First Contribution Chapter: Abdullah Saleem

Extracting MFCC acoustic features from the datasets

Feature extraction on the training dataset

In this segment of our project report, we delve into the intricacies of extracting Mel-Frequency Cepstral Coefficients (MFCCs) from the training data, a foundational step in initialising our prototype Hidden Markov Models (HMMs) for each word in the vocabulary.

The task at hand involved processing a collection of 390 MP3 audio files, representing various speakers uttering one of 13 words.The key to our feature extraction was the preliminary phase, where we set up arrays for storing file paths, audio signals, sampling rates, and the extracted feature sequences. A crucial aspect of this phase was defining the frame and hop sizes for feature extraction, which was set at 30 ms and 10 ms, respectively. This setup ensured a 20 ms overlap, aligning with the tasks requirements.

The core of our process, the feature extraction loop, was a sequence of steps for each audio file in the dataset:
- Utilising audioread, each audio file was parsed to extract the audio signal and its sampling rate.
- We then initiated an audio feature extractor, specifically configured for MFCC extraction. This extractor operated with a Hamming window, aligning with the tasks guidelines.
- The extraction of MFCC features formed the crux of this phase. Here, we focused on extracting 13 MFCC coefficients per frame, intentionally omitting the delta and delta-delta features, to form a 13-dimensional feature vector for each frame.

The culmination of this process was the organisation of the extracted MFCC features into a cell array. This structure was chosen to manage the variability in sequence lengths across the audio files efficiently.This approach to feature extraction was instrumental in capturing the essential characteristics inherent in the spoken words, while adhering to the tasks parameters. The MFCCs encapsulated the nuances in speaker variations and pronunciation styles, which are important for speech recognition systems.

The seq variable, displayed below, is a 390 x 1 vector created using MATLAB's audioread() and audioFeatureExtractor() functions. It encapsulates the MFCC features extracted from each audio file in our dataset, showcasing the depth of our feature extraction process.

Variables - seq

seq

390x1 cell

	1	2	3
1	55x13 doub...		
2	63x13 doub...		
3	52x13 doub...		
4	63x13 doub...		
5	52x13 doub...		
6	73x13 doub...		
7	59x13 doub...		
8	59x13 doub...		
9	63x13 doub...		
10	55x13 doub		

In the process of consolidating the extracted MFCC features, we employed the vertcat() function in MATLAB on the seq variable. This operation resulted in a comprehensive matrix, encapsulating the entirety of the extracted features across all audio files. The resulting matrix, named all_mfccCoeffs, is a substantial 17394 x 13 matrix. This structure not only embodies the full spectrum of our feature set but also visually represents the robustness of the data we processed. Below is an illustrative image of `all_mfccCoeffs`, showcasing its structure and dimensionality.

Variables - all_mfccCoeffs

all_mfccCoeffs

17394x13 double

	1	2	3	4	5	6	7	8	9	10
1	-73.1349	2.3238	0.9684	0.6628	-0.1378	-0.3929	0.4625	0.3280	-0.0760	0.6770
2	-71.0738	1.8652	1.2422	0.6984	0.2762	-0.3519	0.1891	0.1013	0.1023	0.4754
3	-66.5496	3.1643	1.4710	1.3738	0.6327	0.0396	0.1505	0.0556	0.0542	0.1233
4	-62.2887	3.2058	1.3591	1.1012	1.1266	0.5532	0.8709	0.2971	-0.0355	0.6093
5	-61.1160	3.1881	1.1969	0.7653	1.2729	0.4348	0.7564	0.5261	0.3817	0.6600
6	-59.5215	3.7951	1.3285	0.7535	0.8430	-0.0748	0.1391	0.5874	0.6121	0.4325
7	-55.8388	3.9751	-0.0156	0.0325	0.5061	-0.2313	-0.8341	0.0337	-0.1934	-0.1113
8	-52.1203	1.9496	-1.0020	0.7873	-0.6919	0.2904	-1.5838	-0.2876	-0.2438	-0.3182
9	-47.7964	-0.3320	-0.1897	1.1332	-1.5132	0.8065	-0.9829	0.5980	0.2522	-0.1814
10	-44.9247	-1.8119	0.0895	1.2539	-1.9094	1.3388	-0.6505	1.4440	0.6116	-0.2814

Feature extraction on the validation dataset

In the validation phase of our speech recognition project, we focused on the crucial task of extracting Mel-Frequency Cepstral Coefficients (MFCCs) from a smaller dataset. This validation dataset, distinct from the more extensive training set, consists of 55 audio files with a single speaker articulating 11 different words, each repeated five times. Our approach to feature extraction for this dataset mirrored the methodology applied to the training set.

This methodical approach to feature extraction on the validation dataset was instrumental in setting the stage for the model validation. By maintaining consistency in feature extraction parameters across datasets, we ensured a reliable and accurate evaluation of our Hidden Markov Models performance. The insights gained from this phase were pivotal in assessing the models generalisation capabilities and its effectiveness in recognizing speech under varying conditions.

Feature extraction on the test dataset

This dataset consists of audio recordings made by our team, tailored to assess our speech recognition model. For consistency with the training data, each recording was manually trimmed to remove silences and set to the correct sampling and bit rates. Using a MATLAB script, we read the audio files to extract the signal and sampling rate. The script was configured for a 30 ms frame size and 20 ms overlap, using a Hamming window for each frame, to mirror our training dataset's parameters.

The extraction focused on obtaining 13 MFCC coefficients per frame, maintaining the dimensional consistency of our training data. This approach ensured that the extracted features from our test dataset were directly comparable to those in the training phase. The verification of the extraction output confirmed the suitability of these features for our speech recognition task

Label extraction helper function

The function's objective was to classify unique labels corresponding to each spoken word. This classification was foundational for initialising individual HMMs. We commenced by defining necessary variables, including the number of classes (30 unique words), the number of states in each HMM (8), and the dimensionality of feature vectors (13).

We then created a cell array, classFeatureVectors, to store Mel-Frequency Cepstral Coefficients (MFCC) features for each word, and an array, unique_labels, to track the distinct words. The script processed each audio file, extracting word labels from filenames and aligning them with our predefined ClassNames array. It then stored corresponding features in classFeatureVectors and updated unique_labels to ensure a comprehensive collection of unique words.

This methodical approach facilitated the organised representation of features, crucial for the accurate initialization of HMMs. By categorising features based on word labels, we ensured a robust foundation for effective model training.

Chapter Summary

In this chapter, my primary focus was on Task 2, dedicated to feature extraction and model initialization. My contributions were mostly in the preliminary stages, where I spearheaded the extraction of Mel-Frequency Cepstral Coefficients (MFCCs) from the various datasets. This foundational work was critical for the initial setup of Hidden Markov Models (HMMs) for each word in the training set. Additionally, I tackled a significant technical challenge in the calculate_transition_and_occupation.m script, resolving '-inf' errors in transition matrices caused by log space conversion issues. Furthermore, I played a supportive role in calculating the global statistics and flat start model parameters, essential for the initialization of HMMs for each word in our vocabulary