3. Data Problem

This document outlines the specific instructions for preparing the provided database of human voice recordings for training a machine learning model capable of distinguishing between authentic and synthetic voices.

1. Data Exploration and Analysis:

- ✓ Utilize tools such as Matplotlib and Seaborn for in-depth data analysis and visualization.
- ✓ Begin with a comprehensive exploration of the database, understanding characteristics, and assessing the distribution of authentic and synthetic samples.
- ✓ Identify and address imbalanced samples in the dataset.

2. Imbalance Handling:

- ✓ Enhance model performance by employing techniques such as oversampling or undersampling, e.g., using SMOTE or Imblearn.

3. Data Cleaning:

- ✓ Address variations in sample wav length by finding the mean of total sample lengths.
- ✓ Utilize padding techniques to standardize each sample to the fixed mean length.
- ✓ Handle misclassified samples within the dataset.

4. Feature Engineering:

- ✓ Extract relevant acoustic features like MFCCs, spectrograms, and pitch from audio recordings.
- ✓ Experiment with different feature sets to identify the most discriminative ones.
- ✓ Normalize and standardize features for consistent scaling, facilitating model training.

5. Speaker Embeddings:

- ✓ Consider incorporating speaker embeddings to capture individual characteristics, enhancing the model's ability to generalize across diverse voices.
- ✓ Implement suitable methods for extracting speaker embeddings, such as pre-trained models or training on the dataset.

6. Data Splitting:

- ✓ Split the data into training, validation, and test sets, ensuring a stratified split.
- ✓ Evaluate model performance on the validation set, minimizing loss before final testing on the test samples.

7. Data Augmentation:

- ✓ Apply data augmentation techniques to increase model robustness against variations in recording conditions.
- ✓ Techniques may include random pitch shifts, time-stretching, or introducing background noise.

8. Quality Control:

- ✓ Conduct a rigorous quality control check to identify and address anomalies or outliers in the dataset.
- ✓ Verify that data preprocessing steps do not introduce artifacts negatively affecting model performance.

Once the data is prepared following these guidelines, the transition into the model development phase will focus on selecting an appropriate architecture, training the model, and fine-tuning it for optimal performance.