Final Consonant Segmentation for Thai syllable by Using Vowel Characteristics and Wavelet Packet Transform

Kosin Chamnongthai[1], Member,
Wudthipong Pichitwong[2], and **Piyasawat Navaratana Na Ayudhya**[1], Non-members

ABSTRACT

Since Thai final consonant is unique comparing with other languages and plays key role in recognizing the Thai syllables, segmentation of the final consonant phoneme from the vowel is needed and capable of decreasing the amount of recognition patterns and also improving the recognition accuracy. This paper presents a technique to separate the final consonant phoneme from Thai syllable by exploiting the vowel characteristics and Wavelet packet transform. In this method, ending of the vowel phoneme (starting of the final consonant) is considered by vowel characteristic, which has the highest energy in the syllable. The frequency range having this qualification is selected as vowels. It is then employed to determine the filter for vowel signal. The Wavelet packet transform that is appropriate for discriminating vowel (high frequency and long period) from final consonant phoneme (low frequency and short period) is used as the filter. And the ending of vowels frequency signal component is considered to be the segmentation point of the final consonant. The experiments have been performed by 4,350 samples of syllable recorded from 15 males and 15 females. The experimental results gained the 92.89 % accuracy.

Keywords: Segmentation , Phoneme , Final Consonant , Wavelet Packet Transform , Vowel Characteristics

1. INTRODUCTION

According to Thai grammar, a Thai syllable consists of an initial consonant (in this paper, called consonant), a vowel, a tone and a final consonant. Segmenting a syllable into these components can be considered as an sufficient approach for the syllable recognition,intonation and accent analysis, and so on. Thai speech consists of 21 consonants, 18 vowels, five tones and nine final consonants. Therefore, 17,020 patterns (21 consonants × 18 vowels × 5 tones × 9 final consonants) of syllables are possible to be employed. In case of using segmented components for recognition, segmenting into consonant, vowel, tone and final consonant needs to register only 53 speech patterns (21 consonants + 18 vowels + 5 tones + 9 final consonants) in the memory. However, if we do not segment final consonant from syllable, or segment a syllable only into a consonant, a vowel and a tone with one of nine final consonants, the patterns for registering may increase up to 7.5 times approximately, or 396 (= (21 consonants +18 vowels + 5 tones) × 9 final consonants)patterns. The more patterns we have to classify, the more recognition errors we have to meet with. In case of intonation and accent analysis, since Thai final consonant is an important key to determine the meaning, the correct intonation and accent of final consonant specially in term of signal is required in order to train the intonation and accent. Therefore, segmenting vowel and final consonant clearly makes bigger tolerance in classifying and takes advantage in Thai speech recognition. However, since the signal component of final consonant is a very small signal that normally seem to merge as a tail of vowel in time domain, it is hard to correctly detect a segmenting point between vowel and final consonant. We should seriously consider the appropriate segmentation techniques in order to prevent the recognition errors caused by mistaken final consonant segmentation. Moreover, for Thai speech, final consonant is located in the frequency range below the one of vowel, and has more overlap frequency range with vowel comparing with other languages like English. Therefore, in viewpoints of signal in time-frequency domain, it is complicated problem for exactly determining the border of vowel and final consonant in Thai language. Nowadays, there is no appearance of research report contributing in Thai final consonant detection yet.

The related researches pursuing the speech segmentation are as follows. Hanes et al.[1] proposed a technique using formant contour mapping. In this work, the whole phoneme is divided into three overlapped parts i.e. initial consonant, vowel and final consonant, and then these divided parts are processed in recognition process. Although this method can recognize each part of phoneme, it can not absolutely

SP3R24: Manuscript received on March 30, 2004 ; revised on May 7, 2004.
[1]The authors are with King Mongkut's University of Technology Thonburi (KMUTT),Bangkok, Thailand.
[2]The author is with Rachmangala Institute of Technology, North-East campus, Nakornrachsima, Thailand.

segment the final consonant from the vowel. Beng et al.[2] proposed a method for segmenting final consonant by sampled continuous Wavelet transform. He applied this method for segmenting final consonant /s/ in English which is constantly in the same high-frequency region. However, this method is considered to be unable to apply for Thai language, since Thai final consonants are in low frequency, and located in the same frequency range as some parts of vowels. Thanwa Sripramong et al.[3] used the harmonic frequencies to plot the spectrogram and then investigated the duration of each phoneme. By this method, the final consonant can be categorized as either the open or closed syllable. However, all above approaches can only recognize the final consonant as open or closed groups. They still employ a lot of patterns in training and recognition phases, which causes large overheads. Dumrongpati et al.[4] proposed a method for segmenting consonant from vowel by using discrete Wavelet. In this method, the differences between consonant and vowel in low frequency range are employed. However, it is not suitable and easy for Thai vowel and final consonant segmentation, because there is much information of vowel in the same range. The authors [5] also tried to segment continuous speech into syllables by using Quasi-periodic reconsidering based on zero-crossing. This method can segment the syllables of fast-speaking continuous speech. However, by the limitation of window size and processing time, it is inconvenient for final consonant segmentation.

In this paper, we tried to find appropriate technique for Thai final consonant segmentation. A new technique used to separate the final consonant phoneme from Thai syllable by exploiting the vowel characteristics and Wavelet packet transform is presented. In this technique, since amplitude of final consonant is minute comparing with those of vowel, and the neighboring component in a syllable, we approach to detect the final consonant by extracting vowel, and determining the ending point of vowel as the beginning of final consonant. In extracting vowel, since energy of vowel is physically the highest in the syllable, the highest energy component is picked up as vowel component, and the highest frequency of vowel component is then selected as the main of frequency range of the clearest voice component of vowel. To filter the frequency range of vowel, in this paper, a filter designed by selected frequency range using Wavelet packet is applied as an appropriate tool in term of detail analysis in small frequency range. The signal component filtered by the designed filter is converted into the one in time domain as vowel component, and the ending point of vowel is scanned in order to determine the beginning of final consonant. Finally, the signal of final consonant is extracted from the syllable from the determined starting point.

The paper proceeds as follows. Section 2 mentions about analysis of Thai final consonant. Section 3 describes an idea for segmenting final consonant. Section 4 discusses the proposed method. Then the experiments and results are mentioned in Section 5. And the experimental results including evaluation are discussed in Section 6.

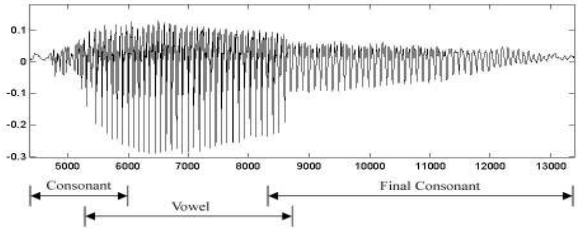

Fig.1: *A sample of "?a:/n/" signal in time domain*

2. ANALYSIS OF THAI FINAL CONSONANT

In Thai language [6], a syllable (S) consists of a consonant (C), vowel (V), tone (T) and final consonant (F) as follows.

$$S = CVTF \tag{1}$$

In term of speech, the final consonant is a speech component as same as a consonant or vowel but located at the tail of syllable. There are nine types of final consonants in Thai language; /p/, /t/, /k/, /?/, /m/, /n/, /ng/, /w/ and /j/. They can be divided into two groups; open and closed final consonants. The first group consists of /m/, /n/, /ng/, /w/ and /j/ that can be pronounced in long-length speech. And the later one which consists of /p/, /t/, /k/ and /?/ is pronounced in short speech with stopper. Fig. 1 and 2 show two sample signals of "?a:/n/" as open-group and "?a:/t/" as closed-group representatives in time domain. We can see the different length of final consonants between open and closed ones. The longer length of final consonant seems to be easy for detection. And we can also see that the final consonant occupies small frequency range comparing with vowel, and the overlapped frequency range between vowel and final consonant gets large ratio size. Moreover, the energy of vowels is the highest part in syllable comparing with others, as shown by dark grey in Fig. 3 and 4.

By physically analyzing voice generating, voice of vowel is generated by blowing air from lung through thyroid cart and outputting from the mouth without interfering by other organs. Then the highest frequency component of the syllable can be assumed as vowel.

3. BASIC CONCEPT FOR SEGMENTING FINAL CONSONANT

Since the Thai final consonant is located in lower frequency range comparing with vowel, and has over-

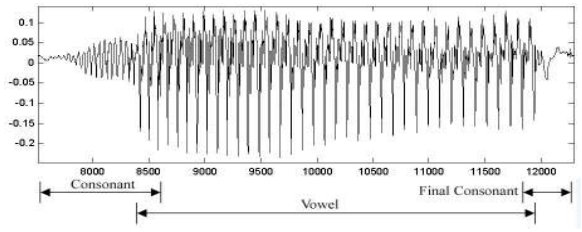

Fig.2: *A sample of "?a:/t/" signal in time domain*

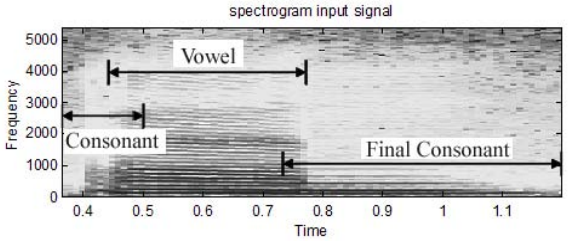

Fig.3: *Spectrogram of "?a:/n/" signal*

lapped frequency range with vowel, it is hard to determine the vowel-final consonant border as shown in A and B of Fig.5. However, if we detect the frequency ranges of both final consonant and vowel as shown in A of Fig. 5, and convert these signals into time domain as shown in C and D of Fig.5, the segmented final consonant may consist of large vowel component comparing with the amplitude of final consonant as shown in C of Fig. 5. But since the final consonant itself has small size, the segmented vowel component might pick up a bit of final consonant part as shown in D of Fig.5. Therefore, the approach of detecting vowel in order to determine the vowel-final consonant border seems to be easier than directly detecting the final consonant.

To detect the vowel component, since vowel has the highest energy in syllable, as shown in Fig. 6, we can pick up it by searching the frequency range for the highest energy as shown in Fig. 7. Then the vowel in time domain is converted from selected frequency range of the highest energy, and the starting point in time domain of final consonant is obtained as shown in Fig. 8.

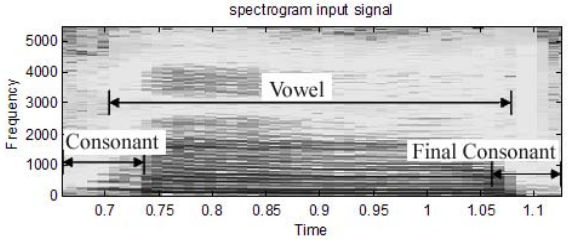

Fig.4: *Spectrogram of "?a:/t/" signal*

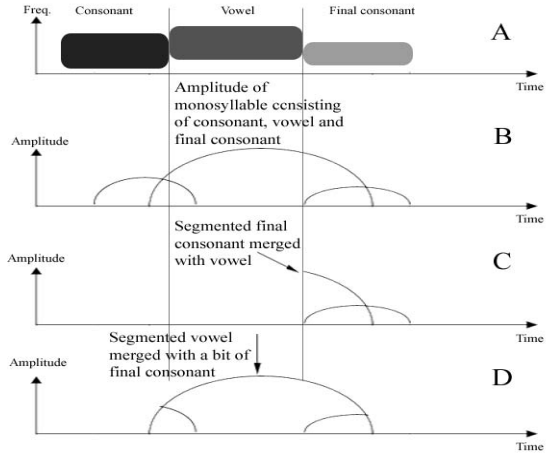

Fig.5: *The difference between detecting vowel and final consonant*

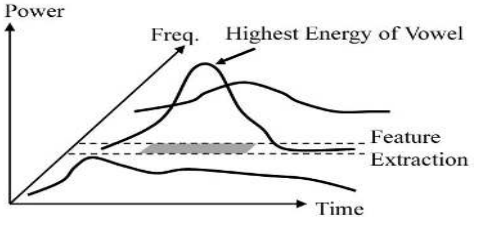

Fig.6: *Components of syllable in power*

4. METHOD

To segment the final consonant in a syllable, in this paper, we propose a method as shown in the following steps.

1. Detection of clear voice of vowel
2. Extracting vowel-frequency component
3. Determining the end point of vowel
4. Final consonant cutting

The signal after syllable segmentation and noise reduction is input for detecting the vowel component of clear voice in step 1. In step 2, the frequency component of vowel is extracted from the vowel component. Then the end point of vowel is determined in the step 3. And final consonant is segmented in step 4. The detail of the processes in these steps are described in the following subsections.

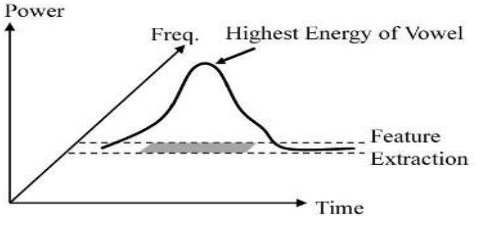

Fig.7: *Picking up frequency range of highest energy as vowel component*

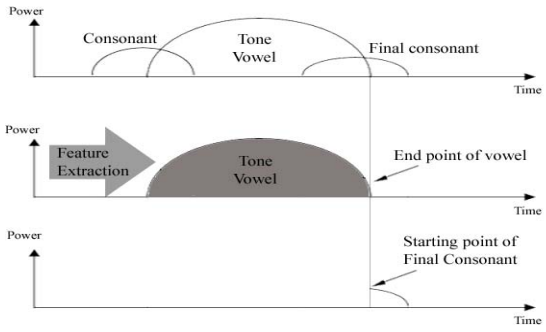

Fig.8: *Selecting vowel component by highest energy and the peak as the most clear voice of vowel*

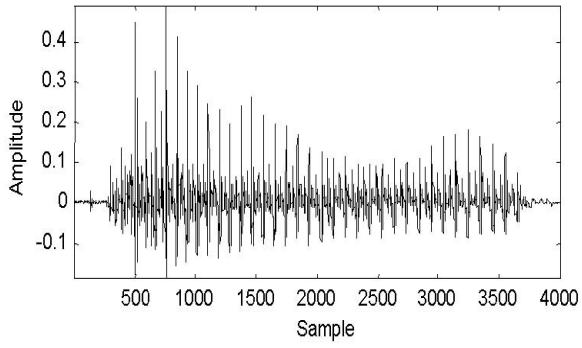

Fig.9: *"?a:/t/" as a sample of input signal*

4.1 Detection of clear voice of vowel

Since the vowel component has the highest energy comparing with other parts of a syllable, this paper proposes to apply power spectrum density (PSD) for detecting the highest energy component as vowel. The power spectrum density is obtained by the method of Welch [7].

For example, the 40 ms data frame was used, and overlapping was perform 75% with sampling frequency 11,025 Hz. The syllable of "?a:/t/" is employed as shown in Fig. 9, and the power spectrum density is obtained as shown in Fig. 10. It shows the peak of the signal is 882 Hz. Then this frequency-range component is proceeded to extract vowel-frequency component.

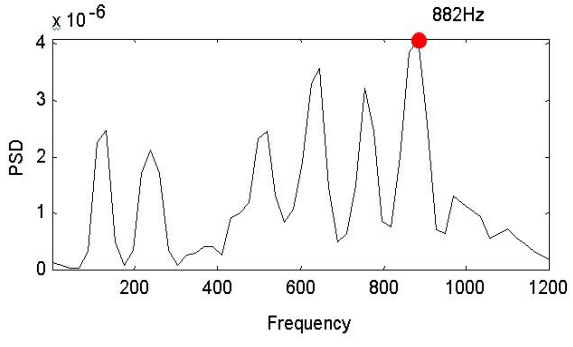

Fig.10: *Power spectrum density of "?a:/t/"*

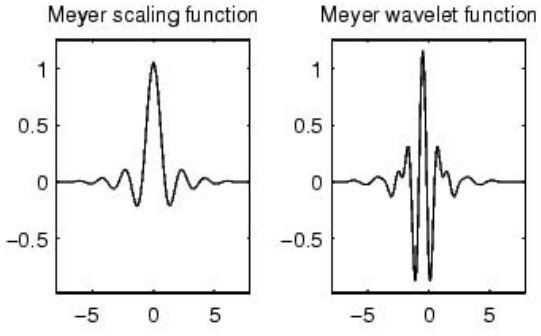

Fig.11: *Mayer Wavelet used as mother Wavelet*

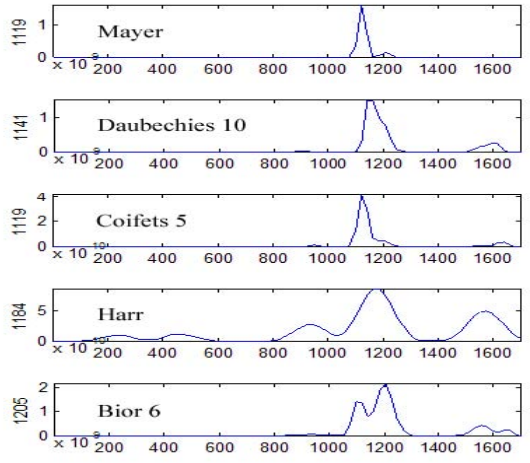

Fig.12: *Comparison of frequency responses by other mother Wavelets*

4.2 Extracting vowel-frequency component

To extract vowel-frequency component, this paper applies Wavelet packets as filter, since the Wavelet packets can analyze the signal into same ranges of frequency comparing with Wavelet transform. The mother Wavelet employed in this paper is Discrete Meyer Wavelet [8], because it is appropriate for FIR based approximation and frequency domain usage. The Discrete Meyer Wavelet is shown in Fig. 11, and the comparison among other mother Wavelets is shown in Fig. 12. Since response of Meyer mother Wavelet is in the smallest frequency range, Meyer mother Wavelet seems to be the most appropriate for this problem.

To determine the appropriate level of Wavelet packets used in this paper, at first, we have to consider the formants of vowels. Since all vowels have much energy in the first formant which occupies frequencies in the range of 0-1,400 Hz.[9], it is very convenient to detect vowel by concentrating only in this range. However, the range of one formant is about 85-100 Hz. wide. That means the resolution of frequency range should be a number within 85-100 Hz. Since the frequency range that human being can pronounce is up to around 5,000 Hz., in this paper, we

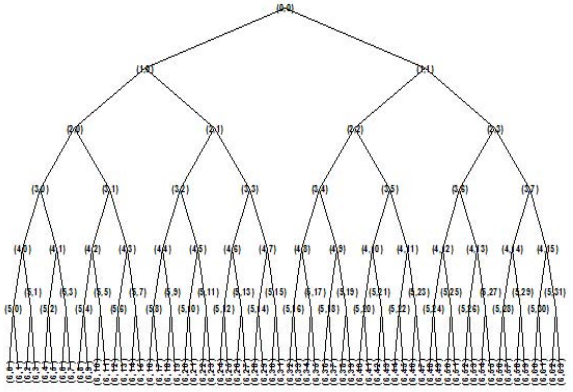

Fig.13: *A tree of 6-level-Wavelet packet nodes*

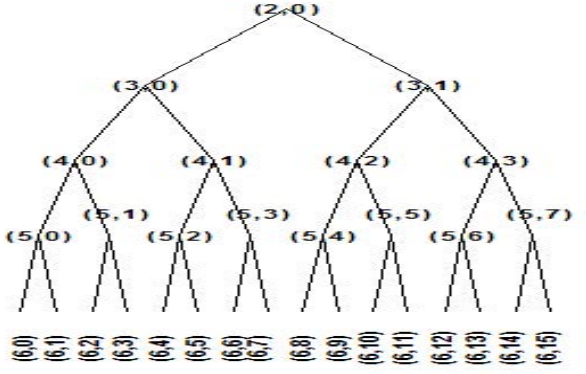

Fig.14: *A group of nodes in the frequency range of a vowel*

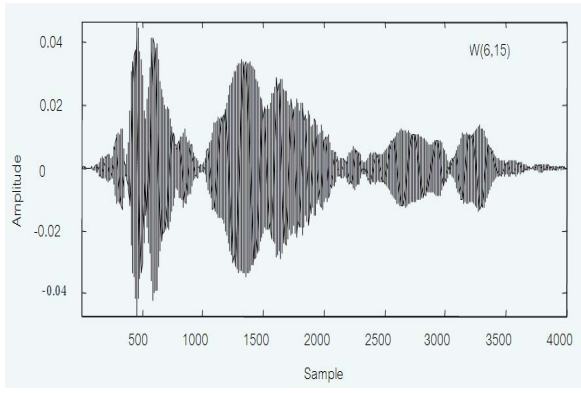

Fig.15: *A sample of segmented vowel*

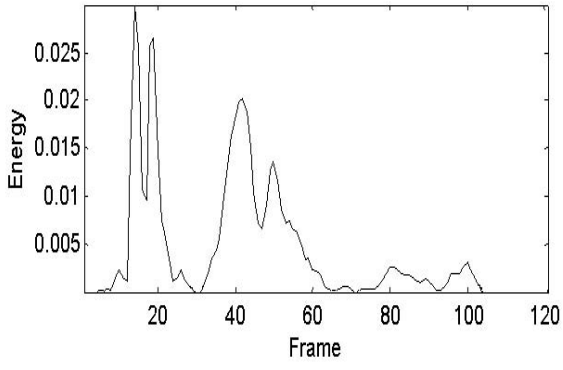

Fig.16: *Energy of vowel component*

sample at 11,025 Hz. for covering up to 5,512 Hz. frequency. Therefore, 86 Hz. frequency range which is within 85-100 Hz. range is selected as appropriate one for the mentioned resolution. The tree of Wavelet packet in this case is shown in Fig. 13, and the nodes that are used for detecting vowels are (6,0) through (6,15) in level six as shown in Fig. 14. The frequency ranges of all nodes are shown in Table 1. For example, if the clearest voice of vowel is in 882 Hz. which is within 861-947 Hz. shown in Table 1, (6,15) node is selected. Then the vowel component is obtained as shown in Fig. 15.

4.3 Determining the ending point of vowel

To determine the ending point of vowel, since vowel can be represented by energy, we can detect the vanishing point of energy in the clear voice of vowel as the ending point of vowel. In this paper, we apply square energy [7] as shown in Fig. 16 to obtain the energy of vowel.

To detect the ending point of vowel (or the starting point of final consonant), since energy does not perfectly vanish at the end of vowel, we have to set an appropriate threshold value to determine the point with lower energy as ending point. Then the ending point of vowel can be found by scanning with the set threshold value from the tail of frame as shown in Fig. 17.

4.4 Final consonant cutting

The number of the frames of the ending point of vowel as shown in Fig.17 is used as reference location for cutting the final consonant from syllable. Fig. 18 shows a sample of cutting final consonant.

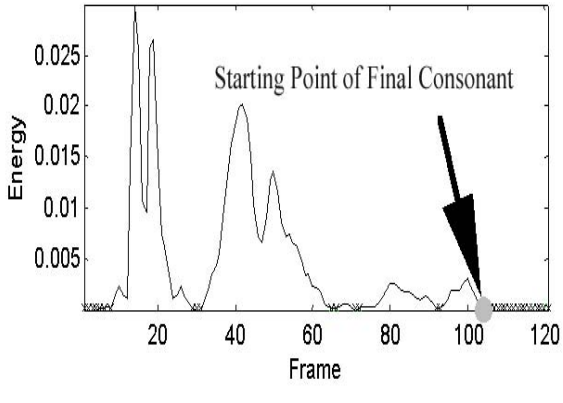

Fig.17: *Scanning the ending point of vowel as starting point of final consonant*

Table 1: *The frequency range of each node of 6-level Wavelet packet*

Node	Frequency	Node	Frequency
(6,0)	0-86	(6,8)	1292-1378
(6,1)	86-172	(6,9)	1206-1292
(6,2)	258-345	(6,10)	1034-1120
(6,3)	172-258	(6,11)	1120-1206
(6,4)	603-689	(6,12)	689-775
(6,5)	517-603	(6,13)	775-861
(6,6)	345-431	(6,14)	947-1034
(6,7)	431-517	(6,15)	861-947

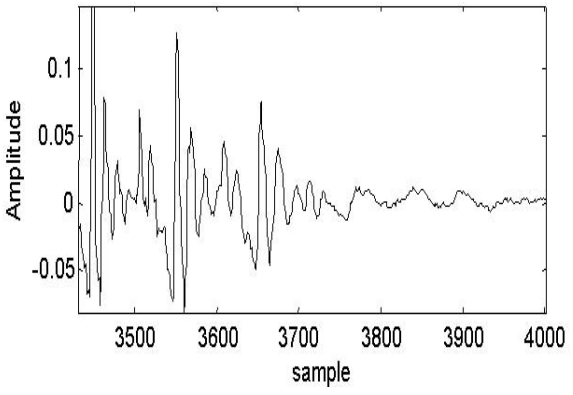

Fig.18: *Segmented final consonant*

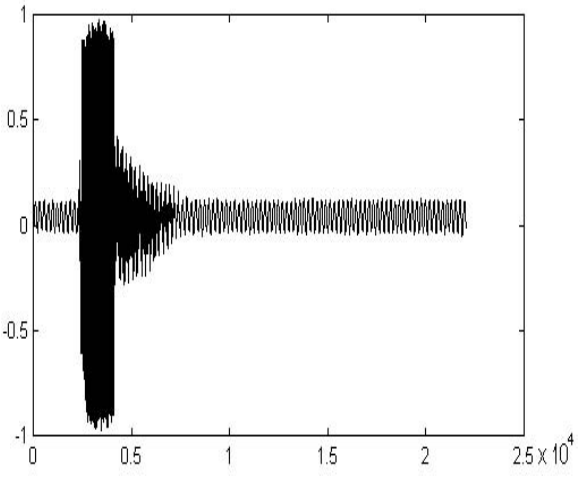

Fig.19: *A speech signal with 24.28 dB.*

5. EXPERIMENTS AND RESULTS

To evaluate the performance of the proposed method, we sample 30 native Thai people during 20-30 years old consisting of 15 males and 15 females for testing. In testing, we used "?" as consonant with 18 vowels and nine final consonants as shown in Table 2 and 3. The speakers see syllable shown by Thai characters on the CRT, listen to the grammatically correct sample-speech recorded by a 26-year-career Phonetic veteran, then pronounce for recording. In recording, the recorder uses 11,025 Hz. sampling frequency with 16 bit data. A testing speaker pronounces 145 syllables. Therefore, 4,350 syllables are collected by 30 testing speakers. The evaluation are performed by four Thai-teaching veterans as subjective way and by neural networks used as a tool of objective way. The results are shown in the following subsections.

5.1 Evaluation results by Thai-teaching veterans

The 4,350 final consonants segmented from 4,350 syllables by the proposed method are evaluated by four Thai-teaching veterans. The evaluation is undertaken in the view point of classifying pure final consonants as good results and final consonants merged with some part of vowel speech as bad ones. The results are shown in the Table 4, 5, 6, 7 and 8.

5.2 Evaluation results by neural networks usage

In evaluation by neural networks, 68 samples of final consonants evaluated by Thai-teaching veterans as correct ones and other 68 samples as wrong ones are employed as learning samples. Totally 136 speeches are used to calculate 100 points of LPC coefficients and fed to back-propagation neural networks. The neural networks consist of 100 input nodes, 50 first hidden nodes, 10 second hidden nodes and one output node. The testing results with 4,350 speeches are shown in Table 9, 10, 11, 12 and 13.

The evaluation results by veterans and neural networks reveal the average accuracy 92.89% and 96.79% respectively.

6. DISCUSSIONS

To analyze the merit and demerit of this method, we divide into several view points such as gender of speakers, length of vowel, changes of final consonant, dialect, consonant and comparison with Fourier transforms as mentioned in the following subsections.

6.1 Gender of Speakers

The experimental results between males and females reveal that the accuracy of male speaking and

Table 2: *Tested syllables consisting of short vowels and final consonant*

	Final Consonant								
Vowel	**p**	**t**	**k**	**ng**	**n**	**m**	**w**	**j**	**?**
a	ap	at	ak	ang	an	am	aw	aj	a?
i	ip	it	ik	ing	in	im	iw	-	i?
v	vp	vt	vk	vng	vn	vm	-	-	v?
u	up	ut	uk	ung	un	um	-	uj	u?
e	ep	et	ek	eng	en	em	ew	-	e?
x	xp	xt	xk	xng	xn	xm	xw	-	x?
ǂ	ǂp	ǂt	ǂk	ǂng	ǂn	ǂm	ǂw	-	ǂ?
o	op	ot	ok	ong	on	om	-	-	o?
@	@p	@t	@k	@ng	@n	@m	-	-	@?

Table 3: *Tested syllables consisting of long vowels and final consonant*

	Final Consonant								
Vowel	**p**	**t**	**k**	**ng**	**n**	**m**	**w**	**j**	**?**
a:	a:p	a:t	a:k	a:ng	a:n	a:m	a:w	a:j	a:?
i:	i:p	i:t	i:k	i:ng	i:n	i:m	i:w	-	i:?
v:	v:p	v:t	v:k	v:ng	v:n	v:m	-	v:j	v:?
u:	u:p	u:t	u:k	u:ng	u:n	u:m	u:w	u:j	u:?
e:	e:p	e:t	e:k	e:ng	e:n	e:m	e:w	e:j	e:?
x:	x:p	x:t	x:k	x:ng	x:n	x:m	x:w	-	x:?
ǂ:	ǂ:p	ǂ:t	ǂ:k	ǂ:ng	ǂ:n	ǂ:m	-	-	ǂ:?
o:	op	o:t	o:k	o:ng	o:n	o:m	o:w	o:j	o:?
@:	@p	@:t	@:k	@:ng	@:n	@:m	-	@:j	@:?

Table 4: *Segmentation results(1)*

	Final Consonant								
		Accuracy (%)			**Accuracy(%)**			**Accuracy (%)**	
Vowel	**p**	**Male**	**Female**	**t**	**Male**	**Female**	**k**	**Male**	**Female**
a	ap	100.00	100.00	at	100.00	93.33	ak	100.00	93.33
i	ip	100.00	100.00	it	100.00	100.00	ik	100.00	100.00
v	vp	100.00	80.00	vt	100.00	86.67	vk	100.00	86.67
u	up	100.00	100.00	ut	100.00	93.33	uk	100.00	93.33
e	ep	100.00	93.33	et	100.00	100.00	ek	100.00	100.00
x	xp	100.00	100.00	xt	100.00	100.00	xk	100.00	100.00
ǂ	ǂp	100.00	86.67	ǂt	100.00	93.33	ǂk	100.00	86.67
o	op	93.33	93.33	ot	100.00	100.00	ok	100.00	100.00
@	@p	100.00	100.00	@t	100.00	93.33	@k	100.00	100.00
a:	a:p	100.00	86.67	a:t	100.00	93.33	a:k	100.00	86.67
i:	i:p	100.00	100.00	i:t	100.00	93.33	i:k	100.00	100.00
v:	v:p	100.00	60.00	v:t	100.00	40.00	v:k	100.00	66.67
u:	u:p	100.00	46.67	u:t	100.00	80.00	u:k	100.00	66.67
e:	e:p	100.00	66.67	e:t	100.00	86.67	e:k	100.00	73.33
x:	x:p	100.00	80.00	x:t	100.00	93.33	x:k	100.00	80.00
ǂ:	ǂ:p	100.00	73.33	ǂ:t	100.00	100.00	ǂ:k	100.00	60.00
o:	op	100.00	80.00	o:t	100.00	93.33	o:k	100.00	86.67
@:	@p	100.00	93.33	@:t	100.00	86.67	@:k	100.00	80.00

female speaking are 98.61% and 86.97% respectively. Since the experiments have been undertaken in laboratory environment with noise from computer and air conditioning, the voice of speakers should be much louder than noise. Normally, men speak louder than women, the signal to noise (S/N) ratio is big. In case of the proposed method that assumes vowel as the highest energy component, lower voice like female speaking sometimes makes the system confused with noise. For example, the speech signal with 24.28 dB S/N ratio as shown in Fig. 19 can be segmented correctly. However, the signal with 10.59 dB S/N ratio

Table 5: *Segmentation results(2)*

	Final Consonant								
		Accuracy (%)			**Accuracy(%)**			**Accuracy (%)**	
Vowel	**ng**	**Male**	**Female**	**n**	**Male**	**Female**	**m**	**Male**	**Female**
a	ang	100.00	93.33	an	100.00	93.33	am	100.00	93.33
i	ing	100.00	100.00	in	100.00	100.00	im	100.00	100.00
v	vng	100.00	73.33	vn	100.00	80.00	vm	100.00	100.00
u	ung	100.00	86.67	un	100.00	100.00	um	100.00	100.00
e	eng	100.00	80.00	en	100.00	100.00	em	100.00	93.33
x	xng	100.00	100.00	xn	100.00	100.00	xm	100.00	100.00
ɨ	ɨng	93.33	80.00	ɨn	100.00	93.33	ɨm	93.33	80.00
o	ong	100.00	100.00	on	100.00	100.00	om	100.00	100.00
@	@ng	100.00	93.33	@n	93.33	93.33	@m	93.33	93.33
a:	a:ng	100.00	80.00	a:n	100.00	86.67	a:m	100.00	80.00
i:	i:ng	100.00	93.33	i:n	100.00	86.67	i:m	100.00	86.67
v:	v:ng	100.00	66.67	v:n	100.00	86.67	v:m	100.00	73.33
u:	u:ng	100.00	46.67	u:n	100.00	73.33	u:m	100.00	40.00
e:	e:ng	80.00	53.33	e:n	86.67	73.33	e:m	93.33	53.33
x:	x:ng	100.00	86.67	x:n	100.00	100.00	x:m	100.00	86.67
ɨ:	ɨ:ng	93.33	53.33	ɨ:n	93.33	53.33	ɨ:m	93.33	73.33
o:	ong	93.33	86.67	o:n	93.33	86.67	o:m	93.33	86.67
@:	@ng	100.00	86.67	@:n	100.00	73.33	@:m	100.00	73.33

Table 6: *Segmentation results 3)*

	Final Consonant								
		Accuracy (%)			**Accuracy(%)**			**Accuracy (%)**	
Vowel	**w**	**Male**	**Female**	**j**	**Male**	**Female**	**?**	**Male**	**Female**
a	aw	93.33	80.00	aj	100.00	86.67	a?	100.00	100.00
i	iw	100.00	100.00	-	-	-	i?	100.00	100.00
v	-	-	-	-	-	-	v?	100.00	93.33
u	-	-	-	uj	100.00	86.67	u?	100.00	100.00
e	ew	100.00	86.67	-	-	-	e?	100.00	100.00
x	xw	100.00	100.00	-	-	-	x?	100.00	100.00
ɨ	ɨw	100.00	93.33	-	-	-	ɨ?	100.00	100.00
o	-	-	-	-	-	-	o?	100.00	100.00
@	-	-	-	-	-	-	@?	100.00	93.33
a:	a:w	93.33	80.00	a:j	86.67	80.00	a:?	100.00	80.00
i:	i:w	100.00	93.33	-	-	-	i:?	100.00	93.33
v:	-	-	-	v:j	100.00	60.00	v:?	100.00	73.33
u:	u:w	100.00	73.33	u:j	100.00	66.67	u:?	100.00	86.67
e:	e:w	93.33	80.00	e:j	86.67	73.33	e:?	93.33	86.67
x:	x:w	100.00	80.00	-	-	-	x:?	100.00	86.67
ɨ:	-	-	-	-	-	-	ɨ:?	100.00	80.00
o:	ow	93.33	100.00	o:j	86.67	86.67	o:?	100.00	100.00
@:	-	-	-	@:j	100.00	73.33	@:?	100.00	80.00

as shown in Fig. 20 is hard to segment, since amplitudes of signal and noise are similar. The experiments performed with 30 people reveal that average S/N ratio of males and females are 17.49 dB and 13.72 dB respectively. Therefore, male speaking takes advantage, and the accuracy rate of male is higher.

6.2 Length of Vowel

In Thai language, there are two groups of vowel in term of vowel length; short and long vowels. The experiments show that syllables with long and short vowels get 88.74% and 96.84% accuracy respectively. In case of long vowel as shown in Fig. 21, since the pronunciation of long vowel is not constantly maintained in the same frequency level, sometimes the fre-

Table 7: *Summary of Segmentation results by proposed method (1)*

	Closed Syllable		**Open Syllable**		**Open Syllable with Half Vowel**		**With Final Consonant /?/**	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
Short Vowel a, i, v, u e, x, ɨ, o, @	99.75%	95.31%	99.01%	92.59%	99.05%	90.48%	100.00%	98.52%
Long Vowel a:, i:, v:, u: e:, x:, ɨ:, o:, @:	99.75%	79.75%	97.04%	75.06%	95.00%	78.89%	99.26%	85.19%

Table 8: *Summary of Segmentation results by proposed method (2)*

Type	**Accuracy**
Male	98.61%
Short Vowel	96.84%
Long Vowel	93.64%

Type	**Accuracy**
Female	86.97%
Short Vowel	88.74%
Long Vowel	92.51%

Table 9: *Segmentation results (1) by neural networks*

	Final Consonant								
		Accuracy (%)			**Accuracy(%)**			**Accuracy (%)**	
Vowel	**p**	**Male**	**Female**	**t**	**Male**	**Female**	**k**	**Male**	**Female**
a	ap	100.00	100.00	at	100.00	93.33	ak	100.00	100.00
i	ip	100.00	80.00	it	100.00	100.00	ik	80.00	80.00
v	vp	100.00	80.00	vt	100.00	80.00	vk	100.00	80.00
u	up	100.00	80.00	ut	100.00	60.00	uk	100.00	80.00
e	ep	100.00	100.00	et	100.00	100.00	ek	100.00	40.00
x	xp	100.00	60.00	xt	100.00	80.00	xk	100.00	100.00
ɨ	ɨp	100.00	100.00	ɨt	100.00	80.00	ɨk	100.00	80.00
o	op	100.00	80.00	ot	100.00	100.00	ok	100.00	100.00
@	@p	100.00	60.00	@t	100.00	80.00	@k	100.00	80.00
a:	a:p	100.00	80.00	a:t	100.00	100.00	a:k	100.00	100.00
i:	i:p	80.00	100.00	i:t	100.00	80.00	i:k	100.00	80.00
v:	v:p	100.00	80.00	v:t	100.00	100.00	v:k	100.00	100.00
u:	u:p	100.00	80.00	u:t	100.00	100.00	u:k	100.00	100.00
e:	e:p	100.00	60.00	e:t	100.00	100.00	e:k	100.00	100.00
x:	x:p	100.00	60.00	x:t	100.00	100.00	x:k	100.00	100.00
ɨ:	ɨ:p	100.00	60.00	ɨ:t	100.00	80.00	ɨ:k	80.00	80.00
o:	op	80.00	80.00	o:t	100.00	80.00	o:k	100.00	80.00
@:	@p	100.00	100.00	@:t	100.00	60.00	@:k	100.00	80.00

quency in tail part of vowel falls off a little bit. For an example as shown in Fig. 22, the peak of PSD is 430 Hz. in Fig. 21(B), this frequency range is selected for filter to extract vowel signal as shown in Fig. 21(C), 21(D). However, some parts of vowel fall off a little bit as mentioned above, they might not be picked up as vowel parts, and finally the segmented final consonant includes these vowel parts as shown in Fig. 21(E) and 21(F).

6.3 Type of Final Consonant

In experimental results between closed and open final consonants, the segmenting accuracy are 93.64% and 92.51% respectively. To find the cause of the errors, since by phonetics /?/, any tones of closed syllable has possibility to change into low tone, and open syllable will become rising tone or falling tone as shown in Fig. 23. If we select the appropriate frequency range for high, middle and low tones as shown by gray color in Fig. 24, 25 and 26 respectively, the extraction will be performed correctly. However, as shown in Fig. 23, curves of rising and falling tones have big changes in frequency levels. Therefore, basically open syllable that has possibility to change into these tones gets risk for error occurring. Though the experiments have been performed in middle tone that is nearly constant in frequency level, the change from middle tone to low tone of closed syllable cause errors. However, the errors of open syllables are a bit bigger than those of closed ones, because falling and

Table 10: *Segmentation results (2) by neural networks*

	Final Consonant							
		Accuracy (%)			**Accuracy(%)**			**Accuracy (%)**
Vowel	**ng**	**Male**	**Female**	**n**	**Male**	**Female**	**m**	**Male**	**Female**
a	ang	80.00	80.00	an	100.00	80.00	am	80.00	80.00
i	ing	100.00	100.00	in	100.00	100.00	im	100.00	100.00
v	vng	100.00	80.00	vn	100.00	100.00	vm	100.00	80.00
u	ung	100.00	80.00	un	100.00	80.00	um	100.00	60.00
e	eng	100.00	80.00	en	100.00	100.00	em	100.00	93.33
x	xng	100.00	100.00	xn	100.00	100.00	xm	100.00	60.00
ɨ	ɨng	80.00	80.00	ɨn	100.00	80.00	ɨm	100.00	60.00
o	ong	100.00	60.00	on	100.00	100.00	om	100.00	60.00
@	@ng	80.00	60.00	@n	100.00	100.00	@m	100.00	80.00
a:	a:ng	100.00	60.00	a:n	100.00	80.00	a:m	100.00	100.00
i:	i:ng	100.00	40.00	i:n	100.00	80.00	i:m	100.00	100.00
v:	v:ng	100.00	80.00	v:n	100.00	80.00	v:m	100.00	100.00
u:	u:ng	100.00	60.00	u:n	100.00	80.00	u:m	100.00	80.00
e:	e:ng	100.00	100.00	e:n	100.00	80.00	e:m	80.00	80.00
x:	x:ng	100.00	60.00	x:n	100.00	100.00	x:m	100.00	80.00
ɨ:	ɨ:ng	80.00	80.00	ɨ:n	100.00	80.00	ɨ:m	100.00	60.00
o:	ong	100.00	100.00	o:n	100.00	60.00	o:m	100.00	100.00
@:	@ng	100.00	60.00	@:n	100.00	40.00	@:m	100.00	40.00

Table 11: *Segmentation results (3) by neural networks*

	Final Consonant							
		Accuracy (%)			**Accuracy(%)**			**Accuracy (%)**
Vowel	**w**	**Male**	**Female**	**j**	**Male**	**Female**	**?**	**Male**	**Female**
a	aw	100.00	80.00	aj	100.00	100.00	a?	80.00	80.00
i	iw	100.00	100.00	-	-	-	i?	100.00	100.00
v	-	-	-	-	-	-	v?	80.00	60.00
u	-	-	-	uj	100.00	60.00	u?	100.00	100.00
e	ew	100.00	100.00	-	-	-	e?	100.00	100.00
x	xw	100.00	80.00	-	-	-	x?	100.00	100.00
ɨ	ɨw	100.00	80.00	-	-	-	ɨ?	100.00	80.00
o	-	-	-	-	-	-	o?	100.00	80.00
@	-	-	-	-	-	-	@?	100.00	80.00
a:	a:w	100.00	100.00	a:j	100.00	40.00	a:?	80.00	60.00
i:	i:w	100.00	100.00	-	-	-	i:?	100.00	100.00
v:	-	-	-	v:j	100.00	100.00	v:?	80.00	100.00
u:	u:w	100.00	100.00	u:j	100.00	80.00	u:?	100.00	80.00
e:	e:w	100.00	100.00	e:j	100.00	80.00	e:?	100.00	80.00
x:	x:w	100.00	80.00	-	-	-	x:?	100.00	80.00
ɨ:	-	-	-	-	-	-	ɨ:?	100.00	60.00
o:	ow	100.00	100.00	o:j	100.00	60.00	o:?	100.00	100.00
@:	-	-	-	@:j	100.00	60.00	@:?	100.00	40.00

rising tones themselves that are the changed tones of open syllables naturally have big changes in frequency level.

6.4 Dialects

For dialects spoken in other parts of Thailand, the main difference between Thai dialects and standard Thai is the change of tones and the length of speech. Therefore, the system error will occur in case of dialect speaking, specially, in case that the tones change into rising and falling ones. However, some parts of error can be rescued by using same dialect samples for learning.

Table 12: *Summary of Segmentation results by neural networks (1)*

	Closed Syllable		**Open Syllable**		**Open Syllable with Half Vowel**		**With Final Consonant /?/**	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
Short Vowel a, i, v, u e, x, ɨ, o, @	99.75%	94.57%	99.01%	94.32%	100.00%	95.24%	98.52%	95.56%
Long Vowel a:, i:, v:, u: e:, x:, ɨ:, o:, @:	99.26%	95.31%	99.51%	92.10%	100.00%	94.44%	98.52%	92.59%

Table 13: *Summary of Segmentation results by neural networks (2)*

Type	**Accuracy**
Male	99.32%
Short Vowel	97.12%
Long Vowel	97.22%

Type	**Accuracy**
Female	94.27%
Short Vowel	96.47%
Long Vowel	96.65%

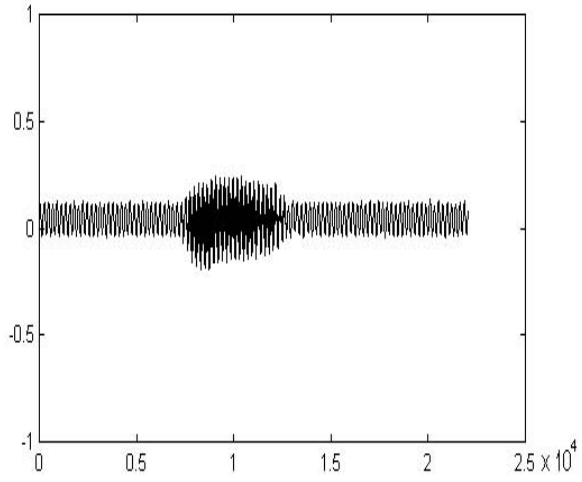

Fig.20: *A speech signal with 10.59 dB.*

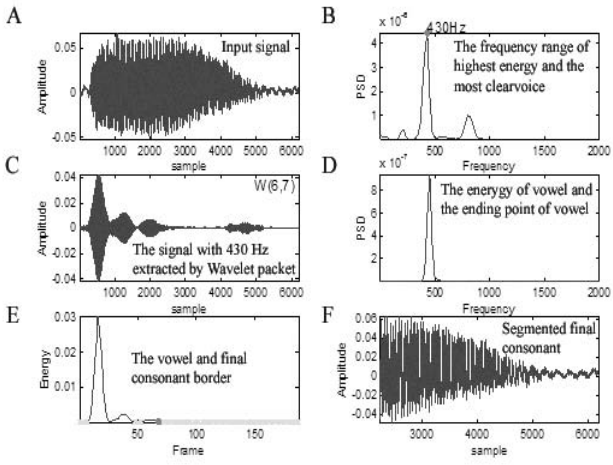

Fig.21: *A mistaken sample of final consonant segmentation*

6.5 Changes of Consonants

Since this method scan for the end of vowel (starting point of final consonant) in time domain from the end of frame, the change of consonant that is located in the beginning of signal in time domain may not cause error.

6.6 Comparison of Filters Designed by Fourier Transforms and Wavelet Packet

Since the filters designed by Fourier transforms depend on the number of orders, the accuracy varies on the selected number of orders. And since the frequency ranges of Thai vowels are not constant, the appropriate numbers of orders are not unique. Therefore, the filters designed by Fourier transforms are considered not to be suitable for this problem. However, the Wavelet packet transform is more appropriate for analysis in detail and can be suitable filter for this problem. In this paper, we compared the filters designed by FIR-100-orders Fourier transform as conventional way with Wavelet packet as shown in Fig. 27. We can see that final-consonant segmentation by Wavelet packet as shown by 1 in Fig. 27 is performed at better location than the one by FIR-100-orders Fourier transforms as shown by 2 in the same figure. Though we design the filter by lower order, the filtered signal includes other frequency ranges. Therefore, the usage of Wavelet packet in designing filter of this problem is more appropriate.

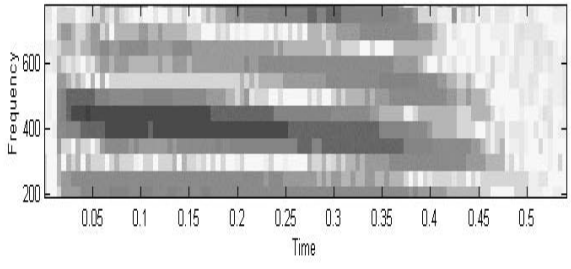

Fig.22: *Spectrogram of syllable with falling*

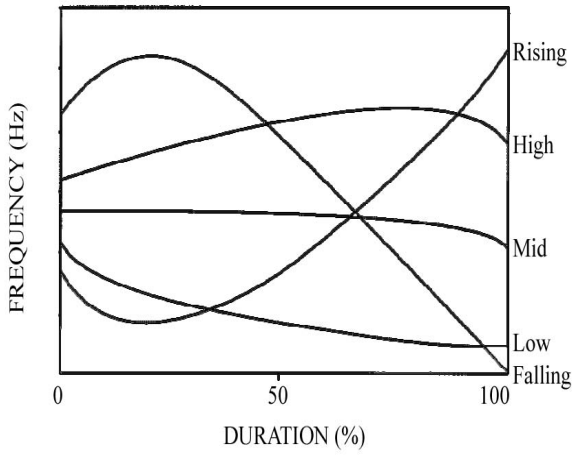

Fig.23: *Thai tone patterns*

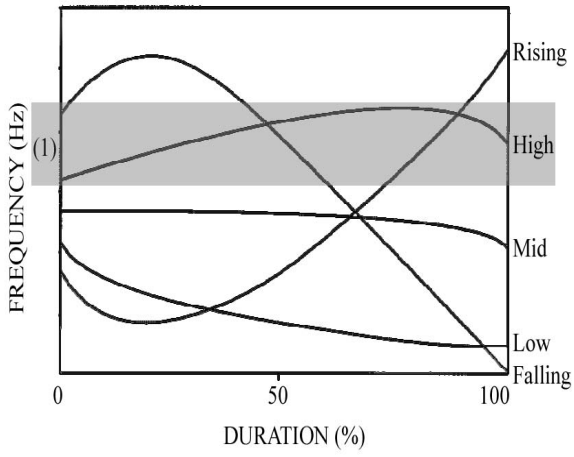

Fig.24: *A frequency range of high tone*

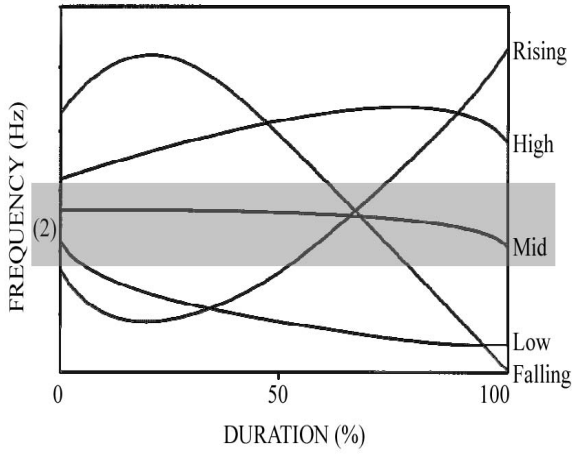

Fig.25: *A frequency range of low tone*

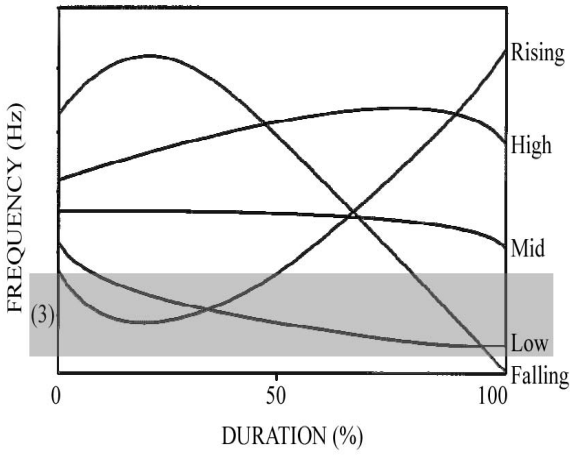

Fig.26: *A frequency range of mid tone*

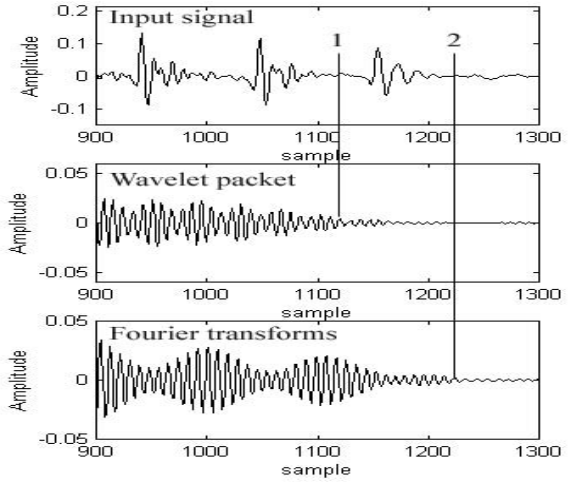

Fig.27: *Comparison of segmentation by Fourier transforms and Wavelet packet*

7. CONCLUSION

This paper proposes final consonant segmentation for Thai syllable by using vowel characteristics and Wavelet packets. The experimental results with 4,350 syllables reveal that the results of final consonant segmentation are 92.89% and 96.84% accurate evaluated by Thai-teaching veterans and neural networks respectively.

ACKNOWLEDGEMENT

The authors wish to thank to Ms. Supapun Mekarat, Thai language and Phonetics lecturer at Rajabhat Institute Suansunanda for making samples, four Thai teaching veterans for evaluation, Dr. Pinit Kumhom (King Mongkut's University of Technology Thonburi) and Ms. Rachada Kongkachandra (Thamasat University) for reviewing.

References

[1] M.D. Hanes, "Acoustic-to-phonetic mapping using recurrent neural networks," in *Proceedings of neural networks*, 1994, pp. 659-662.

[2] T.T. Beng, "The use of wavelet transform in phoneme recognition," in *Proceeding of International Conference on Spoken Language*, 1996, pp. 2431-2434.

[3] T. Sripramong et al., "Thai speech analysis in harmonic frequency domain," in *The 15th Electrical Engineering Conference*, 1992.

[4] K. Dumrongpati et al., "Thai speech segmentation to phoneme using wavelet transform," in *Proceedings of the 1999 National Computer Science and Engineering Conference*, 1999, pp. 357-363.

[5] O. Jarungpronsawad et al., "Speech segmentation using quasi-periodic reconsidering based on zero-crossing," in *Proceedings of International Conference on Intelligent Technologies*, Assumption University, Thailand, 2000, pp. 244-247.

[6] S. Laksaniyawin, *Phonetics and Linguistics (in Thai)*, Chulalongkorn University, Bangkok,Thailand, 1986.

[7] P.D. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modi?ed periodograms," in *Proceedings of Audio Electroacoustics*, Assumption University, Thailand, 1967, pp. 70-73.

[8] M.C. Misiti, *Matlab Wavelet Toolbox, User's Guide Version 1, Natuck*, The MathWorks Incoporated, 1996.

[9] C.S. Burrus, *Introduction to Wavelet Transform*, Prentice-Hall, Texas, 1998.

Kosin Chamnongthai was born in Bangkok Thailand in 1960. He received the B.Eng. in Applied Electronic Engineering from the University of Electro-communication, Tokyo in 1985, M.Eng. in Electrical Engineering from Nippon Institute of Technology, Saitama in 1987 and D.E.E. in Electrical Engineering from Keio University, Tokyo Japan in 1991. He has worked at Electronic and Telecommunication Engineering Department of King Mongkut's University of Technology Thonburi(KMUTT), Bangkok, Thailand, as lecturer in 1991,assistant professor in 1993 and associate professor in 1996 until now. His research interests include image processing,computer vision, robot vision, and natural language processing.He is a member of IEEE, Information Processing Society and Thai Robotics Society.

Wudthipong Pichitwong was born in Nakornrachsima province, Thailand in 1973. He received the B.E. from Rachamangala Institute of Technology, and M.E. degree in Electronic and Telecommunication Engineering from King Mongkut's University of Technology Thonburi (KMUTT), Bangkok, Thailand in 1999 and 2001, respectively. He has worked as a lecturer at Computer Center and Computer Technology Department of Rachmangala Institute of Technology, North-East campus, Nakronrachsima from 1997 until now. His research interests include signal processing and image processing.

Piyasawat Navaratana Na Ayudhya was born in Bangkok, Thailand in 1976. He received the B.E. and M.E. degree in Electronic and Telecommunication Engineering from King Mongkut's University of Technology Thonburi (KMUTT), Bangkok, Thailand in 1997 and 2000, respectively. He has worked as a lecturer at Electrical Engineering Department,King Mongkut's University of Technology Thonburi from 2000 until now. His research interests include electronic circuit, signal processing, voice recognition and image processing.