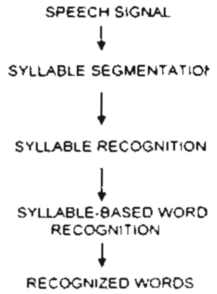


Figure 1: Our Conceptual Connected Speech Recognition System.

2. Thai syllable structure

In Thai language, a syllable consists of a set of phonemes. Each Thai syllable sound is comprised of four phonemes: leading consonant, vowel, ending consonant, and tone.

There are 38 different leading consonants. Actually, 33 of them are from Thai vocabularies, and the other five are borrowed from English vocabularies. These 38 leading consonants can also be classified into two groups: non-cluster and cluster. There are 21 non-clusters and 17 clusters. The cluster is the combination of two different leading consonants.

There are 24 vowels, which can be divided into two groups regarding to the length of vowel sound: 12 short-vowels and 12 long-vowels. Each group can be subdivided into nine major vowels (pure vowel) and three minor vowels (diphthong or the combination of two different major vowels).

There are only nine ending consonants in Thai phonemes. Not all ending consonants can occur with all vowel phonemes. Some syllables may not have ending consonant. Null ending consonant is also counted as one ending consonant class.

Only five tones are present in Thai syllable sounds.

3. Thai syllable analysis

The syllable sounds that consist of different phoneme classes are recorded using 16-bit quantization level, and sampling rate at 11.025 KHz. These syllable signals are then analyzed. By conducting this analysis on these syllable signals, we observed that each of three phoneme classes that constitute the syllable have the distinct signal sound patterns. These phoneme classes are leading consonant, vowel and ending consonant. From our analysis, the sound patterns of leading consonant, vowel and ending consonant can be located at the beginning, middle and ending parts of syllable signal respectively. It is difficult to locate the sound patterns of pure leading consonant and ending consonant in the syllable signal. But it is easy to locate the sound pattern of leading consonant in combination with vowel, and ending consonant in combination with vowel. When each of these sound patterns is played, we heard its sound like a single-syllabled word. And there is limited number of these sound patterns. From these reasons, we decided to treat them, as they are isolated words. Hence, the HMM technique, which gives the best recognition rate for isolated word recognition from our reviews, is used for recognizing these phonemes.

For tone phonemes, we can distinguish them using the fundamental frequency contours of the syllables. The fundamental frequency contours of all five Thai tone phonemes from [4] are given in figure 2. Therefore, we decided to use the Neural Network technique in recognizing tone phonemes.

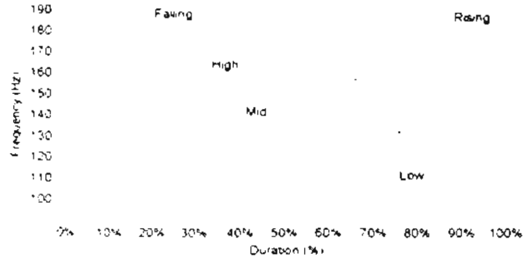


Figure 2: Fundamental Frequency Contours of All Five Tones.

4. Proposed system

Each Thai syllable sound comprises four different types of phoneme, namely leading consonant, vowel, ending consonant, and tone. Actually, there are approximately several thousands of Thai syllable. It is not practical to recognize them all. In the other hand, there are altogether approximately 76 phonemes in Thai language. Therefore, the components of a syllable should be recognized instead.

The overall system comprises five processes namely leading consonant, vowel, and ending consonant (LVET) feature extraction process, leading consonant recognition process (LRP), vowel recognition process (VRP), ending consonant recognition process (ERP) and tone recognition process (TRP). A block diagram of the overall system is given in figure 3. The details of five processes are described in the following subsections.

The speech feature extraction process has a duty to extract the needed speech features for all four recognition processes. Note that each recognition process requires its own set of speech features in order to function. And each recognition process is responsible for recognizing each phoneme part of the syllable as corresponding to its name. For example, the leading consonant recognition process is responsible for recognizing leading consonant phonemes.

For each unknown syllable signal *s*, which is to be recognized, the processing in figure 3 must be carried out. The steps in the processing are as follows:

1) The LVET feature extraction process extracts four different sets of speech features from the speech signal for leading consonant, vowel, ending consonant and tone recognition process.
2) The speech features for vowel and tone recognition processes are then processed by vowel recognition and tone recognition processes simultaneously. The indices of both recognized core-vowel *cv* and vowel *v* are generated as the output of vowel recognition. Then the core-vowel index *cv* is passed to leading and ending consonant recognition process. The tone recognition process generates the index of recognized tone *t* as its output.
3) The leading and ending consonant speech features and an index of recognized core-vowel *cv* are then passed to leading consonant and ending consonant recognition process simultaneously. The output of leading consonant and ending consonant recognition are the indices of recognized leading consonant *lc* and ending consonant *ec* respectively.
4) Finally, all indices of recognized leading consonant *lc*, vowel *v*, ending consonant *ec*, and tone *t* altogether present a recognized syllable *rs*.

Note that leading and ending consonant recognition process depends on the core-vowel recognition from the vowel recognition process.

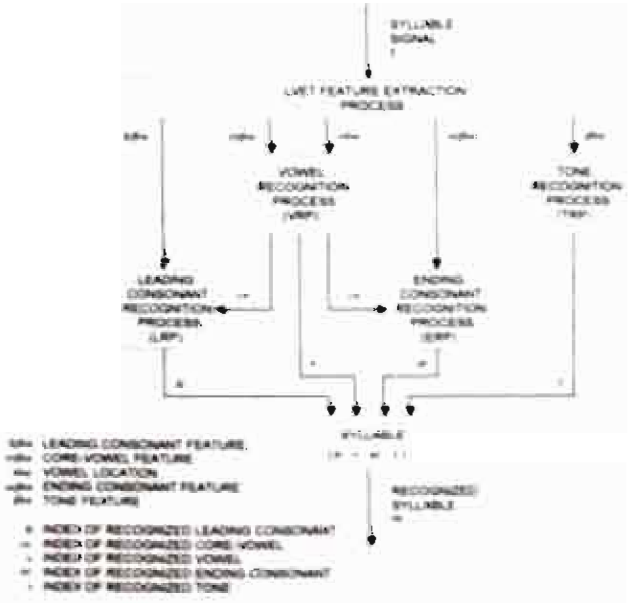

SYLLABLE SIGNAL s
LVET FEATURE EXTRACTION PROCESS
lcfea
cvfea
vloc
ecfea
tfea
VOWEL RECOGNITION PROCESS (VRP)
TONE RECOGNITION PROCESS (TRP)
LEADING CONSONANT RECOGNITION PROCESS (LRP)
cv
cv
ENDING CONSONANT RECOGNITION PROCESS (ERP)
lc
v
ec
t
SYLLABLE (lc v ec t)
RECOGNIZED SYLLABLE rs
lcfea LEADING CONSONANT FEATURE
cvfea CORE-VOWEL FEATURE
vloc VOWEL LOCATION
ecfea ENDING CONSONANT FEATURE
tfea TONE FEATURE
lc INDEX OF RECOGNIZED LEADING CONSONANT
cv INDEX OF RECOGNIZED CORE VOWEL
v INDEX OF RECOGNIZED VOWEL
ec INDEX OF RECOGNIZED ENDING CONSONANT
t INDEX OF RECOGNIZED TONE

Figure 3: A Block Diagram of the Proposed System.

4.1. LVET feature extraction process

This process is responsible for extracting all features needed for the four recognition processes (LRP, VRP, ERP, and TRP). This process is separated into two parts leading consonant, vowel and ending consonant features extraction and tone feature extraction. LRP, VRP, and ERP all use the same feature type but TRP use the different one.

4.1.1. Leading consonant, vowel, and ending consonant feature extraction

The following steps are carried out in order to extract features for leading consonant, vowel and ending consonant recognition.

LPC analysis and energy measurement. A block diagram of this step is given in figure 4. LPC analysis and energy measurement are performed on the syllable signal. The details of LPC analysis and energy measurement are described in [8][9]. The result from LPC analysis and energy measurement is a sequence of cepstral coefficients and energy vectors representing each signal frame. We call them together as *cep_e*.

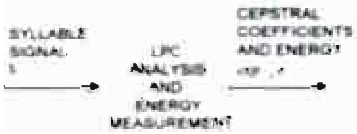

SYLLABLE SIGNAL s
LPC ANALYSIS AND ENERGY MEASUREMENT
CEPSTRAL COEFFICIENTS AND ENERGY
cep_e

Figure 4: A Block Diagram of LPC Analysis and Energy measurement.

Vowel location detection. The feature vector *cep_e* is used in determining the vowel location of the syllable signal. The vowel location *vloc* contains two values. They are the starting and ending frame numbers of vowel location in a given syllable signal. In order to determine the vowel location, the following steps are performed.

1) First step is to determine the Euclidean distance between cepstral feature vector frames. Then a series of difference cepstral values is generated.
2) The next one is to find the starting frame of vowel. Two thresholds are used, namely cepstral and energy threshold (the average energy of the syllable signal). The starting frame of vowel is the first frame in a sequence that lies between two frames, which have difference cepstral value higher than the cepstral threshold. Moreover, the signal energy of the searched frame must higher than the energy threshold. Searching of this frame number must be done in forward direction starting from the first frame in the sequence to the last one.
3) To determine the ending of the frame, same step in 2) is processed, but, this time, opposite direction starting from the last frame to the first one.

LVE Feature Segmentation. After the vowel location *vloc* has been determined. The *cep_e* feature vector is segmented into three parts based on vowel location. These three parts are *lcfea*, *cvfea*, and *ecfea*. They are used as the feature vectors of LRP, VRP, and ERP respectively. A block diagram of this segmentation is given in figure 5. The following cases are applied in segmentation.

- For leading consonant feature *lcfea*, all feature frames between its first frame of *cep_e* and the starting frame number of vowel location, are segmented and used as *lcfea* for LRP.
- For core-vowel feature *cvfea*, all feature frames between the starting and the ending frame numbers of vowel location, are segmented and used as *cvfea* for VRP.
- For ending consonant feature *ecfea*, all feature frames between the ending frame number of vowel location and the last frame of *cep_e* feature vector, are segmented and used as *ecfea* for ERP.

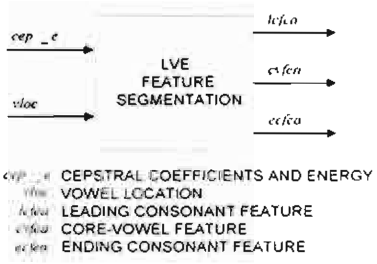

cep _e
vloc
LVE FEATURE SEGMENTATION
lcfea
cvfea
ecfea
cep_e CEPSTRAL COEFFICIENTS AND ENERGY
vloc VOWEL LOCATION
lcfea LEADING CONSONANT FEATURE
cvfea CORE-VOWEL FEATURE
ecfea ENDING CONSONANT FEATURE

Figure 5: A Block Diagram of LVE Feature Segmentation.

4.1.2. Tone feature extraction

A tone feature vector *tfea* is computed over a syllable signal. The extraction steps are as follows.

1) *Cepstral Pitch Detection:* A fundamental frequency contour f_0 is computed over a syllable signal using the cepstral pitch detection method. The method has already been described in [10].
2) *Tone Feature Normalization:* The fundamental frequency contour f_0 is normalized to have the same fundamental frequency contour level (reference point) and a specific number of elements as required in the TRP. That is because male and female have different fundamental frequency contour level and also the number of element in tone feature vector is depended on the length of the syllable sound. In order to normalize fundamental frequency contour to have N values, the following steps are carried out.
   a) Block the fundamental frequency contour values into N+1 frames with 50% overlap.
   b) Compute the mean value of each frame. These values are the elements of a new fundamental frequency.
   c) Normalize the new fundamental frequency contour values to have the same reference point that is the first value. The following formula is used.

$$f_n(n) = \left(\frac{f_0(n)}{f_0(1)} \times r\right) - r, \quad n = 2,3,\ldots, N+1$$

   where $f_n(n)$ is the n^th value of new fundamental frequency contour and r is any positive integer number.
   d) Remove the first value of normalized fundamental frequency contour out. The rest values are altogether a tone feature vector *tfea*.

4.2. Tone recognition process (TRP)

In this process, tone phoneme is recognized. A neural network is employed as recognition engine. A block diagram of this process is given in figure 6. A tone feature vector *tfea* from the LVET feature extraction process is processed. Finally, an index of recognized tone *t* is returned. In order to do tone recognition, the following steps are performed.

1) A neural network is configured and it is trained to classify all five tones.
2) For each unknown tone *t*, its tone feature vector *tfea* is used as the input of the neural network from a previous step.

We call the neural network used in TRP as Tone Neural Network (TONENN). The TONENN is a three layers feed forward neural network consisting of input, hidden, and output layer. There are requirements that input layer must have the same size as a tone feature vector *tfea*, the hidden layer must have a sufficient number of nodes for the network to perform tone classification and the output layer must have five nodes in the output layer corresponding to five tone phonemes. Each output node produces the value between 0 and 1. In this context, each output node value represents the probability for each tone phoneme class. The TONENN is supposed to be trained with a sufficient number of sample data. For each unknown tone to be recognized, its tone feature vector *tfea* is given to TONENN. After a tone feature vector *tfea* is passed to the TONENN, the network will generate five values as the number of output nodes. Each value implicitly represents a probability for one tone phoneme class. The index of the output node that gives the maximum value will be selected as the index of recognized tone *t*.

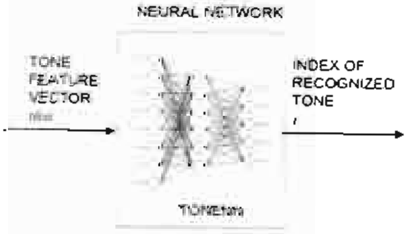


Figure 6: A Block Diagram of TRP.

4.3. Vowel recognition process (VRP)

In this process, vowel phonemes are recognized. 12 different vowel phonemes are called core-vowel in this process. We can recognize vowel by recognizing core-vowel and its length. A block diagram of this process is given in figure 7.

To recognize an unknown vowel, first its core-vowel feature vectors from LVET feature extraction process are processed by core-vowel recognition and an index of recognized core-vowel is returned. Next, its vowel location *vloc* is determined whether it is short or long vowel in vowel length determination step. Finally, these two results are used to determine an index of recognized vowel out of 24 vowel indices. The details of core-vowel recognition and vowel length determination are described as the followings.

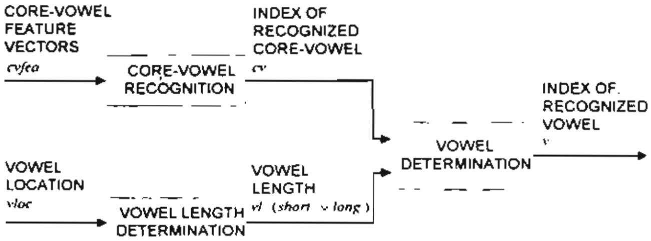

CORE-VOWEL FEATURE VECTORS
cvfea
CORE-VOWEL RECOGNITION
INDEX OF RECOGNIZED CORE-VOWEL
cv
INDEX OF. RECOGNIZED VOWEL
v
VOWEL DETERMINATION
VOWEL LOCATION
vloc
VOWEL LENGTH DETERMINATION
VOWEL LENGTH
vl (short v long)

Figure 7: 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 required. A block diagram of core-vowel recognition is given in figure 8. The type of HMM used in this process is Continuous Density Hidden Markov Model (CDHMM). The details of CDHMM are described in [8][9].

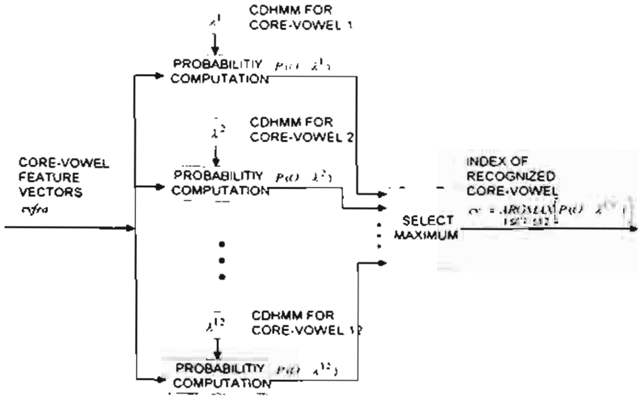


Figure 8: A Block Diagram of Core-Vowel Recognition

In order to do core-vowel recognition, the following steps are performed.

1) For each core-vowel class *cv*, a CDHMM λ^{cv} is built. And then the parameter set of each CDHMM is estimated in order to optimize the likelihood of the training set observation sequences.
2) For each unknown core-vowel, which is to be recognized, its core-vowel feature vectors *cvfea* are used as the observation sequence in the computation of model probabilities for all possible CDHMM models $P(O \mid \lambda^{cv})$. Finally, the index of CDHMM model, which has the highest probability, is selected as the index of recognized core-vowel *cv*.

Vowel length determination: The vowel location *vloc* contains the frame numbers of the starting and the ending points of vowel in a syllable signal. The vowel length is computed by subtracting the ending frame number with the starting frame number and then adding one to the result of subtraction because we want to include the starting and the ending frames. 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.

4.4. Leading consonant recognition process (LRP)

In this process, leading consonant feature vectors from LVET feature extraction process and an index of recognized core-vowel for VRP are processed. Finally, the index of recognized leading consonant is returned as the output. This means that the recognition of leading consonant is based on the recognition of core-vowel in VRP. A block diagram of LRP is given in figure 9.

For unknown leading consonant, which is to be recognized, the following steps are performed.

1) The index recognized core-vowel from VRP is used to select the leading consonant HMM bank.
2) Leading consonant feature vectors are used as the input of the selected LCHMM bank from the previous step. LCHMM bank processes the leading consonant feature vectors and returns the index of recognized leading consonant as its output.

The details of the LCHMM bank are described as the following.

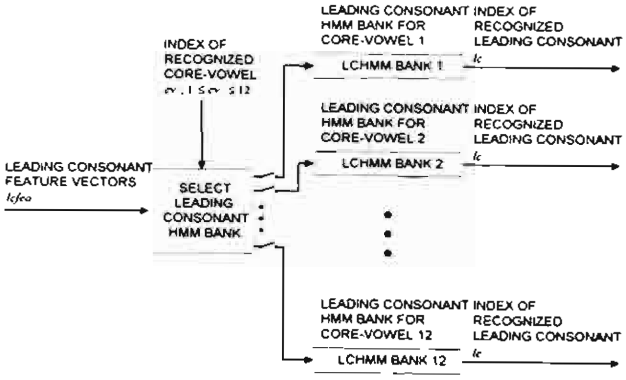
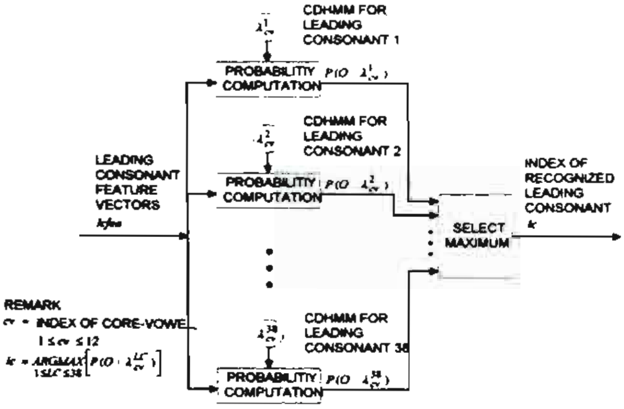

INDEX OF RECOGNIZED CORE-VOWEL
$cv, 1 \le cv \le 12$
LEADING CONSONANT FEATURE VECTORS
lcfea
SELECT LEADING CONSONANT HMM BANK
LEADING CONSONANT HMM BANK FOR CORE-VOWEL 1
LCHMM BANK 1
INDEX OF RECOGNIZED LEADING CONSONANT
lc
LEADING CONSONANT HMM BANK FOR CORE-VOWEL 2
LCHMM BANK 2
INDEX OF RECOGNIZED LEADING CONSONANT
lc
LEADING CONSONANT HMM BANK FOR CORE-VOWEL 12
LCHMM BANK 12
INDEX OF RECOGNIZED LEADING CONSONANT
lc

Figure 9: A Block Diagram of LRP.

LCHMM bank: Figure 10 shows a block diagram of a LCHMM bank. Each LCHMM bank is designed to cover all 38 leading consonant classes. Each LCHMM consists of up to 38 HMMs. This means that for each leading consonant class, there is a HMM corresponding to it. Each HMM in LCHMM bank is a CDHMM. In order to do leading consonant recognition, the following steps are performed.

1) For each leading consonant class *lc* of core-vowel *cv* class, a CDHMM λ_{cv}^{lc} is built. And then the parameter set of each HMM is estimated in order to optimize the likelihood of the training set observation sequences.
2) For each unknown leading consonant and known core-vowel *cv*, which is to be recognized, its leading consonant feature vectors *lcfea* are used as the observation sequence in the computation of model probabilities for all possible HMM models $P(O \mid \lambda_{cv}^{lc})$. Finally, the index of HMM model, which has the highest probability, is selected as the index of recognized leading consonant *lc*.

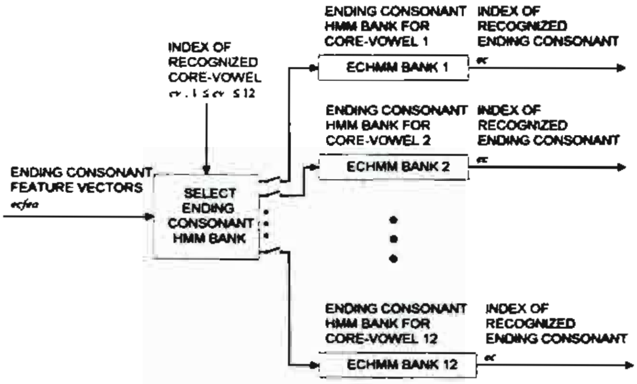


CDHMM FOR LEADING CONSONANT 1
λ_{cv}^{1}
PROBABILITY COMPUTATION
$P(O \mid \lambda_{cv}^{1})$
CDHMM FOR LEADING CONSONANT 2
λ_{cv}^{2}
PROBABILITY COMPUTATION
$P(O \mid \lambda_{cv}^{2})$
LEADING CONSONANT FEATURE VECTORS
lcfea
SELECT MAXIMUM
INDEX OF RECOGNIZED LEADING CONSONANT
lc
REMARK
cv = INDEX OF CORE-VOWEL
$1 \le cv \le 12$
$lc = \underset{1 \le LC \le 38}{ARGMAX}\left[P(O \mid \lambda_{cv}^{LC})\right]$
CDHMM FOR LEADING CONSONANT 38
λ_{cv}^{38}
PROBABILITY COMPUTATION
$P(O \mid \lambda_{cv}^{38})$

Figure 10: A Block Diagram of a LCHMM Bank.

4.5. Ending consonant recognition process (ERP)

In this process, the ending consonant feature vectors and an index of recognized core-vowel from VRP are processed. Finally, the index of recognized ending consonant is returned as the output. This means the recognition of ending consonant is based on the recognition of core-vowel in VRP. A block diagram of ERP is given in figure 11.



INDEX OF RECOGNIZED CORE-VOWEL
$cv, 1 \le cv \le 12$
ENDING CONSONANT FEATURE VECTORS
ecfea
SELECT ENDING CONSONANT HMM BANK
ENDING CONSONANT HMM BANK FOR CORE-VOWEL 1
ECHMM BANK 1
INDEX OF RECOGNIZED ENDING CONSONANT
ec
ENDING CONSONANT HMM BANK FOR CORE-VOWEL 2
ECHMM BANK 2
INDEX OF RECOGNIZED ENDING CONSONANT
ec
ENDING CONSONANT HMM BANK FOR CORE-VOWEL 12
ECHMM BANK 12
INDEX OF RECOGNIZED ENDING CONSONANT
ec

Figure 11: A Block Diagram of ERP.

For unknown ending consonant, which is to be recognized, the following steps are performed.

1) The index recognized core-vowel is used to select the ending consonant HMM (ECHMM) bank. There are 12 ECHMM banks as the number of core-vowels. Each ECHMM bank has the different number of ending consonant classes to be recognized. That is because not all ending consonant can occur with all core-vowels. The details of each ECHMM are described in the next section.
2) Ending consonant feature vectors are used as the input of the selected ECHMM bank from the previous step. The selected ECHMM bank processes the ending consonant feature vectors and returns the index of recognized ending consonant as its output.

ECHMM bank: each ECHMM bank consists of at most 9 HMMs because not all ending consonant can occur with all core-vowels. This means that for each ending consonant class, there is a HMM representing it. Each HMM in ECHMM bank is a CDHMM. A block diagram of an ECHMM bank is given in figure 12. In order to do ending consonant recognition, the following steps are performed.

1) For each ending consonant class *ec* having core-vowel class *cv*, a CDHMM λ_{cv}^{ec} is built. And then the parameter set of each HMM is estimated in order to optimize the likelihood of the training set observation sequences.
2) For each unknown ending consonant having core-vowel *cv*, which is to be recognized, its ending consonant feature vectors *ecfea* are used as the observation sequence in the computation of model probabilities for all possible HMM models $P(O \mid \lambda^{ec})$. Finally, the index of HMM model, which has the highest probability, is selected as the index of recognized ending consonant *ec*.

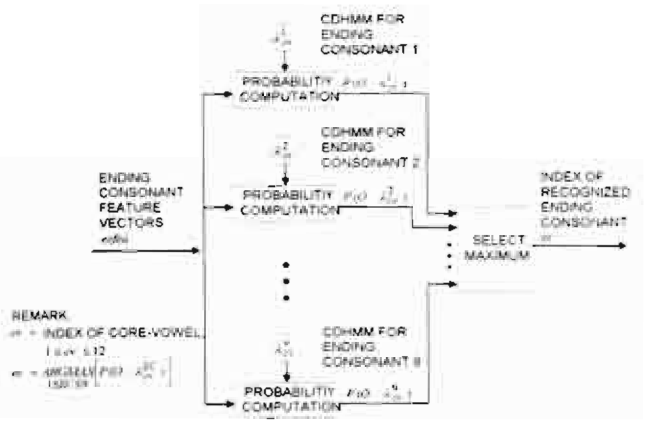


Figure 12: A Block Diagram of an ECHMM Bank.

5. Experimental results

There are three sets of syllables. They are all possible combinations of 38 leading consonants and 24 vowels used in LRP and VRP experiments, all possible combinations of 12 vowels and 9 ending consonants for ERP experiment, and randomly selected syllables comprises all five tone phonemes for TRP and VRP core-length determination experiments. All syllable sounds are gathered for two male and two female speakers. Each speaker spoke each syllable of all syllable sets 10 times. The syllable sounds were recorded using 16-bit quantization level, and sampling rate at 11.025 KHz. Four sets of experiments were conducted for TRP, VRP. LRP, and ERP respectively.

For TRP, there are two experiments conducted. The TONENNs for both experiments are configured to have 9 input nodes, 60 hidden nodes, and 5 output nodes according to 5 tone phonemes. The training algorithm is gradient descent with variable learning rate. Goal and maximum number of training epochs are set to 0.005 and 1000 respectively. The first TONENN is trained using 20% of data and the second one is trained using 50% of data. The recognition result of TRP is shown in table 1.

For VRP, two sets of experiments are conducted. The first set is for core-vowel recognition and the second set is for vowel length determination.

Two experiments are conducted for VRP core-vowel recognition, LRP, and ERP. Both experiment are almost the same except the number of training data. The number of training data for both experiments is 20% and 50% of data respectively.

For core-vowel recognition, LRP, and ERP experiments, the same parameters are used to extract the *cep e* features. They are 30ms frame size, 10ms frame rate, 14 LPC orders, and 12 cepstral coefficients orders. Each CDHMM used in these experiments is configured to have 3 states and 3 mixtures. Their experimental results are shown in table 2, 3, and 4 respectively.

For vowel length determination experiments, the data are divided into two sets: 20 % tuning set and 80% testing set. The vowel length threshold is determined from the tuning set. Finally, the experimental result of vowel length determination is shown in table 5.

Tone No.	Amount	Train 20%		Train 50%	
		Train (%)	Test (%)	Train (%)	Test (%)
1.[Mid]	1920	89 58	88.54	89 48	92 29
2.[Low]	1800	91.39	93 89	91 56	94 67
3.[Falling]	1120	88 84	87 28	85 89	90.00
4.[High]	1560	94 23	91 59	92 82	94 10
5.[Rising]	800	100 00	100.00	100.00	99.75
Total	7200	92.08	91.61	91.33	93.75

Table 1: TRP Experimental Result.

Vowel No.	Amount	Train 20%		Train 50%	
		Train (%	Test (%)	Train (%)	Test (%)
1.[/i/,(ì Ọ	1680	95.83	92 78	97.02	95.12
2.[/w/,(ì Ọ	1680	92.56	87 20	83 81	83 33
3.[/u/,(ì Ọ	1680	91 37	86 83	90.12	90 95
4.[/e/,(ầ Ð)]	1680	95.83	92.49	95 95	96 31
5.[/3/,(ầ ǀ Ð)]	1680	94 05	88 32	91 31	90.48
6.[/o/,(ầl Ð)]	1680	88 39	86.24	89.17	88.45
7.[/ae/,(ál Ð)]	1680	97 92	94.20	96 67	95 24
8.[/)/,(ầ ŒÐ)]	1680	95.83	91 89	94 29	94 52
9.[/a/,(ì Ð)]	1680	95 54	94 57	95.60	94 76
10.[/ia/,(ầ ŒÐ)]	1680	95.83	93.60	95.24	95 95
11.[/wa/,(ầ ⊧Ð)]	1680	93 15	94.05	92.86	94 52
12.[/ua/,(ì ŒÐ)]	1680	96.43	90 55	94 76	96 07
Total	20160	94,39	91,06	93.07	92.98

Table 2: VRP Core-Vowel Experimental Result.

Leading No.	Amount	Train 20%		Train 50%	
		Train (%)	Test (%)	Train (%)	Test (%)
1.[/ph/ ,(¾)]	960	99.48	70.31	91.88	88.96
2.[/th/ ,(·)]	960	98 44	73.57	91 04	93 13
3.[/kh/ ,(¤)]	960	98 96	77 60	95 83	90 63
4.[/p/ ,(»)]	960	98 96	57.68	91 67	86 88
5.[/t/ ,(µ)]	960	97 92	68 23	96 46	91 04
6.[/k/ ,(¡)]	960	98 96	64 06	94 17	90.00
7.[/?/ ,(ǀ)]	960	98 44	62 50	92 92	87 29
8.[/b/ ,(°)]	960	98 96	71.88	95.42	93 75
9.[/d/ ,(´)]	960	98 44	64.32	93 75	88 75
10.[/f/ ,(½)]	960	98 44	69.27	94 58	91 04
11.[/s/ ,(É)]	960	100 00	78 91	97.71	94.38
12.[/h/ ,(É)]	960	96.35	59 64	90.83	84.38
13.[/ch/ ,(*)]	960	100 00	85.81	97 08	96 67
14.[/c/ ,(¨)]	960	98 44	79.30	96 88	95 42
15.[/m/ ,(Á)]	960	99 48	83.59	96.67	94.58
16.[/n/ .(')]	960	97 40	77 99	96 04	92 71
17.[/nj/ ,(§)]	960	98.44	80.60	97 08	97.92
18.[/l/ ,(Å)]	960	99.48	73.57	94 58	91.04
19.[/j/ ,(Â)]	960	99 48	87 50	98.13	98.13
20.[/w/ ,(Ç)]	960	96 35	79.56	95 83	90 42
21.[/r/ ,(Ã)]	960	98 96	61 07	85.63	71.25
22.[/phl/,(¾Å)]	960	95 83	65 76	83.75	82 71
23.[/phr/.(¾Ã)]	960	96.88	59.51	90 21	85 21
24.[/thr/,(· Ã)]	960	97 40	69 92	91 46	87.29
25.[/khl/,(¤Å)]	960	95 31	65 10	90.83	89.17
26.[/khw/,(¤Ç)]	960	98 44	77.08	93 33	90 21
27.[/khr/,(¤Ã)]	960	97 92	67 58	89.17	86 88
28.[/pl/ ,(»Å)]	960	96.88	53.52	84.58	81.46
29.[/pr/ ,(»Ã)]	960	94 79	52 08	84 79	73.33
30.[/tr/ ,(µÃ)]	960	97 92	52 47	85 83	80.00
31.[/kl/ ,(¡ Å)]	960	97 40	47 53	84 79	80 21
32.[/kw/ ,(¡ Ç)]	960	97 92	74.48	93 75	92.71
33.[/kr/ .(¡ Ã)]	960	97 92	60 55	91 67	83 13
34.[/dr/ ,(´ Ã)]	960	98.44	59.51	87 08	81.67
35.[/fl/ ,(¿ Å)]	960	96 88	66 15	86 25	90.21
36.[/fr/ ,(¿ Ã)]	960	100 00	75.00	90.42	88 33
37.[/bl/ ,(° Å)]	960	98.44	64 19	86 25	86 04
38.[/br/ ,(° Ã)]	960	94.79	54.56	86.04	77 29
Total	36480	98.01	68.21	91.69	88.00

Table 3: LRP Experimental Result.

Ending No.	Amount	Train 20%		Train 50%	
		Train (%)	Test (%)	Train (%)	Test (%)
1.[/p/,(°)]	480	100 00	79 69	99 58	95 83
2.[/t/,(ˊ)]	480	100 00	78 39	100 00	95 83
3.[/k/,(¡)]	480	100 00	69 53	98 75	95 00
4.[/?/,(Ð)]	480	100 00	83 07	100 00	97 08
5.[/m/,(A)]	480	100 00	96 35	99 58	97 92
6.[/n/,(')]	480	100 00	94 01	100 00	98 33
7.[/nj/,(§)]	480	100 00	85 16	100 00	97 92
8.[/j/,(Ä)]	280	100 00	97 77	100 00	100 00
9.[/w/,(Ç)]	200	100 00	95 63	100 00	100 00
Total	3840	100.00	85.38	99.74	97.24

Table 4: ERP Experimental Result.

Vowel Length	Amount	Tuning 20%	
		Tuning(%)	Testing(%)
1.|Short|	2680	96 08	95.01
2.|Long|	4520	86 06	86 56
Total	7200	89.79	89.70

Table 5: Vowel Length Experimental Result.

6. Conclusion

This paper has presented the phoneme-based Thai syllable recognition system using Continuous Density Hidden Markov Model (CDHMM) and Neural Network (NN) techniques. The system consists of five processes. The first process is responsible for doing feature extraction. The rest four processes are responsible for doing phoneme recognition. As a syllable consists of four phonemes: leading consonant, vowel, ending consonant and tone. They are namely as leading consonant recognition process (LRP), vowel recognition process (VRP), ending consonant recognition process (ERP) and tone recognition process (TRP). Cepstral coefficients and signal energy frames extracted from a syllable signal are the base feature for LRP, VRP, and ERP. The vowel location is detected using the differences of cepstral coefficients frame and energy and then used to partition the base feature into three parts for LRP, VRP, and ERP respectively. The fundamental frequency contour extracted using cepstral pitch detection techniques is used as feature for TRP. The CDHMM technique is applied as the recognition engine in LRP, VRP, and ERP. The NN technique is applied as the recognition engine of TRP.

The best recognition rates of the leading consonant, core-vowels, and ending consonants phonemes are 88.00%, 92.98%, and 97.24% respectively. Note that in order to get these recognition rates, 50% of data are required as the training set of each CDHMM. The recognition rate of vowel length is 89.70%. From our experimental results, the correctness of syllable segmentation, the correctness of vowel location detection, the size of training set, and the quality of training set play the important role. For tone phonemes, the best average recognition rate is 93.75%. Note that 50% of data are required as the training set of NN. The NN used in TRP has 9 input nodes, 60 hidden nodes, and 5 output nodes according to 5 tone phonemes. The training algorithm is gradient descent with variable learning rate. Goal and maximum number of training epochs are set to 0.005 and 1000 respectively.

7. Future researches

In the future, there are three challenging works for us. The first one is to find the ways to improve the recognition results of our system. The higher recognition result can be achieved by adding more HMM models to each phoneme class. This causes increasing of the recognition time. Second, this system should be integrated with syllable segmentation system. Finally, the syllable-based word recognition system is the next challenging step to fulfill our conceptual Thai connected speech recognition.

8. References

[1] Ahkuputra, V., Jitapunkul, S., Maneenoi, E., Kasuriya, S., and Amornkul, P., *Comparison of Different Techniques On Thai Speech Recognition*, The 1998 IEEE Asia-Pacific Conference on Circuits and Systems, pp. 177-180 1998.

[2] Ahkuputra, V., Jitapunkul, S., Pornsukchandra, W., and Luksaneeyanawin, S., *A Speaker-Independent Thai Polysyllabic Word Recognition Using Hidden Markov Model*, Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, pp. 593-599, 1997.

[3] Jitapunkul, S., Luksaneeyanawin, S., Ahkuputra, V., Maneenoi, E., Kasuriya, S., and Amornkul, P., *Recent Advances of Thai Speech Recognition in Thailand*, The 1998 IEEE Asia-Pacific Conference on Circuits and Systems, pp. 173-176, 1998.

[4] Luksaneeyanawin, S., *Linguistics Research and Thai Speech Technology*, International Conference on Thai Studies, 1993.

[5] Maneenoi, E., Jitapunkul, S., Wutiwuwatchai, C., and Ahkuputra, V., *Modification of BP Algorithm for Thai Speech Recognition*, Proceeding of the 1997 International Symposium on Natural Language Processing, 1997.

[6] Pensiri, R. and Jitapunkul, S., *Speaker-Independent Thai Numerical Voice Recognition by using Dynamic Time Warping*, Proceedings of the 18th Electrical Engineering Conference, pp. 977-981, 1995.

[7] Pornsukchandra, W. and Jitapunkul, S., *Speaker-Independent Thai Numeral Speech Recognition using LPC and the Back*

Propagation Neural Network, Proceedings of the 19th Electrical Engineering Conference, pp. 977-981, 1996.

[8] Rabiner, L. R., *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceeding of IEEE, Vol. 77, No.2, pp. 257-286, February, 1989.

[9] Rabiner, L. R. and Juang, B. H., *Fundamentals of Speech Recognition*, A. Oppenheim, Series Editor, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[10] Rowden, C., *Speech Processing*, London: McGraw-Hill, 1992.

[11] Wutiwiwatchai, C., *Speaker Independent Thai Numeral Speech Recognition Using Neural Network and Fuzzy Technique*, Master's thesis, Chulalongkorn University, 1997.

Thai Syllabic Correction in Connected Thai Speech Recognition

Nunmanus Dachapratumvan
Intelligent Systems Laboratory
Faculty of Science and Technology
Assumption University
Bangkok, Thailand
E-mail: g4219712@au.ac.th

Pratit Santiprabhob
Intelligent Systems Laboratory
Faculty of Science and Technology
Assumption University
Bangkok, Thailand
E-mail: pratit@s-t.au.ac.th

Abstract: The basic syllable rules for Thai language can be defined by use a phoneme encoding method. Thai characters can be classified into the following sets: leading consonant, vowel symbols, Ending consonant and tone makers. Based on this character set and their appearances, the rule of syllabic boundary and ambiguous phonetic sound can be formulated to created Thai Ambiguous Phonetic Dictionary. Word Segmentation algorithm used for separating connected syllables in Thai Speech Recognition to be the possible word segmentation. Some of each word segment will be know or unknown in Thai meaning. Average Likelihood gives the correction unknown Thai words to be Thai word meaning system.

Key words: Thai Phonetic Rule Base, Thai Phoneme Ambiguous Dictionary, Word Segment Algorithm, and Average Likelihood Thai word correction.

1. Introduction

Thai syllabic correction has become one of the most essential things for integrating with the empirical result of Phoneme-base Thai speech recognition system [1]. From the empirical result of Phoneme-base Thai Speech recognition system [1] representation Thai phonetic: Leading Consonant, Vowel, Ending Consonant and Tone. All four parts represent to a syllable.

From the review of Thai Word decoding or Thai Word Segmentation approaches. Thai Word decoding [5,6,9,11] almost approaches use for text retrieved to find the unknown word meaning to word correction. Word Segmentation approaches [7,8,10] are important for separating syllables connected from Thai speech recognition to be each word segments. A detail survey of this technique almost found with Word Segmentation from Thai text sentence. Another approach is important for Thai word properties are Thai Rule-Base System [3,4] use for created Thai Regular Grammar to apply with Thai Syllable Speech Recognition error.

In this research, the proposed system presents a way to correction Thai syllabic connected speech recognition using Word Segmentation algorithm search in Thai Ambiguous Phonetic Dictionary and correction a word with Average Likelihood. This paper is organized as follows: In Section 2, the detail of Thai ambiguous Phonetic Word Model Section 3 and 4, the detail of propose system show the algorithm of Word Segment Algorithm and Average Likelihood Algorithm to find the correction word in Thai word system.

2. Proposed system

The propose system Thai Syllabic Correction by classifying them to correction with Thai Rule base system. The system consists of three processing Thai Ambiguous Phonetic Word Model, Word Segment Algorithm and Average Likelihood Algorithm.

A Thai syllable format can be divided to 4 parts. Those are leading consonant, vowel, ending consonant and Tone. A Thai syllable format using properties of each leading consonant, vowel and ending consonant. And also properties of Thai ambiguous sound modify with Phoneme-Base Thai Speech recognition system using Fuzzy system and Neural Network to create the probability of Thai Syllabic model.

2.1 Thai phoneme word dictionary

The format of Thai syllables properties and the properties of ambiguous phonetic sound can be generated to be Thai Phoneme Word Dictionary. The basic Thai Syllable Format Rules can define to be Thai Rule base Regular Grammar. The syllables building in Thai phonetic system consists of Leading Consonant, Vowel, and Ending Consonant. Thai Phoneme Words have to verify with Thai Rule base Regular Grammar. Thai Phoneme Word Dictionary keeps the words that can possible to occurred error phoneme from Phoneme-base Thai speech recognition system [1].

2.1.1 Modify leading, vowel marker properties

There are many approaches to derive vowel properties. Some approaches consider only the appearance of each vowel symbol. The following rules are used to determine their segmentation.

- The vowels, '–ะ /a/, -   /ua/, โ -ะ /o/, แ -ะ /æ/, เ -าะ /)/, เ -ะ /e/' always require at least one leading consonant and no final consonant follows.
- The vowel all form 'เ -ะ /e/, แ -ะ /æ/, โ -ะ /o/, เ -าะ /)/' always has a leading consonant.

- The vowels เ -/ee/, แ -/ææ/ always precedes consonants.

2.1.2 Modify Final Consonant Properties

The following letters are never used as final consonants ข /kh/,ฑ /t/,ฉ /ch/,ฌ/n/,ฝ /f/,ผ /ph/

2.1.3 Rule-based Method

Although Thai grammar has many exceptions, the majority of syllable usage still follows these rules.

Characters	Meaning
|	Matches either the preceding or the following regular expression.
C, V	Consonant, Vowel

Table1: Character Meaning

Thai characters set symbols
C = {Ci:1<=`I<= 44 set of Thai Leading Consonants}
C_f = { c_i | c_j = letters which consider as a Final Consonant, except /kh/,ฑ /t/,ฉ /ch/,ฌ/n/,ฝ /f/,ผ /ph/}
V = { v_i|v_j = vowels which always place after consonant }

Table2: Thai Characters set symbols

In summary the basic regular expressions were compiled and emphasized to the consonant rules.

	Regular Expression	Vowel Patterns	Example
R1	C (V)(T)(Cf)	ิ ี ุ ู ึ ื	กิน ชุด
R2	C (T)า(Cf)	-า	กล้า ค้า ยา ขาด
R3	C (T)ำ(Cf=ม)	ำ	น้ำ
R4	C (T)อะ C ั (T)(Cf=ว) C อะ(T) (Cf=น) C ัวะ	อะ ัว ั ัวะ	จะ คะ ตัว กัน ผัวะ
R5	โC (T)Cf โC (T)ะ	โอ โ-ะ	โต โกรธ โต๊ะ
R6	ไ|C (T) (Cf=ย) ใ|C (T) (Cf=ย)	ไ ใ	ไกล ใจ
R7	แC (V)(T)(Cf) แC (V)(T)ะ	แ แอะ	แก แกะ แกล้ง แนะ
R8	เC (T)(Cf) เC (T)า(Cf=ว) เC (T)อ(Cf) เC (T)าะ เC (T)ะ เC(V)(T)ีย(Cf|ะ) เC(V)(T)ือ (Cf|ะ)	เ- เ-า เ-อ เ-าะ เ-ะ เอีย เอียะ เอือ เอือะ	เป เกล้า เทอม เกาะ เตะ เมีย เปี๊ยก เสือ เสือก
R9	C (T)วCf	ว	ม้วน

Table3: Thai Regular Expression

2.1.4 Thai ambiguous phonetic system

From the empirical result of syllable recognition of Phoneme Base speech recognition [1] can classified the groups of leading consonant ambiguous phonetic sound base on vowel phoneme. The results of Phoneme base speech recognition can analyst by percent of sound matching show in Table 4.

Leading	2nd Match	3rd Match	4th Match	[illegible]
/ph/ (พ)	/th/ (ท)	/kh/ (ค)	/h/ (ห)	
	0.8421	0.1052	0.0526	0.0001
/th/ (ท)	/ph/ (พ)	/ch/ (ช)	/h/ (ห)	
	0.6176	0.2941	0.0882	0.0001
/kh/ (ค)	/k/ (ก)	/h/ (ห)	/ph/ (พ)	
	0.8235	0.0882	0.0882	0.0001
/p/ (ป)	/w/ (ว)	/l/ (ล)	/t/ (ต)	
	0.3637	0.3181	0.3181	0.0001
/t/ (ต)	/s/ (ส)	/r/ (ร)	/l/ (ล)	
	0.5	0.2812	0.2187	0.0001
/k/ (ก)	/kh/ (ค)	/th/ (ท)	/h/ (ห)	
	0.8285	0.0857	0.0857	0.0001
/?/ (อ)	/h/ (ห)	/nj/ (ง)	/k/ (ก)	
	0.4285	0.3809	0.1905	0.0001
/b/ (บ)	/d/ (ด)	/w/ (ว)	/m/ (ม)	
	0 6486	0.2702	0.0811	0.0001
/d/ (ด)	/n/ (น)	/b/ (บ)	19.[/j/ .(ย)]	
	0.5	0.4117	0.0882	0.0001
/f/ (ฝ)	/s/ (ส)	/th/ (ท)	/ch/ (ช)	
	0 8055	0.1389	0.0555	0.0001
/s/ (ส)	/ch/ (ช)	/f/ (ฝ)	/c/ (จ)	
	0.3749	0.3125	0.3125	0.0001
/h/ (ห)	/th/ (ท)	/kh/ (ค)	/k/ (ก)	
	0.3448	0.3448	0.3103	0.0001
/ch/ (ช)	/c/ (จ)	/s/ (ส)	/kh/ (ค)	
	0.6857	0.1714	0.1428	0.0001
/c/ (จ)	/ch/ (ช)	/s/ (ส)	/th/ (ท)	
	0.5588	0.3529	0.0882	0.0001
/m/ (ม)	/nj/ (ง)	/n/ (น)	/w/ (ว)	
	0.4594	0.4324	0.1081	0.0001
/n/ (น)	/d/ (ด)	/nj/ (ง)	/m/ (ม)	
	0.4324	0.3243	0.2432	0.0001
/nj/ (ง)	/n/ (น)	/m/ (ม)	/l/ (ล)	
	0.5526	0.3947	0.0526	0.0001
/l/ (ล)	/r/ (ร)	/s/ (ส)	/d/ (ด)	
	0.7586	0.1379	0.1034	0.0001
/j/ (ย)	/h/ (ห)	/l/ (ล)	/d/ (ด)	
	0.5	0.3461	0.1538	0.0001
/w/ (ว)	/r/ (ร)	/m/ (ม)	/b/ (บ)	
	0.5312	0.3125	0.1562	0.0001
/r/ (ร)	/l/ (ล)	/s/ (ส)	/w/ (ว)	
	0.4166	0.3333	0.25	0.0001

Table4: Thai Ambiguous Phoneme Matching Values with Vowel /i/ อิ and /ii/ อี

From the empirical result of syllable recognition process [1], the system has leading consonants training 100 times.

Let S = Time of second match
T = Time of third match
F = Time of fourth match

Percent of First match (1st) $= 1.0$

Percent of Second match (2nd) $= \dfrac{S}{\sum \{S, T, F\}}$

Percent of Third match (3rd) $= \dfrac{T}{\sum \{S, T, F\}}$

Percent of Forth match (4th) $= \dfrac{F}{\sum \{S, T, F\}}$

Percent of other match = 1-(%2nd+%3rd+%4th)

2.1.5 Create Thai ambiguous phonetic dictionary (AMPD)

From the result of Thai Ambiguous Phonetic System can generate the words into Thai ambiguous phonetic dictionary. The groups of word model have to separate to 4 groups.

- First Group for monosyllable
- Second group for two syllables
- Third group for three syllables
- Fourth group for four syllables

Example of word model in third group:

Sentence:	สวัสดี		
Phonetic:	สะ	หวัด	ดี

From this word, generate to

สะ =

/s/ (ส)	/c/ (จ)	/f/ (ฝ)	/ch/ (ช)	
	0.5151	0.303	0.1818	0.0001

หวัด =

/w/ (ว)	/h/ (ห)	/b/ (บ)	/k/ (ก)	
	0.4736	0.3684	0.1579	0.0001

ดี =

/d/ (ด)	/n/ (น)	/b/ (บ)	/j/ (ย)	
	0.5	0.4117	0.0882	0.0001

The word model of สวัสดี is metric of [n, 4], where n is number of syllables.

	1	2	3
1	สะ 1	หวัด 1	ดี 1
2	จะ 0.5151	หัด 0.4736	นี 0.5
3	ฝะ 0.303	บัด 0.3684	บี 0.4117
4	ชะ 0.1818	กัด 0.1579	ยี 0.0882

Table 5: Word Model Matrix

3. Word Segmentation Algorithm

Word segmentation algorithm use for separating words from connected speech in a sentence. Word segmentation algorithm is searching by the group of word model. Let

N = # of word model.
T = a given sentence to be word segmented.
Tij = a word of T starting at first position of word to end position of word.
PD = a word model in Thai Ambiguous Phonetic Dictionary (AMPD)
G = syllable group (length of syllables)

The algorithm consists of three steps as follows:

```
1.  if G not in 1st group ,
        For G = 4 to 2, searching from four syllables to two syllables
            For each Ti, i=1,...,n, find a word, in G group of PD satisfying the following conditions:
                •  Si syllables matches found. Let i' = i-1+length of Si. Therefore Wi=Si,i'
                •  Wi is not a word in PD. Go to next Ti
    If G equal 1 syllable, separate all syllables segment by don't search in PD.
```

For example,

1	2	3	4	5	6	7
สะ	หวัด	ดี	นัก	เรียน	ตอน	เช้า

Phonetic speech recognition error:

1	2	3	4	5	6	7
ชะ	บัด	ดี	นัก	เรียน	คอน	เช้า

Generate Word Segment to:

N	G	Word	Wi = Si,i'	AWM
1	3	ชะ-บัด-ดี	S1,3	0.1567
2	2	นัก-เรียน	S4,5	1
3	2	คอน-เช้า	S6,7	0.2948
4	1	ชะ	S1,1	0.1818
5	1	บัด	S2,2	0.3684
6	1	ดี	S3,3	1
7	1	นัก	S4,4	1
8	1	เรียน	S5,5	1
9	1	คอน	S6,6	0.1785
10	1	เช้า	S7,7	1

Let

WM = value of each syllable from word model matrixes (table 4)
AWM = Average Word Model

$$AWM = \frac{\sum WM_{i,i'}}{G}$$

2. From the example in step 1, we can construct the corresponding overlapping graph as shown in Figure1.

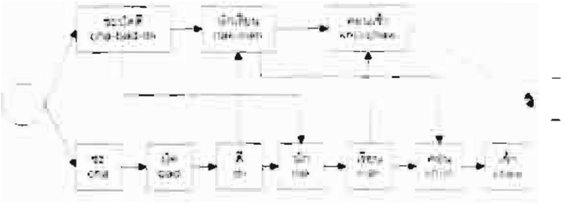


Figure1: Word Segment Overlapping Graph

3. For each component of the graph, find the average value for all paths in graph.

No.	Word Segmented	NWS	AWS
1	ชะบัดดี | นักเรียน |คอนเช้า	3	0.701983333
2	ชะบัดดี | นักเรียน |คอน | เช้า	4	0.6292
3	ชะบัดดี | นัก เรียน |คอนเช้า	4	0.7764875
4	ชะบัดดี | นัก , เรียน |คอน | เช้า	5	0.70336
5	ชะ| บัด| ดี | นักเรียน |คอนเช้า	5	0.51789
6	ชะ| บัด| ดี | นักเรียน |คอน | เช้า	6	0.50005
7	ชะ| บัด| ดี | นัก | เรียน |คอนเช้า	6	0.598241667
8	ชะ| บัด| ดี | นัก | เรียน |คอน |เช้า	7	0.571471429

Table6: Average values of Word Segment

Let NWS = Number of word segment

$$\textit{Average Word Model (AWS)} = \frac{\sum_{1}^{NWS} AWM}{NWS}$$

4. Averages likelihood

From the word segmentation algorithm can obtained a set of words that know and unknown meaning. So Average Likelihood give the solution for correction unknown word meaning to be the real word in Thai meaning system.

```
if G not in 1st group
    for each Word Segmented (Table6) in path
        if AWM = 1
            • Give word correction value (WCV) to be 1
        Else
          For i=1 to 4
            • Change wrong syllable to be syllable of [1,i] in word model matrixes
            • Get the correction syllable
            • Give word correction value (WCV) to be 1
          Next
        Next
Else
    If AWM =1,
            • Give word correction value (WCV) to be 1
    Else
       For i=1 to 4
         if [i,1] = syllable value of [1,1] with a syllable word model matrixes
            • Give word correction value (WCV) to be AWM of [i,1]
            • Get the correction syllable
         Else not exist in [i,1]
            • Give the word correction value (WCV) to be 0.0001 for unknown word meaning
       Next
```

$$\textit{Average Word Correction (AWC)} = \frac{\sum WCV}{NWS}$$

$$\textit{Average Max Likelihood (AML)} = Max(AWC * AWS)$$

So AML gives a Maximum value of Word Segmentation path that the syllables have to correction unknown word to be the know word in Thai meaning system.

5. Experimental Result

The experimental result is mean by using word segmentation accuracy rate (WSAR), Thai word correction accuracy rate (WCAR) and correction rate (CR) is shown the value of performance system.

The initial experiment is base on the following condition:

- 500 Thai word model in Thai word meaning system
- 1,500 ((500*4)-500) for Thai Phoneme word model
- 4 maximum syllables for each word
- 4 groups of word model
- 9 rules of Thai Rule Base System.

The experiment was conducted according to the following steps:

1. Maximum 15 Syllables for each sentence.
2. Thai Ambiguous Phonetic Dictionary was setup the word matrixes model in the discussion of AMPD in section 2.1.5.
3. The Word Segment Algorithm was applied for separating all syllables to be segment of words and created likelihood word segment overlap graph.
4. User Average Likelihood Algorithm to correction unknown Thai words meaning to be know Thai word meaning.

5. The experimental result is shown in Table 7.

#of Syllables	#of Thai Word	# of Word Segment	WSAR(%)	WCAR(%)	CR(%)
15	9	8	88.89	60	100
10	10	10	100	80	90
10	5	4	80	60	100
6	2	2	100	50	100
6	3	2	66.67	66.7	100
3	1	2	50	33.34	100

Table7: The Experimental Result

6. Conclusion

The proposed system is an attempt to provide Thai Syllabic Connected to integrated with the empirical result of syllable recognition process error.

The system was corrected unknown Thai word to be Thai word meaning by Thai Ambiguous Phonetic word model and get the maximum value of Average Likelihood to separated the connected Thai syllables to be segments of a word.

The correction rate is dependent upon the process of determining appropriate sample cycles, which plays the most important role in this system. Applying Thai Rule-Base System with Thai Ambiguous Phonetic System to create the matrixes of word model, can give the knowledge base of unknown word model in each matrixes. Word Segment algorithm and Average Likelihood is also important to cutoff the connected syllable in Thai speech recognition to be segment of Thai word meaning.

All the processes are being tested with data representing subtle features of Thai word. Once the system is accomplished, the overall accuracy of approximately 90% is expected.

7. References

[1] Cheirsilip Ronnarit, Santiprabhob Pratit, *Phoneme-based Thai Speech Recognition System Using Fuzzy System and Neural Network*, IC-AI'2000, July 2000.

[2] Chalireerat Jirawat, Santiprabhob Pratit, *Thai Syllable Segmentation For Connected Speech With Fuzzy System*, IC-AI'2000, July 2000.

[3] Jaruskulchai Chuleerat, *An Automatic Indexing For Thai Text Retrieval*, The School of Engineering and Applied Science, George Washington University, July, 1998.

[4] "อภิษา ศาลา, ลลิตา นฤปิยะกุล บุญเจริญ ศิริเนาวกุล", *การออกเสียงคำหลายพยางค์ในภาษาไทยในคอมพิวเตอร์*, วารสารวิจัยและพัฒนา มจธ ปีที่23 ฉบับที่ 1 มกราคม-เมษายน 2543

[5] Zheng Fang, Wu Jian, Song Zhanjiang, *The Syllable-Synchronous Network Search Algorithm For Word Decoding In Continuous Chinese Speech Recognition*, IEEE ICASSP, Volume(2), 601-604, September, 1999.

[6] Zheng Fang, Wu Jian, Song Zhanjiang, *Improving The Syllable-Synchronous Network Search Algorithm For Word Decoding In Continuous Chinese Speech Recognition*, J. Computer Science & Technology, 15(5), 461-471, September, 2000.

[7] Witoon Kanlayanawat and Somchai Prasitjutrakul, *Automatic indexing for Thai text with unknown words using trie structure*, In Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'97), pages 115--120, Phuket, Thailand, 1997.

[8] S. Charnyapornpong, *A Thai Syllable Separation Algorithm*, M.Eng Thesis, Asian Institute of Technology, Aug. 1983.

[9] A. Kawtrakul, C. thumkonon, and S. Seriburi, *A Statistical Approach to Thai Word Filtering*, Proc. Of the second Symposium on Natural Language Processing, pp. 398-406, 1995.

[10] Y. Poovoranwan and V. Imarom, *Dictionary-base Thai Syllable Segmentaion (in Thai)*, 9th Electrical Engineering Conference, 1986.

[11] Karoonboonyanan, T., Somlertlamvanich, V. and Meknavin, S., *A Thai Soundex System for Spelling Correction*, Proceeding of the National Language Processing Pacific Rim Symposium 1997, 1997, pp. 633-636.

Thai Word Decoder Based on Genetic Algorithm

Wanna Supasirirojana
Intelligent Systems Laboratory
Faculty of Science and Technology
Assumption University
Bangkok, Thailand
E-mail: g4219711@au.ac.th

Pratit Santiprabhob
Intelligent Systems Laboratory
Faculty of Science and Technology
Assumption University
Bangkok, Thailand
E-mail: pratit@s-t.au.ac.th

Abstract: Word decoding plays an important role in connected speech recognition. There are many decoding techniques based on a language model. This paper proposes to use an alternative technique that does not require a language model. The authors employ genetic algorithm as decoding technique to identify the most likely word sequence from the Thai syllable sequence input. The fitness function used in this genetic algorithm is based on the Thai language rule on ambiguous phonetic and empirical result of syllable recognition process.

Key words: Thai Word, Decoding, Genetic Algorithm

1. Introduction

In the field of speech recognition, the decoder is a core of recognition and it is time consuming process. There are many techniques of decoding algorithm develop in the connected speech recognition system.

From the review of decoding technique, many researchers used language model as a knowledge resource input into the decoder. Syllable network and word search tree are used in [2,.3], N-Gram base language model is use in [1,2,3]. The viterbi-decoding algorithm [1,2,3] and network search algorithm [1,2,3] are used as a search algorithm.

This paper proposes a new framework of decoding technique base on genetic algorithm [5] that does not require a language model. The propose decoding technique finds the most likely word sequence from an input syllable sequence that is a result of syllable recognition process. The process generates a set of candidate word sequence to be a solution by picks up a sequence of random word that satisfies the initial population condition. A selection and crossover operation are applied to create a new candidate word sequence in a next generation. The evaluation function for the candidate is fitness function. Fitness function is calculated as by using matrix base on ambiguous degree and Thai language rule.

2. Thai word domain

The proposed algorithm in this research is applied within the boundary of Thai word domain. Word domain composes of words in Thai dictionary and word pronunciation. Each word has 4 maximum syllables. An ID is assigned to each word in order to make it easy to reference form decoding algorithm. An example of Thai words is shown in table 1.

Thai syllable composes of 4 phonemes:

- Leading (L): there are 38 different Thai leading consonants
- Vowel (V): there are 24 different Thai vowel
- Ending (E): there are 9 different Thai ending consonants
- Tone (T): there are 5 different Thai tone

Id	Word	Pronounce (LVET)
7	กำลัง (kam -la ŋ)	6 9 5 1; 18 9 7 1
8	กำลังใจ (kam -lɜ ŋ-cai)	6 9 5 1; 18 9 7 1; 14 9 8 1
11	เก็บ (kep)	6 4 1 2
61	ตลอด (ta-l)t)	5 9 4 2; 18 20 2 2
62	ตลาด (ta-la:t)	5 9 4 2;18 21 2 2
158	เศรษฐกิจ (se:t-tha-kit)	11 16 2 2;2 9 4 2;6 1 2 2
163	สนับสนุน (sa-nap-sa-nun)	11 9 4 2;16 9 1 2;11 9 4 2;16 3 6 5
187	เหตุผล (he:t-phon)	12 16 2 2;1 6 6 5

Table 1: Example of Thai Word Domain.

3. Ambiguous matrix

Thai Ambiguous matrix is base on the Thai language rule on ambiguous phonetic and empirical result of syllable recognition process [6]. There are 4 groups of Thai ambiguous phonetic: ambiguous leading (ALM), Ambiguous vowel (AVM), ambiguous ending (AEM) and ambiguous tone (ATM)

In this research, there are 16 ambiguous leading matrix based on a couple of vowels and the other 3 matrixes for ambiguous vowel, ambiguous ending and ambiguous tone.

	/th/ (ท)	/kh/ (ค)	/k/ (ก)	/h/ (ห)
/ph/ (พ)	32	5	1	2

Table 2: Value of Ambiguous Leading /ph/(พ).

	/ph/ (พ)	/kh/ (ค)	/k/ (ก)	/h/ (ห)	/ch/ (ช)
/th/ (ท)	25	5	2	3	10

Table 3: Value of Ambiguous Leading /th/ (ท).

The partial results of ambiguous leading /ph/ (พ) and /th/ (ท) with vowel /i/ (อิ) and /i:/ (อี) from syllable recognition process [6] with 40 number of test set are shown in table 2 and table 3 respectably. The row in the table is value of ambiguous leading corresponding to the leading in the leftmost column. From talble 2, there are 40 number of test set of recognition the leading "/ph/ (พ)" 32 of 40 recognize as /th/ (ท) and 4 of 40 recognize as /p/ (ป).

Base on the empirical result of ambiguous leading consonant in table 2 and 3, the ambiguous degree (AD) of each phoneme is calculated by using equation (1).

$$AD = \frac{Value\ of\ ambiguous\ \ phoneme}{total\ number\ of\ test\ set} \qquad (1)$$

From the equation (1), AD 1 is set to 1 for an ambiguous degree of phoneme itself.

The ambiguous degree is a value between 0 and 1. The partial ambiguous matrix for leading consonant /ph/ (พ) and /th/ (ท) with vowel /i/ (อิ) and /i:/ (อี) is shown in table 4.

	/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

Table 4: The partial Ambiguous Matrix for Leading Consonant.

4. Word decoding

The authors propose the decoding technique based on genetic algorithm (GA) to find the most likely Thai word sequence given the Thai syllable sequence. Thai syllable sequence from the syllable recognition process is a string of syllable consists of phonemes number. There are many characteristics of GA used in this research.

4.1. Decoding process

The decoding process is started with a set of word sequences to be solution by initial population (refer to sub-section 4.3). The number of generation and acceptable fitness value are set as a condition to stop a decoding process. The fitness value (refer to sub-section 4.4) is calculated for each word sequence in the population. If fitness value of the current word sequence is not good to be a solution then the new generations of word sequences is generated from selecting two parents (refer to sub-section 4.5) and apply crossover operation (refer to sub-section 4.6).

There are 2 steps of stopping decoding process in this research.

First step is assumed that the input syllable from the syllable recognition process has a 100% recognition rate. The acceptable fitness value is set to 1. The fitness function is using the normal degree as a degree of fitness. The process is stopped when there is a fitness value of word sequence (result) equal to the acceptable fitness value. If more than 80% of word sequence in the current population has the same fitness value then assume that the input syllable from the syllable recognition process has a recognition rate less than 100% so the decoding process is restart and second step is applied

For second step, the fitness function is using the ambiguous degree as a degree of fitness. Second step is stopped when more than 80% of word sequence in the current generation has the same fitness value. Word sequence that has a maximum fitness value is picking up as a result.

4.2. Encoding of word sequence

The authors used permutation encoding as a encoding method to represent each word sequence by a string of word ID. A variable-length encoding scheme is applied because the number of word in word sequence is varying according to the number of syllable of input syllable. An example of encoding of word sequence is shown in figure 1.

Word sequence A: "ลูกจ้าง มี เหตุผล ที่ ไม่ เห็น ด้วย กับ นายจ้าง"
(lu:k-ca:ŋ mi: he:t-phon thi: maj hen dua:j kap na:j-ca:ŋ)

144	44	128	187	78	130	188	54	4	90

Word sequence B: "โรงเรียน ผลิต นักเรียน ที่ มี คุณภาพ"
(ro:ŋ-ria:n pha-lit nak-ria:n thi: mi: khun-na-phap)

144	44	128	187	78	130	188	54	4	90

Figure 1: Encoding of Word Sequence.

4.3. Initial population

Initial population generates a set of candidate word sequence. Word sequence is created by pick up a random word that satisfy the initial population condition and add that word to the end of current word sequence then pick up the next word and repeat random process until the length of syllable in word sequence is equal to the length of input syllable.

The initial population conditions are

- The random word has either vowel or tone that equal to vowel or tone on the same position of input syllable.
- Length of the word sequence must be equal to the length of input syllable.

4.4. Fitness function

$$FN = \sum_{i=1}^{w}\left(\sum_{j=1}^{n} \text{aDegree}_j\right) \Big/ NS \qquad (2)$$

where w = # of word in chromosome
n = # of syllable in each word
NS = # of syllable in chromosome

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

The fitness function is calculated by using equation (2).

From the equation (2), the ambiguous degree of each phoneme can be obtained from ambiguous matrix as describe in section 3. The normal degree is the set to 1 if syllable at j is equal to syllable at the same position of input syllable, otherwise it is set to 0. The ambiguous degree is used to measure the similarity of the input syllable and candidate word sequence.

4.5. Selection and crossover point

Two word sequences (father, mother) are random selected from the population by using roulette wheel as a selection method. The random mother is not accepted as a new mother until it is verified as a good mother.

Mother is verified by choosing the crossover points of mother and father that preserve the syllable length of the new word sequences (offspring) as mention in the initial population condition and apply crossover operation (refer to sub-section 4.6) to produce the new word sequences. The fitness function is applied to new word sequences. The mother is good if any of the new word sequence has a better fitness value compare with the parent. A flowchart of selection is shown in figure 2.

4.6. Crossover

Swapping words between two word sequences at the crossover points performs the crossover operation. New word sequences from crossover operation are applied with fitness function. The new word sequence with the better fitness value is put in a new population. An example of crossover is shown in figure 3.

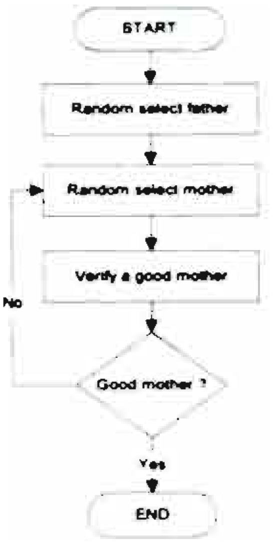


Figure 2: Selection Flowchart.

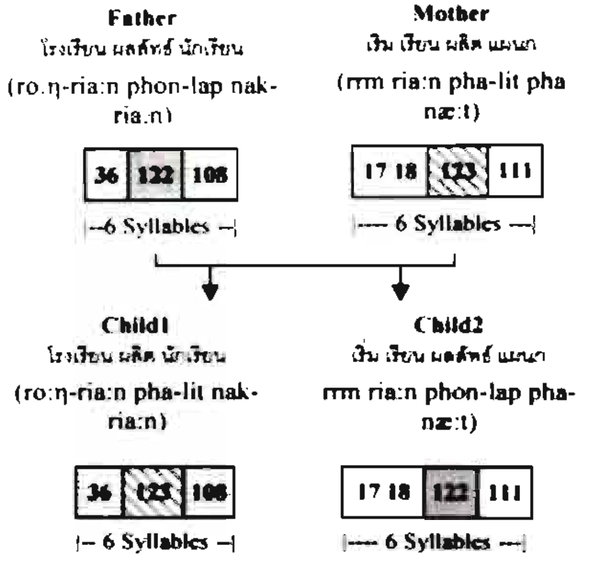


Figure 3: An Example of Crossover Operation.

5. Experimental result

The experimental result is measure by using decoding accuracy rate (DAR) and correction rate (CR). The number of word that the system decodes correctly measures DAR. Correction rate (CR) is number of word that the system makes it correct from the error of input.

The initial experiment is based on the following condition:

- 220 Thai words in Thai word domain.

- 4 maximum syllables for each word.
- 4 group of ambiguous matrix base on [6].

The authors perform 2 experiments from the above conditions.

First experiment is using 200 input syllables with 100% recognition rate from syllable recognition process [6]. The example graph display fitness value for each generation is shown in figure 4. Table 5 is shown the DAR of the experiment.

Initial Condition	DAR
400 initial population size, 12 generations.	100%
200 initial population size, 12 generations.	97%
200 initial population size, 5 generations.	80%

Table 5: Accuracy Rate of Experiment 1.

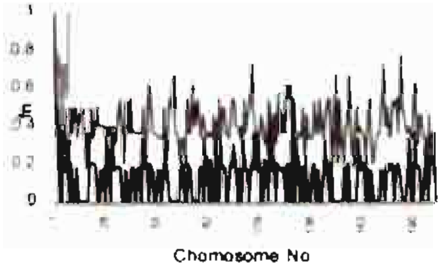


Figure 4: Fitness Value of Experiment 1.

Second experiment is using the different 200 input syllable sequences with a recognition rate from syllable recognition process [6] less than 100%. From the experiment, 85% of word sequence in the 6th-generation has the same fitness value as shown in figure 5. The system recognizes that there is some error with input syllable so the second step of decoding process is used. DAR is 95% and CR is 85%.

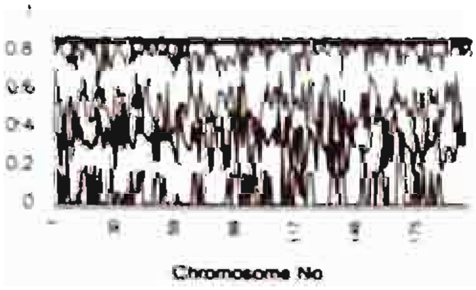


Figure 5: Fitness Value of Experiment 2.

6. Conclusion

This paper proposed the alternative decoding technique that does not require a language model by using empirical result of syllable recognition process.

From the decoding technique being test, the population size, number of generation and recognition rate of the input syllable are factors that effect to the accuracy rate. The population size should be set according to the number of word in word domain. A small number of generations cause the low accuracy rate because in some case the solution is found in the next generation from the maximum generation. The accuracy rate is decrease when an input syllable from a syllable recognition process has recognition rate less than 75% per syllable.

The decoding technique used in this paper is a time-consuming process. Further research should be focus on this problem by applied the mutation operator and considering the condition of selecting a candidate solution in the initial population.

7. References

[1] Jie. Zhao, *Network and N-Gram Decoding in speech recognition*, Mississippi State University, October, 2000.

[2] Zheng Fang, Wu Jian, Song Zhanjiang, *The Syllable-Synchronous Network Search Algorithm For Word Decoding In Continuous Chinese Speech Recognition*, IEEE ICASSP, Volume(2), 601-604, September, 1999.

[3] Zheng Fang, Wu Jian, Song Zhanjiang, *Improving The Syllable-Synchronous Network Search Algorithm For Word Decoding In Continuous Chinese Speech Recognition*, J. Computer Science & Technology, 15(5), 461-471, September, 2000.

[4] Neeraj Deshmukh, *Decoder Strategies, Institute for Signal and Information Processing*, Mississippi State University, 1997.

[5] Lawence Davis, *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York , 1991

[6] Cheirsilip Ronnarit, Santiprabhob Pratit, *Phoneme-based Thai Speech Recognition System Using Fuzzy System and Neural Network*, IC-AI'2000, July, 2000.

[7] K T Lua , *Part of Speech Tagging of Chinese Sentences Using Genetic Algorithm*, Chinese Computing Conference, 45-49,June, 1996.

[8] Hugo M. Ayala, *Natural Language Generation Using Genetic Algorithms*, MIT Department of Mechanical Engineering,1995

A Framework for Connected Speech Recognition for Thai Language*

Pratit Santiprabhob[1], Jirawat Chaiareerat[2], Ronnarit Cheirsilp[3],
Nunmanus Dachapratumvan[4], and Wanna Supasirirojana[5]
Intelligent System Laboratory
Department of Computer Science
Faculty of Science and Technology
Assumption University
Bangkok, 10240,Thailand

Email: pratit[1], jirawat[2], ronnarit[3] @s-t.au.ac.th, g4219712[4], g4219711[5]@au.ac.th

Abstract: Connected speech recognition problem for Thai language, like the similar problem in other languages, involves three sub-problems: 1) syllable segmentation, 2) syllable recognition and 3) 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 tone. 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: Speech Recognition, 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 Figure 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 [1], Conventional Neural Network [2], Modified Back Propagation Neural Network [3], Neural Network with Fuzzy MF Preprocessor [4] and Hidden Markov Model [5]. From the studies in [6] and [7], the Hidden Markov Model (HMM) as used in [5] 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.

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

1) 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.
2) The approach outlined in [5] 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.
3) 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.

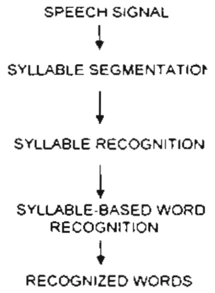


Figure 1: Conceptual Framework for Connected Speech Recognition

2. Framework Architecture

The proposed framework consists of three parts: firstly, syllable segmentation which is described in Section 2.1, secondly, syllable recognition which is outlined in Section 2.2, and thirdly, syllable-based word recognition whose two alternatives are discussed in Section 2.3.

2.1. Syllable Segmentation

The segmentation algorithm used in our framework is based on the concepts of energy and Different Ceptral as explained in [8]. 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.

2.1.1. Parameters Computation

First, the speech signal is pre-processed by means of signal pre-emphasizing technique as described in [9]. The signal is then en-framed into 30 milliseconds long frames with 20 millisecond 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 [8]. 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 [10].

2.1.2. 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.

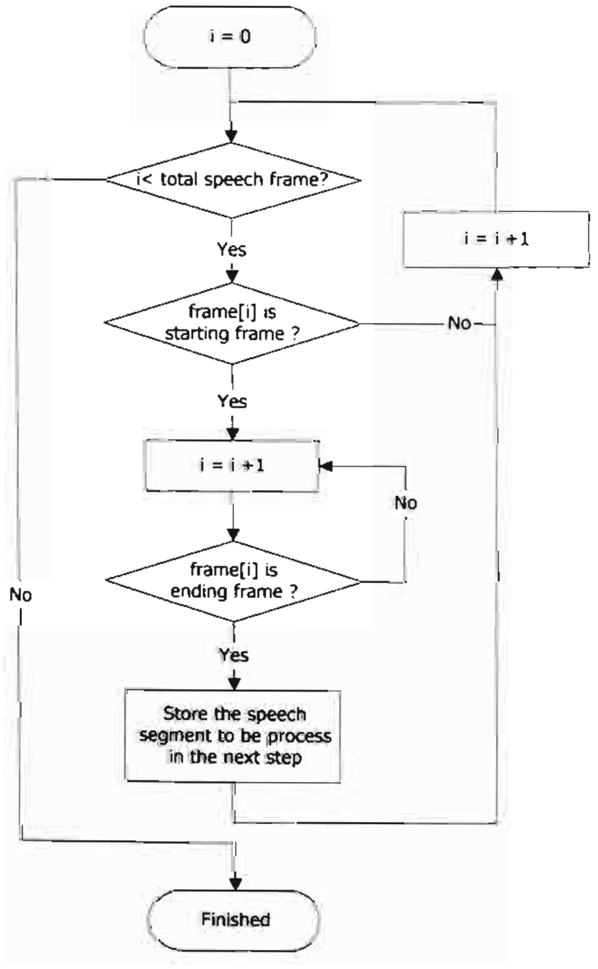
The algorithm is depicted in Figure 2. The results obtained in this step are the speech segments, which will further be segmented into syllables in next step.

2.1.3. 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 [8].

For each Emin frame, five fuzzy input variables are defined, namely 1) the absolute energy of the current Emin frame – EM, 2) the minimum ZCR between the two surrounding PeakE frames, 3) the difference between the EM and the absolute energy of the preceding PeakE frame – DEL, 4) the difference between the EM and the absolute energy of the following PeakE frame – DER, and .5) 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 [8]. Here, Mamdani-type FIS with centroid defuzzification method [11] 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.



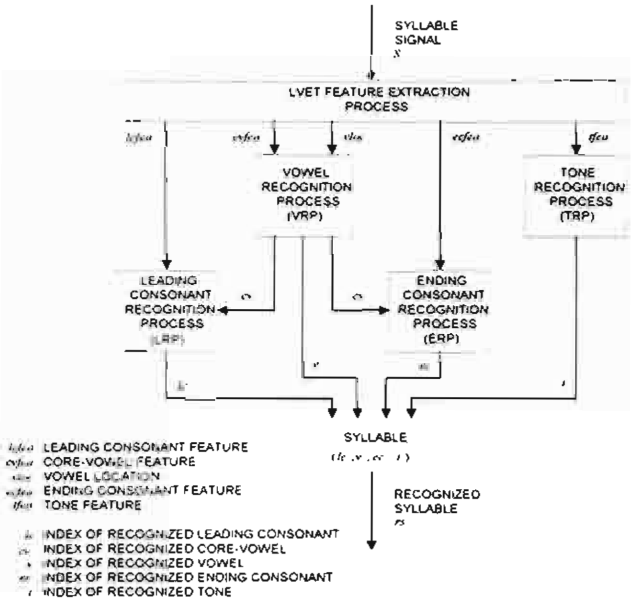
i = 0
i< total speech frame?
Yes
No
frame[i] is starting frame ?
No
i = i +1
Yes
i = i +1
No
frame[i] is ending frame ?
Yes
Store the speech segment to be process in the next step
Finished

Figure 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 1) leading consonant, vowel, ending consonant and tone (LVET) feature extraction process, 2) leading consonant recognition process (LRP), 3) vowel recognition process (VRP), 4) ending consonant recognition process (ERP) and 4) tone recognition process (TRP). A block diagram of the overall recognition system which is described in detail in [12] is given in Figure 3.



SYLLABLE SIGNAL
x
LVET FEATURE EXTRACTION PROCESS
lcfea
cvfea
vlos
ecfea
tfea
VOWEL RECOGNITION PROCESS (VRP)
TONE RECOGNITION PROCESS (TRP)
LEADING CONSONANT RECOGNITION PROCESS (LRP)
ENDING CONSONANT RECOGNITION PROCESS (ERP)
cv
cv
v
ec
lc
t
SYLLABLE
RECOGNIZED SYLLABLE
rs
lcfea LEADING CONSONANT FEATURE
cvfea CORE-VOWEL FEATURE
vlos VOWEL LOCATION
ecfea ENDING CONSONANT FEATURE
tfea TONE FEATURE
lc INDEX OF RECOGNIZED LEADING CONSONANT
cv INDEX OF RECOGNIZED CORE-VOWEL
v INDEX OF RECOGNIZED VOWEL
ec INDEX OF RECOGNIZED ENDING CONSONANT
t INDEX OF RECOGNIZED TONE

Figure 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 [9] and [13] is conducted to determine the Ceptral Coefficient and energy feature vector, so-called CEP_E feature vector. In addition, fundamental frequency contour [14] 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 [12]

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 Figure 4.

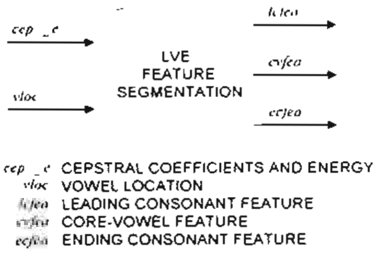

cep _e
vloc
LVE FEATURE SEGMENTATION
lcfea
cvfea
ecfea
cep _e CEPSTRAL COEFFICIENTS AND ENERGY
vloc VOWEL LOCATION
lcfea LEADING CONSONANT FEATURE
cvfea CORE-VOWEL FEATURE
ecfea ENDING CONSONANT FEATURE

Figure 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 Figure 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.

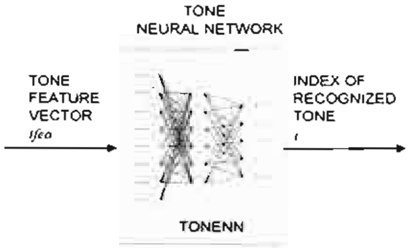


Figure 5: A block diagram of TRP.

2.2.3. Vowel Recognition Process (VRP)

In this process, vowel phonemes are recognized. In order to recognize vowels two elements must be determined, 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 Figure 6. The process is divided into the core vowel recognition part and the vowel length determination part, which are further discussed below.

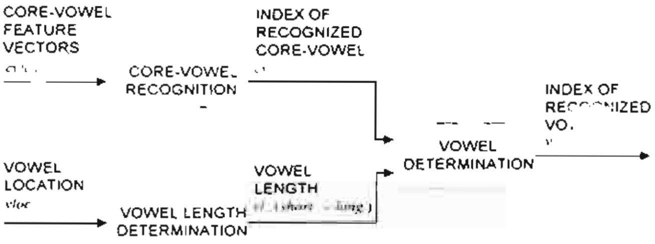


Figure 6: A block diagram of VRP.

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 [9][13].

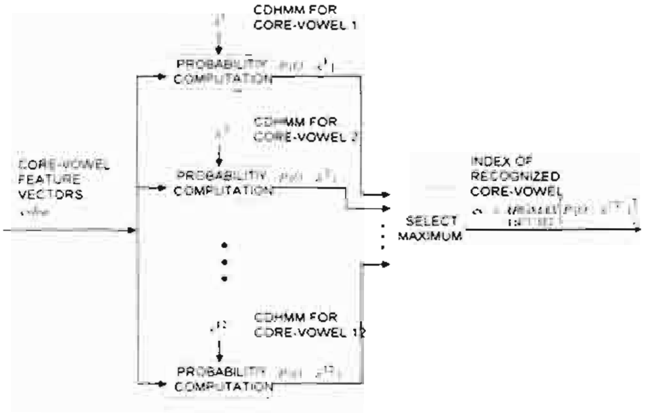


Figure 7: A block diagram of core-vowel recognition

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 Figure 8.

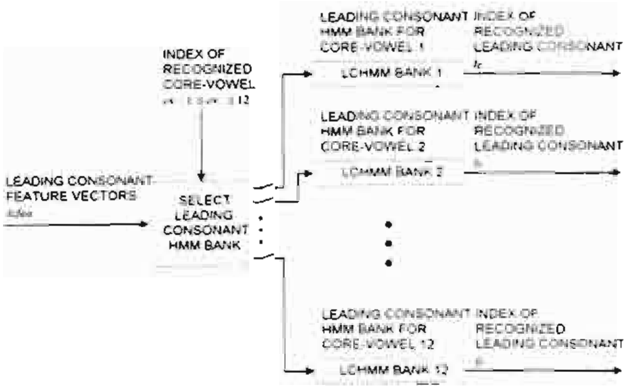


Figure 8: A block diagram of LRP.

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. Figure 9 shows a block diagram of an LCHMM bank.

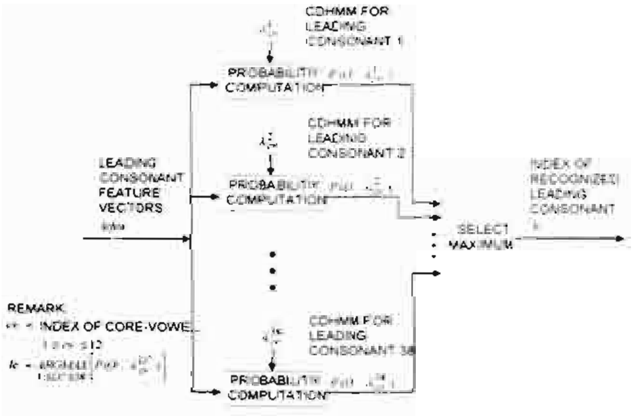


Figure 9: A block diagram of an LCHMM bank.

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 Figure 10.

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 Figure 11.

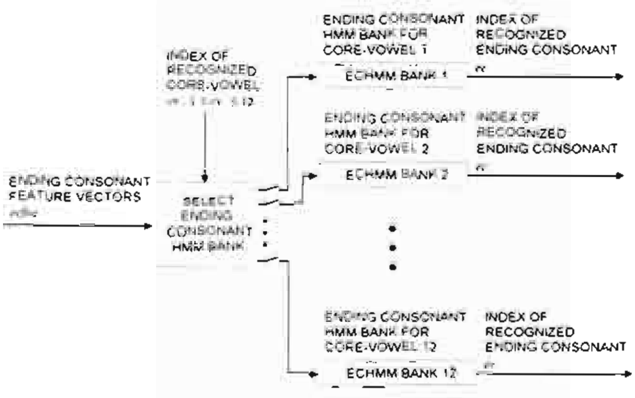


Figure 10: A block diagram of ERP.

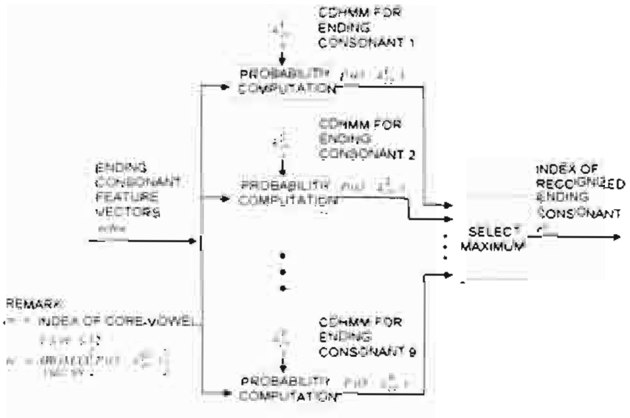


Figure 11: A block diagram of an ECHMM bank.

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 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 Figure 12.

9 /a/ ะ 21 /aa/ า				
[redacted]				
1.[/ph/ ,(พ)]	2.[/th/ ,(ท)] 30 0.7894	3.[/kh/ ,(ค)] 6 0.1578	12.[/h/ ,(ห)] 2 0.0527	38 0.0001
2.[/th/ ,(ท)]	1.[/ph/ ,(พ)] 21 0.5675	3.[/kh/ ,(ค)] 8 0.2162	13.[/ch/ ,(ช)] 8 0.2162	37 0.0001
3.[/kh/ ,(ค)]	2.[/th/ ,(ท)] 26 0.7878	6.[/k/ ,(ก)] 5 0.1515	1.[/ph/ ,(พ)] 2 0.0606	33 0.0001
4.[/p/ ,(ป)]	8.[/b/ ,(บ)] 20 0.6896	10.[/f/ ,(ฝ)] 5 0.1724	1.[/ph/ ,(พ)] 4 0.1379	29 0.0001
5.[/t/ ,(ต)]	13.[/ch/ ,(ช)] 10 0.4545	18.[/l/ ,(ล)] 7 0.3181	2.[/th/ ,(ท)] 5 0.2273	22 0 0001
6.[/k/ ,(ก)]	3.[/kh/ ,(ค)] 12 0.4615	17.[/nj/ ,(ง)] 8 0.3076	21.[/r/ ,(ร)] 6 0.2308	26 0 0001
7.[/?/ ,(อ)]	17.[/nj/ ,(ง)] 13 0.4643	5.[/t/ ,(ต)] 9 0.3213	3.[/kh/ ,(ค)] 6 0.2143	28 0.0001
8.[/b/ ,(บ)]	15.[/m/ ,(ม)] 20 0.5882	4.[/p/ ,(ป)] 9 0.2647	9.[/d/ ,(ด)] 5 0.1470	34 0.0001
9.[/d/ ,(ด)]	19.[/j/ ,(ย)] 13 0.5652	14.[/c/ ,(จ)] 5 0.2174	16.[/n/ ,(น)] 5 0.2173	23 0.0001
10.[/f/ ,(ฝ)]	11.[/s/ ,(ส)] 32 0.8649	20.[/w/ ,(ว)] 4 0.1080	13.[/ch/ ,(ช)] 1 0.0270	37 0.0001
11.[/s/ ,(ส)]	14.[/c/ ,(จ)] 17 0.5151	10.[/f/ ,(ฝ)] 10 0.3030	13.[/ch/ ,(ช)] 6 0.1818	33 0.0001
12.[/h/ ,(ห)]	3.[/kh/ ,(ค)] 10 0.3571	2.[/th/ ,(ท)] 9 0.3214	1.[/ph/ ,(พ)] 9 0.3214	28 0.0001
13.[/ch/ ,(ช)]	14.[/c/ ,(จ)] 15 0.4688	2.[/th/ ,(ท)] 11 0.3437	11.[/s/ ,(ส)] 6 0.1875	32 0.0001
14.[/c/ ,(จ)]	11.[/s/ ,(ส)] 23 0.6216	13.[/ch/ ,(ช)] 10 0.2702	19.[/j/ ,(ย)] 4 0.1081	37 0.0001
15.[/m/ ,(ม)]	17.[/nj/ ,(ง)] 21 0.7000	4.[/p/ ,(ป)] 5 0.1666	8.[/b/ ,(บ)] 4 0.1333	30 0.0001
16.[/n/ ,(น)]	17.[/nj/ ,(ง)] 23 0.6388	9.[/d/ ,(ด)] 11 0.3056	21.[/r/ ,(ร)] 2 0.0555	36 0.0001
17.[/nj/ ,(ง)]	16.[/n/ ,(น)] 18 0.5625	15.[/m/ ,(ม)] 9 0.2812	9.[/d/ ,(ด)] 5 0.1562	32 0.0001
18.[/l/ ,(ล)]	21.[/r/ ,(ร)] 12 0.5217	16.[/n/ ,(น)] 6 0.2608	9.[/d/ ,(ด)] 5 0.2174	23 0.0001
19.[/j/ ,(ย)]	9.[/d/ .(ด)] 37 0.9249	14.[/c/ .(จ)] 2 0.0500	17.[/nj/ .(ง)] 1 0.0250	40 0.0001
20.[/w/ ,(ว)]	4.[/p/ .(ป)] 10 0.4166	8.[/b/ .(บ)] 7 0.2917	17.[/nj/ .(ง)] 7 0.2916	24 0.0001
21.[/r/ ,(ร)]	18.[/l/ .(ล)] 16 0.6399	8.[/b/ .(บ)] 5 0.2000	6.[/k/ .(ก)] 4 0.1600	25 0.0001
22.[/ph/ ,(พล)]	1.[/ph/ .(พ)] 21 0.7241	18.[/l/ .(ล)] 5 0.1723	2.[/th/ .(ท)] 3 0.1034	29 0.0001
23.[/ph/ ,(พร)]	12.[/h/ .(ห)] 21 0.7241	3.[/kh/ .(ค)] 5 0.1723	6.[/k/ .(ก)] 3 0.1034	29 0.0001
24.[/kh/ ,(คล)]	3.[/kh/ .(ค)] 26 0.7878	18.[/l/ .(ล)] 5 0.1515	26.[/kh/ .(คว)] 2 0.0606	33 0.0001
25.[/kh/ ,(คล)]	3.[/kh/ .(ค)] 26 0.7878	18.[/l/ .(ล)] 5 0.1515	26.[/kh/ .(คว)] 2 0.0606	33 0.0001
26.[/kh/ ,(คว)]	3.[/kh/ .(ค)] 26 0.7878	20.[/w/ .(ว)] 5 0.1515	6.[/r/ .(ร)] 2 0.0606	33 0.0001
27.[/kh/ ,(คร)]	3.[/kh/ .(ค)] 26 0.7878	2.[/th/ .(ท)] 5 0.1515	6.[/r/ .(ร)] 2 0.0606	33 0.0001
28.[/p/ ,(ปล)]	4.[/?/ .(ป)] 9 0.3913	8.[/b/ .(บ)] 8 0.3477	18.[/l/ .(ล)] 6 0.2609	23 0.0001
29.[/p/ ,(ปร)]	4.[/?/ .(ป)] 9 0.3913	8.[/b/ .(บ)] 8 0.3477	2.[/th/ .(ท)] 6 0.2609	23 0.0001
30.[/t/ ,(ตร)]	5.[/t/ .(ต)] 16 0.5000	11.[/s/ .(ส)] 9 0.2812	21.[/r/ .(ร)] 7 0.2187	32 0.0001
31.[/k/ ,(กล)]	33.[/k/ .(กร)] 12 0.4615	27.[/kh/ .(ค)] 8 0.3076	18.[/l/ .(ล)] 6 0.2308	26 0.0001
32.[/k/ ,(กว)]	6.[/k/ .(ก)] 12 0.4615	27.[/kh/ .(ค)] 8 0.3076	20.[/w/ .(ว)] 6 0.2308	26 0.0001
33.[/k/ ,(กร)]	27.[/kh/ .(ค)] 12 0.4615	17.[/nj/ .(ง)] 8 0.3076	21.[/r/ .(ร)] 6 0.2308	26 0.0001

Figure 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 Figure 13.

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

Figure 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 [15]. It can be summarized in three steps as follows.

1) 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.
2) Construct a graph containing all possible combinations of word models in the given sentence. An example of such a graph is shown in Figure 14.
3) For each path (combination) of word models in the graph, calculate the average recognition probability of the path.

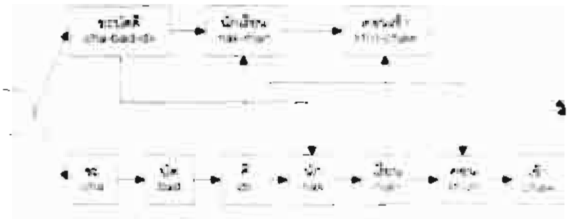


Figure 14: An example of a graph containing all possible combinations of word models

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 [15]. 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) [16].

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 Figure 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

Figure 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

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

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

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

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 Figure 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 [17].

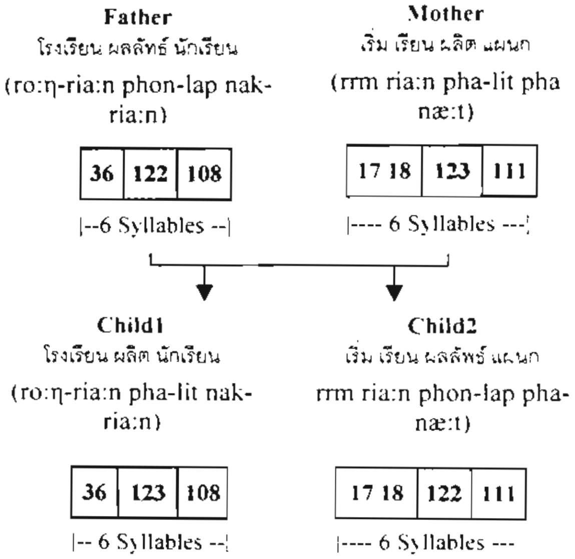


Figure 16: An example of the crossover operation

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

[1] Pensiri, R. and Jitapunkul, S., *Speaker-Independent Thai Numerical Voice Recognition by using Dynamic Time Warping*, Proceedings of the 18th Electrical Engineering Conference, pp. 977-981, 1995.

[2] Pornsukchandra, W. and Jitapunkul, S., *Speaker-Independent Thai Numeral Speech Recognition using LPC and the Back Propagation Neural Network*, Proceedings of the 19th Electrical Engineering Conference, pp. 977-981, 1996.

[3] Maneenoi, E., Jitapunkul, S., Wutiwuwatchai, C., and Ahkuputra, V., *Modification of BP Algorithm for Thai Speech Recognition*, Proceeding of the 1997 International Symposium on Natural Language Processing, 1997.

[4] Wutiwiwatchai, C., *Speaker Independent Thai Numeral Speech Recognition Using Neural Network and Fuzzy Technique*, Master's thesis, Chulalongkorn University, 1997.

[5] Ahkuputra, V., Jitapunkul, S., Pornsukchandra, W., and Luksaneeyanawin, S., *A Speaker-Independent Thai Polysyllabic Word Recognition Using Hidden Markov Model*, Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, pp. 593-599, 1997.

[6] Ahkuputra, V., Jitapunkul, S., Maneenoi, E., Kasuriya, S., and Amornkul, P., *Comparison of Different Techniques On Thai Speech Recognition,* The 1998 IEEE Asia-Pacific Conference on Circuits and Systems, pp. 177-180 1998.

[7] Jitapunkul, S., Luksaneeyanawin, S., Ahkuputra, V., Maneenoi, E., Kasuriya, S., and Amornkul, P., *Recent Advances of Thai Speech Recognition in Thailand,* The 1998 IEEE Asia-Pacific Conference on Circuits and Systems, pp. 173-176, 1998.

[8] Chaiareerat, J. and Santiprabhob P., *Fuzzy-based Thai Syllable Segmentation for Connected Speech using Energy and Different Ceptral,* Proceedings of InTech/VJFuzzy, pp.334-337, December, 2002.

[9] Rabiner, L. R. and Juang, B. H., *Fundamentals of Speech Recognition,* A. Oppenheim, Series Editor, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[10] Jittiwarangkul N., Jitapunkul S., Luksaneeyanawin S., Ahkuputra V., Wutiwiwatchai C., *Thai Syllable Segmentation for Connected Speech based on Energy,* Proceedings of the IEEE APCCAS, pp. WP1-8.1, Nov. 1998.

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

[12] Cheirsilp, R. and Santiprabhob P., *Phoneme-Based Thai Syllable Recognition by Means of Soft Computing,* Proceedings of InTech/VJFuzzy, pp.325-333, December, 2002.

[13] Rabiner, L. R., *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,* Proceeding of IEEE, Vol. 77, No.2, pp. 257-286, February, 1989.

[14] Rowden, C., *Speech Processing,* London: McGraw-Hill, 1992.

[15] Dachapratumvan, N. and Santiprabhob P., *Thai Syllablic Correction in Connected Thai Speech Recognition,* Proceedings of InTech/VJFuzzy, pp.314-319, December, 2002.

[16] Lawrence D., *Handbook of Genetic Algorithms,* Van Nostrand Reinhold, New York, 1991.

[17] Supasirirojana, W. and Santiprabhob P., *Thai Word Decoder Based on Genetic Algorithm,* Proceedings of InTech/VJFuzzy, pp.320-324, December, 2002.

โครงการ “การเรียนรู้จำเสียงพูดภาษาไทยโดยใช้นิวโรฟัซซี่”

Thai Speech Recognition using NeuroFuzzy