AU J.T. 8(3): 113-123 (Jan. 2005)

A Framework for Connected Speech Recognition for Thai Language*

Pratit Santiprabhob, Jirawat Chaiareerat, Ronnarit Cheirsilp, Nunmanus Dachapratumvan and Wanna Supasirirojana

Faculty of Science and Technology, Assumption University
Bangkok, Thailand

Abstract

Connected speech recognition problem for Thai language, like similar problem in other languages, involves three sub-problems: (i) syllable segmentation, (ii) syllable recognition, and (iii) syllable-based word recognition. This paper presents a framework upon which a speech recognition system can be built. The approach taken in our framework differs from a so-called word-based approach in which whole words are trained to be later recognized. Our approach attempts to recognize syllables based on their constituent phonemes; the recognized syllables are then grouped into words within a given context of discourse. The four constituent phonemes of Thai syllables are leading consonant, vowel, ending consonant and tones. The proposed framework utilizes several soft computing techniques in different parts. As for the signal-processing portion of the framework, Fuzzy System (FS) is used in the syllable segmentation part while the Neural Network (NN) and Hidden Markov Model (HMM) are used in the syllable recognition part. On the other hand, Genetic Algorithm (GA) and rule-based system techniques are used to develop alternative methods to recognizing words from given set of syllables

***Keywords**: Hidden Markov Model, neural network, fuzzy system, genetic algorithm, rule-based system.*

1. Introduction

Speech is a primary means of human communications. It is the most natural way for humans to convey ideas, to exchange information, to give instruction, etc. A speech is an intelligible group of words. Thus, the foundation for the understanding of human speech is the understanding of spoken words, which in turn requires the recognition of spoken words to first be achieved. Our proposed framework outlines methods that can be used to solve this spoken words (or speech) recognition problem. This is indeed an exciting yet challenging research area. Speech is seen as the way humans will interact with computers in the future. In general, humans can speak about two times faster than a proficient typist can type. In addition, this mode of man-machine interaction allows for hand-free operation such as giving on-board computer an instruction while driving a car.

Techniques for recognizing words as trained are widely commercially available. These words are not connected, individual words that can be encoded as templates. On the other hand, recognizing connected speech is a totally different problem with a magnitude of difficulty. Our proposed framework is conceptually depicted in Fig. 1. In the first step, the given speech is segmented into syllables. Then, in the second step, each syllable is attempted a recognition from its constituent phonemes. Eventually, in the third step, the recognized syllables are decoded into words within a given context of discourse.

Various researchers have developed different alternatives to the problem of Thai speech recognition. Different techniques are used such as Dynamic Time Wrapping (Penisiri and Jitapunkul 1995), Conventional Neural Network (Porsukchandra and Jitapunkul

* *This research is supported in part by the Thailand Research Fund.*

19960), Modified Back Propagation Neural Network (Maneenoi *et al.* 1997), Neural Network with Fuzzy MF Preprocessor (Wutiwiwatchai 1997) and Hidden Markov Model (Ahkuputra *et al.* 1997). From the studies in (Ahkuputra *et al.* 1998 and (Jitapunkul *et al.* 1998), the Hidden Markov Model (HMM) as used in Ahkuputra *et al.* (1997) is identified as the technique that yields the best recognition rate.

However, there are a number of limitations observed with regard to the research works cited.

(i) All of the research works utilizes the word-based speech recognition approach. The whole words are trained/encoded. Hence, the approaches can recognize only a small set of vocabularies such as numbers, names and commands.

(ii) The approach outlined in (Ahkuputra *et al.* 1997) is not readily applicable to the connected speech recognition problem in general since the number of syllables has to be determined before the recognition can be undertaken.

(iii) In all of the approaches, computational requirement grows in proportion to the number of vocabularies they are trained/encoded to recognize.

In order to overcome the limitations discussed above, our proposed framework attempts to recognize connected speech in terms of syllabic units. This requires that words in a given connected speech be segmented into syllables before recognition can be achieved.

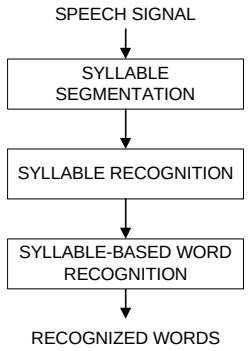

Fig. 1. Conceptual Framework for Connected Speech Recognition

Framework Architecture

The proposed framework consists of three parts: syllable segmentation, syllable recognition and syllable-based word recognition.

Syllable Segmentation

The segmentation algorithm used in our framework is based on the concepts of energy and Different Ceptral as explained in (Chaiareerat and Santiprabhob 2002. The segmentation algorithm consists of three steps: parameters computation, threshold based segmentation and fuzzy based segmentation. First, the speech signal is pre-processed to enhance the signal quality. Then, necessary parameters are calculated. These parameters are used in segmenting the speech signal. Finally, a fuzzy inference system is used to identify the ending point and starting point of each syllable in each resulting segment.

Parameters Computation: First, the speech signal is pre-processed by means of signal pre-emphasizing technique as described in Rabiner and Juang 1993). The signal is then en-framed into 30 milli-seconds long frames with 20 milli-second overlapping factor between frames. For each frame, fours parameters are computed: High Amplitude Rate (HAR), Absolute Energy, Zero Crossing Rate (ZCR), and Different Ceptral (DC). Detailed descriptions of these parameters can be found in Chaiareerat and Santiprbhob (2002). A graph representing each of the four parameters is respectively constructed. Finally, the contours of each graph are then smoothed according to the Moving Average Smoothing algorithm (Jitiwarangkul *et al*. 1998).

Threshold based Segmentation: In this step, the threshold-based segmentation algorithm eliminates the silent portions of a given speech using a set of threshold values calculated from the beginning part of the speech. Here, the original speech signal is segmented into groups of syllables called speech segments.

The algorithm works as follows. The speech signal is searched from the first frame to find the pairs of starting frames and ending frames. The following rules are then applied to

determine whether a frame *I* is starting frame or ending frame or neither.

If Energy[i]>E_th1 or HAR[i]>HAR_th1 then frame I is staring frame.
If Energy[i]<E_th2 or ZCR[i]=0 or HAR[i]<HAR_th2 then frame I is ending frame.

Where:
Energy[i] is the ABS Energy at frame i
HAR[i] is the HAR at frame i
ZCR is the ZCR at frame i
E_th1 and E_th2 are Energy thresholds calculated from the background noise at the beginning of the speech signal.
HAR_th1 and HAR_th2 are HAR thresholds calculated from the background noise at the beginning of the speech signal.

The algorithm is depicted in Fig. 2. The results obtained in this step are the speech segments, which will further be segmented into syllables in next step.

Fuzzy based Segmentation: Each speech segments resulted from the threshold-based segmentation algorithm is once again segmented in this step. The ultimate results are syllables to be recognized. There are four steps in segmentation. First, local peak energy frames, so-called PeakE frames are identified in each speech segment. Then, local minimum energy frames between two PeakE Frames, so-called Emin frames are also identified. The identification rules for these frames are given in (Chaiareerat and Santiprbhob (2002).

For each Emin frame, five fuzzy input variables are defined, namely, (i) the absolute energy of the current Emin frame – EM, (ii) the minimum ZCR between the two surrounding PeakE frames, (iii) the difference between the EM and the absolute energy of the preceding PeakE frame – DEL, (iv) the difference between the EM and the absolute energy of the following PeakE frame – DER, and (v) the maximum DC between the two surrounding PeakE frames – DCMAX. Finally, a Fuzzy Inference System (FIS) is constructed to determine the frame whether it is a boundary frame or not based on these five input variables. Fuzzy terms for each of the parameters and the fuzzy rules are defined in (Chaiareerat and Santiprbhob (2002). Here, Mamdani-type FIS with centroid defuzzification method (MathWorks 1999). is employed.

After the boundary frames are located, the center speech signal sample of each frame is used as a boundary point to demarcate the boundary between syllables.

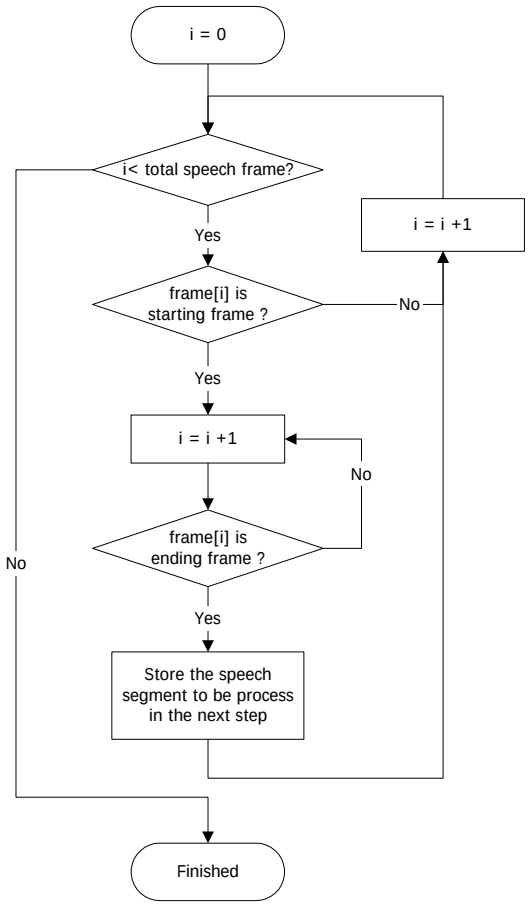

Fig. 2: The algorithm to detect the starting and ending frames.

2.2 Syllable Recognition

Each Thai syllable sound comprises four different types of phoneme, namely leading consonant, vowel, ending consonant, and tone. In order to recognize a Thai syllable, all these four constituents of that syllable must be recognized.

The proposed syllable recognition system comprises five processes namely: (i) leading consonant, vowel, ending consonant and tone (LVET) feature extraction process, (ii) leading consonant recognition process (LRP), (iii) vowel recognition process (VRP), (iv) ending consonant recognition process (ERP), and (v) tone recognition process (TRP). A block diagram of the overall recognition system that is described in detail in Cheirsilp and Santiprabhob 2002) is given in Fig. 3.

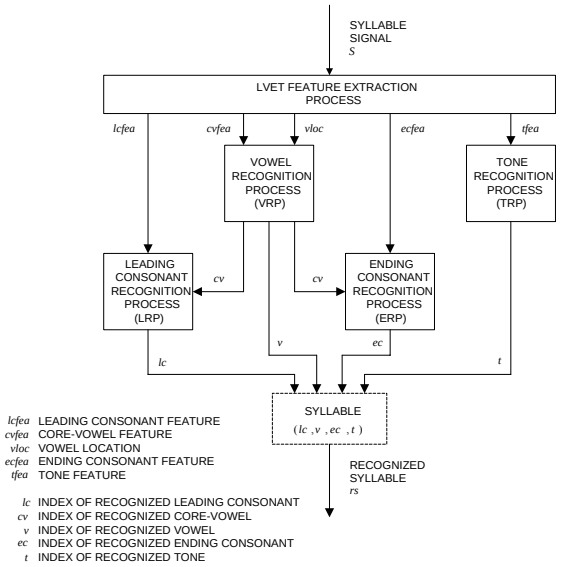

Fig. 3. A block diagram of the syllable recognition system

2.2.1 LVET Feature Extraction Process: This process is responsible for extracting all the features needed from each segmented syllable signal for the four following recognition processes, i.e. LRP, VRP, ERP, and TRP.

The Linear Predictor Coefficient (LPC) analysis as defined in Rabiner and Juang 1993) and (Rabiner 1989) is conducted to determine the Ceptral Coefficient and energy feature vector, so-called CEP_E feature vector. In addition, fundamental frequency contour Rowdern 2003) is extracted from each segmented syllable signal.

Then, vowel location is detected based on the differences between CEP_E feature vectors of the frames of that syllable. Ceptral and energy thresholds are used to determine the beginning and ending frames of the vowel part. Details of this vowel location detection algorithm is given in Cheirsilp and Santiprabhob (2002).

Subsequently, the CEP_E feature vector is segmented into three feature vectors, *lcfea*, *cvfea* and *ecfea*, according to the vowel location. These three feature vectors are to be used as the inputs of LRP, VRP, and ERP, respectively. A Block diagram of this particular process, so-called LVE feature segmentation, is given in Fig. 4.

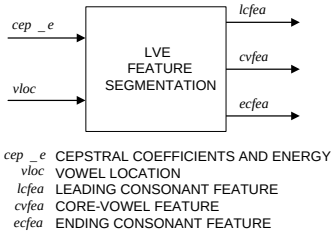

Fig. 4. A block diagram of LVE feature segmentation

On the other hand, the fundamental frequency contour is used to construct a tone feature vector, *tfea,* which becomes an input into TRP.

2.2.2 Tone Recognition Process (TRP): In this process, the tone phoneme is recognized. A neural network is employed as the recognition engine. A block diagram of this process is given in Fig. 5. The tone feature vector, *tfea* from the LVET feature extraction process is processed. As a result of the recognition for each syllable, an index of a recognized tone *t* is returned.

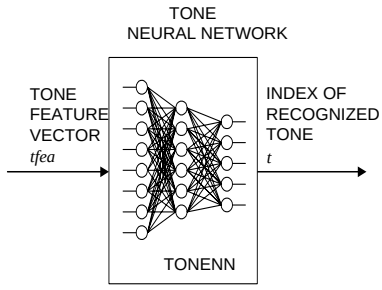

Fig. 5: A block diagram of TRP

2.2.3 Vowel Recognition Process (VRP): In this process, vowel phonemes are recognized. In order to recognize vowels two elements must be determined. They are type of vowel and vowel length. There are 12 different types of vowel, so-called core vowels and two vowel lengths, short and long, in Thai language. A block diagram of this VRP is given in Fig. 6. The process is divided into the core vowel recognition part and the vowel length determination part, which are further discussed below.

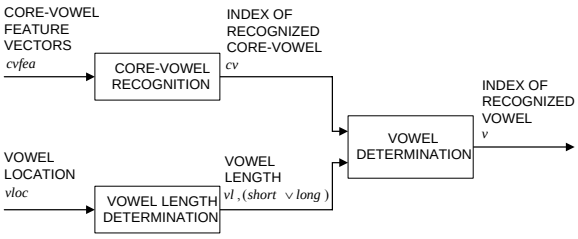

Fig. 6. A block diagram of VRP

Core-vowel recognition: A Hidden Markov Model (HMM) is used to represent each core-vowel class. Hence, 12 HMMs are included. A block diagram of core-vowel recognition is given in Figure 7. The type of HMM used in this process is Continuous Density Hidden Markov Model (CDHMM) whose details are described in Rabiner and Juang (1993).

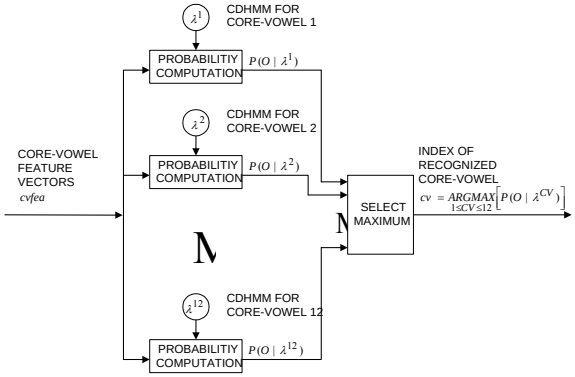

Fig. 7. A block diagram of core-vowel recognition

Vowel length determination: The vowel location has been identified with the frame numbers of the starting and the ending points of the vowel with respect to each segmented syllable signal. The vowel length can easily computed in terms of number of frames from these starting and ending points. A simple threshold method is then used to determine whether the vowel is short or long. If the vowel length exceeds the threshold, it is long vowel. Otherwise, it is short vowel.

2.2.4 Leading Consonant Recognition Process (LRP): Here, leading consonant feature vectors, *lcfea* from LVET feature extraction process and the index of recognized core-vowel from VRP are processed. As a result of this LRP process, an index of recognized leading consonant is returned. This means that the recognition of leading consonant depends on the recognized core-vowel type from the VRP. A block diagram of LRP is given in Fig. 8.

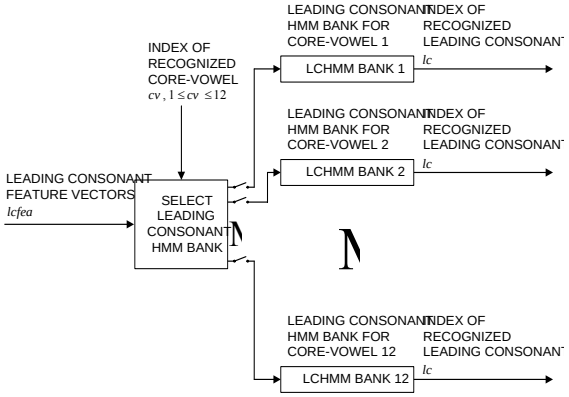

Fig. 8. A block diagram of LRP

For each core-vowel type, there is an LCHMM bank. Each LCHMM bank is designed to cover all possible 38 leading consonant classes of Thai language. Each LCHMM consists of 38 HMMs. This means that for each leading consonant class, there is a HMM corresponding to it. Each HMM in LCHMM bank is also a CDHMM. Fig. 9 shows a block diagram of an LCHMM bank.

2.2.5 Ending Consonant Recognition Process (ERP): In this process, the ending consonant feature vectors, *ecfea* and the index of recognized core-vowel from VRP are similarly processed. An index of recognized ending consonant is returned as the output. Observe that the recognition of ending consonant is also based on the core-vowel recognized in VRP. A block diagram of this ERP is shown in Fig. 10.

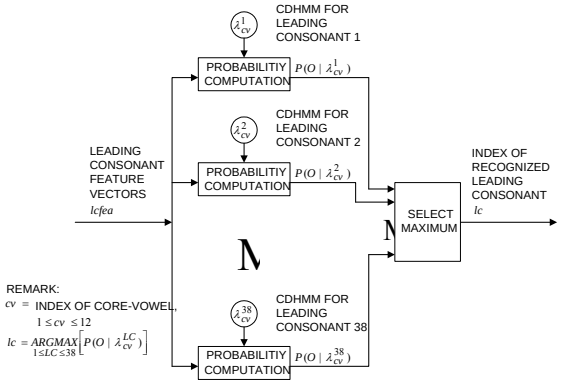

Fig. 9. A block diagram of an LCHMM bank

Like in the case of the leading consonant, there is a corresponding ECHMM bank for each core-vowel. Each ECHMM bank consists of at most 9 HMMs because not all ending consonants can be associated with every core-vowel. An HMM in each ECHMM bank represents an ending consonant class associated with the corresponding core-vowel. Each HMM in ECHMM bank is also a CDHMM. A block diagram of an ECHMM bank is depicted in Fig. 11.

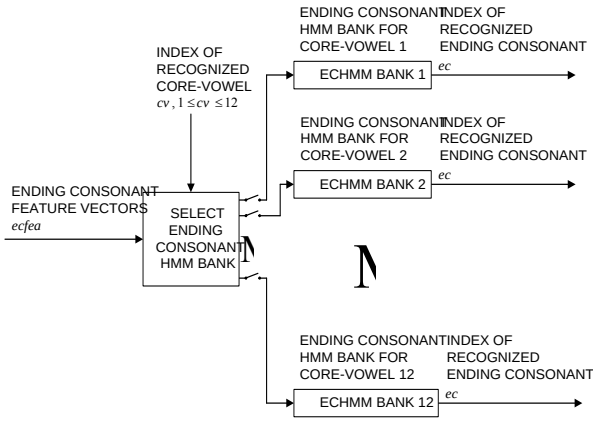

Fig. 10. A block diagram of ERP

2.3 Syllable-Based Word Recognition

After the syllables are recognized, one more difficult task awaits us. This is the grouping of the recognized syllables into meaningful words. We propose two approaches to accomplish this particular task. They are Rule-based Word Recognition, and Genetic Decoding Algorithm as outlined in Sections 2.3.1 and 2.3.2, respectively. Both of these approaches require the context of discourse in

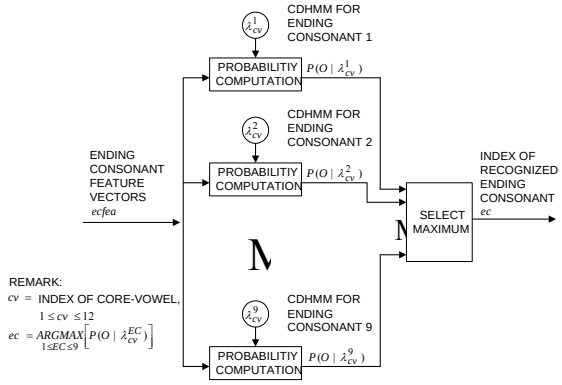

Fig. 11. A block diagram of an ECHMM bank

order to associate syllables into appropriate vocabularies. Note that the syllable recognition as described in Section 2.2 does not always produce perfect results. These two word recognition approaches attempt to make appropriate corrections while trying to recognize the words.

2.3.1 Rule-based Word Recognition: In this first alternative, an ambiguous phonetic dictionary of Thai language is constructed for words within the context of discourse. Words are considered according to the number of syllables contained.

For each word, a word model matrix is constructed. This matrix contains potential variations to the pronunciation of the word as possibly recognized by the processes described in Section 2.2. Probabilities of the alternative variations to the pronunciation are calculated for each syllable of the correct word. These alternative variations are called second, third, forth and other matches. The recognition probabilities of the variations for each syllable are summed to 1. An example of a table showing potential alternative variations to different leading consonants of a vowel /a/ is shown in Fig. 12.

9 /a/ อะ 21 /aa/ อา				
Leading	**Second Match**	**Third Match**	**Fouth Match**	**Other**
1.[/ph/ ,(พ)]	2.[/th/ ,(ท)] 30	3.[/kh/ ,(ค)] 6	12.[/h/ ,(ห)] 2	38
	0.7894	0.1578	0.0527	0.0001

2.[/th/ ,(ท)]	1.[/ph/ ,(พ)] 21	3.[/kh/ ,(ค)] 8	13.[/ch/ ,(ช)] 8	37
	0.5675	0.2162	0.2162	0.0001
3.[/kh/ ,(ค)]	2.[/th/ ,(ท)] 26	6.[/k/ ,(ก)] 5	1.[/ph/ ,(พ)] 2	33
	0.7878	0.1515	0.0606	0.0001
4.[/p/ ,(ป)]	8.[/b/ ,(บ)] 20	10.[/f/ ,(ฝ)] 5	1.[/ph/ ,(พ)] 4	29
	0.6896	0.1724	0.1379	0.0001
5.[/t/ ,(ต)]	13.[/ch/ ,(ช)] 10	18.[/l/ ,(ล)] 7	2.[/th/ ,(ท)] 5	22
	0.4545	0.3181	0.2273	0.0001
6.[/k/ ,(ก)]	3.[/kh/ ,(ค)] 12	17.[/nj/ ,(ง)] 8	21.[/r/ ,(ร)] 6	26
	0.4615	0.3076	0.2308	0.0001
7.[/?/ ,(อ)]	17.[/nj/ ,(ง)] 13	5.[/t/ ,(ต)] 9	3.[/kh/ ,(ค)] 6	28
	0.4643	0.3213	0.2143	0.0001
8.[/b/ ,(บ)]	15.[/m/ ,(ม)] 20	4.[/p/ ,(ป)] 9	9.[/d/ ,(ด)] 5	34
	0.5882	0.2647	0.1470	0.0001
9.[/d/ ,(ด)]	19.[/j/ ,(ย)] 13	14.[/c/ ,(จ)] 5	16.[/n/ ,(น)] 5	23
	0.5652	0.2174	0.2173	0.0001
10.[/f/ ,(ฝ)]	11.[/s/ ,(ส)] 32	20.[/w/ ,(ว)] 4	13.[/ch/ ,(ช)] 1	37
	0.8649	0.1080	0.0270	0.0001
11.[/s/ ,(ส)]	14.[/c/ ,(จ)] 17	10.[/f/ ,(ฝ)] 10	13.[/ch/ ,(ช)] 6	33
	0.5151	0.3030	0.1818	0.0001
12.[/h/ ,(ห)]	3.[/kh/ ,(ค)] 10	2.[/th/ ,(ท)] 9	1.[/ph/ ,(พ)] 9	28
	0.3571	0.3214	0.3214	0.0001
13.[/ch/ ,(ช)]	14.[/c/ ,(จ)] 15	2.[/th/ ,(ท)] 11	11.[/s/ ,(ส)] 6	32
	0.4688	0.3437	0.1875	0.0001
14.[/c/ ,(จ)]	11.[/s/ ,(ส)]	13.[/ch/ ,(ช)]	19.[/j/ ,(ย)] 4	37
	23	10		
	0.6216	0.2702	0.1081	0.0001
15.[/m/ ,(ม)]	17.[/nj/ ,(ง)] 21	4.[/p/ ,(ป)] 5	8.[/b/ ,(บ)] 4	30
	0.7000	0.1666	0.1333	0.0001
16.[/n/ ,(น)]	17.[/nj/ ,(ง)] 23	9.[/d/ ,(ด)] 11	21.[/r/ ,(ร)] 2	36
	0.6388	0.3056	0.0555	0.0001
17.[/nj/ ,(ง)]	16.[/n/ ,(น)] 18	15.[/m/ ,(ม)] 9	9.[/d/ ,(ด)] 5	32
	0.5625	0.2812	0.1562	0.0001
18.[/l/ ,(ล)]	21.[/r/ ,(ร)] 12	16.[/n/ ,(น)] 6	9.[/d/ ,(ด)] 5	23
	0.5217	0.2608	0.2174	0.0001
19.[/j/ ,(ย)]	9.[/d/ ,(ด)] 37	14.[/c/ ,(จ)] 2	17.[/nj/ ,(ง)] 1	40
	0.9249	0.0500	0.0250	0.0001
20.[/w/ ,(ว)]	4.[/p/ ,(ป)] 10	8.[/b/ ,(บ)] 7	17.[/nj/ ,(ง)] 7	24
	0.4166	0.2917	0.2916	0.0001
21.[/r/ ,(ร)]	18.[/l/ ,(ล)] 16	8.[/b/ ,(บ)] 5	6.[/k/ ,(ก)] 4	25
	0.6399	0.2000	0.1600	0.0001
22.[/ph/ ,(พล)]	1.[/ph/ ,(พ)] 21	18.[/l/ ,(ล)] 5	2.[/th/ ,(ท)] 3	29
	0.7241	0.1723	0.1034	0.0001
23.[/ph/ ,(พร)]	12.[/h/ ,(ห)] 21	3.[/kh/ ,(ค)] 5	6.[/k/ ,(ก)] 3	29
	0.7241	0.1723	0.1034	0.0001
24.[/kh/ ,(คล)]	3.[/kh/ ,(ค)] 26	18.[/l/ ,(ล)] 5	26.[/kh/ ,(คว)] 2	33
	0.7878	0.1515	0.0606	0.0001
25.[/kh/ ,(คล)]	3.[/kh/ ,(ค)] 26	18.[/l/ ,(ล)] 5	26.[/kh/ ,(คว)] 2	33
	0.7878	0.1515	0.0606	0.0001
26.[/kh/ ,(คว)]	3.[/kh/ ,(ค)]	20.[/w/	6.[/r/ ,(ร)] 2	33

	26	,(ว)] 5		
	0.7878	0.1515	0.0606	0.0001
27.[/kh/ ,(ค ร)]	3.[/kh/ ,(ค)] 26	2.[/th/ ,(ท)] 5	6.[/r/ ,(ร)] 2	33
	0.7878	0.1515	0.0606	0.0001
28.[/p/ ,(ป ล)]	4.[/?/ ,(ป)] 9	8.[/b/ ,(บ)] 8	18.[/l/ ,(ล)] 6	23
	0.3913	0.3477	0.2609	0.0001
29.[/p/ ,(ป ร)]	4.[/?/ ,(ป)] 9	8.[/b/ ,(บ)] 8	2.[/th/ ,(ท)] 6	23
	0.3913	0.3477	0.2609	0.0001
30.[/t/ ,(ต ร)]	5.[/t/ ,(ต)] 16	11.[/s/ ,(ส)] 9	21.[/r/ ,(ร)] 7	32
	0.5000	0.2812	0.2187	0.0001
31.[/k/ ,(ก ล)]	33.[/k/ ,(ก ร)] 12	27.[/kh/ ,(ค)] 8	18.[/l/ ,(ล)] 6	26
	0.4615	0.3076	0.2308	0.0001
32.[/k/ ,(ก ว)]	6.[/k/ ,(ก)] 12	27.[/kh/ ,(ค)] 8	20.[/w/ ,(ว)] 6	26
	0.4615	0.3076	0.2308	0.0001
33.[/k/ ,(ก ร)]	27.[/kh/ ,(ค)] 12	17.[/nj/ ,(ง)] 8	21.[/r/ ,(ร)] 6	26
	0.4615	0.3076	0.2308	0.0001

Fig. 12. An example of potential alternative pronunciation variations

Using data from appropriate tables, a word model matrix for any give word in the context of discourse can be constructed. An example of such a matrix is given in Fig. 13.

	1	2	3
1	สะ (sa)	หวัด (wat)	ดี (di:)
	1	1	1
2	จะ (ca)	ปัด (pat)	นี (ni:)
	0.5151	0.4166	0.5
3	ฝะ (fa)	บัด (bat)	บี (bi:)
	0.303	0.2917	0.4117
4	ชะ (cha)	งัด (njat)	ยี (ji:)
	0.1818	0.2916	0.0882

Fig. 13. An example of word matrix

Utilizing the word matrices, each given sentence is run through an algorithm called Word Segmentation. This algorithm separates words contained in a sentence of a connected speech. Details of this algorithm are described in Dachapratumvan and Santiprabhob (2002). It can be summarized in three steps as follows:

(i) Determine all possible word models of different lengths; say one syllable to four syllables, according to the recognized syllables of a sentence. Calculate the recognition probability of each word model based on the probabilities of its syllables.

(ii) Construct a graph containing all possible combinations of word models in the given sentence. An example of such a graph is shown in Fig. 14.

(iii) For each path (combination) of word models in the graph, calculate the average recognition probability of the path.

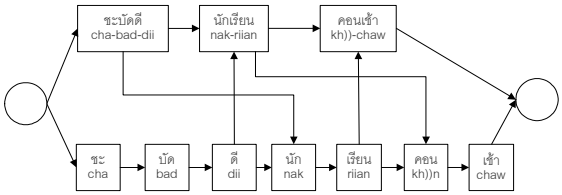

Fig. 14. An example of a graph containing all possible combinations of word models

Among the resulting paths (combinations) of word models, the one with the highest average recognition probability is chosen as a candidate. This candidate is then subject to another algorithm called Averages Likelihood, which attempts to correct errors left over from the syllable recognition process. Details of this algorithm are also given in Dachapratumvan and Santiprabhob (2002). In essence, this algorithm basically looks at each word model in that candidate sentence, for any word model with a recognition probability lower than 1, an attempt is made to change its syllable(s) whose recognition probability is lower than 1 to the corresponding syllable(s) of the correct word for the model. Note that this correction can only be done for words defined in the context of discourse, i.e. the words need to be included in the dictionary of the system.

2.3.2 Genetic Decoding Algorithm: Unlike the very structured rule-based algorithm described in the previous section, the decoding process presented in this section is based on the concept of Genetic Algorithms (GA) (Lawrence 1991).

First of all, appropriate ambiguous matrices need to be constructed for the four types of phoneme, i.e. leading consonant, vowel, ending consonant and tone. Each of these matrices contains possible variations of incorrect recognition of concerned phonetic value, e.g. a given leading consonant together with corresponding ambiguous degrees. Each ambiguous degree is basically a probability of that particular incorrect recognition among all the incorrect recognitions. A partial ambiguous matrix for two leading consonants is shown in Fig. 15.

	/ph/(พ)	**/th/(ท)**	**/kh/(ค)**	**/k/(ก)**	**/h/(ห)**	**/ch/(ช)**
/ph/(พ)	1	0.8	0.125	0.025	0.05	0
/th/(ท)	0.625	1	0.125	0.05	0.075	0.25

Fig. 15. A partial ambiguous matrix for leading consonants

The decoding process then starts with a set of potential word sequences as the initial population. These word sequences for the initial population are selected with the following two conditions.

- Words are randomly selected in such a way that they have the same vowel or tone to those in the same positions in the input word sequence.
- The number of syllables in a word sequence is equal to the number of syllables in the input word sequence.

The fitness value is calculated for each word sequence according to the fitness function below.

$$FN = \sum_{i=1}^{w}\left(\sum_{j=1}^{n} \text{aDegree}_j\right) \Big/ NS$$

where w = # of words in chromosome
n = # of syllables in each word
NS = # of syllables in chromosome

$$\text{aDegree}_j = \begin{cases} \text{ambiguous degree} \\ \\ \text{normal degree} \end{cases}$$

If the fitness value of the current word sequence is not good enough to be a solution, then a new generation of word sequences is generated by selecting two parents and applying the crossover operation. An example showing the crossover operation is given in Fig. 16.

The number of generations and acceptable fitness value are set as a condition to stop the decoding process. There are two alternatives to stopping the decoding process.

In the first alternative, it is assumed that the input syllable from the syllable recognition process has a 100% recognition rate. The acceptable fitness value is then set to 1. The fitness function uses the normal degree as a degree of fitness. The process is stopped when there is a fitness value of a word sequence equal to the acceptable fitness value. If, however, more than 80% of word sequences in the current population have the same fitness value then it can be concluded that the input syllable from the syllable recognition process has a recognition rate less than 100%, i.e. there are some errors. In such a case, the decoding process is restarted with the second alternative.

For second alternative, the fitness function uses the ambiguous degree from a corresponding ambiguous matrix as a degree of fitness. This alternative is stopped when more than 80% of word sequences in the current generation have the same fitness value. The word sequence that has a maximum fitness value is picking up as the result of this decoding process. The details of this genetic word-decoding algorithm can be found in Supasirirojana and Santiprabhob (2002).

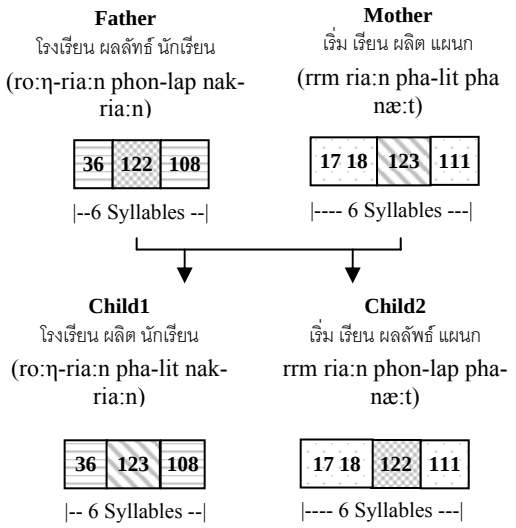

Fig. 16. An example of the crossover operation

3. Conclusion

This paper presents an overall framework upon which a connected speech recognition system for Thai language can be built. This speech recognition problem is divided into three research problems, namely syllable segmentation problem, syllable recognition problem and syllable-based word recognition problem.

The first two problems tackle the signal-processing portion. As for the syllable segmentation process, it needs to be tuned to fit the speaking style of representative speakers. The one used in our experiments is tuned for moderate speaking speed with typical loudness. It has also been observed that the quality of syllable recognition depends on the quality of the training set. The system tends to perform better when recognizing speeches of speakers whose sample words are included in the training set. In addition, even though we attempt to recognize syllables based on their phonemes, words that are included in the training set tend to be recognized better than those that are not. An important factor here is on the varying pattern of the speech signal when two syllables are connected. The syllables are recognized better when the training set contains their connecting pattern.

It should clearly be seen that the signal processing portion alone cannot achieve a high recognition rate in most cases. Two techniques to improve the recognition rate by means of associating recognized syllables with words from a given context of discourse are proposed. The rule-based word recognition approach is quite traditional and very well-structured, while the genetic word decoding algorithm follows a soft computing paradigm. Both show promising results. However, it should be observed that both techniques rely on the empirical result concerning the probability of incorrect recognition with respect to each phonetic value.

With the current stage of advancement in computing platform, it can be concluded that the connected speech recognition can still only be practically achieved within a given context of discourse. Enough words from the context need to be included in the training set for the syllable recognition process as well as in the dictionary for the syllable-based word recognition process in order to obtain reasonably high recognition rate.

4. References

Ahkuputra, V.; Jitapunkul, S.; Pornsukchandra, W.; and Luksaneeyanawin, S. 1997. A Speaker-Independent Thai Polysyllabic Word Recognition Using Hidden Markov Model, Proc. IEEE Pacific Rim Conf. Communications, Computers and Signal Processing, pp. 593-9.

Ahkuputra, V.; Jitapunkul, S.; Maneenoi, E.; Kasuriya, S.; and Amornkul, P. 1998. Comparison of Different Techniques On Thai Speech Recognition, The 1998 IEEE Asia-Pacific Conference on Circuits and Systems, pp. 177-80.

Chaiareerat, J.; and Santiprabhob P. 2002. Fuzzy-based Thai Syllable Segmentation for Connected Speech using Energy and Different Ceptral, Proc. InTech/VJFuzzy, pp. 334-7.

Cheirsilp, R.; and Santiprabhob, P. 2002. Phoneme-Based Thai Syllable Recognition by Means of Soft Computing, Proc. InTech/VJFuzzy, pp.325-33

Dachapratumvan, N.; and Santiprabhob, P. 2002. Thai Syllablic Correction in Connected Thai Speech Recognition, Proc. InTech/VJFuzzy, pp.314-9.

Jitapunkul, S.; Luksaneeyanawin, S.; Ahkuputra, V.; Maneenoi, E.; Kasuriya, S.; and Amornkul, P. 1998. Recent Advances of Thai Speech Recognition in Thailand, The 1998 IEEE Asia-Pacific Conference on Circuits and Systems, pp. 173-6.

Jittiwarangkul, N.; Jitapunkul, S.; Luksanee-yanawin S., Ahkuputra V., Wutiwiwatchai, C. 1998. Thai Syllable Segmentation for Connected Speech based on Energy. Proc. IEEE APCCAS, pp. WP1-8.1.

Lawrence, D. 1991. Handbook of Genetic Algorithms, Van Nostrand Reinhold, New York, NY, USA.

Maneenoi, E.; Jitapunkul, S.; Wutiwuwatchai, C.; and Ahkuputra, V. 1997. Modification of BP Algorithm for Thai Speech Recognition, Proc.1997 Int. Symp. Natural Language Processing,

MathWorks, Inc. 1999. Fuzzy Logic Toolbox for use with MATLAB User Guide, Version 2.

Pensiri, R.; and Jitapunkul, S. 1995. Speaker-Independent Thai Numerical Voice Recognition by using Dynamic Time Warping, Proc. 18th Electrical Engineering Conference, pp. 977-81.

Pornsukchandra, W.; and Jitapunkul, S. 1996. Speaker-Independent Thai Numeral Speech Recognition using LPC and the Back Propagation Neural Network. Proc. 19th Electrical Engineering Conference, pp. 977-81.

Rabiner, L.R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. IEEE. 77(2): 257-86.

Rabiner, L.R.; and Juang, B.H. 1993. Fundamentals of Speech Recognition, A. Oppenheim, Series Editor, Prentice-Hall, Englewood Cliffs, NJ, USA.

Rowden, C. 1992. Speech Processing. McGraw-Hill, London, England.

Supasirirojana, W.; and Santiprabhob, P. 2002. Thai Word Decoder Based on Genetic Algorithm. Proc. InTech/VJFuzzy, pp.320-4.

Wutiwiwatchai, C. 1997. Speaker Independent Thai Numeral Speech Recognition Using Neural Network and Fuzzy Technique, Master's thesis, Chulalongkorn University, Bangkok, Thailand.