- **mfcc_01_mean** to **mfcc_20_mean**, **mfcc_01_std** to **mfcc_20_std**:
  MFCCs capture the short-term power spectrum of the audio and are widely used to represent the timbral and tonal texture of the speech signal. Higher MFCC values typically correspond to more energetic or resonant parts of the audio, while lower values reflect flatter or softer spectral content. Both positive and negative values are common and expected.

Chroma Features

- **chroma_01_mean** to **chroma_12_mean**, **chroma_01_std** to **chroma_12_std**:
  Chroma features quantify the energy distribution across the twelve pitch classes of the musical octave (C, C#, ..., B). Higher chroma values indicate stronger tonal or harmonic content, while lower values are associated with noise-like or atonal speech. Values are non-negative and typically normalized within [0, 1].

Spectral Features

- **spec_centroid_mean**, **spec_centroid_std**:
  The spectral centroid represents the center of mass of the spectral energy. Higher centroids are associated with brighter, higher-frequency content (e.g., sibilant sounds), while lower centroids suggest darker, lower-frequency speech.

- **spec_bandwidth_mean**, **spec_bandwidth_std**:
  Spectral bandwidth measures the spread of frequencies around the centroid. Higher values indicate wider spectral spread and typically noisier speech; lower values imply a more concentrated, purer tone.

- **spec_contrast_band_1_mean** to **spec_contrast_band_7_mean**, **spec_contrast_band_1_std** to **spec_contrast_band_7_std**:
  Spectral contrast quantifies the difference between peaks and valleys in the spectrum across multiple frequency bands. Higher contrast values reflect richer, more structured sounds (e.g., vowel-consonant alternations), while lower values indicate flatter, more

noise-like spectra.

- **spec_rolloff_mean**, **spec_rolloff_std**:
  Spectral roll-off is the frequency below which a fixed percentage (typically 85%) of the total spectral energy is contained. Higher roll-off values suggest greater energy in the high-frequency range (e.g., presence of fricatives), while lower values imply concentration in the lower frequencies.

Temporal Features

- **zcr_mean**, **zcr_std** (Zero Crossing Rate):
  The zero crossing rate measures how often the audio waveform crosses the zero amplitude axis. Higher rates are indicative of noisier or unvoiced sounds (e.g., "s," "sh"), while lower rates correspond to voiced, smoother speech. Values are always positive.

- **rmse_mean**, **rmse_std** (Root Mean Square Energy):
  RMSE quantifies the short-term energy of the signal and reflects overall loudness. Higher RMSE values correspond to louder speech segments, while lower values reflect quieter or more subdued recordings. RMSE values are positive and typically small.

Interpretation of Feature Values

- **High MFCCs**: Energetic, dynamic tonal characteristics.
- **Low MFCCs**: Flatter, less resonant tonal properties.
- **High Chroma Energy**: Strong harmonic or tonal elements.
- **High Spectral Centroid**: Bright, high-frequency dominated speech.
- **High Spectral Bandwidth**: Noisy, broadband signals.
- **High Spectral Contrast**: Rich alternation of voiced/unvoiced components.
- **High Zero Crossing Rate**: Noisy, unvoiced sounds.
- **High RMSE**: Loud and energetic speech.