Tile name: **Thai Speech Corpus for Speech Recognition**

The name of authors: **Sawit Kasuriya, Virach Sornlertlamvanich, Patcharika Cotsomrong, Supphanat Kanokphara, and Nattanun Thatphithakkul**

Affiliations**: Information Research and Development, National Electronics and Computer Technology Center (NECTEC)**

Mailing address: **National Electronics and Computer Technology Center, 112 Phahon Yothin Rd., Klong 1, Klong Luang, Pathumthani 12120, THAILAND**

Email address for correspondence:
sawitk@nectec.or.th, virach@nectec.or.th, aye@nectec.or.th, supphanat_k@nectec.or.th, nattanun_t@nectec.or.th

Thai Speech Corpus for Speech Recognition

Abstract

Nowadays, the improvement of speech recognition technology is growing fast and many techniques are presented. However, even the best algorithm with carefully designed system cannot accomplish good performance speech recognition if the system is trained from poor corpus. Therefore, the speech corpus is the basic research that is necessary and should be consistent, balanced, as well as covering all possible phonemes in the language. This paper indicates how our corpus is built in order to preserve all those properties: consistency, balance, and containing possible phoneme combination.

1. Introduction

Work on the speech recognition task, the speech corpus is very important in order to build such a speech recognition application system. Then, the speech corpus is basic necessary research of those tasks.

This corpus was developed in the scope of the speech recognition task project. The main objective of this project is to collect and prepare the speech database for the acoustic models and some language models construction. Also, the design for Thai speech corpus is considered for large vocabulary continuous speech recognition system development. These are two principle sets of this corpus: the phonetically distributed (PD) set and the 5,000 vocabularies set.

This paper is organized as follows. The second section presents speech corpus for speech recognition. There are defining the transcription for Thai language and currently Thai speech corpus for speech recognition. The next section is corpus tools and the last one is problem of word variation in Thai language.

2. Speech Corpus for Speech Recognition

This section describes the speech corpus in particular of speech recognition. Some topic of this section does not go into the detail because it was published in the previous papers [C. Wutiwatchai et al. 2002, R. Thongprasirt et al. 2002]. Hence, defining the transcription for Thai language, currently Thai speech corpus for speech recognition, corpus tools, and problem of word variation in Thai language are narrated as follows.

2.1 Defining the transcription for Thai language

The forms of Thai syllables are C_iV and C_iVC_f and the tone is marked onto each syllable. Five different tones in Thai is divided into two groups: (1) the static group--high, middle, and low tone, and (2) the dynamic group--rising and falling tone. Thai phonetic system has 21 single consonants, 12 double consonants, 24 vowels, and more than 5 double consonants that use for pronouncing the foreign word. The single and double Thai consonants are shown in Table 1 and 2 respectively. And Table 3 indicates differences between the initial consonant and final consonant whereas Table 4 is Thai vowel symbols. These symbols of Thai phonemes are a little bit different from those of another Thai linguists. But in the case of Thai tone, the digits 0 to 4 are used to represent the five tones, which are middle, low, falling, high and rising, respectively.

Thai transcription are used in this research, in forms of either / C_i_V_T/ or / C_i_V_C_f_T/, where C_i denotes the initial consonant (including single and double consonant), V denotes the vowel (both short and long vowel), C_f denotes the final consonant (some single consonants), and T denotes the tone. For example Thai word, "กล่อง" (means "box") is /kl_@_ng^_1/, "คุณครู" (means "teacher") is /kh_u_n^_0/khr_uu_z^_0/. Because same symbols stand for many initial and final consonant, '^'symbol is represented for final consonant in order to differentiate from initial consonant.

Table 1. Single Thai consonants (21 phonemes)

Place and Manner		***Labial***	***Alveolar***	***Palatal***	***Velar***	***Glottal***
Stop	Voiceless Unaspirated	p (ป)	t (ฏ,ต)	c (จ)	k (ก)	z (อ)
	Voiceless Aspirated	ph (พ ภ ผ)	th (ฐ,ฑ,ฒ,ถ,ท,ธ)	ch (ฉ,ฌ,ช)	kh (ข,ค,ฆ)	
	Voiced	b (บ)	d (ฎ,ด)		ng (ง)	
Non-stop	Nasal	m (ม)	n (ณ,น)			h (ห,ฮ)
	Fricative	f (ฝ,ฟ)	s (ซ,ศ,ษ,ส)			
	Trill		r (ร ฤ)			
	Lateral		l (ล,ฬ)			
	Approximant	w (ว)		j (ญ,ย)		

Table 2. Double Thai consonant (12 phonemes)

Double Consonant	***Phoneme Symbol***	***Double Consonant***	***Phoneme Symbol***
ปร	pr	กร	Kr
ปล	pl	กล	Kl
พร	phr	กว	kw
พล	phl	คร	khr
ตร	tr	คล	khl
ทร	thr	ขว	khw

Table 3. Initial and final consonant symbol (26 phonemes and 12 phonemes)

Consonant	***Phoneme***		***Consonant***	***Phoneme***	
	Initial (Ci)	Final (Cf)		Initial (Ci)	Final (Cf)
ก	k	k^	บ	b	p^
ข,ค,ฆ	kh	k^	ป	p	p^
ง	ng	ng^	ผ,พ,ภ	ph	p^
จ	c	t^	ฝ,ฟ	f	p^
ฉ,ฌ,ช	ch	t^	ม	m	m^
ซ,ศ,ษ,ส	s	t^	ร	r	n^
ญ,ย	j	j^	ล,ฬ	l	n^
ฎ,ด	d	t^	ว	w	w^
ฏ,ต	t	t^	ห,ฮ	h	-
ฐ,ฑ,ฒ,ถ,ท,ธ	th	t^	อ	z	-
ณ,น	n	n^	Foreign lang.	br,bl,fr,fl,dr	f^,s^,ch^,l^

Table 4. Thai vowels symbol (24 phonemes)

Tongue Advancement / ***Tongue Height***	***Front (short/long)***	***Central (short/long)***	***Back (short/long)***
Close	i, ii (อิ, อี)	v, vv (อึ, อือ)	u, uu (อุ, อู)
Mid	e, ee (เอะ, เอ)	q, qq (เออะ, เออ)	o, oo (โอะ, โอ)
Open	x, xx (แอะ, แอ)	a, aa (อะ, อา)	@, @@ (เอาะ, ออ)
Diphthongs	ia, iia (เอียะ, เอีย)	va, vva (เอือะ, อัวะ)	ua, uua (อัวะ, อัว)

2.2 Currently Thai speech corpus for speech recognition

Recent the progress of Thai speech research, we talked about the phonetically balance (PB) and the phonetically distributed (PD) selection [C. Wutiwatchai et al. 2002] and our speech corpus project with universities cooperation [R. Thongprasirt et al. 2002]. Previous papers have been presented the detail of some process of our corpus development such as, PD selection, corpus design, corpus plan, and distribution of recording. At present, text processing and all of sentence selection have been finished. We are in state of recording and speech alignment. Therefore, the summary of text processing, corpus design, corpus tools, and

recording conditions are briefly presented in this paper.

2.2.1 Text Processing

Thai language is the one of the alphabetic language. There is no sign or space between words or sentences. Sometime, the space is placed between only adjacent sentences but it is very ambiguous rule and depends on the writer. The complication of Thai language is how to separate the sentences from any paragraph and segment the words from a sentence. That means word and sentence definition are major problems of Thai language. In consequence, we have to manually handle the text corpus that takes long time and needs many linguists. This text corpus that used in this development, takes more than a year to arrange Thai text corpus for speech corpus.

In this section, the shortly detail of grapheme-to-phoneme (G2P) that are the principle of text processing, are described as follows. The G2P is a routine that converts an input word sequence into their corresponding phonetic transcription. It is one of the essential processes in developing a speech corpus. They have many approaching techniques to implement the G2P such as, dictionary-based, rule-based, and statistical-based. The detail of our latest approach is presented in [P. Tarsaku et al. 2001]. This module has included syllable and word detection. The performance of G2P depends on syllable boundaries because some phonemes in some syllables are not corresponding to their graphemes (depending on Thai words) and syllable detection is not complete accuracy (approximately 80%). Therefore, the G2P module has some error that especially occurs in foreign words. Their phonemes of word were checked by the linguists after they have been passed the G2P process. In addition, the phonemes sequence of each word in the sentence has been checked. Some phonemes or tones are manually corrected.

2.2.2 Corpus Design

The objective of this corpus is to develop a large-vocabulary continuous speech recognition (LVCSR) corpus for Thai language. This corpus aims at 5,000 vocabularies coverage, which is limited by Thai text corpus. Thai text corpus is collected from the Open Linguistic Resources Channeled toward InterDisciplinary research (ORCHID) [V. Sornlertlamvanich et al. 1998], magazines, Thai encyclopaedia, and journals. Only ORCHID has already manually tagged for text corpus. It contains 27,634 sentences. After the others text corpus are included, there are nearly 2,500,000 words (43,255 vocabularies) within 180,504 sentences.

The contents of this corpus consist of two sets: (1) the phonetically distributed (PD) sentences set

and (2) 5,000 Thai vocabulary coverage sentences set. The detail of both sets are described as follows.

(1) Phonetically distributed sentence (PD) set

To initial acoustic model efficiently, phonetically balanced sentences (PB) is usually used for training. PB is the smallest set of sentences covering all phonemic units in the language. In our case, the phonemic unit is biplone. PD is the extension of PB. It does not only cover all biphone, but the text distribution is also similar to the daily used context (ORCHID corpus in this case).

The PD selection process starts from PB construction. In PB construction process, the sentence containing mostly unselected biphone is chosen one by one until all biphones are included in the PB set. Before constructing PD set, the biphone distributions of ORCHID are calculated. Then, some sentences are added to PB to change the distribution as same as distributions of ORCHID. The number of adding sentences should be kept at minimum while the biphone distribution of PD set is closest to ORCHID's distribution. More details of PB/PD construction can be found in [C. Wutiwiwatchai et al. 2002].

(2) 5,000 Vocabularies Set

The objective of this set is to collect the structure of Thai language for language model (LM) construction. This set is divided into three subsets: the training set (TR), the development test set (DT), and the evaluation test set (ET). The TR set is used to train language models. The DT and ET sets are used for testing in development and evaluation phases respectively.

The process of TR, DT, and ET selections are illustrated in Figure 1. Firstly, the words of all sentences are listed and sorted. There are 43,255 vocabularies. The sentences containing the first 5,000 vocabularies that most frequently occurring, are selected. These sentences (11,202 sentences) are chosen to the next step. The TR set (3,007 sentences) is selected by collecting the minimum amount of sentences that pertains 5,000 vocabularies. The remaining sentences are divided into two sets: set A and B, for language model construction (5,000 sentences) and DT, ET selection (3,195 sentences), consecutively. In addition, the set B is selected by calculating the sentence scores (defined in (1)) of each sentences and choosing the 3,195 sentences that are the highest sentence scores. On the other set, the tri-gram language model is created by 5,000 sentences and 3,007 sentences (TR set). There are 8,007 sentences that use for LM construction. And LM is used for calculating the perplexity of each sentence in set B. The next procedure, the 1,000 sentences that have the medium perplexity (around 100 to 300), are selected. The last step is to randomly divided into DT set and ET set.

$$SC = \frac{\left(\sum_{i}^{N_w} \left\{ \frac{1}{Wf_i} \right\} \right)}{N_w} \quad (1)$$

Where: SC denotes the sentence scores

N_w denotes the number of words in each sentences

Wf_i denotes the i^{th} word frequency of 8,195 sentences

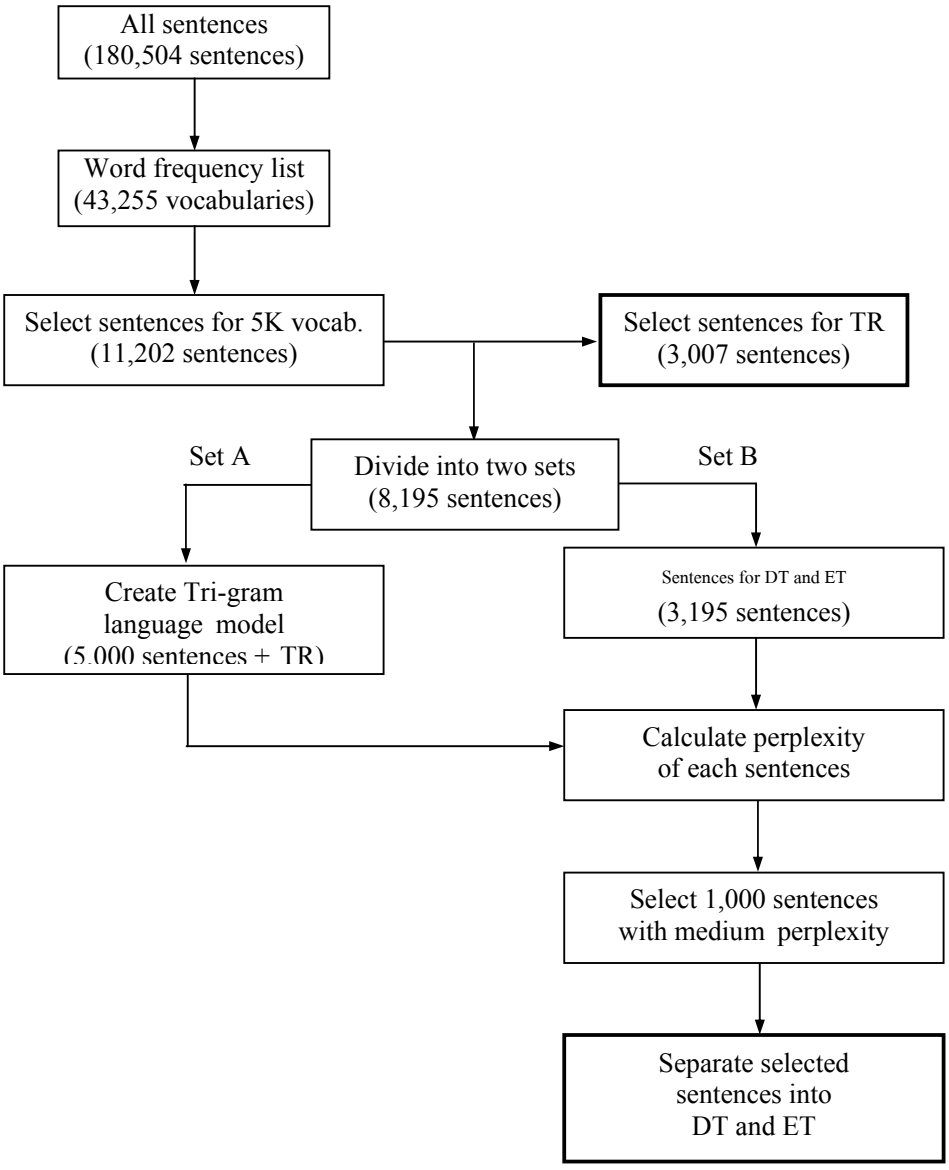

Figure 1. TR, DT and ET selection

Table 5. Summary of Phonetically distributed sentences set

Attribute	PD set
No. of sentences	802
No. of vocabularies	2,269
No. of words	7,847
No. of syllables	12,702
No. of phonemes	38,106

Table 6. Classified the syllables

Attribute	***Tone 0***	***Tone 1***	***Tone 2***	***Tone 3***	***Tone 4***
No. of syllables	4,198	2,953	2,373	2,080	1,098

Table 7. Summary of 5,000 Thai vocabulary coverage sentences set

Attribute	***TR set***	***DT set***	***ET set***
No. of sentences	3,007	500	500
No. of vocabularies	5,000	1,622	1,630
No. of words	55,504	8,076	8,290
Difference from TR	0	3,378	3,370
Difference from DT	0	0	609
Difference from ET	0	617	0

Table 8. Summary of a number of n-gram

n-gram	***1-gram***	***2-gram***	***3-gram***
LM (8,007 sentences)	5,000	47,354	98,423

(3) Distribution of each set

Our corpus project cooperates with two universities. We provide a fund and texts used for recording. Prince of Songkha University and Mahanakorn University of Technology have collaborated this project. We recorded 48 speakers while both universities have recorded 200 speakers. Thus, the total of speakers is 248 speakers. The distribution of each set is shown in Table 9.

Table 9. The sentence distribution of this corpus

Institute and group	***No. of speakers***	***No. of sentences per speaker***			
		PD	***TR***	***DT***	***ET***
PSU_1	60	20	101	-	-
PSU_2	20	20	-	50	-
PSU_3	20	20	-	-	50
MUT_1	60	20	101	-	-
MUT_2	20	20	-	50	-
MUT_3	20	20	-	-	50
NEC_1	24	35	126	-	-
NEC_2	12	35	-	42	-

		No. of sentences per speaker			
NEC_3	12	35	-	-	42

Each set of this corpus has been distributed to universities because the speakers, who come from the different place, usually pronounce dissimilar utterance in the same text. Also, the speakers who utter the PD and the TR set, read neither DT set nor ET set. That means each group will contain the PD set and only one set of TR, DT, or ET set.

(4) Record conditions

The utterances are recorded in two environments: the clean speech environment (CS) and the office environment (OF). These environments are separated by the signal to noise ratio (SNR) Moreover, the SNR of CS and OF are around 30 dB and 20 dB respectively. The accessories of sound recorder used in these two environments are the same, except the microphones. The microphone used in CS, is a high quality head set (Senheiser HMD-410 close-talk). For the OF, the lower quality ones, are a close-talk (TELEX H-41) and a dynamic microphone (SONY F-720), are used for recording.

All utterances are recorded according to reading styles. The average time, for reading 35 sentences (PD set), is shown in the following table. From this table, the male take more times than the female and the standard derivation (SD) of male is three times from the SD of female.

Table 10. The average time of speaker's utterance

Gender	***Time average***	***SD***
Female	216.31 second	15.24
Male	238.59 second	44.99
Average	227.45 second	30.12

3 Corpus Tools

This section describes the tools that are used in this corpus. There are automatic segmentation, automatic sentences distributor, and wave cutting tool. These tools are developed during corpus construction. Its detail is described as follows.

3.1 Automatic Phoneme Segmentation

This automatic phoneme segmentation employs Hidden Markov Models Tool Kit (HTK) [S.Young et

al. 2000] as a based system. The Automatic phoneme segmentation requires sentence transcriptions and corresponding speech database as inputs. The first process of this system is Thai Grapheme-to-Phoneme (G2P) that was developed by [P.Tarsaku et al. 2001]. This process generates phonetic transcriptions of those sentences before the flat start process was started. The flat start process constructs the initial acoustic models from the phonetic transcription (but the speech data does not label). Therefore, the speech data will be aligned in the re-label training. The re-label training employed the initial acoustic models from the flat start process to update the phonetic transcriptions. These speech alignments are used in the isolated training process for the acoustic models training again. These confirm the latest acoustic models that are better than the initial ones. After the isolated training produced the acoustic models, there were used in the second re-label training to generate the final phonetic transcriptions. All of the procedure is illustrated in Figure 2(a).

In addition, the flat start process is the acoustic model training technique of HTK in the case that has no time-alignment phonetic transcriptions. This technique starts from calculating global means and variances of all speech parameters by using those parameters as the beginning point of each phone model and retraining those models by using Baum-Welch to obtain the optimum model. This process is used for creating initial acoustic model for first re-label training.

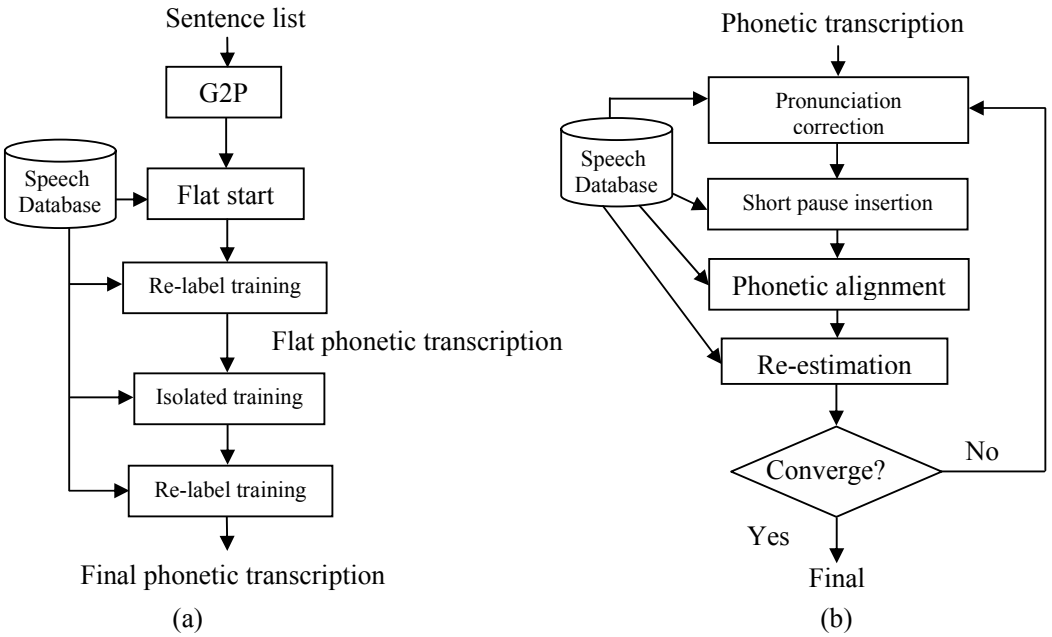

Figure 2. Automatic phoneme segmentation
(a) The procedure of automatic phoneme segmentation
(b) The process of re-label training

Furthermore, the isolated training is another training technique of HTK. Each phone is trained

separately according to the transcriptions. To do this, new segmentations are constrained and cannot go beyond sub-word boundary. Therefore, new segmentations have less error due to unsaturated model. The isolated training can be separated into two parts: non-overlapped and overlapped boundary segmentation. The process starts from non-overlapped to overlapped boundary training respectively in order to reduce error from unsaturated models. The non-overlapped boundary training used the speech utterances are fragmented. That means each speech utterance is separately. On the other hand, the overlapped boundary training allows some probabilistic overlap around the boundaries. It might compensate the co articulation problem.

Re-label training

As there is still some error from G2P, there are more than three processes applied during re-label training: pronunciation correction, short pause insertion, and phonetic alignment. The forced Viterbi algorithm is used for these processes. The re-label training can update transcriptions during training in order to obtain maximum likelihood transcriptions. The re-labeled training process is shown in Figure 2(b). And these processes are shortly described as follows.

Pronunciation Correction: As the pronunciation from G2P still has some error due to the complexity in Thai pronunciation variation, the pronunciation for training may deteriorate the quality of the training model. However, the dictionary generated from G2P limits the variation in pronunciation of each word. Using Viterbi algorithm, the correct pronunciation can be selected from those pronunciation candidates.

Short Pause Insertion: In the database, there are many long words that the speaker cannot say it without breathing. Therefore, there might be a short pause during recording process. The automatic short pause insertion is built to support this error. For the short pause insertion, there is an algorithm searching the possible point of short pause in a word (usually after syllable). Then, the forced Viterbi is employed to find the best pronunciation.

Automatic Phonetic Alignment: After the pronunciation correction, the short pause insert processes, and the correct phonetic transcriptions were generated, the time alignment transcriptions are created by using the forced Viterbi algorithm with the phonetic transcriptions.

Pronunciation correction, short pause insertion, and phonetic alignment employ the acoustic models to update the input transcriptions for the re-estimation process. Then, the re-estimation process builds the acoustic models as the input of those three processes. This process will repeat until the log probability of the

updated models is less than the last ones.

3.2 Automatic Sentences Distributor

In order to make sure the balance of phonetic distribution in our corpus, we separated the corpus with statistical balance. The first purpose is to distribute the sentences that are used for recording, in three places that are two universities and our laboratory. (Note that as the number of speakers in this corpus is large, we cannot record within one place). And the second, the distributed sentences in each place are distributed according to each speaker. The detail of this system is explained in [R. Thongprasirt et al. 2002].

3.3 Wave Cutting Tool

In the recording process, we record the speech utterances of each speaker within one DAT tape. Therefore, we have to cut the concatenated utterances according to the transcription before using in the next step. However, this was manually done so it causes a lot of time consumption. The Automatic Wave Cutting tool has been used for our work. The algorithm, which is used in this research, is modified from Unsupervised Clustering without Averaging (UWA). The basic idea of this system is just the speech recognition system that two phonetic units exist: silence and signal. The reason of using UWA in our system is its simplest and sufficient recognition system requires only small calculation.

We try to modify UWA so that it will be less calculation. Many rules were applied in this tool such as; frame energy which is used as feature vector, mean of signal that is equal to mean of every frame in the database, mean of silence that is equal to zero, and the clustering is done in one iteration. After we cluster the speech utterances into the signal and the silence, all silences that are shorter than the threshold value will be signal. This is because there might be some short pause in each sentence.

4 Problem of word variation in Thai language

Word segmentation is the crucial problem in Thai language processing. In part of the TR, DT, and ET selection [Thongprasirt et al. 2002], the most 5,000 frequent word list has been rechecking due to the problems of words segmentation. The error of words segmentation effects on the frequency words list and words may be added or deleted. The error here does not mean that the words was segmented in the wrong way, but the meanings of sentence are wrong. After all sentences parsing through automatic words segmentation program,

they have to be examined again by human. The way to point out whether they are words or not is not distinguishable even by native speakers. Actually, it depends on individual judgement. For example, most Thai may consider "ออกกำลังกาย" (exercise) a whole word, but some of them may consider "ออกกำลังกาย" a compound: "ออก" (take)+ "กำลัง" (power)+ "กาย" (body) [V. Sornlertlamvanich et al. 2000]. Therefore, the following problems have occurred:

- Compound words were segmented to be isolated words e.g. it should be "เลือกตั้ง" (election) instead of "เลือก" (to select)+ "ตั้ง" (to put).

- In the other way, isolated words were decided to be compound words, e.g. it should be "ให้" (to give) + "การ" (prefix) instead of "ให้การ" (to give an evidence) in some context.

These kinds of problem depend on human judgement using their lexical knowledge base. Words were defined and based on their meaning in the context. To overcome these problems, we try to get through a whole 5000 words list, especially, in words which may be considered in both way and then go back to determine it in sentence again by linguists. Actually, it is time consuming and there are some words that are difficult to make a decision.

References

J.G. Wilpon and L.R. Rabiner, "A modified K-means clustering algorithm for use in isolated word recognition", IEEE Trans. Acoustics, Speech Signal Proc., ASSP-33 (3), pp 587-594, 1985.

L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recogition", In Proc. IEEE, 77 volume 2, pp 257 – 286, 1989.

V. Sornlertlamvanich, N. Takahashi, and H. Isahara, "Thai Part-Of-Speech tagged corpus: ORCHID", Proceedings of the Oriental COCOSDA Workshop, pp.131-138, 1998.

K. Sjölander, J. Beskow, "Wavesurfer – An Open Source Speech Tool" In Proc. *ICSLP*, volume 4, pp. 464 – 467, 2000.

M.J. Makashay, C.W. Wightman, A.K. Syrdal, A. Conkie, "Perceptual Evaluation of Automatic Segmentation in Text-to-speech Synthesis", In Proc. *ICSLP*, volume 2, pp. 431– 434, 2000.

S. Young D. Kershaw, J. Odell, D. Ollason, V. Valchev, P. Woodland, "The HTK book", http://htk.eng.cam.ac.uk/docs/docs.shtml, 2000.

P. Tarsaku, V. Sornlertlamvanich, R. Thongprasirt, “Thai Grapheme-to-Phoneme using Probabilistic GLR Parser” In Proc. Eurospeech, volume 2, pp. 1057 – 1060, 2001.

S. Nefti., O. Boeffard, “Acoustical and Topological Experiments for an HMM-based Speech Segmentation System”, In Proc. *Eurospeech*, pp. 1711 – 1714, 2001.

C. Wutiwatchai, P. Cotsomrong, S. Suebvisai, S. Kanokphara, “Phonetically Distributed Continuous Speech Corpus for Thai Language” Third International Conference on Language Resources and Evaluation (LREC2002), pp.869-872, 2002.

R. Thongprasirt, V. Sornlertlamvanich, P. Cotsomrong, S. Subevisai, S. Kanokphara, “Progress Report Corpus Development and Speech Technology in Thailand” COCOSDA, pp. 300-306, 2002.