



Digit Recognizer Using Single and Average Template Matching Techniques

Ms. Savitha and S Upadhya
EXTC Department
Fr. C Rodrigues Institute of Technology
Sector 9 A , Vashi
INDIA

Abstract: Template matching approach to speech recognition is wherein a collection of processed speech patterns (the templates) are carefully selected and stored as reference patterns representing the dictionary of candidate utterance. Recognition is then carried out by matching an unknown spoken utterance with each of these pre-stored patterns and selecting the category of the best matching pattern to be the spoken utterance. Dynamic Time Warping (DTW) is one such algorithm used for template matching. This paper presents the two template matching techniques namely the Single and Average template matching techniques that were developed to recognize the English digits spoken in isolation. The algorithm was implemented both for speaker dependent and speaker independent type of isolated digit recognition and a comparison of recognition accuracies for the two template matching techniques was made.

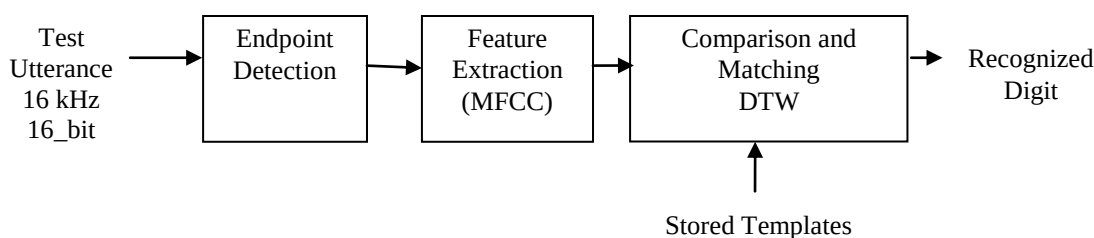
Keywords: Dynamic Time Warping; Global Distance; MFCC; Speaker Independent; Speaker Dependent.

I. Introduction

A template matching speech recognition system relies on an efficient mechanism for comparing two templates (speech patterns) and obtaining a measure of the “distance” between them. Problems that arise are the dynamic stretching or compressing of the templates during the matching process to cope with different speaking rates. The algorithm must also be as fast and as accurate as possible. This task can be achieved by a technique called Dynamic Time Warping (DTW); which is an efficient dynamic programming procedure for achieving an optimum nonlinear time alignment between two templates.

The block diagram of the speech recognition system using template matching is shown in Figure 1. It has the following components: Sound Recording and end point detection, feature extraction, and pattern recognition. The speech signal generated from microphone for the test utterance is sampled and digitized for storing. The end point algorithm detects the boundary of the spoken utterance by segmenting the stored utterance input into frames and calculating the frame energy. Feature Extraction component generates feature vectors for the speech utterance. It generates Mel Frequency Cepstrum Coefficients (MFCC) as the feature that should be used to uniquely identify the given sound signal. In the Recognition component template matching recognition technique is used where reference and test templates are compared using dynamic time warping algorithm for recognition.

Figure 1 Block Diagram of Speech Recognition System



The accuracy of the Template (DTW) based speech recognition system greatly relies on the quality of the reference template prepared. In Single template matching technique the reference template is generated using a single utterance of the given word. This has been known to cause problems due to the fluctuation in which words are uttered by the user of the system. Words can be distorted by the speed in which they are spoken as well as by physical irregularities in the speaker's voice.

The problem of preparing reliable reference templates to improve the recognition rate was solved by using the Average template matching technique. Here the reference template is extracted from a set of utterances and then averaged out after time normalization.

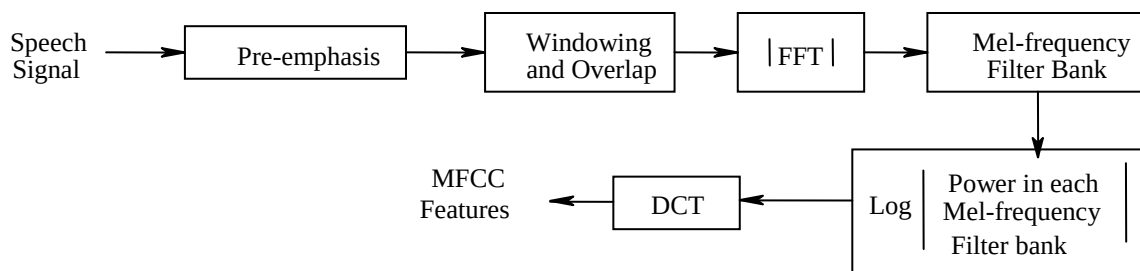
II. Feature Extraction

A wide range of possibilities exist for parametrically representing the speech signal for the speech recognition task, such as Linear Prediction Coding (LPC), Mel- Frequency Cepstral Coefficients (MFCC), and others. MFCC is perhaps the best known and most popular and is used here.

MFCC's are based on the known variation of the human ear's critical bandwidths with frequency; filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the *mel-frequency* scale, which has linear frequency spacing below 1 kHz and a logarithmic spacing above it. Speech wave is the response of the vocal tract to one or more excitation signals. In spectral terms, the speech spectrum can be treated as the product of excitation spectrum (rapidly time varying) and a vocal tract spectrum (relatively slowly time varying). A logarithmic operation on this signal reduces the product into sum. This operation is similar to deconvolution. The inverse cosine transform of this sum gives us MFCC coefficients.

Figure 2. shows the block diagram to generate the MFCC coefficients. The general procedure for producing the Mel cepstrum starts with computing the DFT with the selected frame length and overlap. The power spectrum is estimated for each frame using Fast Fourier Transform (FFT) and squaring it. Then the power spectrum is converted to Mel spectrum with the help of a Mel filter bank and the Discrete Cosine Transform (DCT) is used to get the cepstrum coefficients. The Mel filter bank that has been developed has 40 equal area triangular band pass filters with 50% overlap. 13 filters are spaced linearly for center frequencies less than 1kHz starting from 133.33 Hz and 27 filters spaced logarithmically for center frequencies greater than 1 kHz up to 8 KHz.

Figure 2 MFCC Extraction Algorithm [1, 2 and 3]



III. Dynamic Time Warping

An input template and a reference template can be expressed as a sequence of feature vectors obtained from each frame. The feature vector type used here is the Mel Frequency Cepstral Coefficients (MFCCs) as discussed in section II. Since the feature vectors could possibly have multiple elements, a means of calculating the distance is required. The idea of DTW is to find a distance measure between these two feature vectors. This distance in DTW is calculated using the Euclidean distance metric. Therefore the local distance (Euclidean distance) between feature vector x of signal 1 and feature vector y of signal 2 is calculated using equation (1).

$$d(x, y) = \sqrt{\sum_j (x_j - y_j)^2} \quad (1)$$

To obtain a global distance between two speech patterns (represented as a sequence of vectors) a time alignment must be performed [1].

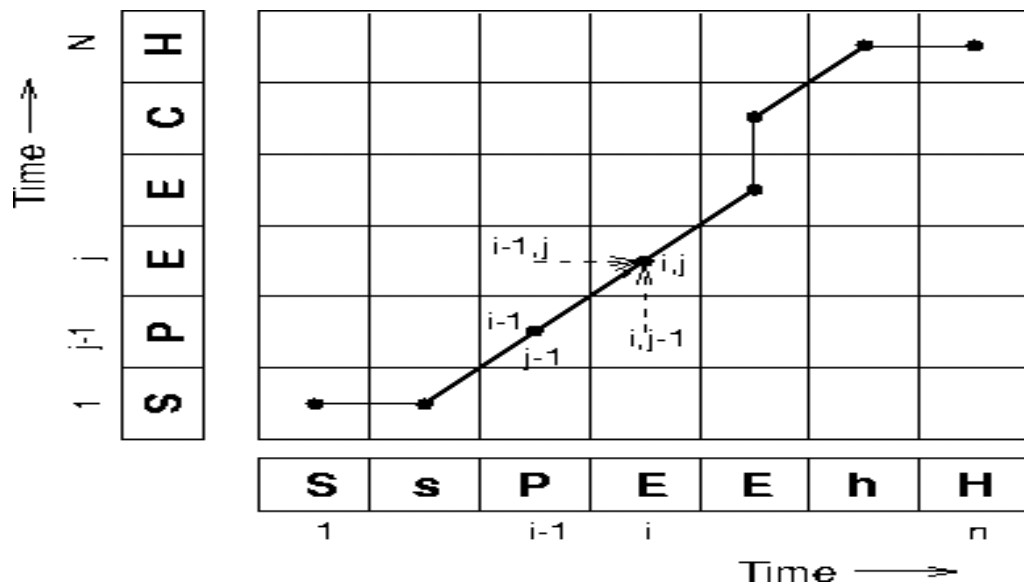
This problem is illustrated in Figure. 3, in which a "time-time" matrix is used to visualize the alignment. As with all the time alignment examples, the reference pattern (template) goes up the side and the input pattern goes along the bottom.

In this illustration the input "SsPEEhH" is a 'noisy' version of the template "SPEECH". The idea is that 'h' is a closer match to 'H' compared with anything else in the template. The input "SsPEEhH" will be matched against all templates in the system's repository. The best matching template is the one for which there is the lowest distance path aligning the input pattern to the template. A simple global distance score for a path is simply the sum of local distances that go to make up the path.

To find the best-matching (= lowest global distance) path between an input and a template, evaluate all possible paths - but this is extremely inefficient as the number of possible paths is exponential in the length of the input. Instead, we will consider what constraints there are on the matching process (or that we can impose on that process) and use those constraints to come up with an efficient algorithm. The constraints imposed are

- Matching paths cannot go backwards in time;
- Every frame in the input must be used in a matching path;
- Local distance scores are added to give the global distance.

Figure 3 Dynamic Time Warping [4]



This algorithm is known as Dynamic Programming (DP). When applied to template-based speech recognition, it is often referred to as Dynamic Time Warping (DTW). DP is guaranteed to find the lowest distance path through the matrix, while minimizing the amount of computation. The DP algorithm operates in a time-synchronous manner: each column of the time-time matrix is considered in succession (equivalent to processing the input frame-by-frame) so that, for a template of length N , the maximum number of paths being considered at any time is N .

Every frame in the template and the input must be used in a matching path means that if we take a point (i, j) in the time-time matrix (where i indexes the input pattern frame, j the template frame), then previous point must have been, $(i-1, j-1)$, $(i-1, j)$ or $(i, j-1)$. The key idea in dynamic programming is that at point (i, j) we just continue with the lowest distance path from $(i-1, j-1)$, $(i-1, j)$ or $(i, j-1)$.

If $D(i, j)$ is the global distance up to (i, j) and the local distance at (i, j) is given by $d(i, j)$ then

$$D(i, j) = \min [D(i-1, j-1), D(i-1, j), D(i, j-1)] + d(i, j) \quad (2)$$

Given that $D(1, 1) = d(1, 1)$ (this is the initial condition),

Thus we have the basis for an efficient recursive algorithm for computing $D(i, j)$. The final global distance $D(n, N)$ gives us the overall matching score of the template with the input. The input word is then recognized as the word corresponding to the template with the lowest matching score. (Note that N will be different for each template.) [4]. For basic speech recognition DP has a small memory requirement; the only storage required by the search (as distinct from the templates) is an array that holds a single column of the time-time matrix.

IV. Template Matching Techniques

A. Single Template Matching

In single template matching, templates from a single speaker are considered for recognition. The template of a particular speaker is taken as the reference template and the test input of the same or different speaker is compared with this reference template.

Selecting the reference templates must be done carefully, as the chosen template must match all the test templates for the word it represents with a small DTW distance, as well as matching all the other test templates for the alternative words with the largest possible DTW distance. The traditional way of preparing the reference templates is by judiciously selecting one example for each word (needed to be recognized) and considering it as a reference template for that word.

B. Speaker Dependent

For speaker-dependent recognition, the speaker trains the system to recognize his/her voice by speaking each of the words in the inventory several times. Here the reference template of a speaker is compared with the test input of the same speaker and the accuracy in the recognition is recorded.

C. Speaker Independent

A speaker independent system must consider templates from multiple speakers for recognition. As can be imagined the speaker independent task is a more difficult one because of the wide speaker variation. Here reference template of a speaker is compared with the test input of some other speaker and the accuracy of recognition is recorded.

The disadvantage of using a single reference template is that it is not robust to the speech signal variability. That is because it is almost impossible for a person to repetitively speak a word exactly in the same way. The speech signal produced would vary according to many factors. Therefore, if the template created is bad, the user would have to change it until he finds a suitable template for the tested word. It is found that the single template method works poorly across speakers. Hence to improve the performance of the recognizer average template matching technique was adopted.

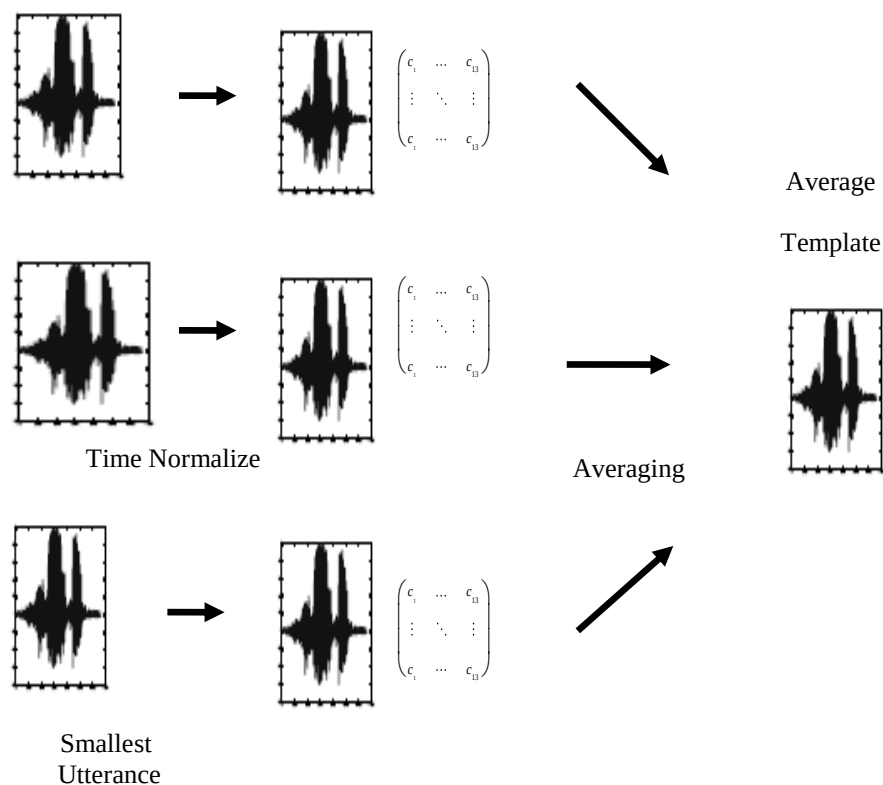
D. Average Template Matching

One of the main problems in single template matching using DTW is the preparation of reliable reference templates for the set of words to be recognized. A simple novel technique for preparing reliable reference templates to improve the recognition rate score is the average template matching technique. It extracts the reference template from a set of utterances rather than one utterance. This technique can be adapted to any DTW-based speech recognition systems to improve its performance.

To overcome this problem without incurring more computations in the recognition phase, a technique is developed to prepare more robust templates, called crosswords reference templates [5] or average reference template. Using these templates has greatly improved the recognition accuracy, as it is prepared from multiple utterances rather than just one utterance.

A few utterances (4 utterances is normally sufficient) for each word have to be prepared beforehand. Then, the shortest length of the extracted templates is calculated. Next, the template with the shortest length is chosen to be the best template. This later template is considered as the initial reference template. Then the other templates are time aligned by the DTW process such that their lengths will be equal to the chosen initial template. Finally, averaging the time-aligned templates across each frame will create the final reference template as shown in Figure 4.

Figure 4 Average Template Matching



The advantages of average template matching are that, as many utterances are taken, the variability of speech signal can be accounted for. The effect of bad utterance can be mitigated out of the good utterances. Also no additional computation cost is required at runtime in recognizing process [5].

V. Results

The performance of speech recognition systems is usually specified in terms of accuracy and speed. Accuracy may be measured in terms of recognition performance, which is usually rated with word recognition rate or Recognition accuracy given in equation (3).

$$\text{Word recognition rate} = \frac{\text{Number of correctly recognised words}}{\text{Total number of test utterances}} \quad (3)$$

The parameters chosen for the template construction process are

- The input vector length for the cepstral coefficients was set at 13.
- Sampling frequency = 16 kHz.
- Window length = 16ms (256 samples)
- Window type = Hamming

The dataset used for testing the templates' performance is the 10 English digits (0-9). Training examples are excluded from the recognition tests in all recognition tests. Number of speakers used in speaker dependent type recognition was 30 and in the speaker independent type recognition was also 30.

Four experiments have been conducted. The first two were speaker dependent cases for the two types of matching techniques and the next two were speaker independent cases for the two types of matching techniques.

A. Speaker Dependent

For speaker-dependent recognition, the speaker trains the system to recognize his/her voice by speaking each of the words in the inventory several times. Here the reference template of a speaker is compared with the test input of the same speaker and the accuracy in the recognition is recorded.

In single template matching technique, the reference template of a speaker was matched with the test input of the same speaker. Several recognition tests have been conducted to select the best template from five samples of the same speaker. In average template matching technique, four utterances for each word of a single speaker were taken and averaged to obtain the reference template and the fifth utterance of the same speaker was taken as the test input.

Table I. shows the recognition accuracies for the speaker dependent case for the two matching techniques implemented.

Table I Recognition accuracy – Speaker Dependent.

Matching Technique	Recognition Accuracy
Single Template	99.3 %
Average Template	100%

B. Speaker Independent

A speaker independent system must consider templates from multiple speakers for recognition. As can be imagined the speaker independent task is more difficult because of the wide speaker variation. In single template technique, the reference template of a speaker is compared with the test input of some other speaker. In average template technique, utterances of four different speakers were taken for each word and averaged to obtain the reference templates and the utterance of the fifth speaker was taken as the test input.

Table II. shows the recognition accuracies for the speaker independent case for the two matching techniques implemented.

Table III Recognition accuracy – Speaker Independent

Matching Technique	Recognition Accuracy
Single Template	79.3 %
Average Template	93.0%

From Tables I and II, it is observed that the speaker dependent accuracies were better than speaker independent accuracies in both the matching techniques. This is mainly because the test input is compared with the reference templates created from utterances spoken by the same speaker. The fact that humans never produce exactly the same acoustic waveform twice serves to explain the reduction in accuracy of speaker dependent recognizer using single template matching technique as compared to the average template matching technique.

In speaker independent recognition, the single template matching technique has reduced accuracy because the variability of the speaker is not accounted for in this technique. Also the fact that of inter-speaker variability, i.e. each individual has a different vocal tract and a different style of speaking explains the reduction in accuracy.

VI. Conclusion

Two template-matching techniques were implemented and their performance was evaluated. From the results presented it can be noticed that the speaker dependent recognizers had better accuracy in comparison to the speaker independent recognizers. This is obvious due to the fact that there are no speaker variations in the former. The average template matching technique had a much better word recognition rate in comparison with the single template matching technique. Also no additional computation cost is required at runtime in the recognizing process in the case of single and average template matching techniques.

References

- [1] Molau S., Pitz M., Schliitel R., and Ney H., "Computing Mel-frequency cepstral coefficients on the power spectrum", *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2001)*, vol. 1, pp. 73-76, May 2001.
- [2] Han W., Chan C. F., Choy C. S., and Pun K. P., "An efficient MFCC extraction method in speech recognition", *Proc. IEEE International Symposium on Circuits and Systems (ISCAS'2006)*, pp. 145-148, May 2006.
- [3] Do M. N., "An automatic speaker recognition system", Audio Visual Communications Laboratory, Swiss Federal Institute of Technology, Lausanne, Switzerland.
http://lcavwww.epfl.ch/~minhdo/asr_project/.
- [4] Stuart N Wrigley, "Speech recognition by dynamic time warping", Speech and Hearing Research Group, University of Sheffield, Sheffield S1 4DP, United Kingdom.
<http://www.dcs.shef.ac.uk/~stu/com326/sym.html>
- [5] Abdulla W.H., Chow D., and Sin G. "Cross-words reference template for DTW-based speech recognition systems", *TENCON 2003, Conference on Convergent Technologies for Asia-Pacific Region*, vol. 4, pp.1576 - 1579, Oct. 2003.
- [6] Vuckovic.V., "Dynamic time-warping method for isolated speech sequence recognition", *Telecommunications in Modern Satellite, Cable and Broadcasting Service, 2001. TELSIS 2001. 5th International Conference on*, vol. 1, 19-21, pp 257 – 260, Sept. 2001.
- [7] Steven W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*, California Technical Publishing, 1997-2006.