Audio Content Analysis

I calculated the byte size of 5,908,598 audio files and grouped them based on their byte sizes. The below Excel file contains two columns: "Byte Size," representing the size of the audio files, and "Group Size," indicating the total number of files that share the same byte size.

group_sizes_based_on_Byte_size

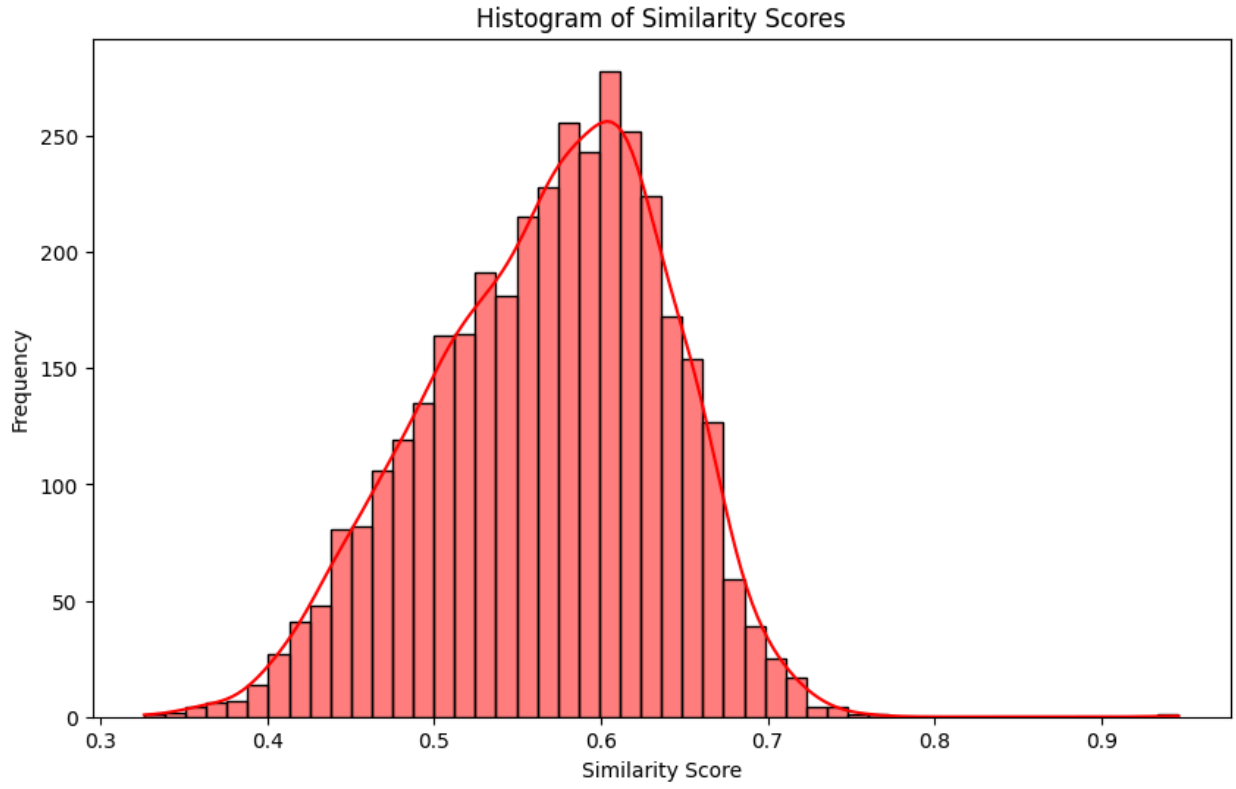
Group with Byte size: 87,314 and 4204 files.

Next, I generated embedding for each audio file with a byte size of 87,314 (a total of 4,204 files) using the "facebook/wav2vec2-base-960h" model. After obtaining the embeddings, I used FAISS (Facebook AI Similarity Search) to identify the most similar audio file for each query within the group

Below are the FAISS algorithm results, which include each filename, the filename of its most similar counterpart within the group, and the similarity score between them.

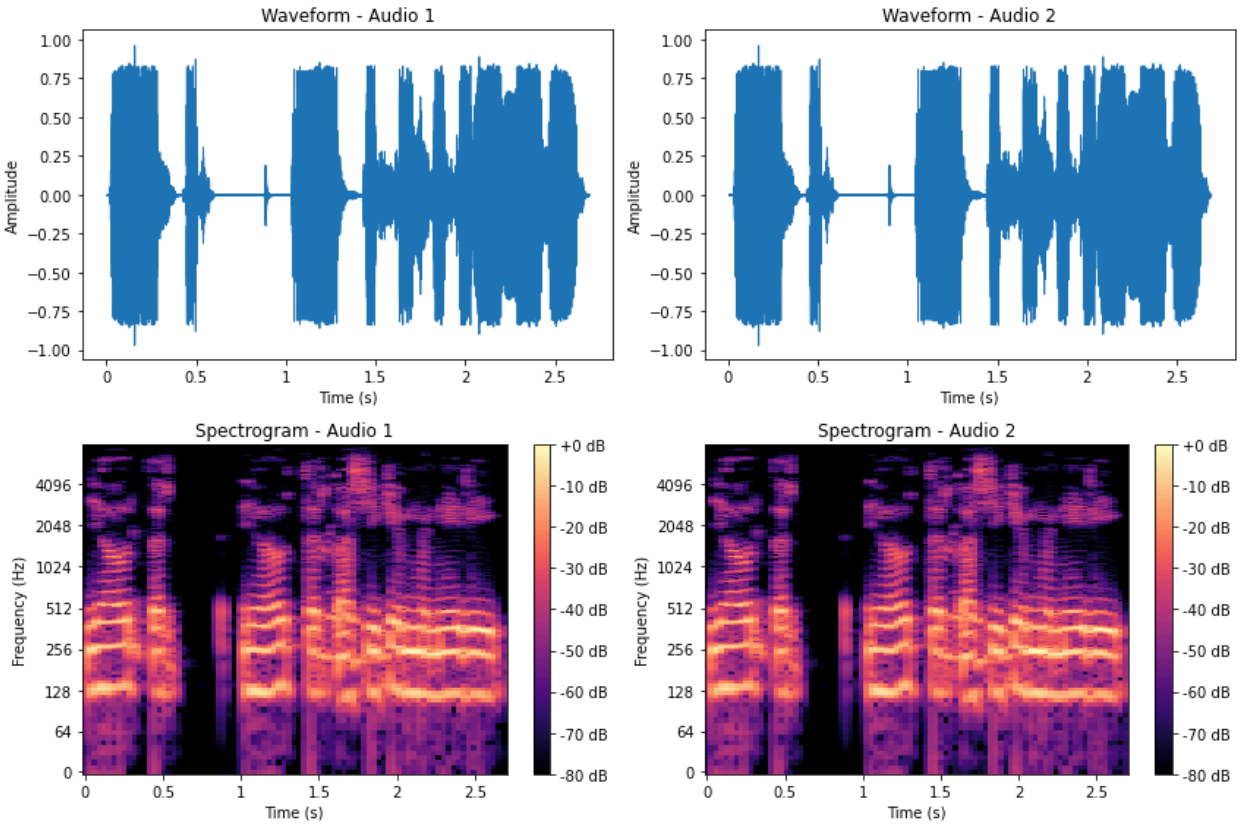
group_similarity_results

The below figure represents the histogram of similarity score from the above xlsx file.



Analysis

After listening to the audio pairs with a similarity score of less than 0.8, I found that they aren't exactly similar. There was only one pair with a similarity score of 0.945 that sounded exactly the same upon listening. I also plotted the waveform and spectrogram for both audios in this pair to better understand the match. As for all other files with a similarity score of 1, they were found to be exactly similar after listening.



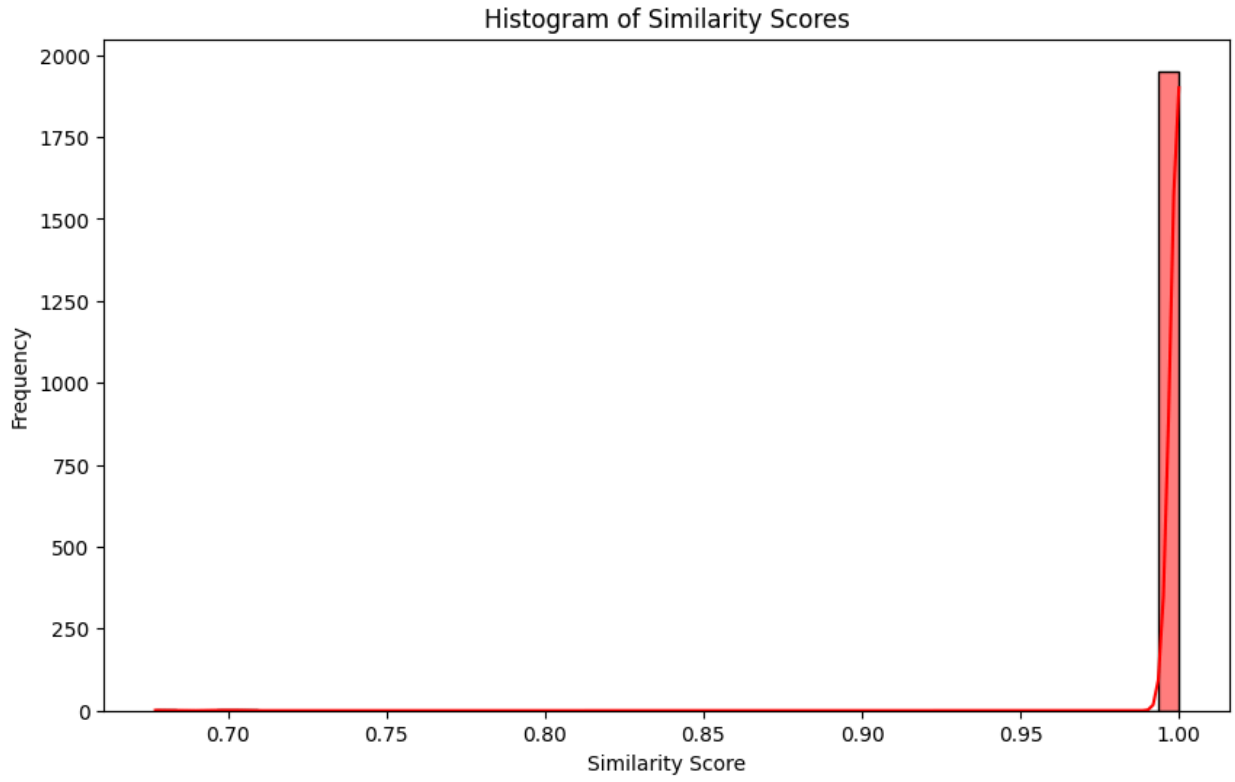
Reference Pair in Megdap file

Using the same procedure as described above, I identified the most similar audio files for the reference files listed in the TSV.

These reference pairs were found to be identical in the Megdap file: `/data1/manual_qc_references/utils/ref_audiomatch/audio_match_allAudios_megdap_2024-06-03_spkrSize.tsv`.

similarity_results

The below figure represents the histogram of similarity score from the above xlsx file.



Analysis

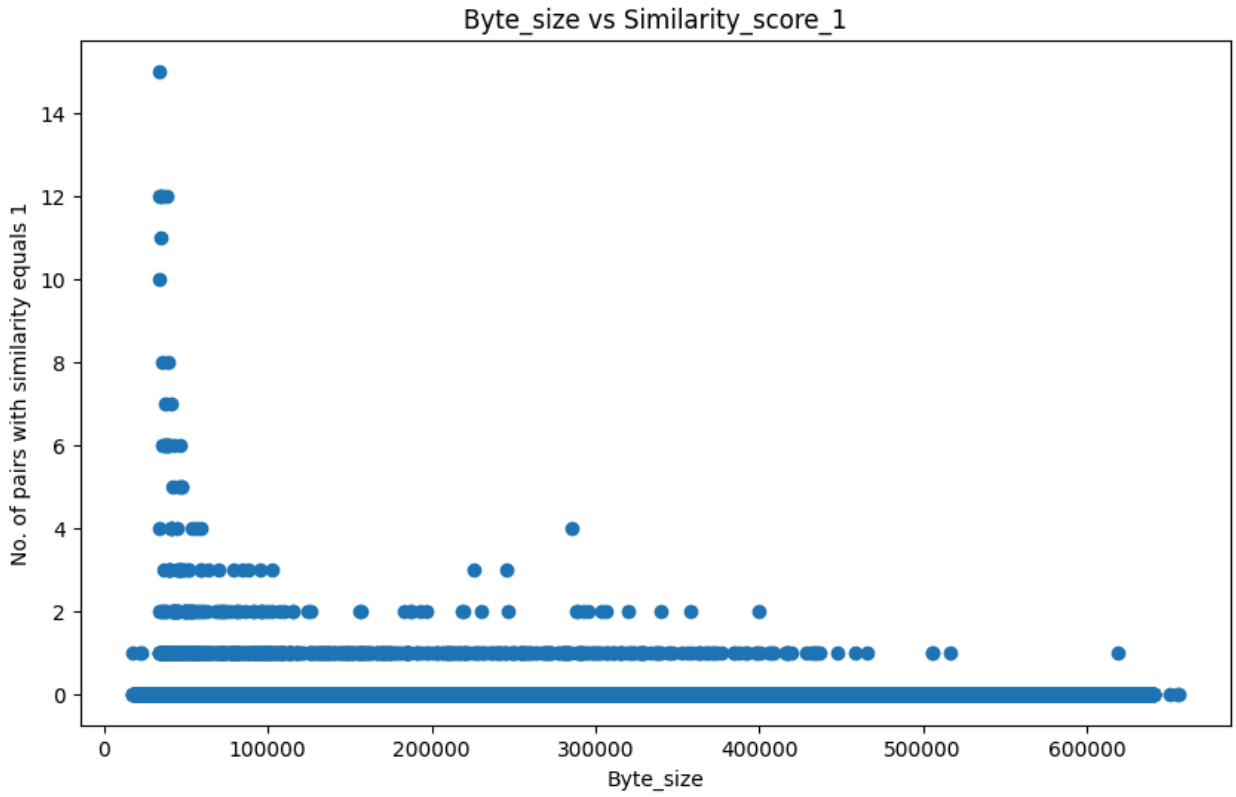
In the above Excel file, there are only three pairs with a similarity score of less than one, approximately 0.7. After listening to the corresponding audio files, I found that they are not similar. As for all other files with a similarity score of 1, they were found to be exactly similar after listening.

All groups

Using the same procedure as described earlier, I've identified the most similar audio files across all groups categorized by different byte sizes. I have compiled this information into an Excel file, which details the number of files with a similarity score of 1 for each group.

similarity_analysis

Below is the scatter plot of Byte_size vs number of files with similarity score of 1:



Analysis:

Some audio files feature mumbling or unclear speech, causing the model to generate identical embeddings for these files, which results in a similarity score of 1, despite them being different. However, for files with clear speech, the similarity scores are accurate. A similarity score above 0.9 is worth considering, as files with such a score are likely to be the same.

Path to code -

My gcp instance - 35.200.234.154

/data/Root_content/Vaani/audio_content_analysis