International Journal on
"Technical and Physical Problems of Engineering"
(IJTPE)
Published by International Organization of IOTPE

ISSN 2077-3528
IJTPE Journal
www.iotpe.com
ijtpe@iotpe.com

March 2013 | Issue 14 | Volume 5 | Number 1 | Pages 59-65

A NEW CONNECTED WORD RECOGNITION USING SYNERGIC HMM AND DTW

M. Mosleh [1] N. Hosseinpour [2]

1. Department of Computer Engineering, Dezful Branch, Islamic Azad University, Dezful, Iran, mosleh@iaud.ac.ir
2. Department of Information Technology, Dezful University of Medical Science, Dezful, Iran
hoseinpour@dums.ac.ir

Abstract- Connected Word Recognition (CWR) is used in many applications such as voice-dialing telephone, automatic data entry, automated banking systems and, etc. This paper presents a novel architecture for CWR based on synergic Hidden Markov Model (HMM) and Dynamic Time Warping (DTW). At first, the proposed architecture eliminates obvious silent times from inputted speech utterance by preprocessing operations. Then, in order to determine boundaries of the existing words in the compressed utterance, a set of candidates for boundary of each word is computed by using the existing capability of the HMM model. Finally, recognition operation is performed by using the synergic between HMM and DTW methods. The architecture has been compared with TLDP method from recognition accuracy and time complexity viewpoints. The evaluation results show that the proposed method significantly improves recognition accuracy and recognition time in comparison with the TLDP method.

Keywords: Connected Word Recognition (CWR), Hidden Markov Model (HMM), Dynamic Time Warping (DTW).

I. INTRODUCTION

Speech recognition has been paid attention by many researchers in recent decades. There are two types of speech recognition: Isolated Speech Recognition (ISR) and Fluent Speech Recognition (FSR). The ISR systems can recognize speech utterances, including enough silent times among their words. In these speech utterances, the beginning and ending of the words recognize easily and there is not any overlapping between neighbor words. In such systems, a single word is recognized as a complete entity with no explicit knowledge for phonetic content of the word. The second type is FSR systems. These systems can recognize speech utterances in which no obvious silent times among its words exists, necessarily. In the other words, the words boundaries are completely unknown and also neighbor words may have been overlapping.

From the view point of speech recognition, there are two kinds of fluent speech utterances. The first category contains speech utterances that are derived from a small and medium vocabulary such as digit strings, spelled letter sequences and the strings for accessing limited databases. These utterances are extracted from small and moderate size vocabularies, which are highly depended on word syntax. Recognition such utterances are called Connected Word Recognition (CWR). The basic speech recognition unit can be the word (or phrase). The second category is derived from a medium-large vocabulary. The basic speech-recognition unit in such systems cannot be the word because of having complex constraints.

In such a case, sub-word speech units, such as the phone, are necessary to implement the speech recognition system. Such these systems are known as Continues Speech Recognizing (CSR) systems [1]. Until now, several algorithms have been presented for the CWR subject. In general, the presented methods can be divided into two general classes. The first class includes Level Building (LB) [2], One Stage (OS) [3] and Two Level Dynamic Programming (TLDP) [4]. These algorithms are based on Dynamic Programming (DP) method and are able to extract the best word string with verifying all of possible regions in the speech utterance. The main disadvantage of these algorithms is high time consumption due to heavy computations. Second class of the CWR algorithms can perform recognition operations based on determining the words boundaries in the speech utterance by using a series of primary operations. Kamashki Prasad and et al by using Minimum Phase Group Delay feature segmented speech utterance, at first, and then applied Hidden Markov Models (HMM) for recognition operation [5].

Perera and et al, applying Dynamic Adaptation Threshold feature in order to divide speech utterance, as well as they utilized Artificial Neural Network (ANN) for recognition process [6]. Furthermore, Kawtrakul and Deemagarn used form MFCC and energy feature for segmentation and applied HMM models for recognition operation [7]. Although these methods are faster than first class methods, because of existing overlapping problem between neighbor words (called co articulation problem), these cannot specify words boundaries properly and as the result their recognition accuracy is low.

This paper presents a novel architecture for the CWR subject, which can perform recognition operation with proper speed and accuracy in comparison with the mentioned methods. In the beginning, the proposed method compresses the speech utterance to a set of active parts by eliminating obvious silent parts. Then, in order to determine existing words boundaries in each active part, a set of candidates for each boundary word is obtained by using existing capability of HMM model. Finally, recognition operation is performed by a synergy between HMM and Dynamic Time Warping (DTW) methods. Also, the proposed architecture can apply grammatical rules in the form of Finite State Automata (FSA). The proposed method has been compared by the TLDP method from recognition accuracy and time complexity viewpoints. The evaluation results show that the proposed method significantly improves recognition accuracy and recognition time.

II. PRELIMINARIES

A. Hidden Markov Model (HMM)

HMM is a statistical model that is used for modeling random sequences in finite state machine form [8]. Just now, the HMM models are widely applied in speech recognition domains. These models are robust against time variants of speech features and variety of speakers. An *N*-state HMM with *K* observations assigned to each state can be defined as $\lambda = \langle A,\ B,\ \prod \rangle$ where, $A = \{\ a_{ij}\ \}_{N\times N} = P(q_i \mid q_j)$ is a transition probability matrix which describes a probability from state q_i to q_j. $B = \{b(k)\}_{K\times N} = P(V_k \mid q_j)$ is an observation probability matrix which explains the probability of obtaining V_k symbol in the q_j state. $\prod = \{\pi_i\}_{1\times N}$ is an initial probability vector.

For an observation sequence $O = \{o_1,\ o_2,\ ...,\ o_T\}$ and a HMM model $\lambda = \langle A,\ B,\ \prod \rangle$, $P(O \mid \lambda)$ probability indicates a likelihood amount between the observation sequence and the HMM model. The likelihood amount is obtained with using Viterbi decoding algorithm:

- Initialization

$$\varphi_1(j) = \pi(j).b_j(1)\ ,\ j = 1,\ 2,\ ...,\ N \tag{1}$$

- Recursion

$$\varphi_{t+1}(j) = \max\left(\varphi_t(i).b_j(o_{t+1})\right) \tag{2}$$

$$t = 1,\ 2,\ ...,\ T\text{-}1\ ,\ 1 \le i \le N$$

- Termination

$$P^V = P(O \mid \lambda) = \max\left(\varphi_T(i)\right) \tag{3}$$

$$i = 1,\ 2,\ ...,\ N$$

where *N* is number of states and *T* is observation sequence length**.**

B. Dynamic Time Warping (DTW)

The DTW method is an algorithm for measuring similarity between two patterns that may vary in time. Consider two patterns P_1 and P_2 with *M* and *N* frames, respectively. Alignment measure between P_1 and P_2 patterns by using dynamic time warping technique is computed according to Equation (4):

$$D(P_1, P_2) = \min_{w(n)} \left\{ \sum_{m=1}^{M} d(P_1(m), P_2(w(m))) \right\} \tag{4}$$

where the distance $d(.,.)$ is a measure and $w(.)$ is warping function.

C. Connected Word Recognition (CWR)

As stated previously, connected word recognizing system can compute the best existing word sequence in the speech utterance which has been extracted from a small and medium vocabulary. In the following, the manner of the CWR systems operation will be presented. Assume a test utterance, *T*, and a set of word reference patterns, R_v , $1 \le v \le V$ **.**

$$T = \{t(1),\ t(2),\ ...,\ t(M)\} = \{t(m)\}_{m=1}^{M} \tag{5}$$

where $t(i)$ is corresponding to a feature vector.

$$R_v = \{r_v(1),\ r_v(2),\ ...,\ r_v(N^v)\} = \{r(m)\}_{m=1}^{N^v} \tag{6}$$

where N^i is the duration of the *i*th word reference pattern.

The CWR subject can now be considered as an optimization problem. The main purpose is to find a sequence from of *L* word reference patterns, R^*, that have best matching with test utterance T. Hence, the best sequence of reference patterns, R^*, is a concatenation of *L* reference patterns, i.e.

$$R^* = \{R_{q^*(1)} \oplus R_{q^*(2)} \oplus ... \oplus R_{q^*(L)}\} \tag{7}$$

in which each index, $q^*(i)$, is in the range [1, *V*].

In order to compute R^*, consider creating an arbitrary pattern R^s of the form

$$R^s = R_{q^s(1)} \oplus R_{q^s(2)} \oplus ... \oplus R_{q^s(L)} = \{r^s(n)\}_{n=1}^{N^s} \tag{8}$$

where N^s is the total duration of the concatenated reference pattern R^s. According to Equation (4), alignment measure between R^s and *T* can be computed

$$D(R^s, T) = \min_{w(n)} \left\{ \sum_{m=1}^{M} d\left(t(m), r^s\left(w(m)\right)\right) \right\} \tag{9}$$

In order to compute the best match, Equation (9) must be optimized over each possible value *L* and each reference word pattern, i.e.

$$D^* = \min_{R^s} \{ D(R^s, T) \} = \tag{10}$$

$$= \min_{L_{\min} \le L \le L_{\max}} \min_{1 \le q(i) \le V} \min_{w(m)} \sum_{m=1}^{M} d\left(t(m), r^s\left(w(m)\right)\right)$$

It should be mentioned that time complexity of direct calculation of Equation (10) is

$$\text{Time Complexity} = O(M \times L \times V^L) \tag{11}$$

where *M* is test pattern length, *L* is number of words in the test pattern and *V* is number of reference words.

III. THE PROPOSED METHOD

This section presents a novel architecture for connected word recognition matter that can apply any word syntax in order to improve recognition operation. Figure 1 shows a block diagram of the proposed method. The proposed method includes two parts: words boundaries analysis in the inputted speech utterance and pattern recognition operations. Identifying the words boundaries is performed in two steps: deleting evident silent times from the speech utterance and finding short pauses inter words for determining words boundaries.

The first one is quite easy and causes the inputted speech utterance is compressed. For this purpose, energy and zero crossing features together are used. In the following, feature extraction and vector quantization operations are performed and as a result observation sequence is obtained.

The second step is a little more difficult. To determine short pauses inter words, we use from existing capability of HMM models. Recall that we use from left-right HMM for modeling all of the words. As can be seen from Figure 1, by entering observation sequence to *Boundary Analysis* unit, decoding operations based on Viterbi decoding algorithm [8], are followed in different models by *Likelihood Computation* modules starting in the first state and finally ending the last state. In order to discover words boundaries, a partial joint likelihood of the last state is only needed. For the model which adapts to the input speech utterance, this likelihood will not change considerable as the word is analyzed from one observation to next observation until the end of the word. For each model, when this likelihood exceeds from a preset threshold value for twice respectively, the pair of the total likelihood along with the length of passed observations, (*S*, *L*), is resulted. Therefore, each passed observations length takes in to consideration as a candidate for the word boundary.

It should be mentioned that the length of passed observations in each model has to be at least equal to its corresponding minimum length of all training observation sequences. The second part of the proposed method refers to recognition operations. This part is able to determine the best boundary for each word in the inputted speech utterance as well as perform recognition operations well. The manner of its operation will be described in the following.

As can be observed from Figure 1, the *Decision Maker* unit that is placed after the *Boundary Analysis* unit receives V pairs (S_i, L_i), $1 \le i \le V$ and results in one of the following outputs:

- The first output is the best pair with its corresponding word index.
- The second output is K% of the best pairs with their corresponding words indexes ($K<V$).

Comparator unit is capable to compare the best likelihood amount to a preset threshold value. If the best likelihood is less than the threshold value, the word along with its boundary will be only recognized by HMM models. Then, *Activation Pattern* unit will only active the *Likelihood Computation* modules corresponding to words which are taken place in Finite State Automaton (FSA) after the recognized word. Otherwise, *Activation Pattern* unit will active all of the *Likelihood Computation* modules for subsequent processing.

Finite State Automaton (FSA): The generated language by a formal grammar is described with a finite state automaton as (S, W, Φ) such that S is the set of states; W is the set of words and Φ describes a set of transitions.

When HMM models can't determine the best word corresponding to the inputted speech utterance, *Pattern Matching* unit is applied. This unit is able to select the best pair among *K%* of resulted pairs by *Boundary Analysis* unit. This unit performs pattern matching operation based on the DTW method.

The Flowchart relevant to the operation manner of the proposed method has been shown in Figure 2. As can be seen, in the beginning, obvious silent times are removed from the inputted speech, and it converted to a set of overlapped frames. Then, the feature extraction and vector quantization operations are performed and an observation sequence is generated.

By entering the observation sequence to the *Boundary Analysis* unit, the *Likelihood Computation* units produce the V pairs (S_i, L_i), $1 \le i \le V$. After normalization the likelihood values, the word with the best likelihood value is selected and sent to *Comparator* unit by the *Decision Maker* unit.

The *Comparator* unit compares the best received likelihood value with a preset threshold corresponding with the word. In order to explain execution process, a conditional variable *Tag* has been used. By considering the value of the conditional variable *Tag* and the output of the *Comparator* unit, one of following cases will be happen:

The first case occurs when the conditional variable *Tag* is zero, and the best likelihood value is less than the preset threshold. In this case, the word which is detected just by using the HMM models, is displayed as well as the proper activation pattern is determined.

The second case occurs when the conditional variable *Tag is* zero and the best likelihood value is more than the preset threshold. In this case, there is no sufficient matching between the current selected words and the observation sequence. Therefore, the observation sequence must be examined by all reference words again. For this purpose, at first, the value of conditional variable *Tag* becomes one and then activation pattern is entirely one.

The third case occurs when the *Tag* variable is one, regardless of the output of the *Comparator* unit. In such a case, the HMM method cannot recognize the word by itself. Here, *K* candidate boundaries for the word will be sent to *Pattern Matching* unit for reexamination. Finally, the best word is selected by synergy HMM and DTW methods and also the proper activation pattern is extracted for next times.

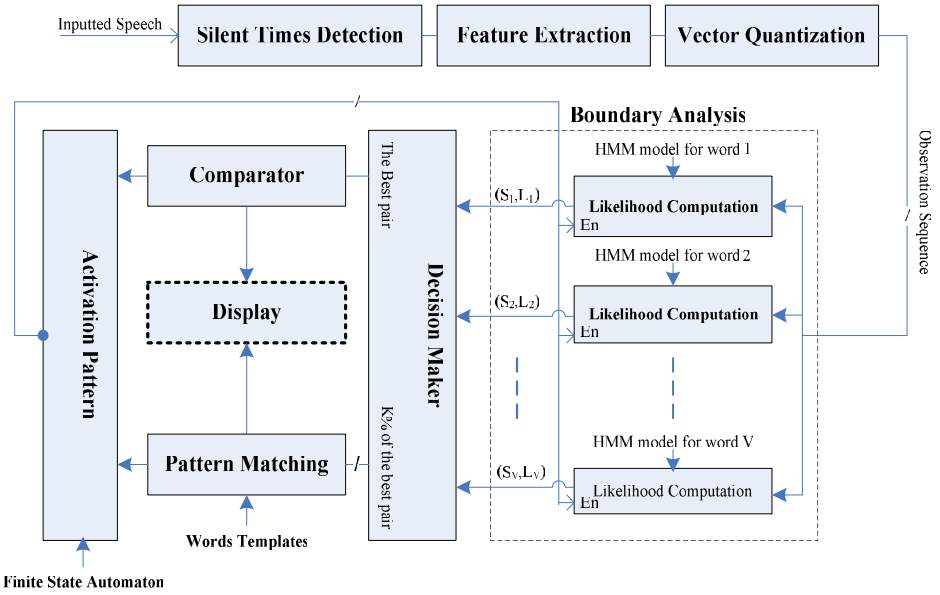


Figure 1. Block diagram of the proposed CWR based on synergic HMM and DTW methods

IV. EXPERIMENTAL RESULTS

In order to evaluate the proposed method, the FSA related to airline booking has been used [9]. This FSA contains 24 states, 30 transmissions and 28 words. It has the ability to produce 18 different strings with various lengths. Training data set includes 2240 speech utterances that have been uttered by 50 different speakers.

In the beginning, the inputted speech utterances were divided into 16 ms frames with 8 ms overlapping, and then evident silent times were removed by using energy and zero crossing rate features. 12MFCC coefficients with their energy and dynamics features have been extracted as features. Training samples have been used for creating left-right HMM models and templates. In order to train discrete HMM models, forward-backward estimation method Baum-Welch has been applied.

Since the number of states in the HMM models have an important role in performance of the proposed method, so to achieve higher performance, Self Adaptive HMM has been used [10]. According to this technique, a model matches its states with correct states automatically. Such technique is based on the principle that a model with correct states has less entropy rather than a model with incorrect ones.

For testing the proposed system, two classes of connected word-utterances have been used. First class contains the utterances which have been uttered in the FSA format. These utterances have different lengths between four to eight words (called 4-FSA to 8-FSA). The second class contains speech utterances which don't have certain format. In the other word, these utterances necessarily have not been uttered in the FSA format and can be random combination of the vocabulary words. Error rate is calculated by using following Equation [11]:

$$ERR(\%) = \frac{\#\text{Sub} + \#\text{Del} + \#\text{Ins}}{\#\text{words in the speech utterance}} \times 100 \qquad (12)$$

where,

Sub: an incorrect word was substituted for the correct word

Del: a correct word was omitted in the recognized sentence

Ins: an extra word was added in the recognized sentence

As was mentioned, the most problem of the CWR systems is inability for proper determining of words boundaries in the inputted speech utterance. Although, until now, many methods have been presented for the CWR subject, but none of them don't have recognition accuracy of the TLDP method. This method is a well-known CWR algorithm that assigns a proper word string to inputted speech utterance. The most important its deficiency is high time complexity, which caused it is not proper for real time applications. A. Agbago and C. Barrier applied a modified TDLP in the CSR systems for recognizing phoneme sequences [12]. Furthermore, Y. Kim and H. Jeong presented hardware solution for real time execution of the TLDP method on Field Programmable Gate Array (FPGA) [13]. Therefore, the proposed method has been compared with this algorithm. The obtained results from comparison have been shown in Tables 1, 2, respectively.

As can be seen from the Table1, 2, the average of error rate in the TLDP method and the proposed method are 27.37% and 12.71% in the formal format utterances as well as 28.11% and 26.08% in the informal format utterances, respectively. Therefore, recognition accuracy of the proposed method in the formal format utterances is more than twice recognition accuracy of the TLDP method while in the informal format utterances is close. The percentage of error rate for the substitution, deletion and insertion errors in the TLDP method and the proposed method have been shown in the Figure 3.

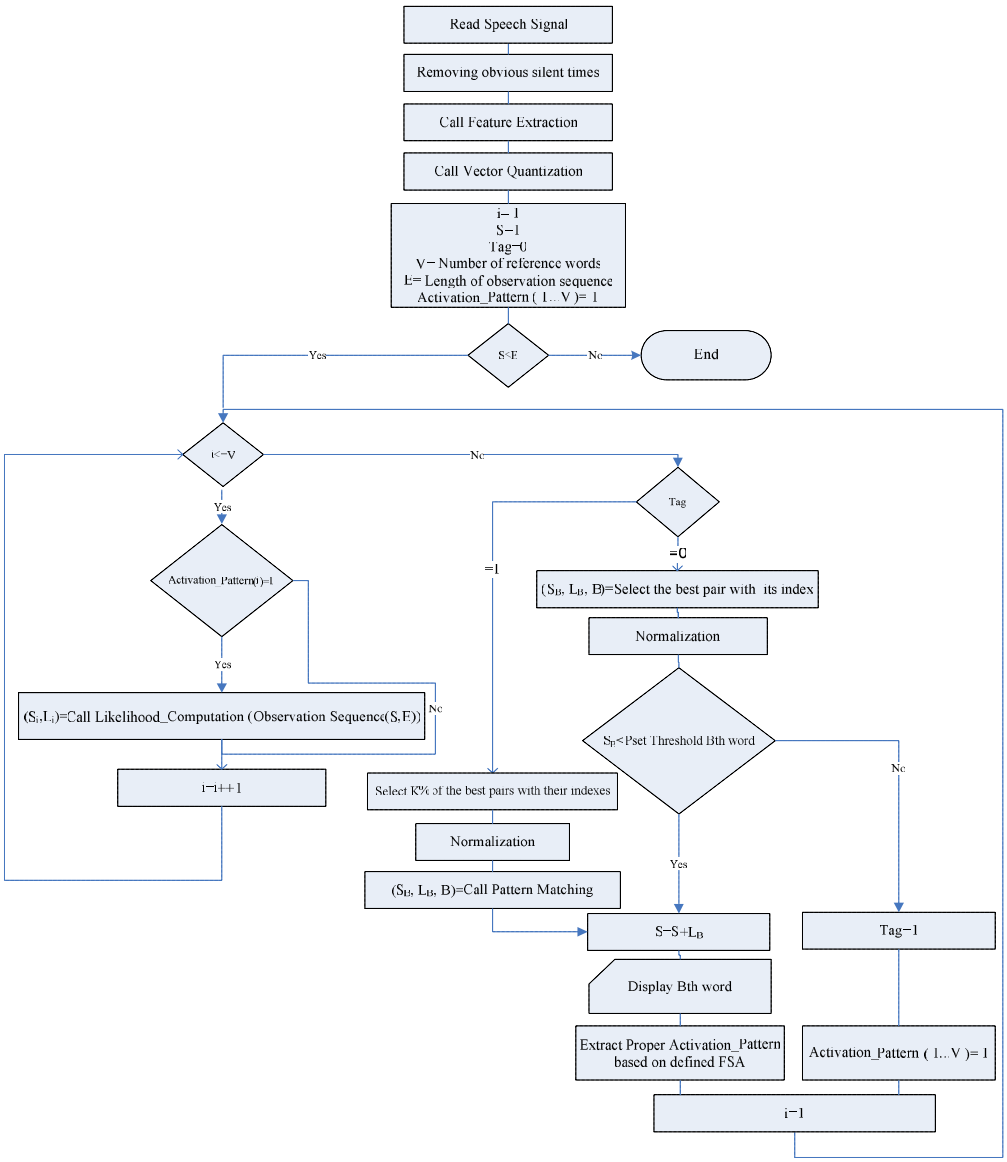


Figure 2. Flowchart the manner of function of the proposed method

Table 1. The obtained results of evaluating the proposed method with the TLDP method for formal connected word-utterances

		Hypothesis			
		5-FSA	6-FSA	7-FSA	8-FSA
No. of utterances		64	31	67	45
No. of Substitution	TLDP method	47	25	77	64
	The proposed method	29	14	43	28
No. of Deletion	TLDP method	21	16	23	22
	The proposed method	5	4	7	10
No. of Insertion	TLDP method	15	13	20	16
	The proposed method	6	7	7	8
The TLDP error rate (%)		25.93	29.03	26.22	28.33
The proposed method error rate (%)		12.50	13.44	12.15	12.77

Table 2. The obtained results of evaluating the proposed method with the TLDP method for informal connected word-utterances

No. of utterances		150
No. of words		690
No. of substitution	TLDP method	125
	The proposed method	129
No. of Deletion	TLDP method	41
	The proposed method	17
No. of Insertion	TLDP method	28
	The proposed method	34
The TLDP error rate (%)		28.11
The proposed method error rate (%)		26.08

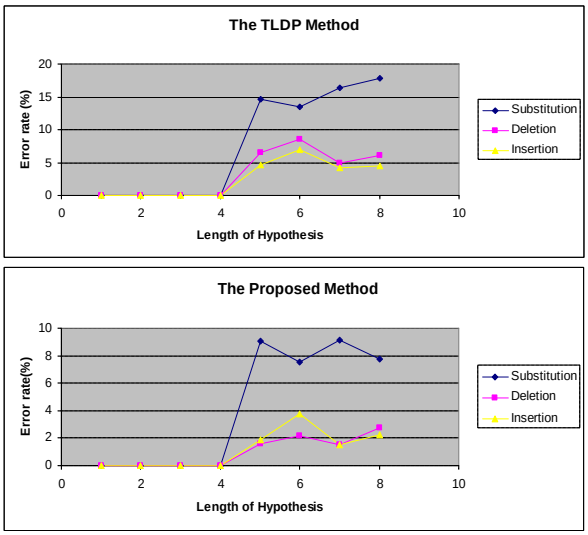
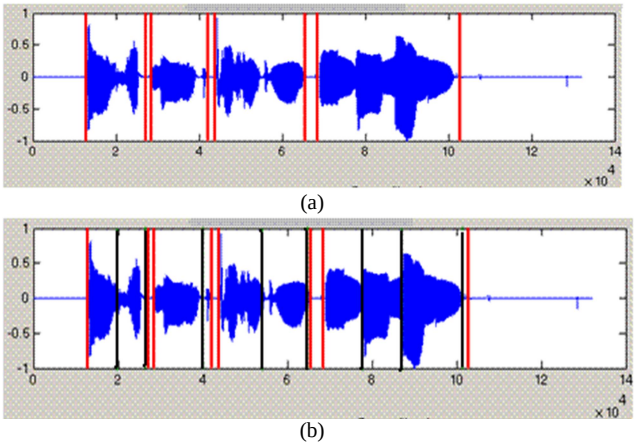


Figure 3. The percentage of error rate for the substitution, deletion and insertion errors in the TLDP method and the proposed architecture

According to Equation (12), the total error rate is related to the substitution, deletion and insertion errors. The substitution error refers to part of the system error that is the result of incorrect pattern matching operations while the insertion and deletion errors are assigned to system inability to determine proper boundary of the words in the inputted speech utterance.

As can be seen from Figure 3, the substitution error rate in the proposed method is much more than the sum of the deletion and insertion errors. So, it can be concluded that the proposed method can overcome to co-articulation problem between adjacent words is able to determine words boundaries in the inputted speech utterance properly. Figure 4 shows deleting of evident silent times and specifying the words boundaries in the inputted speech utterance using the proposed method.



(a)

(b)

Figure 4. (a) Deleting of evident silent times by pre-processing operations (b) Determining of words boundaries by the synergic the HMM and DTW methods

The Time complexity of the TLDP method equals to $O\left(M \times V \times \overline{DTW}\right)$ where M is the inputted speech utterance length, V is number of reference patterns and $\overline{DTW}$ is average cost of the DTW method. Therefore, the proposed method is associated with two important characteristics which can result in significant decreasing of time complexity:

1. Compressing the inputted speech utterance length by using deleting silent times.
2. Finding words boundaries in the inputted speech utterance by evaluating few regions instead of all possible regions by using applying the capability of the HMM method.

Therefore, such capabilities caused numerous decreasing of the DTW calls and so the proposed architecture is able to do the CWR with very appropriate speed in comparison with the TLDP method.

V. CONCLUSIONS

In this paper, a new architecture for recognizing connected words has been presented. At the beginning, the proposed method compresses the input string with eliminating the obvious silent times through preprocessing operations. Then the proposed method extract the best sequence of words by synergy between the HMM and the DTW methods. Also, this proposed system can apply grammatical rules in the form of the FSA. The proposed method has been compared with the TLDP method from views of recognition accuracy and time complexity. The evaluation results show that the proposed method has better recognition accuracy and very lower time complexity than the TLDP method.

REFERENCES

[1] L. Rabiner, B. Juang, "Fundamentals of Speech Recognition", Prentice Hall, 1993.

[2] C. Myers, L. Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition", IEEE Transactions on Acoustics Speech and Signal Processing, Vol. 29, No. 2, pp. 284-297, April 1981.

[3] J.S. Bridle, M.D. Brown, R.D. Chamberlain, "Continuous Connected Word Recognition Using Whole Word Templates", Radio and Electronic Engineer, Vol. 53, No.4, pp. 167-175, April 1983.

[4] H. Sakoe, "Two-Level DP-Matching - A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 27, No. 6 pp. 588-595, December 1979.

[5] V. Kamakshi Prasad, T. Nagarajan, H.A. Murthy, "Automatic Segmentation of Continuous Speech Using Minimum Phase Group Delay Functions", Speech Communication, Issue 3-4, Vol. 42, pp. 429-446, April 2004.

[6] K.A.D. Perera, R.A.D.S. Ranathunga, I.P. Welivitigoda, R.M. Withanawasam, "Connected Speech Recognition with an Isolated Word Recognizer", Proc. International Conference on Information and Automation, pp. 319-323, Colombo, Sri Lanka, December 2005.

[7] A. Deemagarn , A. Kawtrakul, "Thai Connected Digit Speech Recognition Using Hidden Markov Models", Proc. International Conference on Speech and Computer

(SPECOM), pp. 731-735, Saint-Petersburg, Russia, September 2004.
[8] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proc. IEEE, Vol. 77, No. 2, pp. 257-286, February 1989.
[9] F. Owens, "Signal Processing of Speech", Macmillan Publishing Company, 1993.
[10] J. Li, J. Wang, Y. Zhao, Z. Yang, "Self-Adaptive Design of Hidden Markov Models", Pattern Recognition Letters, Issue 2, Vol. 25, pp. 197-210, January 2004.
[11] X. Huang, A. Acero, H. Hon, "Spoken Language Processing", Prentice Hall PTR, 2001.
[12] A. Agbago, C. Barrier, "Fast Two-Level Dynamic Programming Algorithm for Speech Recognition", Proc. ICASSP, Vol. 5, 2004.
[13] Y. Kim, H. Jeong, "A Systolic FPGA Architecture of Two-level Dynamic Programming for Connected Speech Recognition", IEICE Transactions on Information and Systems, Vol. E90-D, No. 2, pp. 562, February 2007.

BIOGRAPHIES

Mohammad Mosleh received the B.Sc. degree in Computer Hardware Engineering from Dezful Branch, Islamic Azad University, Dezful, Iran in 2003, and the M.Sc. Ph.D. degrees in Architecture of Computer Systems from Science and Research Branch, Islamic Azad University, Tehran, Iran, in 2006 and 2010, respectively. He is an Assistant Professor in the Department of Computer Engineering at Dezful Branch, Islamic Azad University. His main research interests are in the areas of speech processing, machine learning, intelligent systems and audio watermarking.

Najmeh Hosseinpour received the B.Sc. degree in Electronics Engineering from Dezful Branch, Islamic Azad University, Dezful, Iran in 2009 and the M.Sc. degree in Architecture of Computer Systems from the same university in 2012. Her main research activities cover the areas of intelligent systems, speech recognition and speaker recognition.