VOL. 10, NO. 1, JANUARY 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

# ADAPTED DTW JOINT WITH WAVELET TRANSFORM FOR ISOLATED DIGIT RECOGNITION

John Sahaya Rani Alex, Thalkari Sanket Shivkumar and Nithya Venkatesan
School of Electronics Engineering, VIT University, Chennai Campus, Chennai, India
E-Mail: jsranialex@vit.ac.in

**ABSTRACT**

Dynamic Time Warping (DTW) is a template matching approach based on dynamic programming algorithm. This paper proposes an adapted DTW algorithm for calculating the global distance matrix. Speech signals are decomposed using Discrete Wavelet Transform (DWT) into various frequency sub-bands and the resulted sub-bands of unknown; template digit utterances are compared using the adapted DTW. The performance of the proposed approach is tested with TIDIGITs data. The results indicate that there is a reduction in the order of complexity compared to DTW and increase in the performance.

**Keywords:** dynamic time warping (DTW), speech recognition, discrete wavelet transform (DWT).

## 1. INTRODUCTION

Isolated digit recognition finds its application in phone banking, Interactive Voice Response (IVR) system and much more. Isolated digit recognition system includes a front end speech signal processing block that extracts meaningful information called feature vectors from a speech signal. Widely used feature extraction methods are Mel-Frequency Cepstral Coefficients (MFCC) method, Perceptual Linear Predictive (PLP) method. These methods use based on Fourier transform. Recently, Discrete Wavelet Transform (DWT) is being used in speech compression and speech enhancement methods. DWT of a signal provides the time-frequency representation of a signal. This transform is used to analyse non-stationary signals like speech signal. Fourier transform only gives frequency information of the speech signal that is the reason DWT is chosen to analyse the speech signal. Wavelet transform defined by a basis function and a scaling function. Here, Daubechies wavelet is used as basis function. Implementation of Wavelet decomposition can be thought of as a successive filtering of low pass and high pass filtering of speech signal.

Different approaches used for implementing Automatic Speech Recognition (ASR) are Template based, Acoustic Model based. Out of all above listed one of the easiest strategy for isolated digit recognition is used in [1, 2] which suggest the use of template based method known as Dynamic Time Warping (DTW) to search for similarity between reference and test speech. But the major problem with DTW based approach is its computational complexity and estimation of threshold [3] and poor modelling of word duration [4], it means that as here actual starting point and ending point is not known of the digit to be searched and there may be silent part in between. Acoustic model based method known as Hidden Markov Model (HMM) based ASR is widely used for isolated digit recognition. The fundamental idea of this technique is to build HMM model for words or phonemes.

Because of the problems associated with the DTW moreover accuracy of HMM model it is used in most of the previous work of isolated digit recognition [5]. On the other hand HMM based isolated digit recognition suffers from number of problem like mainly while formation large amount of illustrated training data has to be designed. This annotation is alone time consuming and it requires language expertise. One more problem associated with HMM is regarding flexibility, since new word we can't add directly without training it. These problem related with the HMM, in recent years leads to use of DTW based ASR [1]. As recent advancement and reduced complexity with matlab, DTW based system is considered.

Because of these reasons, DTW which is based on template matching approach is being revised. In this algorithm first of all both test and reference speech signals are decomposed using DWT. Then sub-bands of both test and reference speech signals are compared by using DTW. If the reference is not same as test signal then the time warping program does not find path which align with the reference, result in higher Global Warping Cost (GWC). If the digit is matched the reference, the time warping function will complete the path with reasonable GWC. If GWC is less than the threshold value, then detect the unknown digit as the reference digit.

The rest of the paper is structured as follows: Section 2 briefly describes the system design which talks about DTW; DWT. Section 3 discusses experimental setup of proposed algorithm. Test and results are also discussed in section 3. Conclusion is discussed in section 4.

## 2. SYSTEM DESIGN

### A. Wavelet analysis

#### a. Statement of the problem

In this paper, the problem of recognizing a small set of prescribed word spoken is investigated. It describes a method for speaker of independent word recognition using wavelet transform (WT) coefficients. Typically, Using MFCC, 13 feature vectors per frame of 25ms are taken. So for duration of 1sec speech signal 400-500 such frames are concatenated, form one dimensional feature

vector. In Dynamic Time Warping (DTW) method, template feature vectors are matched with unknown spoken word feature vectors to identify the word spoken. The differences between the present word recognition method and the typical DTW method lies in the features selected for analysis and in the length of the wavelet coefficients. In this paper, unknown spoken utterance, template words are processed using Discrete wavelet decomposition of different levels and select only sub-band 's' of 'p' level unknown spoken utterance, sub-band 's' of 'p' level template utterance is chosen and matched using DTW method. This is shown in Figure-3. Various levels of wavelet decomposition for speech recognition are proposed.

**b. Daubechies-N wavelet family**

The different families make trade□offs between how compactly the basis functions are localized in space and how smooth they are. In this paper as speech signals are with oscillation we will use Daubechies wavelet. It is named after its inventor, the mathematician Ingrid Daubechies. Daubechies-32 wavelet is used for decomposition of speech Signal.

**c. Discrete wavelet transform**

Wavelet transform does not change the information content present in the speech. Human auditory system is linear in low frequency speech; it gives an identity to a speech. In wavelet analysis[6], we are finding out approximations and details coefficients. The approximations are the high- resolution, low-frequency part of the speech. The details coefficients are the low-resolution, high frequency parts. The DWT is defined by the following equation:

$$W(j,k) = \sum_{j}\sum_{k} x(k)2^{-j/2}\Psi(2^{-j}n-k) \qquad (1)$$

Where ψ (t) is a time function called as mother wavelet Daubechies-32. In the pyramidal algorithm the signal is analysed at different frequency bands with different resolution by decomposing the signal into a coarse approximation and detail information. The coarse approximation is then further decomposed using the same wavelet decomposition step. This is achieved by successive highpass and lowpass filtering of the time domain signal and is shown in the following Figure-1(a) and Figure-1(b).

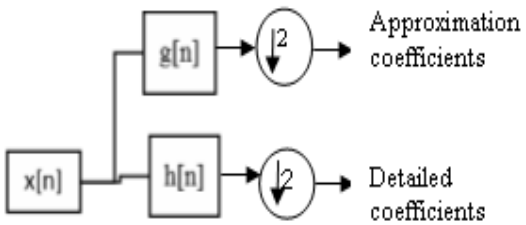


**Figure-1(a).** Discrete wavelet transform.

**B. Dynamic time warping**

Feature extraction represents a crucial step in speech recognition. DWT coefficient of test signal and database are to be computed first, then by the DTW d [.] [.], and then a [.] [.] are determined. Then from the cost matrix a [.] [.], optimal warping path element $W_k$ is calculated. Then the sum of all elements those are belonging to warping path is found out. Then based upon value of threshold it is confirmed that the input digit is detected or not[7].

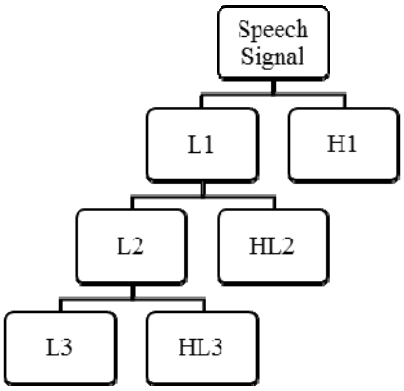


**Figure-1(b).** Decomposition tree.

**C. Hypothesis for DTW alignment**

It is well known that in speech processing, when we record the speech signal, occurrences of the same word, even if that is of same person, are not exactly same. This is the biggest challenge in speech recognition. This could be because of many factors like speaker variation, accent, and pronunciation and further may be because of noise while creating the database. This proposed system is depends on one hypothesis that the distance between the word and template are small when digit is not present.

The DTW algorithm, which is a pattern matching algorithm that can be used for isolated digit recognition, is based on alignment of template with test speech. When we want to compare feature template sequence of different words, then the sequences must be warped in dynamic manner [8, 9]. The algorithm will find out warping path between template and input speech to be tested. Here we have two sequences of feature vectors, let template x [.] with length n and test template speech y [.] with length m. The algorithm can be summarised in the following steps [1].

**a. Standard DTW Algorithm**

**Step-1:** Form the distance matrix, *d* [.] [.], with *n* rows and *m* columns, in which *x* [.] is written from left to right along the horizontal axis, and *y*[.] is written from top to bottom on the vertical axis. Each element [*i*,*j*] of *d*[.][.], is determined as:

$$d[i][j] = |x[i] - y[j]|; \text{ i=1, 2…n , j=1,2…m.} \qquad (2)$$

Where d [.] [.] is a local distance matrix.

**Step-2:** Using DTW algorithm next we will find accumulation matrix between sequences, which is determined by dynamic programming algorithm. Following criteria is used to form the accumulated distance matrix, a[*i*] [*j*] with *n* rows and *m* columns.

$$a[i][j] = d[i][j] + \min(a[i, j-1], a[i-1, j-1], a[i-1, j]) \quad (3)$$

Here a[*i*] [*j*] is the minimum distance between the sequence x [.] and y [.].

**Step-3:** Then next is to find out warping path. Warping path is the path through the minimal distance matrix from a [1, 1] to a [n, m]. Here initial condition is a (1, 1) =d (1, 1).

$$W_{i,j} = \min(a[i, j-1], a[i-1, j-1], a[i-1, j]) \quad (4)$$

**Step-4:** Then the global warp cost is to be find out, Where $W_i$ are those elements which are belonging to warping path and p is number of them.

$$GC = \sum_{i=1}^{p} W_i \quad (5)$$

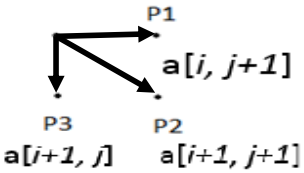


**Figure-2(a).** Standard DTW.

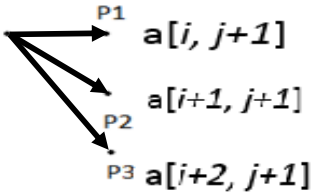


**Figure-2(b).** Modified DTW.

**b. Proposed DTW algorithm**

For the second approach that is with skipping distance following algorithm is used.

**Step-1:** Minimum edit distance is calculated.

**Step-2:** In this method we will not find out accumulation matrix rather we will directly find out the warping path by following formula

$$W_{i,j} = \min(d[i, j-1], d[i-1, j-1], d[i-2, j-1]) \quad (6)$$

In this method (*i-1, j*) is skipped

**Step-3:** Then here sum of global warping path cost as in step 4 in above algorithm is calculated

## 3. PROPOSED APPROACH

**A. The wavelet based DTW approach**

The proposed approach is completely described in following algorithm [1] and its architecture is in Figure-3. Here S-1 is sub-band 1.

**Step-1:** Record the input signal.

**Step-2:** Obtain the $j^{th}$ level DWT of both input speech signal and template speech signal which is stored in database, using family of wavelet filter Daubechies db-32. Where j=5 is preferred because it gives best representation of speech for recognition.

**Step-3:** Consider only sub-band s of each one of the transformed signals (input and templates), where in this paper sub-band used is s=1.

**Step-4:** Consider the input signal sub-band 1 and template sub-band 1 are used as feature vector for DTW approach. Next step is follow DTW algorithm

**Step-5:** After applying it the path which gives warping path cost less than threshold is the recognised speech signal.

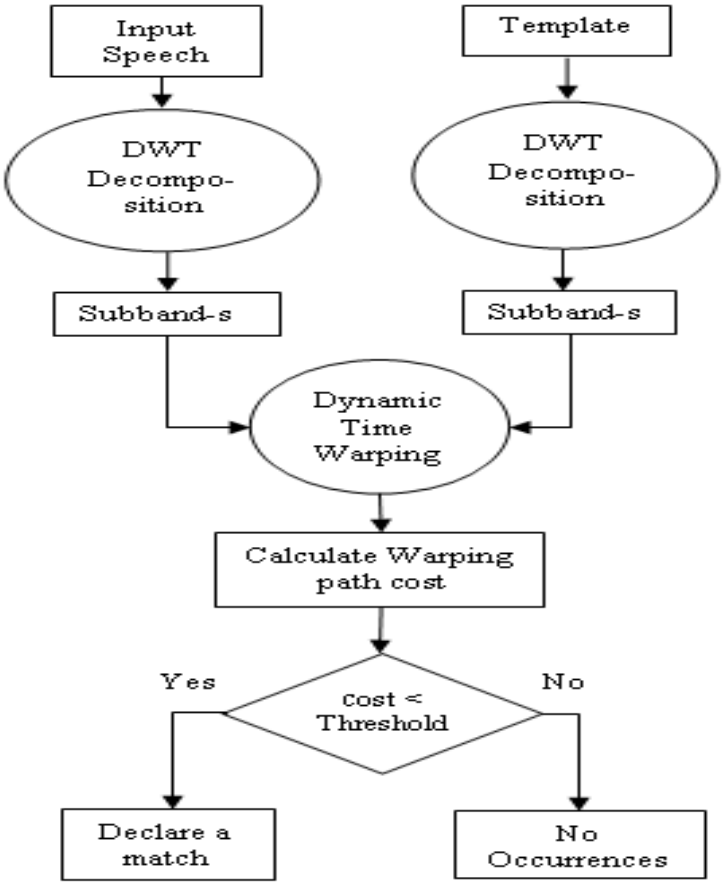


**Figure-3.** Proposed system design for isolated digit recognition.

**B. Implementation and results**

For this experiment, we have considered 10 different samples for each digit, so totally 100 speech samples are used from TIDIGIT corpus. Each digit utterances are decomposed using DWT with 5 levels. As per the proposed approach, sub-bands at the respective levels are compared using modified DTW. Table-1 shows experimental results for DTW-DWT approach implemented in this paper, first column is the number of decomposition levels used in the experiment and second column shows the different digits used in experiment for the reference and test signal. Then next three columns represents respectively warping path cost of DTW algorithm, (m x n) size of accumulation matrix and last one gives internal time of execution of program. Elapsed time is recorded to find computational complexity of DTW algorithm. It is observed that the level 5 decomposition of speech gives best results for the proposed approach.

**Table-1.** Speaker independent speech recognition.

| Decomposition level | Reference-test signal | Warping path cost | Window length | Elapsed time (seconds) |
|---|---|---|---|---|
| DTW original | 1a-1b | 1.0023e+06 | 7360-4320 | 278.92 |
| DTW original | 1a-1a | 0 | 4320-4320 | 189.49 |
| DTW original | 1b-2a | 1.8287e+06 | 7360-3360 | 243.92 |
| DTW original | 1b-4a | 1.7728e+06 | 7360-3280 | 243.23 |
| Level 1 | 1a-1b | 3.8591e+05 | 2191-3711 | 78.4 |
| Level 1 | 1a-1a | 0 | 2191-2191 | 41.44 |
| Level 1 | 1b-2a | 6.5386e+05 | 3711-1711 | 45.73 |
| Level 1 | 1b-4a | 6.6257e+05 | 3711-1671 | 55.25 |
| Level 3 | 1a-1b | 3.5409e+04 | 595-975 | 7.08 |
| Level 3 | 1a-1a | 0 | 595-595 | 4.83 |
| Level 3 | 1b-2a | 7.3067e+04 | 975-475 | 5.73 |
| Level 3 | 1b-4a | 6.2299e+04 | 975-465 | 5.92 |
| Level 5 | 1a-1b | 1.0011e+03 | 196-291 | 2.06 |
| Level 5 | 1a-1a | 0 | 196-196 | 1.77 |
| Level 5 | 1b-2a | 1.8366e+03 | 291-191 | 1.84 |
| Level 5 | 1b-4a | 1474.8 | 291-163 | 1.83 |
| Level 6 | 1a-1b | 17.7327 | 129-177 | 1.88 |
| Level 6 | 1a-1a | 0 | 96-96 | 1.72 |
| Level 6 | 1b-2a | 539.2049 | 177-114 | 1.80 |
| Level 6 | 1b-4a | 102.7533 | 177-113 | 1.72 |

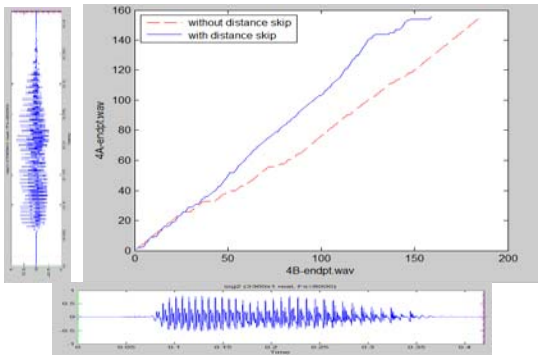


**Figure-4.** Best path, resulted from proposed method, using level j=5 with and without skipping distance.

As it can be seen from Figure-4 which shows warping path for 4A and 4B, where A and B denotes digit uttered by the same person twice. Red colour path shows without skipping distance and blue colour shows the path for with skipping distance. Table-2 shows results for speaker independent speech recognition for 5A of AL and 5A of AW in TIDGIT database.

**Table-2.** Experiment for speaker independent speech recognition.

| Digits | Warping path cost | Window length | Elapsed time |
|---|---|---|---|
| 2 level | 7.9761e+004 | 1247-1447 | 22.022883 |
| 3 level | 1.5836e+004 | 755-655 | 6.521068 |
| 4 level | 4.5882e+003 | 409-359 | 2.387714 |
| 5 level | 839.3939 | 211-236 | 1.436490 |
| 6 level | 45.4698 | 137-149 | 1.230236 |

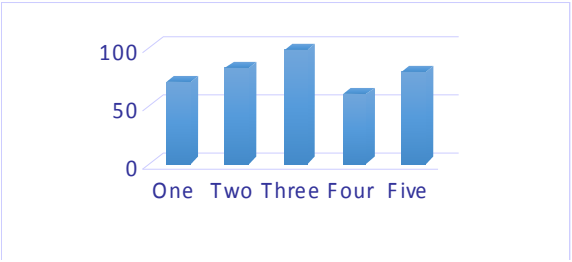


**Figure-5.** Percentage of correct detection of digits one to five.

Table-3 shows the 3 level wavelet decomposition results, for Speaker Dependent. In this experiment two cases for finding warping path are considered, first one is without skipping distances, and second is with skipping distance. By this it was observed that the time required to execute the program is reduced by 20-40%. Figure- 5 shows percentage of accuracy for digits one to five using proposed approach. For each digit, 10 different samples were used [8]. For finding the percentage of correct detection, we have compared with different speakers and the percentage of correct detection is 83%. While comparing with different digits wrong detection has observed for about 30%. Table-4 shows the percentage of accuracy of original DTW and proposed DTW. It can be observe that the accuracy is increased from 66 to 79 percentages as compared to conventional approach.

**Table-3.** Time required executing with and without skipping distances for level-5.

| Test—database | Elapsed time without skip distance | Elapsed time with skip distance |
|---|---|---|
| 1A.wav - 1A.wav | 2.260813 | 1.972174 |
| 1A.wav – 1B.wav | 2.874650 | 2.008905 |
| 1A.wav - 5A.wav | 2.320412 | 1.585857 |

**Table-4.** Performance of the proposed methods.

| | Correct decision | Wrong | Miss | % accuracy |
|---|---|---|---|---|
| DTW Original + DWT level 1 | 16 | 6 | 3 | 66 |
| Proposed DTW + DWT level 5 | 19 | 3 | 2 | 79 |

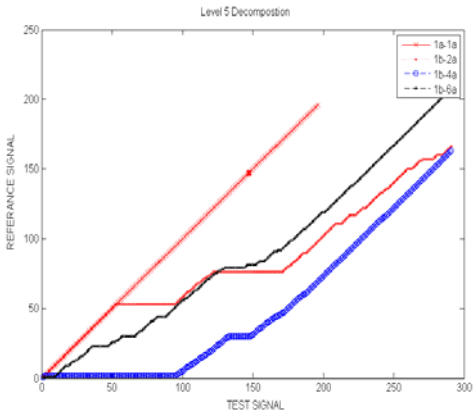


**Figure-6.** Best path, resulted from the proposed method using j=5.

As can be seen in Figure-6 in this set of test, level 5 wavelet decomposition is used to compare one input digit with stored database. Figure-7 shows decomposed waveform of seven.wav using j=1, 3, 5, 6 level wavelet decomposition.

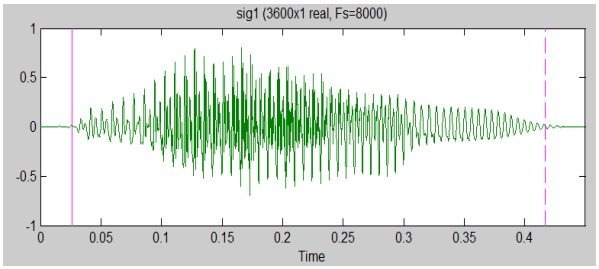


**Figure-7(a).** Original signal seven.wav.

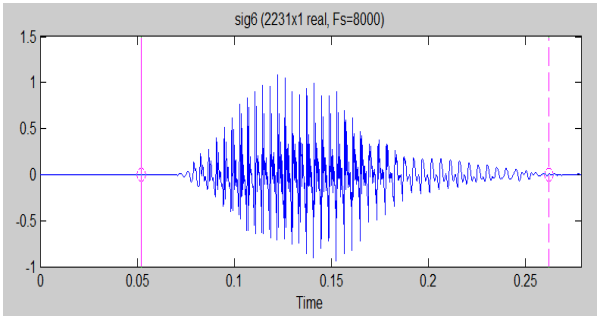


**Figure-7(b).** Approx coeff. of level 1.

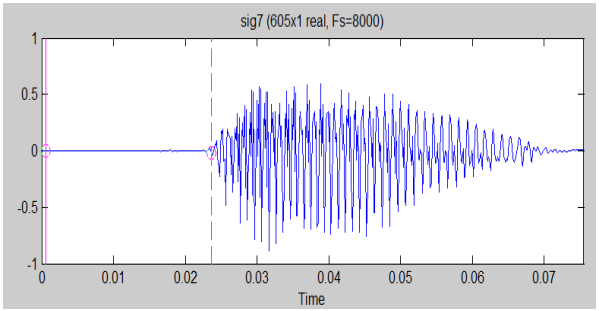


**Figure-7(c).** Approx coeff of level 3.

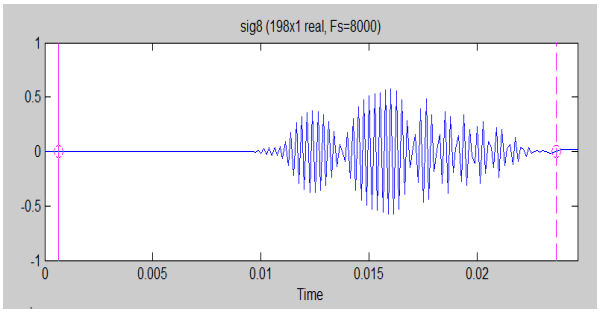


**Figure-7(d).** Approx coeff of level 5.

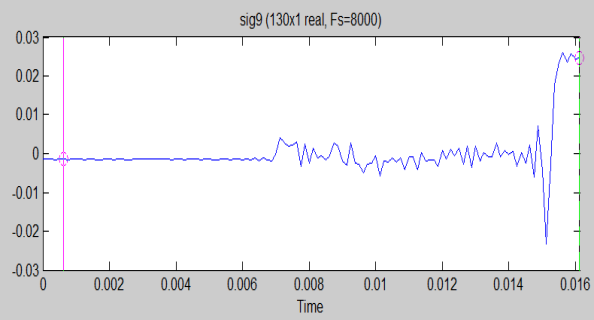


**Figure-7(e).** Approx coeff of level 6.

**Figure-7.** Various decomposing levels of spoken word seven.

## CONCLUSIONS

This paper proposed isolated digit recognition system based on adapted version of DTW along with DWT. Since the complexity of the DTW is $O(N^2)$ where N is the maximum length of the time series sequence considered. Level-5 sub-band comparison of template and unknown digit resulted in less complexity without any loss in recognition accuracy. The proposed digit recognition gives good performance in terms of recognition accuracy and recognition speed. Recognition speed means the time required to execute program and find the match.

## REFERENCES

[1] S. Barbon Jr., R.C. Guido, S.H. Chen, L.S. Vieira and F.L. Sanchez. 2007. Improved Dynamic Time Warping Based on the Discrete Wavelet Transform. 9th IEEE International Symposium on Multimedia, IEEE ISM 2007, Taiwan. 1: 256-261.

[2] R.W. Bossemeyer, J. G. Wilpon, C. H. Lee and L. R. Rabiner. 1988. Automatic speech recognition of small vocabularies within the context of unconstrained input. The Journal of the Acoustical Society of America. 84: S212.

[3] Y. Peng and F. Seide. 2005. Fast Two-Stage Vocabulary-Independent Search in Spontaneous Speech. Proc. ICASSP'05. pp. 481-484.

[4] F. Seide, Y. Peng, M. Chengyuan and E. Chang. 2004. Vocabulary independent search in spontaneous speech. Proc. ICASSP '04. 1: I-253-6.

[5] MarutiLimkar, RamaRao and VidyaSagvekar. 2012. Isolated Digit Recognition Using MFCC and DTW. International Journal on Advanced Electrical and Electronics Engineering, (IJAEEE), ISSN 2278-8948, 1(1).

[6] S.A. Mallat. 1989. A Theory for Multiresolution Signal Decomposition: The wavelet Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 11: 674-693.

[7] Titus Felix FURTUNĂ. 2008. Dynamic Programming Algorithms in Speech Recognition. Revista Informatica Economica nr. 2(46).

[8] TI digit database. http://www.ldc.upenn.edu/Catalog/CatalogEnt ry.jsp?catalogId=LDC93.