International Journal of Information Technology and Knowledge Management
January June 2009, Volume 2, No. 1, pp. 65-68

SPEECH PERCEPTION & ANALYSIS OF FLUENT DIGITS' STRINGS USING LEVEL-BY-LEVEL TIME ALIGNMENT

Rohit Anand*, Balwinder Singh & Nidhi Sindhwani*****

The speech processing and perception of the fluent or connected digits are proposed in this paper in which a pattern recognition based speaker trained connected digits' recognizer is used in which the different isolated digits' templates are used as the synthetic reference patterns so as to compare them with the different interfering templates of the connected test string. An optimal time alignment is used at each digit (i.e., level) to spot & display the digits of the fluent test string with high degrees of success. The procedure was applied to fifteen speakers in the speaker-dependent mode resulting in the recognition accuracies of more than 95 percent for the test strings containing upto six connected digits. As the length of string increases beyond six, the error rates increase somewhat. Also, the recognition accuracy was calculated for the different length test strings containing either digit 2 or digit 8 or both (because digit 2 and digit 8 are the shortest duration digits and hence, they may be skipped in the recognition). Moreover, the average computational time for displaying the whole test string was measured for the variable lengths of the test string.

Keywords: Speech Perception, Time Alignment, Dynamic Time Warping, Connected Digits, Spectogram.

INTRODUCTION

Speech recognition has been a goal of research for almost five decades. Speech perception is concerned with the detection of speech by comparing with the reference patterns of isolated digits. The pattern-comparison based speech perception is used here because the system can adapt to the different speakers by just updating the reference patterns [1].

Fluent or connected digits recognition can be viewed as a restricted connected speech recognition task which is characterized by a relatively small and limited vocabulary size i.e., digits only [2]. Figure 1 shows the spectograph of a connected digits' string (1-4-7) spoken in a fluent manner.

The perception task is not easy if the effects of coarticulation between adjacent words are too strong [2].

There are two reasons why the problem of fluent digits' speech is so much focused. First of all, fluent string improves the input frequency of data and the recognition system operational facility. Second is that the digits' utterance boundaries within the spoken test string are not known i.e., the digits' boundaries are fuzzy or non-unique because of sound coarticulation. Numeric data input on the basis of digit-by-digit utteration would be an irritating work[1,3]. The technique of recognizing an utterance of fluent digits (i.e., connected digits) is very important for the automatic speech recognition tasks.

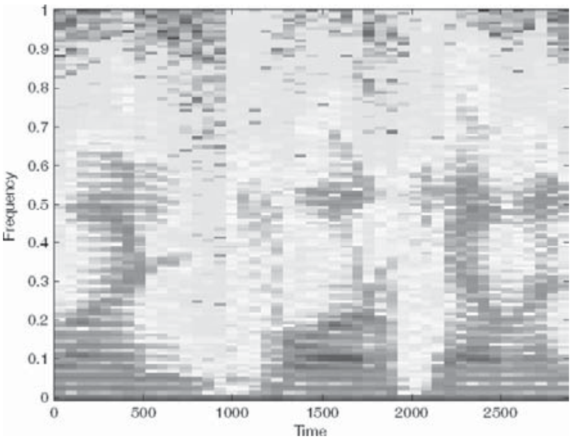

Figure 1: Illustration of the Spectograph of a Connected Digits String Spoken as '1-4-7'

For the connected digits' (or words') recognition problem, it is assumed that reference patterns are available while in continuous speech recognition, there are no fixed reference patterns and a segmentation-labeling scheme is used to provide an estimation of a spoken utterance [4].

The three disciplines are applied in this paper for the speech recognition – Signal Processing (to extract the spectrum), Pattern Recognition (to compare the spectral information of the test and reference patterns) and computer science (to implement the algorithm in software). The approach to be discussed in this paper is dynamic time-alignment for recognizing the digits accurately and the implementation to be discussed here is evolutionary [5,6]

* ECE Department, N. C. College of Engineering, Israna, Panipat (Haryana), INDIA, E-mail: *roh_anand@rediffmail.com*

** ECE Department, Guru Nanak Dev Engineering College, Ludhiana (Punjab), INDIA, E-mail: *bsdhaliwal_79@yahoo.co.in*

*** ECE Department, Amity School of Engineering & Technology, New Delhi, INDIA, E-mail: *nidhiece@indiatimes.com*

and speaker dependent. Further, the recognition task has been performed on 15 different speakers to find the recognition accuracy estimation. For the applications like dialing of telephone numbers or the credit cards or catalog ordering, the connected digits recognition is very much advantageous [4] .

TIME ALIGNMENT PROCEDURE

Speech spectograms can be compared on a short time basis. The simplest time-alignment solution is linear time-alignment but it doesn't specify the true situation for real-time utterance. The most commonly used time-alignment technique is the Dynamic Time Alignment technique which will be employed here[4].

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 is based upon dynamic programming (DP) [7]. Two concepts are important in this - *Local Distance & Global Distance*.

The distance measure between two feature vectors is calculated using the *Euclidean* distance metric [8]. So, the local distance between feature vector x of signal 1 and feature vector y of signal 2 is given by:

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

To obtain the global distance between two speech patterns (having a sequence of vectors), a time alignment must be performed. The problem is to find that sequence of reference patterns which minimizes the time alignment distance between the test pattern & the super reference pattern formed by concatenating the reference patterns [9].

The time alignment distance, D is given by

$$D = \min_{w(m)} \left[\sum_{m=1}^{M} d(m, a(m)) \right] \quad (1)$$

where $a(m)$ is the alignment function which does the mapping from $T(m)$ to R_S and $d(m, a(m))$is the local distance between frame m of test string and frame $a(m)$ of concatenated reference string.

Level-by-Level Implementation

The time alignment algorithm can be implemented in levels. The aim is to implement the speech perception of fluent digits level-by-level where each level is having one digit. For every level, the minimum accumulated distance for all the reference patterns is calculated and accordingly, the best reference that minimizes the distance is calculated.

Using frame-by-frame Dynamic Time Warping, the important point to remember is that the different frames of the test string are interfered to each other. So, the size of the test frame is taken somewhat larger than the corresponding reference digit frame (in this case, 1.5 times larger).

Further, the recognizer will be used for variable length strings operating for radio-quality AM. The reference patterns and the test patterns are recorded in the standard mono mode (rather than stereo mode) having standard 96 dB Signal to Noise (S/N) ratio.

RESULTS & ANALYSIS

After studying the spectograms of the various isolated digits spoken (i.e., 0 to 9) and storing these spectographs, the algorithm was implemented for the various connected digits' sequences of variable lengths. The implementation carried out was evolutionary and speaker trained. While uttering a particular digit in the test mode as well as reference mode, the mode of speaking was not changed so much because the different utterances (as well as spectograms) of the same digit can have different durations and the utterances (as well as spectograms) of even the same digit with the same duration can differ due to the different parts of the digit being spoken at different rates The algorithm was implemented for the various speakers for the different lengths of the string.

The Dynamic Time Alignment graphs were also observed for the various digits of the strings of different lengths. These time alignment graphs can be plotted for every digit in the test string. This graph shows the calculation of distance which is undoubtedly minimum for that digit for which the graph has been plotted. For example, in the case of a connected test string '101010101' uttered, the dynamic time alignment graphs were observed for both the digits '1' & '0'. Both these graphs are shown in figure 2 and figure 3 below:

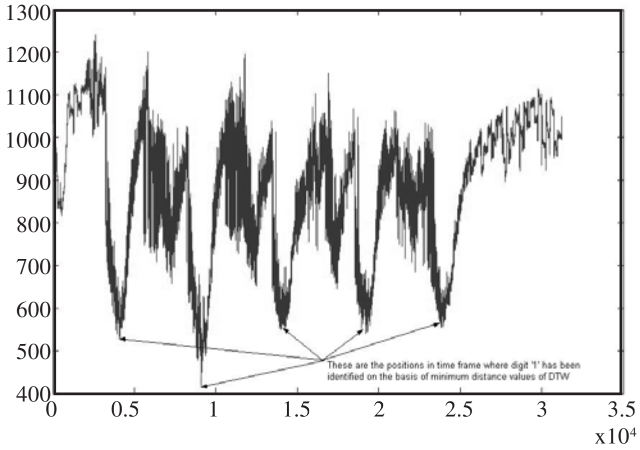

Figure 2: Alignment Graph for the Digit '1' in the String '101010101'

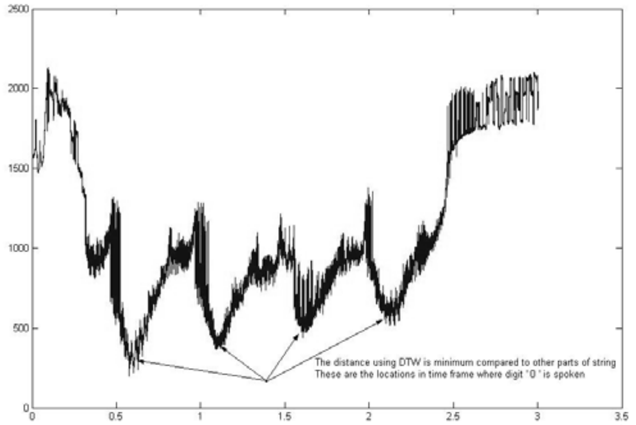

Figure 3: Alignment Graph for the Digit '1' in the String '101010101'

Figure 2 shows that the distance for the digit '1' is minimum at the locations in the time frame where digit '1' is spoken and figure 3 shows the same for the digit '0'.

The main advantage of the above algorithm is that it maintains the great accuracy in estimating the best possible matching string and it requires no heuristic rules and overhead. The procedure is suited for recognizing even the long word sequences and for real time operation. The algorithm is well suited for the security purpose (because it is speaker trained).

The computational time for displaying the whole string was measured for the different lengths of the string (from 2 digits string to 10 digits string) and on the basis of 100 experiments each on the different string lengths, the average time for displaying the whole string was calculated for each string length. The following graph (figure 4) shows the observed values for average time (for displaying the string) with respect to the length of the string.

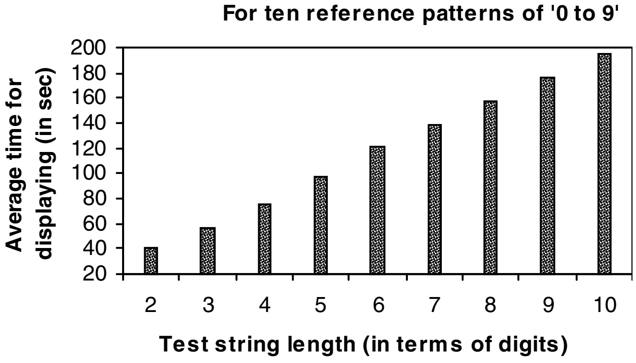

Figure 4: Average Time (of String Display) Plot with Respect to Test String Length (on 100 Experiments Basis)

On the basis of above graph, it can be seen that the computational time taken is proportional to the number of digits in the test pattern i.e., as the length of test string increases by some factor, the computational factor also increases proportionally.

On the basis of 100 experiments each by 15 different speakers on every string (two digits' string to eight digits' string), the recognition accuracies were found. All the speakers had uttered the speech independently. Firstly, the recognition accuracy of each speaker was found for a particular length string on the basis of 100 experiments. After that, the total average recognition accuracy was found corresponding to all the 15 speakers for every length of the test string. The results observed are shown in the table (1) in which L indicates the length of the test string.

Table 1
Recognition Accuracy (in %age) on the basis of 100 Experiments on Each String Length for 15 Individual Speakers and the Average Recognition Accuracy (in %age)

Speaker	*Recognition accuracy (in %age) on the basis of 100 experiments on each string length*						
	2 L	3 L	4 L	5 L	6 L	7 L	8 L
1	100	100	100	98	94	92	91
2	100	100	100	97	96	94	90
3	100	100	98	98	96	95	90
4	100	100	99	99	97	93	89
5	100	100	100	97	98	95	93
6	100	100	99	98	96	94	92
7	100	100	98	98	96	96	93
8	100	100	100	99	96	97	89
9	100	100	99	99	97	93	87
10	100	100	99	98	96	90	92
11	100	100	99	97	96	92	88
12	100	100	98	97	98	94	88
13	100	100	98	98	96	95	87
14	100	100	99	99	95	96	89
15	100	100	100	98	95	93	91
Av.R.A.(%)	100	100	99.1	98	96.1	93.9	89.9

In the above table, L indicates the no. of digits in the test string & Av. R.A.is the Average Recognition Accuracy.

From the above table, it can be seen that the error rates increase with the increase in the length of the test pattern. It can also be seen that percentage recognition accuracies are more than 95% in case of the strings containing upto 6 connected digits. It is the human nature that he rarely speaks more than five digits fluently.

Digit 2 & the digit 8 are the shortest duration digits. So, they may not be recognized sometimes. 100 experiments each by 15 different speakers were performed on the different length test strings containing the digit 2 or the digit 8 or both digits 2 and 8 & on the basis of those experiments, the recognition accuracy was estimated for those strings. The recognition accuracy (in %age) results are plotted w.r.t.

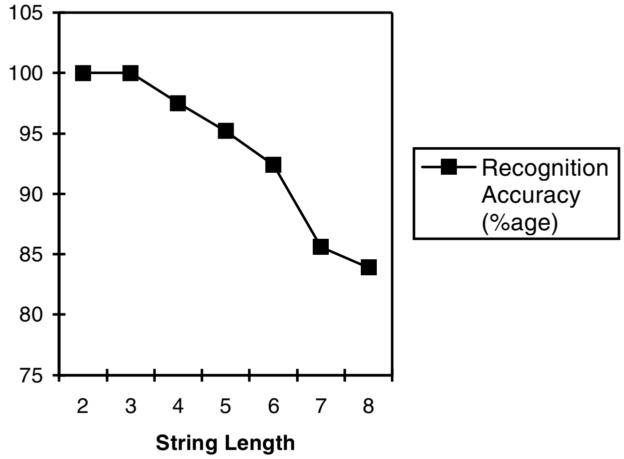

Figure 5: Recognition Accuracy (in %age) of the Test Strings (Containing Digit 2 or Digit 8 or both) w.r.t. Test String Length

length of the string (having the digit 2 or digit 8 or both) in figure 5.

One more observation is that the time taken for the computation is directly proportional to the number of reference patterns stored. So, the computational time taken is less than that in Two-Level Dynamic Programming. Algorithm used for the connected speech recognition [10].

Further, the algorithm has a large scope in future because it can be extended to solve the large vocabulary continuous speech recognition and some post-correction technique can also be used to provide a further improvement in the recognition accuracy.

CONCLUSION

A system for the speech recognition of connected digits was described in this paper. A time-alignment procedure was used which was implemented on every digit. After implementing the algorithm many times on the different length test strings for the different speakers, it is concluded that the overall performance of the approach is very good because of its less computational time, its great accuracy and ease of implementation. The implementation carried out is evolutionary [5,6] and speaker-dependent.

References

[1] Rabiner L. R. & Juang B. H., "*Fundamentals of Speech Recognition,*" Pearson Education (Singapore) Private Limited, International Publications, Delhi, (2003).

[2] Zelinski R. and Class F., "*A Segmentation Algorithm for Connected Word Recognition Based on Estimation Principles,*" *IEEE Transactions on Acoustics, Speech & Signal Processing*, **ASSP-31**, (4), (Aug. 1983), 818–827.

[3] Sakoe Hiroaki, "Two-Level DP-Matching – A Dynamic Programming Based Pattern Matching Algorithm forConnected Word Recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing,* **ASSP-27**, (6), (December 1979), 588–595.

[4] Rabiner L. R. & Levinson S. E., "Isolated & Connected Word Recognition–Theory & Selected Applications," *IEEE Trans. on Communications*, **COM-29**, (5), (May 1981), 621–659.

[5] The Mathworks Inc., "*Communications Toolbox (Ver. 2) For Use With Matlab – User's Guide,*" (July 2002).

[6] *www.mathworks.com*

[7] Rabiner L. R. and Schmidt C. E., "*Application of Dynamic Time Warping to Connected Digit Recognition,*" *IEEE Trans. on Acoustics, Speech & Signal Processing*, **ASSP-28**, (4), (August 1980), 377–388.

[8] *www.speech.iiit.ac.in*

[9] Myers C. S. and Rabiner L. R., "Connected Word Recognition Using a Level-Building Dynamic Time Warping Algorithm," *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, **6**, (April 1981), 951–955.

[10] Agbago A. and Barriere C., "Fast Two-Level-Dynamic-Programming algorithm for Speech Recognition," *IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP-04), NRC*, **5**, (May 2004), 129–132.