CS6300 Assignment 4 Report

Sandeep Kumar Suresh - CS22Z121, Praveen S V - CS21D201

4 March, 2023

1 MFCC Feature Extraction

Mel Scale is a scale that maps frequency of signal to the frequency of perception of the human ear. In Figure 1, we observe that it follows a nearly linear scale upto 1 KHz and a logarithmic curve above 1 KHz.

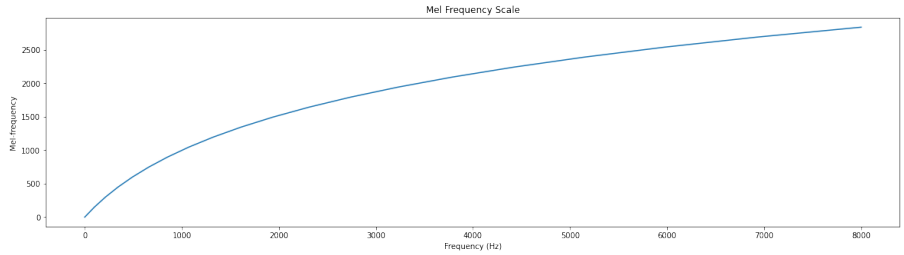

Figure 1: Mel Scale

The triangles in Figure 2 for the Mel Filter Bank correspond to the frequency in the Mel scale. Figure 2 has 40 Filter banks that model the Mel Scale.

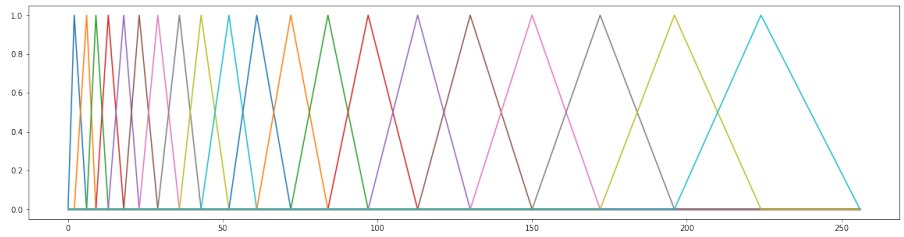

Figure 2: Filter Bank

In Figure 3, the Mel-Spectrogram appears to be like a smoothened version of the Spectrogram. Mel Spectrogram is the Spectrogram multiplied by the Mel Filter Banks. In Figure 2 we can see that there is large number of filters in the lower frequency and less number of filter at the higher frequency .This is due to the fact that the Mel Filter bank is similar to how our ear perceives sounds , i.e. models the perception of the ear. This is intuitive as human hearing is not equally sensitive to all frequency, i.e it is non linear.

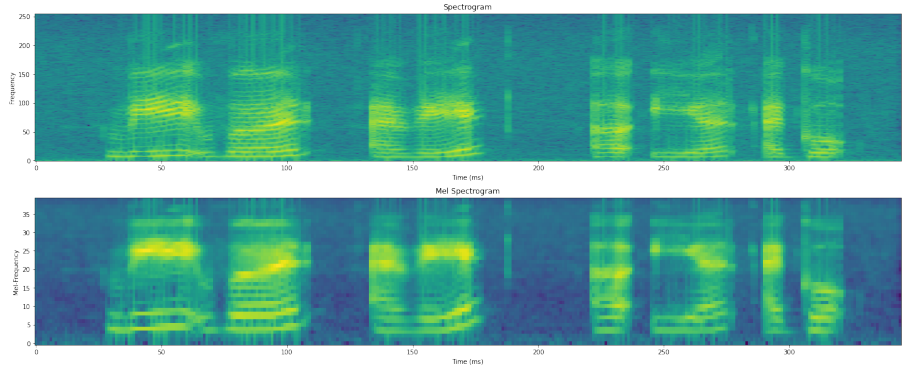

Figure 3: Spectrogram vs MelSpectrogram

2 DTW

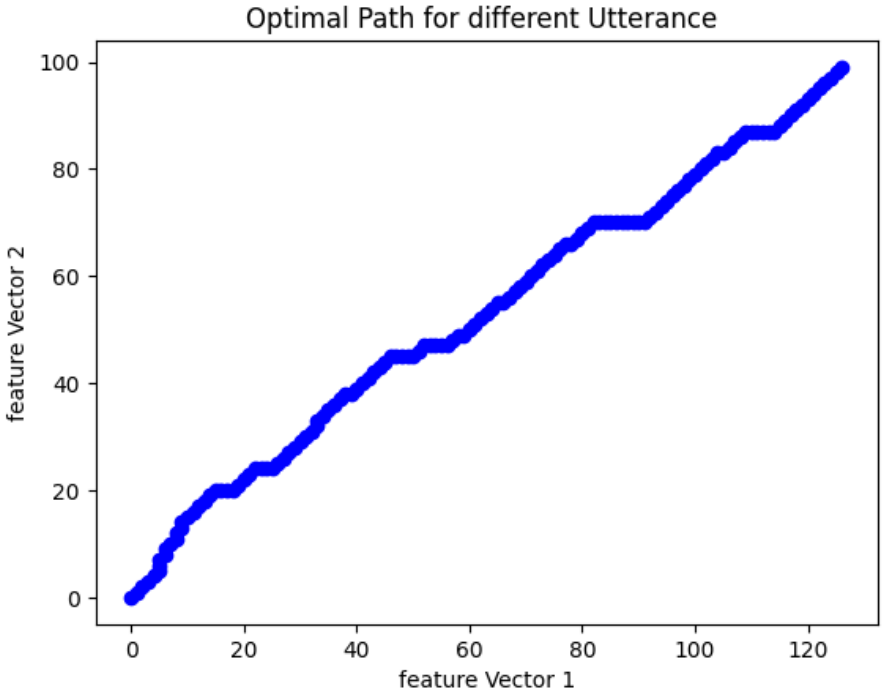

Figure 4: Allignment for utterances of different lengths

2.1 Results for Isolated-Digit Recognition

We perform Isolated Digit Recognition using DTW and achieve a best test-set accuracy of 98.33%.

N_{FFT}	$N_{mel\_filters}$	$N_{DCT\_coefficients}$	**test accuracy (%)**
512	40	13	96.67
1024	40	13	98.33
2048	40	13	98.33

Table 1: Results varying number of DFT points considered for Isolated Digit Recognition

N_{FFT}	$N_{mel\_filters}$	$N_{DCT\_coefficients}$	**test accuracy (%)**
1024	10	13	98.33
1024	20	13	98.33
1024	40	13	98.33

Table 2: Results varying number of mel-filters for Isolated Digit Recognition

- In Table 1, we compare the test-set accuracies obtained by varying the number of FFT points (N_{FFT}) used in the computation. While using a smaller window size of 512 points yields poorer performance, it is sufficient to use 1024 points to expect better results.
- Similarly, in Table 2, we draw a comparison of the choice of the number of mel-filters vs. the test-set accuracy. We observe that by $N_{FFT} = 13$ and using 13 DCT coefficients, there is no impact of varying the number of mel-filters, suggesting that even 10 filters are good enough to solve the problem. However, since we deal with a limited test set, it would also be interesting to see if this observation extends to the remaining digits.
- We also study the choice of number of DCT coefficients, in Table 3, and find that lower numbers show utility, with 8 and 13 yielding 96.67% performance whereas using 20 coefficients hampers test-set accuracy.

N_{FFT}	$N_{mel\_filters}$	$N_{DCT\_coefficients}$	**test accuracy (%)**
512	40	8	96.67
512	40	13	96.67
512	40	20	95.00

Table 3: Results varying number of DCT co-efficients for Isolated Digit Recognition

Next, in Table 4, we study the original DTW algorithm vs. the KNN-DTW. For this, we choose the minimum over the average of the first K distances for each reference digit.

Parameters	**K**	**test accuracy(%)**
n_fft=512, sr=20000, hop_size=0.010, n_mels=40, n_dcts=13	5	78.33
	10	91.66
	15	91.66
n_fft=1024, sr=20000, hop_size=0.010, n_mels=40, n_dcts=13	5	80.00
	10	96.67
	15	95.00

Table 4: Comparison of KNN-DTW test set accuracies for different values of K.

- Since, using an elbow-method equivalent for finding the optimal K was computationally expensive, we study three options of K: 5, 10, and 15. We find that using a $K = 10$ yields the best performance.
- However, the performance of KNN with DTW was not better than vanilla DTW. A possible explanation is that for the same digit, by comparing using each reference template in vanilla DTW, we are able to find a better match, whereas the average-distance obtained match is confusing the model more.

2.2 Results for Keyword-Spotting

Solving Keyword-spotting using DTW, is not as straight-forward. Using unconstrained-endpoint DTW enables us to obtain partial matches. A naive way would be to form all possible concatenations of reference templates upto level L, ($L = 3$ for our dataset), and then perform UE-DTW to find the concatenated reference that has the closest match to the test template. However, this method is computationally expensive.

- We notice that just by employing UE-DTW using an isolated digit with a test of connected digits gives us interesting alignments. Notice in Figure 5, the regions corresponding to the "3"s have monotonically increasing alignments with high slope, whereas the middle region where "z" is pronounced has a flat slope. In this way, the naive UE-DTW does seem to help us spot the keyword "3" in the test utterance. However, this is not an fully-automated process.

- In the second approach as seen in Figures 7, 9, 10, and 11, we utilize VAD to identify voiced segments and try employing DTW to match each of these segments. The idea was to classify a segment as matching the given word, if the DTW distance was below a certain threshold. However, while running this for a few examples, we saw that the setting a threshold became rather difficult. Suppose the keyword given is '3', then we found that some '3's were getting rejected even though they matched due to the high threshold while setting lower thresholds implied segments without '3's would be wrongly misclassified as containing the keyword '3'.
- Finally, another way to approach the problem would have been to uniformly segment the test utterance into L frames, here we know L is either two or three and perform DTW with each of these segments to identify if the keyword is present. However, this approach also seems to have the same limitations as that of using segments obtained from VAD.

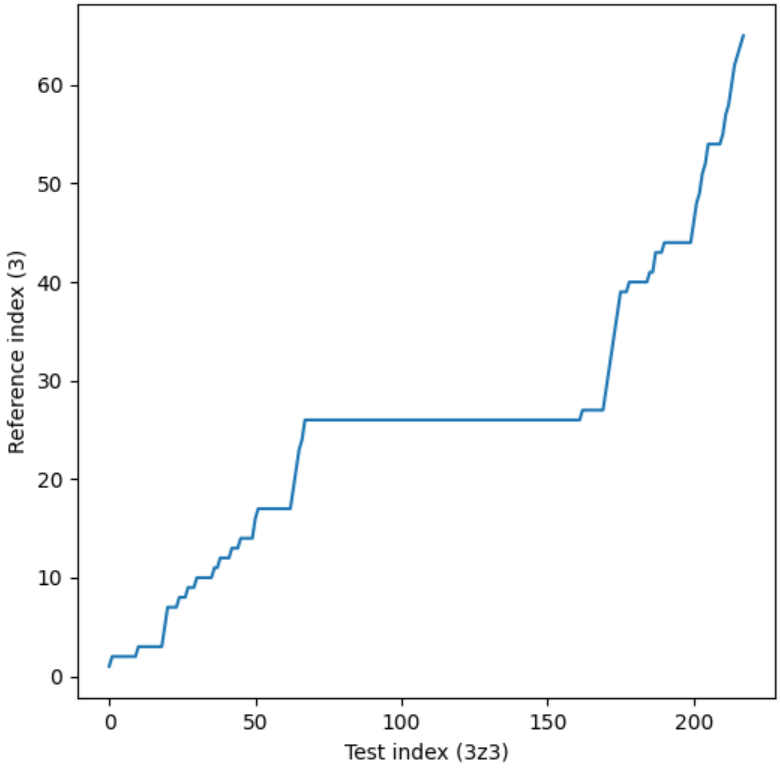

Figure 5: UE-DTW alignments of isolated digit "3" with the connected digits "3z3"

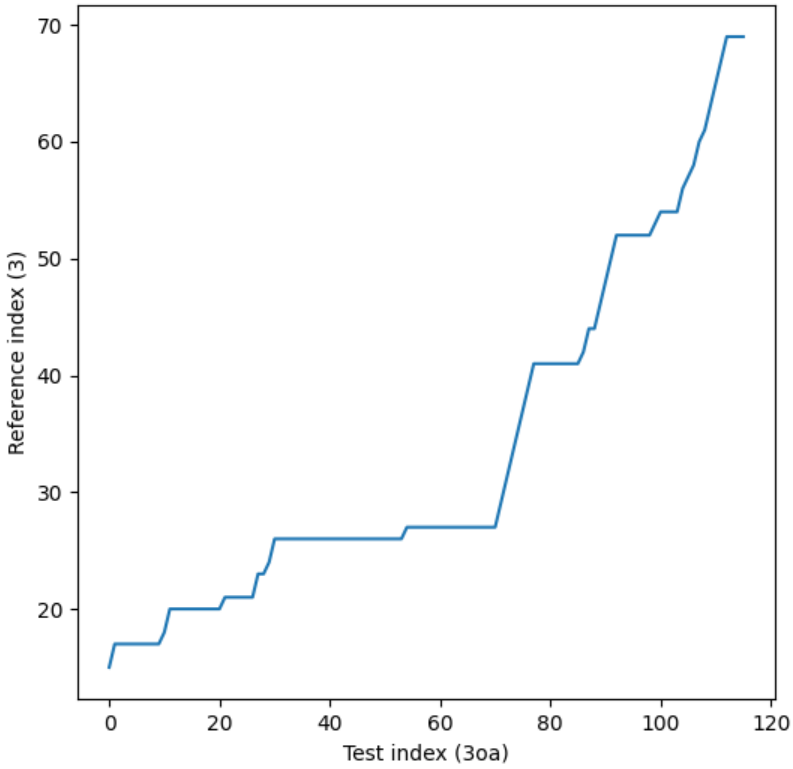

Figure 6: UE-DTW alignments of isolated digit "3" with the connected digits "3o"

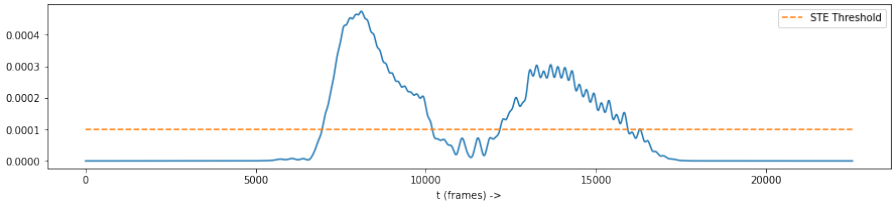

Figure 7: STE

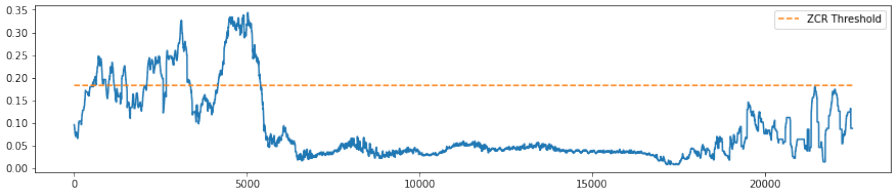

Figure 8: ZCR

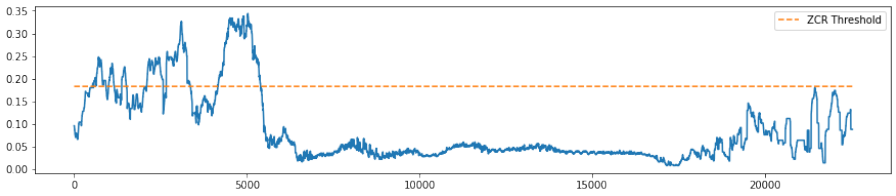

Figure 9: ZCR

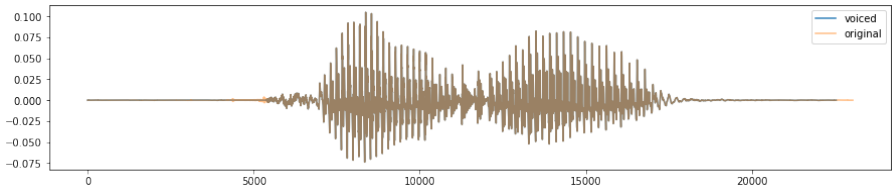

Figure 10: VAD

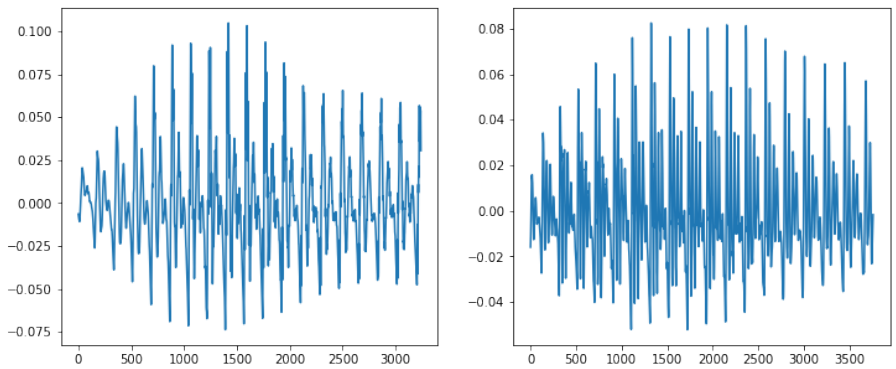

Figure 11: Segments obtained from VAD which is further matched by DTW