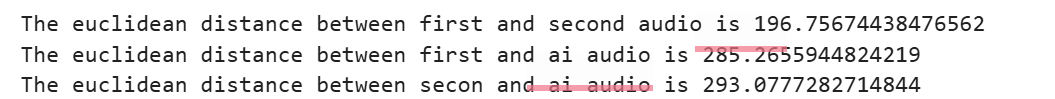
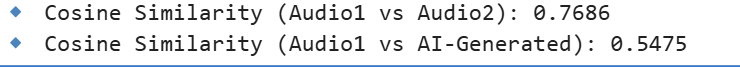
**Voice Identification:**

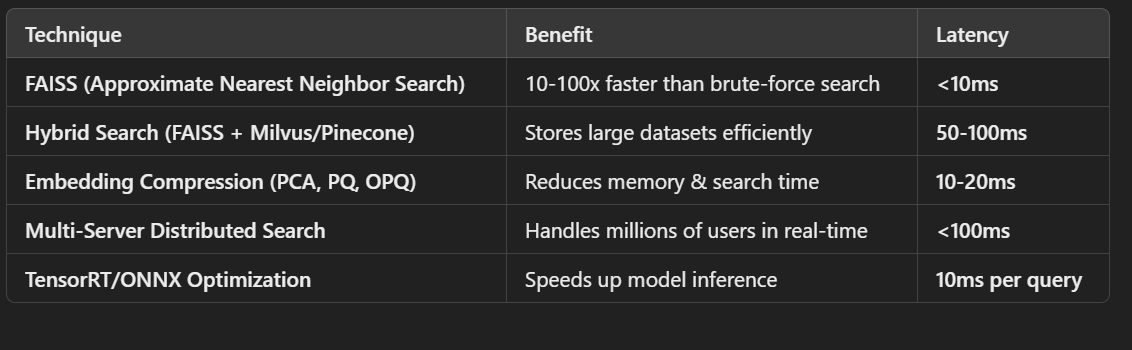
1. **When first user registers maintain a 5-10 sample audio samples of user.**
2. **Extract the features from the audio sampels of the user at the time of registration.**
   1. **The features could be MFCC.**
   2. **Or the features could be the aggregation of many features like although MFCC is recommended by many. Try combining other features if it improves the performance then retain else drop other features**
      1. **MFCC**
      2. **Zero crossing**
      3. **Chroma features**
3. **If the MFCC features doesn’t perform nicely we can use pretrained models like ECAPA-TDNN, Wav2Vec2 to extract features from the audio samples. I tried the ECAPA-TDNN model its size is 250 mb approx. The result with three voice samples:**

****

**We should check with more samples approx. 50 to make a decision whether we can use this model or not.**

**The embedding dimension of this model is 192.**

**To further reduce the latency of the model we can try dimentionality reduction to reduce the embedding length(around 50 or 100 we should try and decide as a hyper parameter) of the voice sample so that it takes less time.**

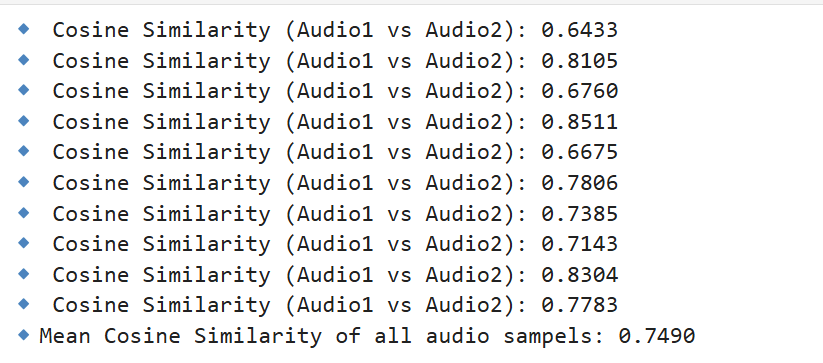
**Few more efficien user searching methods are **

**Need to try and explore which one works better and easy to implement.**

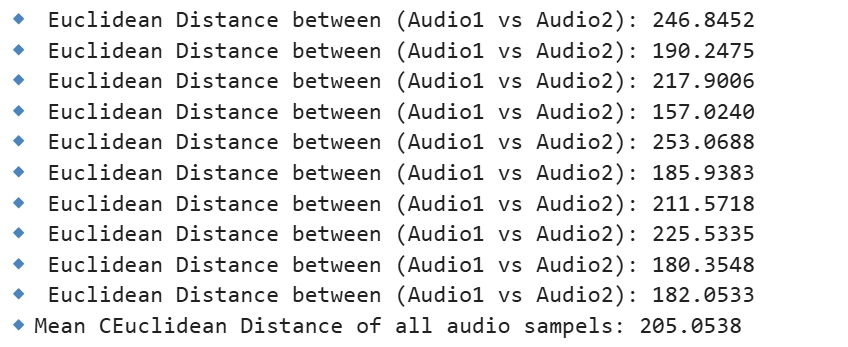
1. **As new users register extract features and append to database.**
   1. **Compute centroid (mean or median embedding of all the sample embeddings for a user in the database).**
2. **During login use the same model and extract feature and compare with all the centroid embeddings. If the similarity exceeds a particular value grant login. (Search using above mentioned techniques for reduced latency)**
3. **Training model is optional for current problem. We can directly grant login through voice similarity**

**For voice identification there are two techniques:**

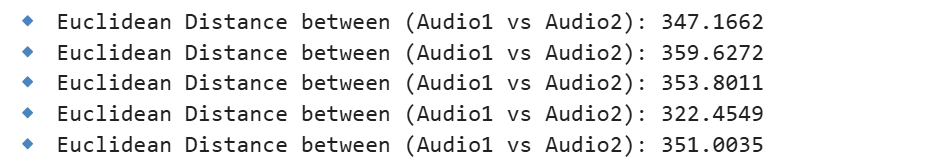
**Text Independent: The user can say any phrase.**

**When I used ECAPA-TDNN and tried text independent I got a huge deviations in the features of the same person’s voice( in this case my voice.) Below is an example when I stored five samples of my voice and tried to find similarity between all of them in pairs.**

**Euclidean**

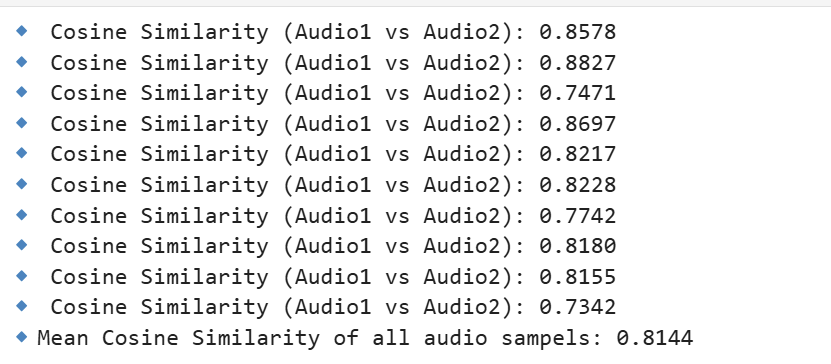
****

**I felt that if I say two different phrases with different lengths the distance or similarity of the pair is low when compared to two phrases with approx. same length.**

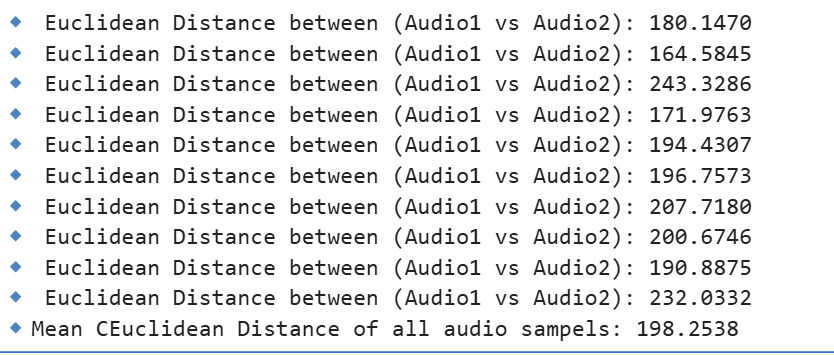
**When I compared 6th voice sample of mine to imitate login of the user these where the results. Very high Euclidean distance**

**Text Dependent: The user has to repeat same password again and again.**

**When I used ECAPA-TDNN and tried text dependent voice authentication the results are below.**

****

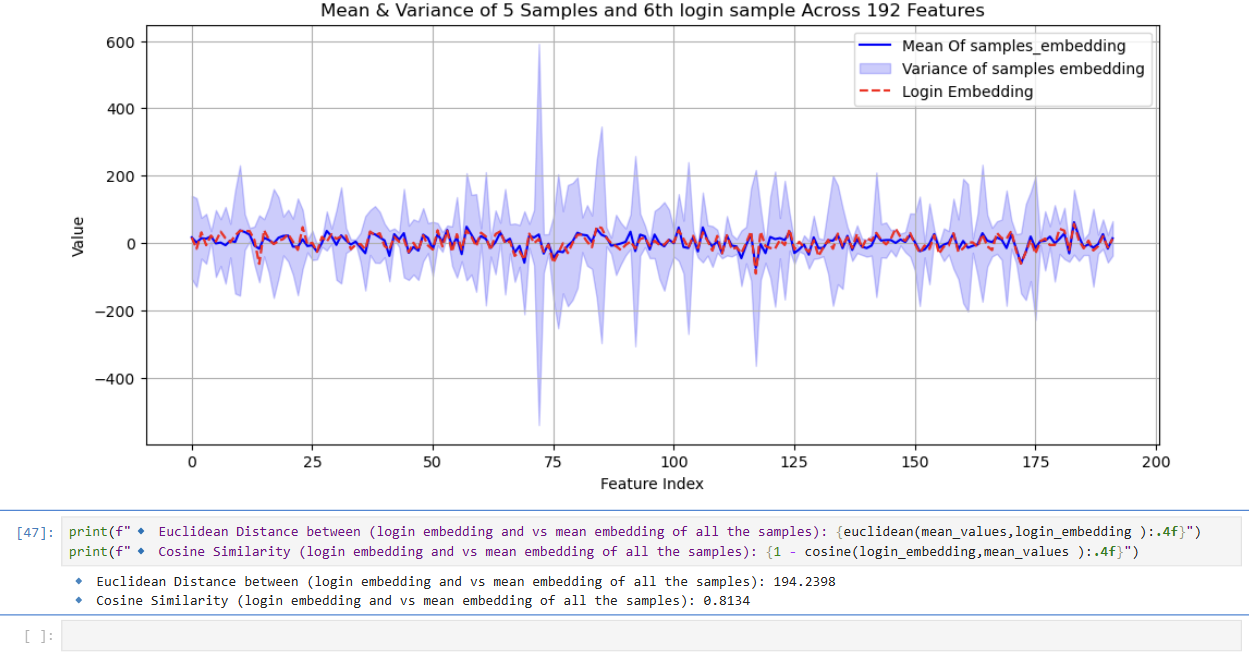
**Euclidean:**

****

**This seems better than text independent technique.**

**When I compare similarity with 6th voice imitating login of a user I got below metrics**

****

**** these are the values I am getting when I compare the login sample with the mean of all the voice samples in the database. Maybe if we try it for the voice samples of 8-10 people we can approximate the values of Euclidean and cosine to identify a particular user.