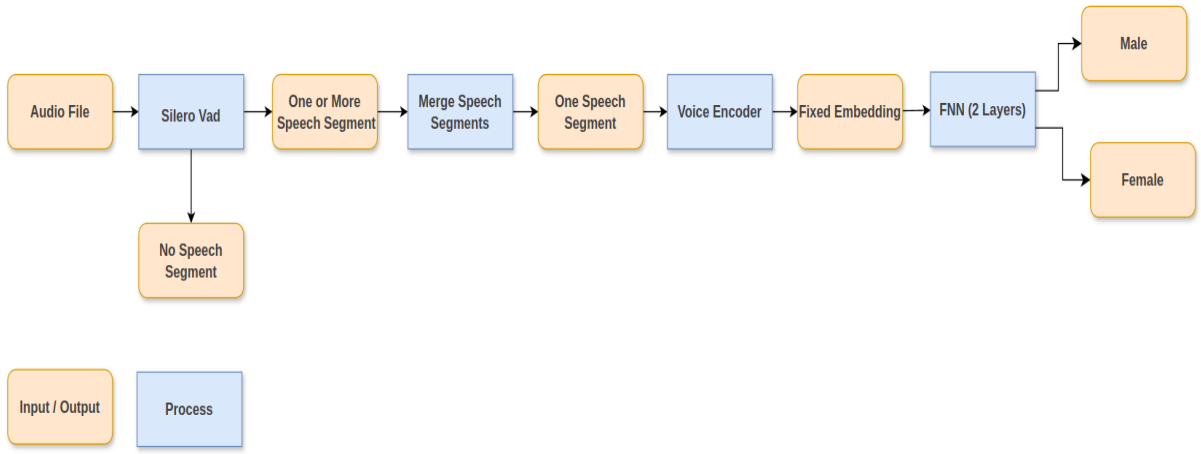
1. Introduction

The goal of this project was to build a system that processes user-recorded speech audio and identifies whether someone is speaking (Voice Activity Detection), and if so, predicts the speaker's gender. The system needed to handle various audio issues, such as background noise, long silences, volume inconsistencies, and varying audio lengths. A pre-trained Silero VAD model was provided for voice activity detection, and a pre-trained voice encoder model of choice could be used for speaker embeddings. All other implementation steps were required to be done from scratch.

2. Pipeline Overview



To address these requirements, three different voice encoder models were investigated:

1. **ECAPA-TDNN**
2. **Resemblyzer**
3. **Wav2Vec**

All three generate a fixed-length embedding regardless of input duration:

- ECAPA-TDNN → 192-dimensional embedding
- Resemblyzer → 256-dimensional embedding
- Wav2Vec → 768-dimensional embedding (after averaging across time frames)

Ultimately, ECAPA-TDNN and Resemblyzer embeddings demonstrated strong performance for gender classification, while Wav2Vec appeared more focused on linguistic features rather than speaker-specific characteristics.

3. Implementation Details

3.1 Voice Activity Detection: Silero VAD

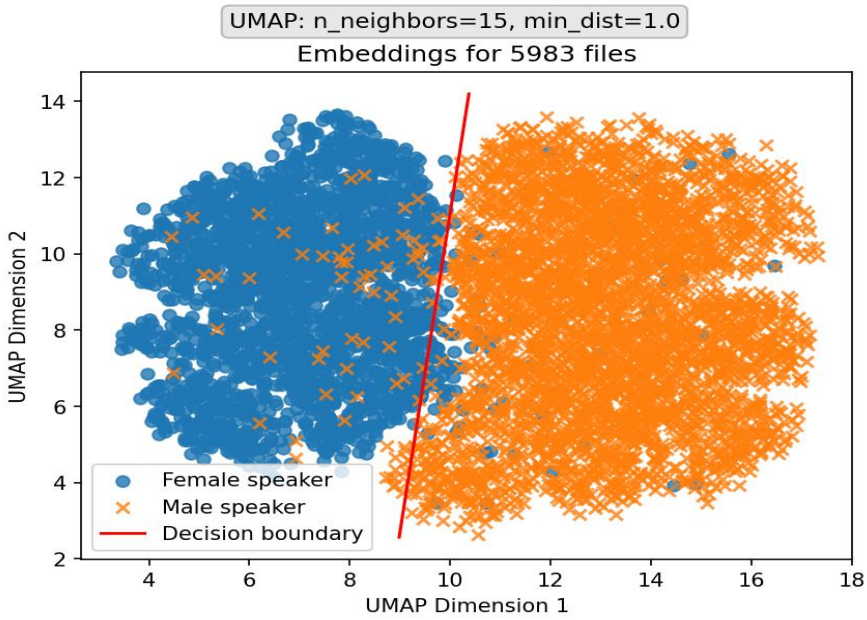
Silero VAD was chosen to detect whether a given audio chunk contains speech. Key points include:

- **Chunking**:
  - Audio is split into chunks of 512 samples at a 16 kHz sampling rate (around 32 ms per chunk).
- **Speech Probability**:
  - Each chunk receives a probability score (0–1) indicating the likelihood of speech.
  - A default threshold of 0.5 is used to identify whether a chunk is "speechy."
  - Once "triggered," the VAD remains active until probabilities drop below 0.35 (threshold − 0.15) for at least 100 ms.
- **Minimum and Maximum Durations**:
  - Short segments (< 250 ms) are discarded.
  - Very long segments are split if they exceed a specified maximum duration.
- **Examples of Speech vs. Non-Speech**:
  - Loud, intelligible audio is classified as speech.
  - Laughter, screams, and singing with discernible words are considered speech.
  - Background bird noises, city sounds, or other non-human audio are considered non-speech.

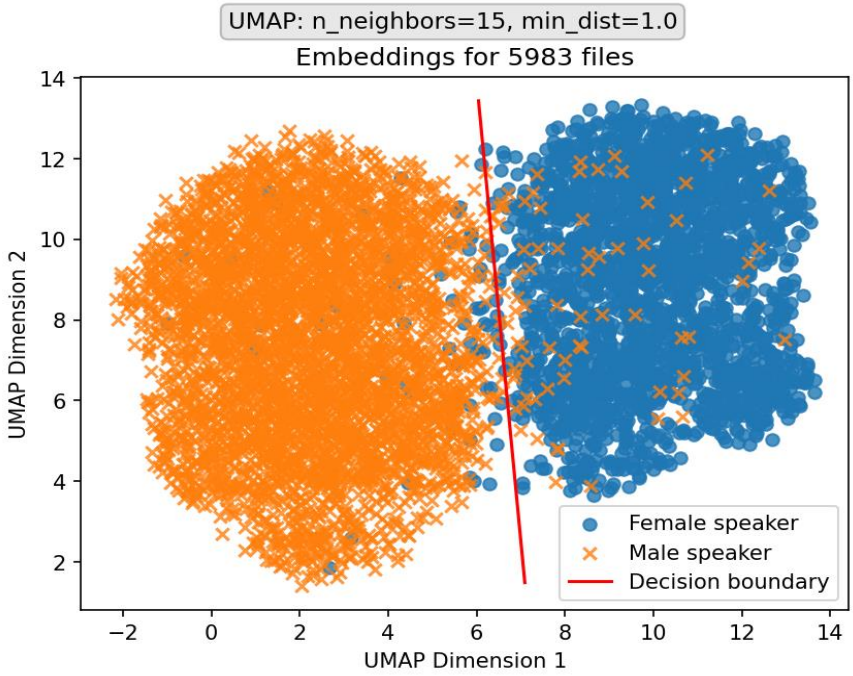
3.2 Voice Encoders

A pre-trained voice encoder was required to extract speaker embeddings. Three encoders were tested:

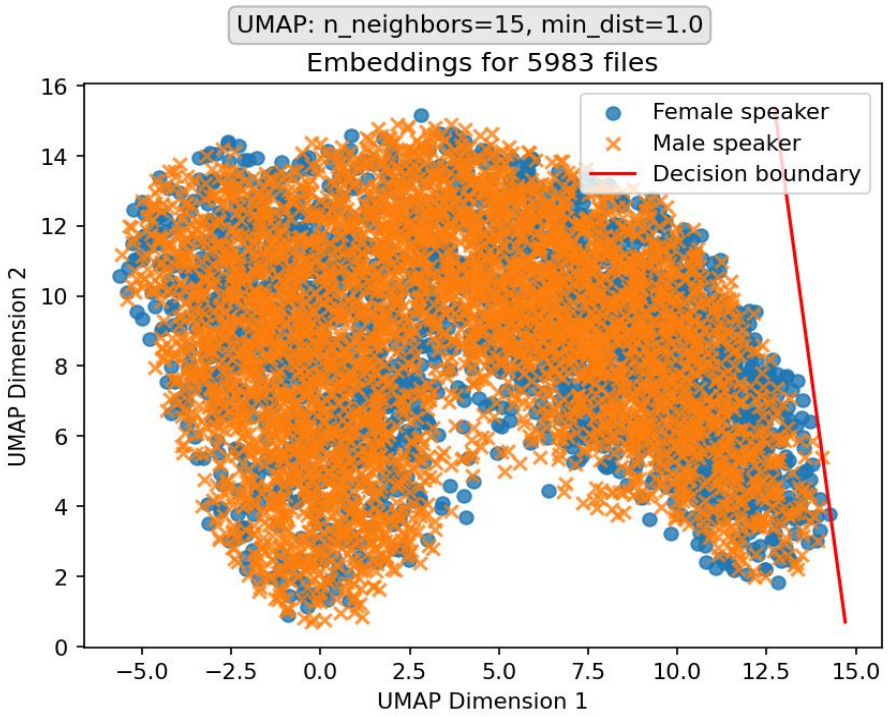
1. **ECAPA-TDNN**
   a. Uses **Attentive Statistical Pooling (ASP)** to handle variable-length inputs.
   b. Produces a **192-dimensional** speaker embedding.
   c. **Preprocessing involves** STFT, Mel filterbanks, and per-feature normalization.



2. **Resemblyzer**
    a. Splits audio into **partial utterances** (default ~1.6 seconds each).
    b. Extracts 40 Mel-frequency bands for each partial, then **averages partial embeddings** into a single **256-dimensional** vector.
    c. **Preprocessing involves** volume normalization (up to a target loudness) and trimming of long silence

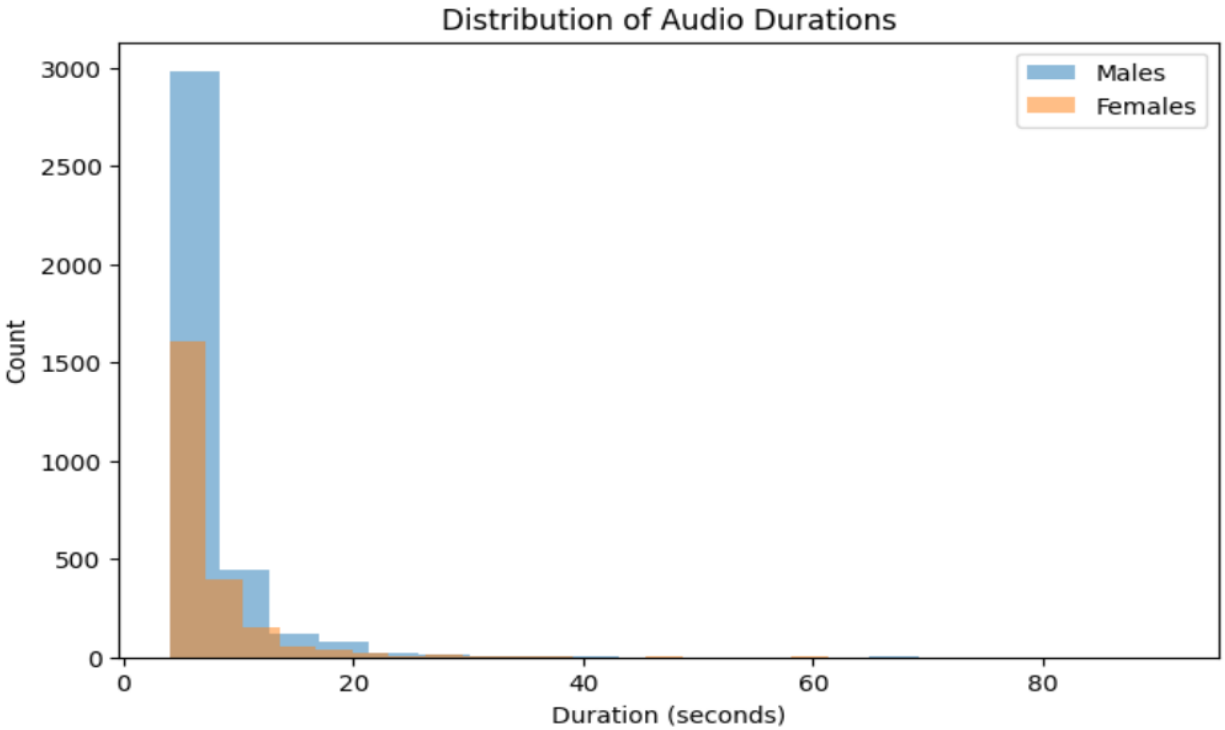


3. **Wav2Vec**
    a. By default, produces variable-length outputs aligned to the input duration.
    b. **Averages frame-level** outputs to yield a single **768-dimensional embedding**.
    c. **Preprocessing involves** zero mean and unit variance normalization.
    d. Embeddings are more strongly **correlated with linguistic information than speaker characteristics**.



4. Experimental Setup

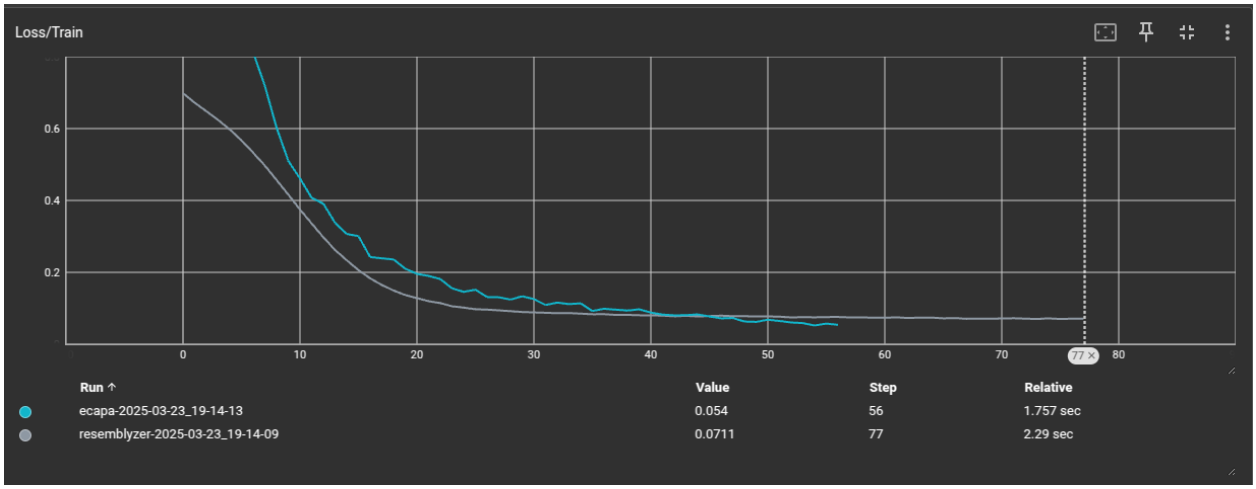
- **Dataset**:
    - Approximately 60% male and 40% female audio samples.
    - Mean audio duration ~7 seconds (distribution skewed to shorter lengths).
- **Data Splits**:
    - 80% training, 10% validation, and 10% testing.

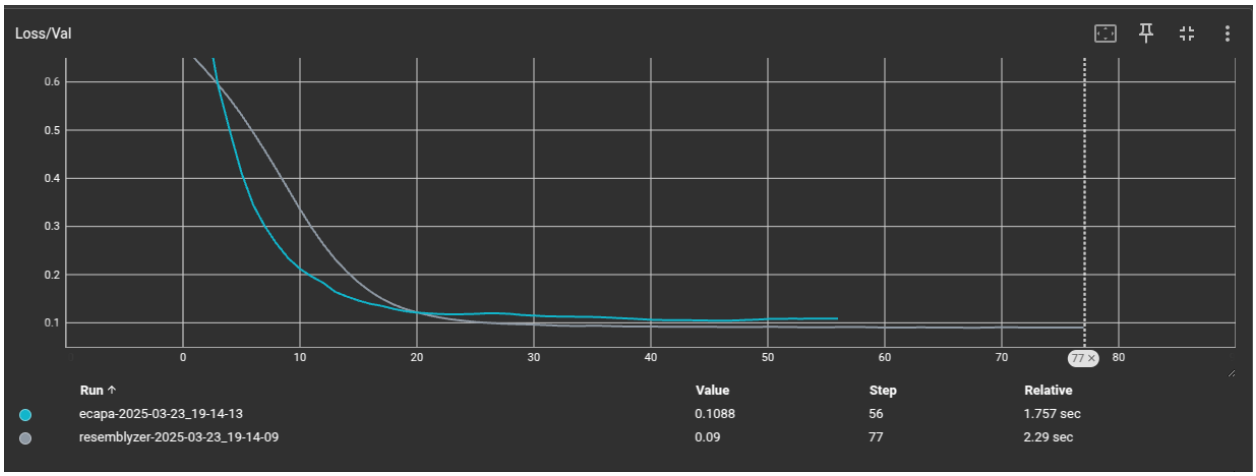


- **Classification Model**:
  - After extracting embeddings (from ECAPA-TDNN or Resemblyzer), a small neural network classifier was trained:
    - One hidden layer: (input_dim, input_dim)
    - Output layer: 2 classes (male/female)
    - Cross-entropy loss function.

5. Results

1. **ECAPA-TDNN**
   a. **Total Epochs:** 57
   b. **Best Validation Loss:** 0.1043
   c. **Test Accuracy:** 0.9783
   d. **Test Loss:** 0.1078
   e. **Early Stopping:** True
2. **Resemblyzer**
   a. **Total Epochs:** 78
   b. **Best Validation Loss:** 0.0898
   c. **Test Accuracy:** 0.9800
   d. **Test Loss:** 0.0718
   e. **Early Stopping:** True





Both models achieved high accuracy on the gender classification task, and there was very little difference in final performance.

6. Analysis and Observations

- Most of the misclassified files were found to be mislabeled in the dataset (e.g., 194.m4a, 457.m4a, and 2263.m4a in the female folder, and 1064.m4a and 2215.m4a in the male folder).
- A few truly challenging samples included one file (386.m4a in the female folder) with background music, which was misclassified by ECAPA-TDNN, and another file (1334.m4a in the female folder) with multiple overlapping voices, which was misclassified by Resemblyzer.
- Overall, the minor discrepancies in these few problematic files were not indicative of significant performance gaps.

7. Conclusion

The combination of Silero VAD for speech detection and a pre-trained speaker encoder (ECAPA-TDNN or Resemblyzer) yielded near-perfect accuracy on the given dataset for gender classification. Instead of implementing a separate custom preprocessing stage, I relied on the recommended preprocessing steps already integrated into these pretrained models. This approach resulted in robust performance, even in the presence of volume inconsistencies or background noise.