Connected Digits Recognition using Distance Calculation at each Digit

Amit Garg, Nikita and Poonam

1,2,3Assistant Professor, Department of CSE, IIET, Kinana (Jind), Haryana, India
1 amit_garg171986@yahoo.com, 2nikitasagar@gmail.com, 3madaan.punam@gmail.com

Abstract
Speech recognition is related to the detection of speech by comparing with some patterns. A key question in the speech recognition is how the speech patterns are compared to determine their similarity (i.e., the distance between patterns). The goal of speech recognition is to separate the acoustic events of interest (i.e., the speech to be processed) in a continuously recorded signal from the other parts (e.g. background). In the art of fluent or connected digits recognition, isolated digits are used as the reference patterns. The technique of Dynamic Time Warping (DTW) is the most reliable technique to identify the digits in the string of connected digits. This technique will be used by applying it at each level or frame where each frame is assumed to have only one digit out of many digits in the test string.

Keywords- Signal Processing, Pattern Processing and Dynamic Time Warping.

I. INTRODUCTION
Speech recognition begins with a user creating a speech signal to accomplish a given task. The spoken output is first recognized in that the speech signal is decoded into a series of words that are meaningful according to the syntax, semantics and pragmatics of the recognition task. A higher-level processor that uses a dynamic knowledge [1, 2] representation to modify the syntax, semantics and pragmatics according to the context of what it has previously recognized obtains the meaning of the recognized word. The feedback from the higher-level processor reduces the complexity of the recognition model by limiting the search for valid (acceptable) input sentences (speech) from the user. The recognition system responds to a user in the form of a voice output or text output displayed.

The most difficult aspect of research in speech recognition is its inter disciplinary nature. The various disciplines that will be applied to our speech recognition problem according to [2, 3].

Signal Processing – The process of extracting the relevant information from the speech signal in an efficient manner. The signal processing includes the type of spectral analysis on the basis of spectrographs which can be used to characterize the time-varying properties of speech signal.

Pattern Recognition – The set of algorithms used to cluster data to create one or more patterns of a data ensemble and to match (compare) a pair of patterns on the basis of feature enhancements of the patterns.

The pattern recognition approach is as shown in figure 1. This process involves the four steps.

Step1: Feature measurement in which a sequence of measurements is made on the input signal to define the test pattern.

Step2: Pattern training in which one or more patterns corresponding to speech sounds of the same class are used to create a pattern representative of the features of that class. The resulting pattern is called a reference pattern.

Step3: Pattern classification in which the unknown test pattern is compared with each class of reference pattern and a measure of the similarity (distance) between the test pattern and each reference pattern is computed.

Step4: Decision Logic in which the reference pattern similarity scores are used to decide which reference pattern best matches the unknown test pattern.

Figure1: Pattern recognition model for speech recognition
II. Speech Recognition of Connected Digits

A key question in the speech recognition is how the speech patterns are compared to determine their similarity. In case of discrete utterance (often called isolated words or phrase) recognition, the basic assumption was that the speech to be recognized comprised a single word or phrase and was to be recognized as a complete entity with no explicit knowledge for the phonetic content of the word or phrase. Another assumption was that each spoken utterance had a clearly defined meaning and ending point that could be found using some type of speech endpoint detector. Pattern recognition problem is solved by most popular K-nearest neighbour (KNN).

On the contrary, in the connected digits, the digit sequence is spoken in a fluent manner. Speaking a sequence of digits in a fluent manner seems to be quite natural and efficient. Figure 2 illustrates the log energy contour of isolated & connected digits recognition problem [2, 4, 5].

The connected speech recognition is different from continuous speech recognition. In connected speech recognition, the spoken input is a sequence of words or digits from the specified vocabulary. Typical examples [3] include connected digits string where the vocabulary is the set of 10 digits (0-9) or connected letters recognition. For Example spell the name, word etc. Where vocabulary is the set of 26 letters i.e. (A-Z). Continuous speech recognition, on the other hand, involves the recognition from basic units of speech e.g., phonemes, syllables demisyllables, diphones etc. and hence, implies the need for some form of the segmentation of the speech into letters and labeling of the units. It is generally acknowledged that continuous-speech recognition is a considerably more difficult problem than connected word recognition.

The main problem to resolve is that we don’t know the utterance boundaries of the different digits within the connected test string except for the beginning of the 1st digit in the test string and the end of last digit in that string. We don’t know precisely where any digit begins or ends. So, it is difficult to specify the digit boundaries because of sound coarticulation. Figure 3 illustrates the connected digits recognition problem.

III. Dynamic Time Warping Algorithm (with time alignment)

According to [4, 7, 8], the heart of the Digit recognition procedure is the unconstrained dynamic time warping (DTW) algorithm. In this technique, the test data is converted to templates. The recognition process then consists of matching the incoming speech with the stored templates. The template with the lowest distance measure from the input pattern is the recognized digit. The best
match that has lowest distance measure is based upon dynamic programming. This best match may truly represent the best time alignment. Further, time alignment in the current context is performed on a short-time basis. This is called a DTW (Dynamic Time Warping) digit recognizer.

To understand DTW, two concepts are important:-
(i) Local Distance - Difference between a feature of one signal and a feature of the other. (ii) Global Distance - The overall difference between an entire signal and another signal possibly has different length.

To obtain the global distance between two speech patterns which is represented as a sequence of vectors, a time alignment must be performed on each level of the test digits string.

Figure 4 illustrates an example of dynamic time alignment between test pattern and super reference pattern.

IV. Results & Discussion

The LB Dynamic Time Warping Algorithm was observed for various strings of length 5 and length 9.

Consider a 9 digit string ‘010010110’ spoken as a test string in which one reference pattern each of digit ‘0’ and ‘1’ was stored as reference pattern. The spectograph observed is shown in figure 5.

Figure 5: Spectrograph of the string ‘010010110’

In the above example, all 0s and 1s of the test string were spoken in almost similar mode, so the string was exactly recognized as ‘010010110’. In figure 6 is shown the DTW graph observed for the digit ‘0’ which shows that the distance using DTW is minimum at the locations in the time frame where digit ‘0’ is spoken (for the string ‘101010101’).

Figure 6: DTW graph for the digit ‘0’ in the string ‘101010101’

V. Conclusions

The conclusion is that the pattern speech recognition of the connected digits by doing the frame-by-frame computation of Dynamic Time Warping Algorithm is very much easy, accurate and less time consuming. Further, the implementation carried out is speaker-dependent (or speaker-trained) and evolutionary. The overall performance of the approach is very good because of its computation time and ease of implementation.

References

[8] www.springerlink.com
[9] www.speech.iiit.ac.in