Digit Recognition Using Dynamic Time Warping

spate_235

Project Overview

This project focuses on recognizing spoken digits using the Dynamic Time Warping (DTW) algorithm. DTW is a powerful technique for measuring similarity between two sequences, which may vary in time or speed. The dataset provided includes 300 spoken digit recordings (30 utterances for each digit from three speakers: Jackson, Nicolas, and Theo). Each recording is stored as a matrix with dimensions M×13M×13, where MM denotes the number of frames and 13 represents the feature dimensions per frame.

The primary objective is to take each test utterance and compute its alignment cost with all other recordings to identify the digit being spoken accurately.

Methodology

Step 1: Dissimilarity Matrix Computation

For each test utterance TT and reference utterance RR, we compute an M×NM×N dissimilarity matrix SS, where MMand NN represent the number of frames in TT and RR, respectively. The elements of SS are calculated using the Euclidean distance between corresponding frames of TT and RR.

Step 2: Accumulated Cost Matrix and Optimal Path Calculation

Using the dissimilarity matrix SS, we construct an accumulated DTW cost matrix DD. Each element D(m,n)D(m,n) is computed by considering the minimum of three neighboring elements, which represent different potential paths:

- D(m−1,n)D(m−1,n): Insertion
- D(m,n−1)D(m,n−1): Deletion
- D(m−1,n−1)D(m−1,n−1): Match

The formula used is:

D(m,n)=S(m,n)+min(D(m−1,n),D(m,n−1),D(m−1,n−1))D(m,n)=S(m,n)+min(D(m−1,n),D(m,n−1),D(m−1,n−1))

Once D(M,N)D(M,N) is calculated, the backtracking process starts from this position, tracing back to (0,0)(0,0) to identify the optimal alignment path.

Step 3: Normalized Alignment Cost

The alignment cost between TT and RR is determined by summing the dissimilarity values along the optimal path and normalizing it by the path length. This normalized cost reflects how similar the two utterances are, with lower costs indicating higher similarity.

Step 4: Confusion Matrix Generation

For each test utterance, the alignment costs with all reference utterances are sorted, and the closest 29 reference recordings are selected. The confusion matrix is then computed based on how often the correct digit is identified among these nearest neighbors.

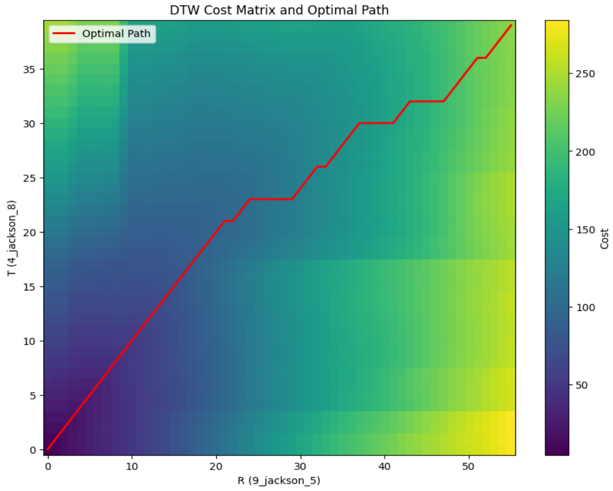
Results

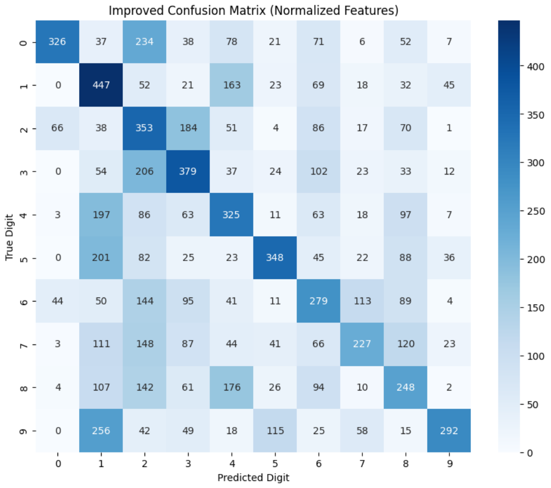
DTW Cost Matrix and Optimal Path

The first image below illustrates a sample DTW cost matrix between two recordings, along with the optimal path (highlighted in red). The optimal path represents the lowest-cost alignment between the test and reference utterances, indicating the areas of highest similarity.

Improved Confusion Matrix

The confusion matrix below demonstrates the model's performance in distinguishing different digits. Each row represents the actual digit spoken, and each column represents the predicted digit. High values along the diagonal indicate successful recognition, while off-diagonal values represent misclassifications.





True Digit \ Predicted Digit	0	1	2	3	4	5	6	7	8	9
0	326	37	234	38	78	21	71	6	52	7
1	0	447	52	21	163	23	69	18	32	45
2	66	38	353	184	51	4	86	17	70	1
3	0	54	206	379	37	24	102	23	33	12
4	3	197	86	63	325	11	63	18	97	7
5	0	201	82	25	23	348	45	22	88	36
6	44	50	144	95	41	11	279	113	89	4
7	3	111	148	87	44	41	66	227	120	23
8	4	107	142	61	176	26	94	10	248	2
9	0	256	42	49	18	115	25	58	15	292