Thai Speech Recognition Corpora

Sawit Kasuriya[1], Virach Sornlertlamvanich[1,2], Patcharika Cotsomrong[1], Supphanat Kanokphara[1], and Nattanun Thatphithakkul[1]
[1]Speech technology section, Information research and development division, National electronics and computer technology center (NECTEC)
[2]Thai Computational Linguistics Laboatory, CRL
112 Phahon Yothin Rd., Klong 1, Klong Luang, Pathumthani 12120, THAILAND
sawitk@nectec.or.th, virach@nectec.or.th, aye@nectec.or.th,
supphanat_k@note.nectec.or.th, nattanun_t@note.nectec.or.th

Abstract

Nowadays, the improvement of speech recognition technology is growing fast and many techniques have been proposed. However, even the best algorithm with carefully designed system cannot accomplish good-performance speech recognition if the system is trained from unreliable corpus. Therefore, the speech corpus is a crucial research area. This paper describes the speech corpus (ORCHID-SPEECH CORPUS and NECTEC-ATR Thai speech corpus), which is developed for Thai speech recognition. It also indicates how the speech corpus is built in order to preserve important properties: consistency, balance, and containing possible phoneme combinations. Therefore, the corpus design, the details of each corpus set, and problem of them are also presented in this paper.

Keywords

speech corpus, speech recognition, corpus design, phonetically balance, phonetically distributed, and Thai language

1. Introduction

The crucial part of speech recognition application development is speech corpus that is used for training the acoustic models and language models, testing and evaluating the system. The essential part of speech corpus construction is the design step for building the speech recognition application system. The objectives of each speech corpus correspond with application tasks, the size of corpus, the speech type (isolated-word or continuous speech), and the channel of speech (clean speech, quasi-quiet, or telephone line). All of there factors have some trade-offs with the objective of application task.

This paper presents the Thai-language speech corpus that has been developed for a speech recognition system. There are two speech corpus: Thai speech corpus for large vocabulary continuous speech recognition (ORCHID-SPEECH CORPUS) and NECTEC-ATR Thai speech corpus. The main purpose of both corpus is to develop both Thai acoustic and language models for some domains. Both of there models were initiated by National Electronics and Computer Technology Center (NECTEC) and cooperative organization which included research institutes and universities. The details of them are described in the following section.

The rest of this paper is organized as follows. The second section presents the basic knowledge of Thai language, Thai phonetic system, defining transcription and text processing. The third section describes the details of both Thai speech corpus, for example, corpus design, structure of each corpus set, and record condition. The last section presents the language variation in Thai language which explains why Thai language is the one of ambiguous language.

2. Overview of Thai Speech Corpus

The common process of designing a speech corpus consists of text selection, text processing, and defining transcription. These processes are briefly described in this section.

2.1 Domain of text corpus

The domain of text corpus that is used for word or sentence selection, is very significant in designing a speech corpus. To be useful in a real-world application, the domain must be specified close to the application task. For example, some corpus were collected from the dialogues in front of hotel reception for developing the automatic hotel reservation system. Example of Thai text corpus include the Open Linguistic Resources CHannelled toward InterDisciplinary research with POS tagged corpus (ORCHID-POS TAGGED CORPUS) (V. Sornlertlamvanich et al. 1998), magazines, Thai encyclopaedia, and journals. Only ORCHID-POS TAGGED CORPUS has already manually tagged for text corpus. It contains 27,634 sentences. In addition to ORCHID-POS TAGGED CORPUS, others text corpus are included, which contain nearly 2,500,000 words. (43,255 vocabularies) from 180,504 sentences. This text corpus is used for selection in phonetically balanced sentence set (PB) and phonetically distributed sentence set (PD).

Therefore, the fist step of corpus creation is the text management phase. This phase is the article selection, sentences selection, word segmentation, and grapheme-to-phoneme (G2P) processing. The main objective is to choose the suitable sentences for the next step (text processing). Therefore reducing the total time of the text processing. This phase is quite important for text processing since the size of original text is enormous.

2.2 Text processing

Thai language is the one of the alphabetic language. There is no explicit use of any word and sentence delimiting symbol or character. Sometimes the space is placed only between adjacent sentences but it is very ambiguous 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. As a consequence, we have to manually handle the text corpus which is a laborious task and requires many linguists. The text corpus that is used in this development, took more than a year to arrange for speech corpus.

In this section, the briefly detail of grapheme-to-phoneme (G2P) that are the principle of text processing, is 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. There are many techniques for implementing the G2P such as, dictionary-based, rule-based, and statistical-based. The detail of our latest approach was presented in (P. Tarsaku et al. 2001). This module 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 completely accurate (approximately 80%). Furthermore, the G2P module has some error due to the difficulty in detecting the word boundaries for rare Thai words and foreign words. Therefore, its performance also strongly depends on word boundaries. These phonemes of word were checked by the linguists after they were passed from the G2P process.

2.3 Defining transcription

Defining transcriptions are necessary during the corpus creation because Thai phonetic system differs from other languages. Some of Thai phonemes are used as the same symbol in another language, for example, the standard phonemes such as /i/, /a/, and etc. Most Thai phonemes are not the same the phonemes (usually are consonants) from other languages. One important thing when we defined the transcriptions is that it must be language-independent. They must not contain any reserved characters. There is no transcription standard in Thai. Thus many researchers have tried to set phoneme symbols for using in their own research. Therefore, this paper has defined most Thai phoneme symbols to be compatible with another Thai researchers or linguists (S. Luksaneeyanawin 1993). But some phonemes are not the same due to some problem in programming. Consequently, we changed it to other forms such as /?/ to /z/ and double characters of vowel are used for the long vowels.

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

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 1. Thai Phonetic Symbol for Thai initial and final consonant

Therefore, 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 consonants (including single and double consonants), V denotes the vowel (both short and long vowel), C_f denotes the final consonants (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_0/. Because same symbols stand for many initial and final consonants, '^'symbol is represented for final consonants in order to differentiate from initial consonants. Finally, a number of Thai phonetic symbols are 74. These symbols do not considered Thai tones. If the tones are considered in each phonetic symbol, the total number of Thai phonetic symbols will approximately increase by five times. The following Table 1, 2, and 3 show the Thai Phonetic Symbol.

Double Consonant	***Phoneme***	***Double Consonant***	***Phoneme***
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 2. Thai Phonetic Symbol for Thai double consonant

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)

Table 3. Thai Phonetic Symbol for Thai vowel

3. Current Thai speech corpus for speech recognition

The acoustic models (AM) and the language models (LM) construction depend on the size of speech database. If the size is large enough, it is very helpful to speaker-independent AM training (S. Kasuriya et al. 2002). The language models are also the same condition with the acoustic model training but they differ from the language models, particularly the text domain of each application. On the other hand, the acoustic models can be constructed from the other speech data domain (same environment and language). But the language models are not trained from the other text data, which differ from the target domain.

This paper covers two speech corpus: Thai speech corpus for large vocabulary continuous speech recognition (ORCHID-SPEECH CORPUS) and NECTEC-ATR Thai speech corpus. The different between both corpus is the corpus size, the purpose of each set in the corpus, and cooperative researcher team. However, the main objective of both corpus

is to construct Thai speech corpus for speech recognition. Both corpus construction includes the sentence selection process, the statistics of sentence in the phonetically balanced sentence set and the phonetically distributed sentence set, the corpus contents, and recording distribution of each set.

3.1 Thai speech corpus for large vocabulary continuous speech recognition

Probably the most prevalent problem in developing Thai speech recognition is the lack of large standardized speech corpus, which makes it impossible for researchers to develop a commercially usable Thai speech recognition module. The main usage of this corpus is for developing a large vocabulary continuous speech recognition system. This project is a cooperation between NECTEC and two universities: Prince of Songkha University (PSU) and Mahanakorn University of Technology (MUT) to record and label the speech data. This corpus is called ORCHID-SPEECH CORPUS. The corpus design, text selection, and recording processes are discussed in this section.

3.1.1 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. This corpus used ORCHID-POS TAGGED CORPUS, magazines, Thai encyclopaedia, and journals as the text domain.

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 details of both sets are described as follows.

(1) Phonetically distributed sentence (PD) set

To construct 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 set is the extension of PB set. It does not only cover all biphone, but the text distribution is also similar to the daily used context (ORCHID-POS TAGGED CORPUS corpus in this case).

The overall PD selection procedure is illustrated in Figure 1. The PD selection process starts from PB construction. The PB set was collected in advance and became the initial set for the PD set. The unit score in PB selection is a reverse of unit frequency, whereas the unit score in PD selection is a constant subtracted by a unit reduction score and multiplied by the number of times it has been selected and included in PD set. The unit reduction score is the reverse of unit frequency. R is defined in the PD selection step as the degree to which the statistical distribution of the units in the selected sentence set is similar to that in the original text. More details can be explored in (Shen 1999).

During the PB construction process of this corpus, 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-POS TAGGED CORPUS are calculated. Then, some sentences are added to PB to change the distribution to the distributions of ORCHID-POS TAGGED CORPUS. The number of adding sentences should be kept at minimum while the biphone distribution of PD set is the closest to the ORCHID-POS TAGGED CORPUS's distribution. More details of PB and PD construction can be found in (C. Wutiwiwatchai et al. 2002). The summary of PD set is shown in Table 4. Table 5 classifies the number of syllables in PD set.

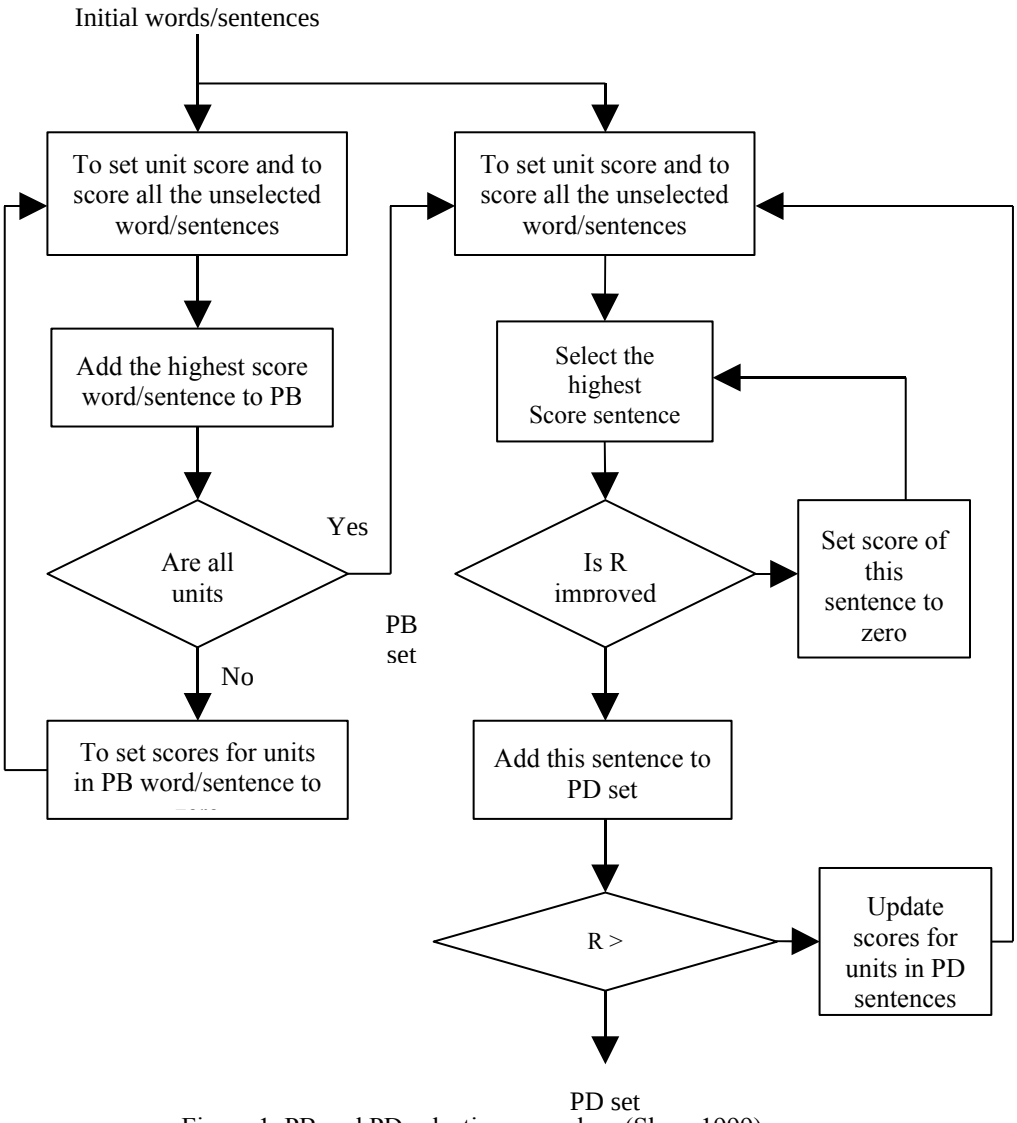

Figure 1. PB and PD selection procedure (Shen, 1999)

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 4. Summary of Phonetically distributed sentences set

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 5. Classified the syllables

(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.

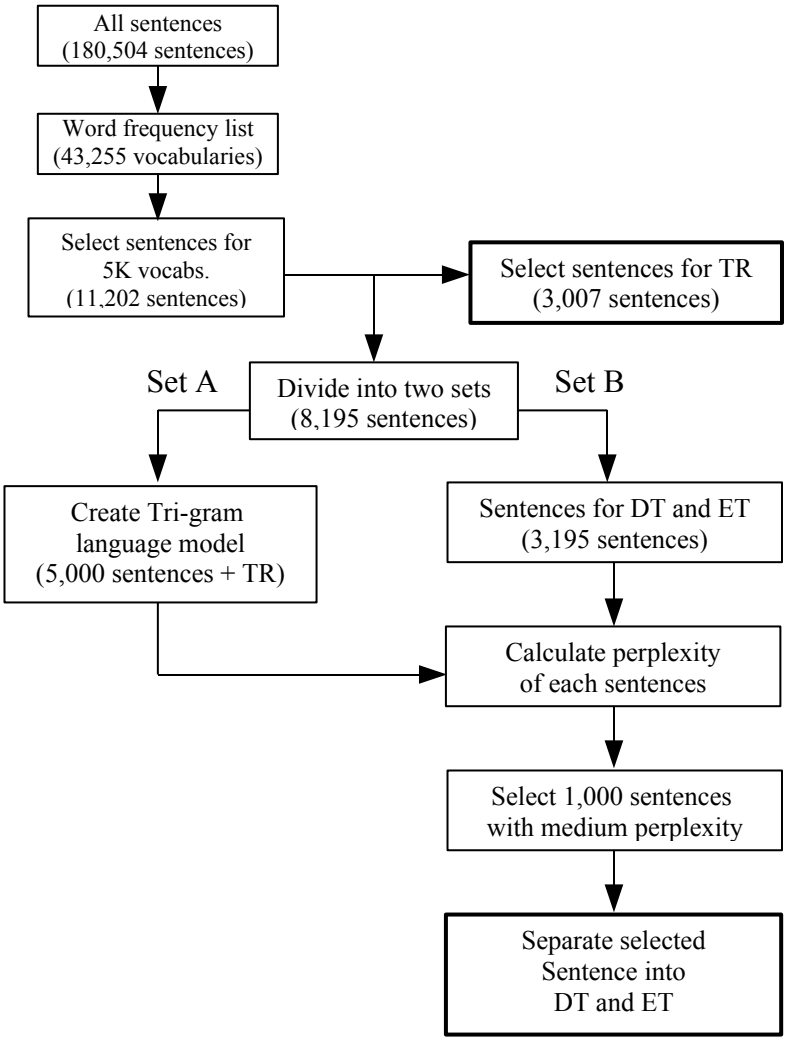

Figure 2. TR, DT and ET selection

The process of TR, DT, and ET selections are illustrated in Figure 2. 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 have 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 procedures, 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. Table 6 and Table 7 are shown the summary of 5,000 vocabulary coverage sentences set and summary of a number of n-gram, respectively.

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

where: SC denotes the sentence scores

N_w denotes the number of words in each sentences

$\mathrm{wf_i}$ denotes the i^{th} word frequency of 8,195 sentences

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 6. Summary of 5,000 Thai vocabulary coverage sentences set

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

Table 7. Summary of a number of n-gram

3.1.2 Distribution of each set

Our corpus project cooperates with two universities, Prince of Songkha University and Mahanakorn University of Technology. We provide a fund and texts for recording. NECTEC 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 8. 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.

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	-
NEC_3	12	35	-	-	42

Table 8. The sentence distribution of this corpus

3.1.3 Phone alignment

Only in the PD set, every phoneme following in the transcription has been label with the time period that was consistent with its utterance. For transcribing the phoneme duration, WAVESURFER program is selected as a tool to mark a phone boundary (K. Sjölander and J. Beskow 2000). The restriction of silence at the beginning and ending of the file should not be less than 300 ms and the restrictions of phoneme alignment are as follow:

(1) At the beginning and ending of each file, a silence phase or /sil/ must be placed. That means the silence phase appears before the first phoneme. And the last phoneme is followed by the silence phase.

(2) Each phoneme separated by waveform changing and spectrogram. Considered by visualization and listening coordinately, phoneme is divided at zero crossing point.

(3) If the silence space between phonemes is less than 20 ms, the following rules will be considered.

(3.1) There is vowel in front of silence space, and it comes up with consonant. The silence space will be a part of consonant.

(3.2) There is a silence space between plosive consonants or not plosive consonants. The silence space will be divided between two consonants in half.

(3.3) There is a silence space between fricative consonant and any consonants except plosive consonant. The silence space will be segmented as a part of fricative consonant.

(3.4) There is a silence space between plosive consonant and fricative consonant. The silence space will be segmented as 70 percents for plosive consonant and 30 percents for fricative consonant.

(4) If the length of silence space between phonemes is more than 20 ms, the silence space will be considered to be a short pause.

(5) In the case that phonemes cannot be differentiated by considering waveform and spectrogram, the following rules will be applied.

- If the vowel is the constituent of the approximant consonant, the listening and the formant changing will be considered.

- If the nasal consonant is followed by the nasal consonant, the criteria of waveform will be segmented as 60 percents for the final consonant and 40 percents for the initial consonant. This is because the final consonant seems to be longer when it has pronounced.

- In the case of the nasal consonant is followed by the approximant consonant, the previous rule (the nasal consonant is followed by the nasal consonant) will be applied.

3.1.4 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 recorders 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.

Firstly, speeches are recording into a DAT tape with a sampling rate of 48 kHz and 16-bit quantizations. Recorded speeches were then played back into PC via an optical connection to a sound card. A waveform is then downsampled into 16 kHz (also use the prefilter to avoid aliasing). Note that a power supply is needed for the OF environment, since the TELEX H-41 headset has a condenser microphone. All utterances are recorded according to reading styles. The average time, for reading 35 sentences (PD set), is shown in the following table (Table 9). 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.

Gender	Time Average	SD
Female	216.21 second	15.24
Male	238.59 second	44.99
Average	227.45 second	30.12

Table 9. The average time of speaker's utterance

3.2 NECTEC-ATR Thai speech corpus

This Thai speech corpus is cooperated by NECTEC, Thailand and spoken language translation research laboratories of Advanced Telecommunications Research Institute International (ATR), Japan. We have designed this corpus for the acoustic models construction and the language models of the hotel reservation (S. Kasuriya et al. 2003). This corpus is quite small in size. But the content of corpus should cover all Thai phones. Therefore, this corpus is divided into three principle parts: isolated-words set (DB1), phonetic balanced sentences set (DB2), and conversational speech set (DB3). The details of these sets are as follows:

3.1.1 Corpus design

First step of database design is the text selection. This process is included in the designing step that should be set during the step of designing database structures. Hence, the details of text selection or creation are described in this section. The domain of text that is used in this database is collected from magazines, journals, encyclopaedias, and newspapers. After we have got the text corpus, original text is classified and segmented. This text corpus is

called ORCHID-POS TAGGED CORPUS. Henceforth, the details of database design in each set are described as follows:

DB1 (Isolated word set)

This set consists of three subsets. The principle of this set is 5,000 vocabularies subset, which is the biggest subset in this set. Therefore, we divided into five subsets (D0 to D4). Each subset contains 1,000 vocabulary words. The other subsets are the phonetic balanced words and the extra words. The details of each set creation is described in the following and the details of them are shown in Table 10.

- *5,000 vocabularies subset (D0-D4):* We have counted the frequency of each vocabulary in the LEXiTRON dictionary with the text corpus (Thai magazines, journals, and encyclopaedias) and the top 5,000 vocabularies of the higher frequencies were selected. Then they are randomly divided into five subsets. The process of 5,000 vocabularies subset is shown in Figure 5.

- *PB word subset (D5):* the 5,000 vocabularies were used to select the PB words. The selection procedure is presented in Figure 1. All phonemes with the least amount of words and balanced occurrence were collected. Hence, the phoneme occurrences in this set equal to phoneme occurrences of 5,000 vocabularies subset. A number of words in this subset are 640. Furthermore, this selection procedure is also used in PB sentences selection.

- *Extra word subset (D5):* The vocabulary word that occurred in conversational speech set and did not occur in 5,000 vocabularies subset or PB word subset is called "Extra words" which contained 131 words.

Attribute	**D0**	**D1**	**D2**	**D3**	**D4**	**PB**	**Extra**
No. of words	1,000	1,000	1,000	1,000	1,000	640	131
No. of syllables	1,933	1,970	1,912	1,938	1,934	1,157	338
No. of phones	5,092	5,223	5,047	5,139	5,119	2,955	857
No. of unique syllables	809	821	792	974	944	788	229
No. of unique phones	63	62	63	64	61	64	59
No. of unique biphones	907	882	886	886	910	1,103	389
No. of unique triphones	3,122	3,132	3,066	3,088	3,032	2,535	688

Table 10. The details of each subset in DB1

DB2 (Phonetically Balanced Sentences set)

The purpose of this set is to collect the sentences that consist of Thai biphones. The Large text corpus is required in order to get various biphone defining. But the text corpus that was used in this database is quite small; all Thai biphones were not collected. The procedure in PB sentences selection is the same as PB words selection as shown in Figure 1. The sentences in this procedure were used as the input. The number of any attributes in ORCHID-POS TAGGED CORPUS and PB sentences set are shown in Table 11. This PB selection process is the one step of PD selection process that is described in the previous section of another Thai speech corpus.

Attribute	**ORCHID-POS TAGGED CORPUS**	**PB sentences**
No. of sentences	27,634	390
No. of words	352,113	3,900

No. of syllables	568,490	5,319
No. of phones	1,398,994	13,440
No. of biphones	N/A	1,628

Table 11. Comparison between ORCHID-POS TAGGED CORPUS and PB sentences set

DB3 (Conversational Speech Set)

A hotel reservation system is the one of well-known speech recognition application. It is useful to the traveling and hotel business. Resources of text are the major problem of this application because there are many speaking styles, several different dialogues, many kinds of hotel, and etc. The transcriptions that were used in this database are translated from 50 dialogues that were used at Spoken Language Translation Research Laboratories (SLT), Advanced Telecommunication Research International Institute (ATR) in Kyoto, Japan. These dialogues have been translated to more than two languages such as English and Japanese.

The details of each corpus set are briefly shown in Table 12 and the size of each set is shown in Table 13. From Table 1, Table 2, and Table 3, there are 74 Thai phones. But the number of phones in each corpus set (as shown in Table 12) is not 74 because some phones have missed in the text corpus such as “ia”, “bl”, “br”, “dr”, “fl”, and “l^” in DB1, “ua”, and “va” in DB2, and “bl”, “br”, “dr”, “fl”, “fr”, “khl”, “ia”, “ua”, “va” in DB3. Most missing phones in DB1 are foreign phones. And DB2 is the most nearly complete Thai phones.

Attribute	**DB1**	**DB2**	**DB3**
No. of sentences	None	398	1,637
No. of words	5,771	3,377	18,787
No. of syllables	11,182	5,501	23,308
No. of phones	29,432	14,472	61,489
No. of unique words	5,771	1,478	736
No. of unique syllables	1,668	1,160	586
No. of unique phones	68	72	65
No. of unique biphones	1,296	953	1,619
No. of unique triphones	8,244	6,032	4,437

Table 12. The details of each corpus set

Time unit	**DB1**	**DB2**	**DB3**
Minutes	890.31	1,388.04	266.20
Hours	14.84	23.13	4.44

Table 13. Time recording of each set

3.1.2 Record conditions

All utterances were recorded in a quasi-quiet room. The qualities of them are around 20 dB. And only dynamic microphone (unidirectional microphone: SONY F720) is used in recording. A number of speakers are 20 males and 20 females (18 to 40 years old). Each speaker uttered all database sets (DB1 to DB3) and uttered only one subset (D0-D4) in DB1, D5 in DB1, DB2, and five dialogues in DB3. Each subset is distributed to balance

recording. All utterance is recorded in reading style and it is middle and official dialect that is spoken in the middle area of Thailand.

4. Language variation

In first step of corpus creation, text classification and text processing are used in text management step as described in 2.1 and 2.2. Many techniques of text analysis were applied in Thai text processing such as, morphology analysis, syntax analysis, and phonology analysis. The details of them are described in the following.

4.1 Morphology analysis

Word segmentation is the crucial problem in Thai language processing. In part of the TR, DT, and ET selection, 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 parse through the automatic words segmentation program, they have to be examined again by human. The way to point out some words is not distinguishable even by native speakers. Actually, it depends on individual judgment. 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. 1998). 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 judgment 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 5,000 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.

4.2 Syntax analysis

For the PD set, the text from ORCHID-POS TAGGED CORPUS is used, for the other sets text from various places such as journal or magazine are used. Furthermore, the text corpus must be tagging with any information that is important in language structure or text analysis such as part-of-speech (POS), word boundary, sentence boundary, and paragraph separation. Text resources or plain texts that are found in any Thai websites, therefore, is not good enough to be used in the text selection step. It's time consumer in phase of text preparation, text tagging and text segmentation. Therefore, it would be more convenient if we used the text corpus, which is completely done those processes. After all usable sentences have been selected, the texts are word-segmented automatically and then manually rechecked. Each selected sentence is then transformed into its equivalent non-verbal form by removing and/or changing any special symbols to be verbal word such as hyphen, question mark and repeater symbol. In addition, any foreign word is changed into its corresponding Thai word, and any parenthetical information is discarded.

4.3 Phonology analysis

The objective of text selection step is to keep all Thai phones and possible phone combinations such as biphones and triphones. After finishing text selection phase, G2P is an important module in corpus building. Although the G2P module has been improved, some errors are unavoidable. Some important rules as described below are regulated in order to transcribe consistently.

1) The phonemes in some syllables are not corresponding to their graphemes. For example, /th-a:-n-2/ (you) is always distorted to be /th-a-n-2/ (vowel shortened when read while its grapheme is still presented as long vowel). In this case, we have chosen the shortened pronunciation which is the more natural speaking style.

2) Some abbreviations can be pronounced in either full or abbreviated form. For example, /ph-@:-0/s-@:-4/ (B.E.) is the pronunciation of the abbreviation of /ph-u-t-3/th-a-3/s-a-k-1/k-a-1/r-a:-t-1/ (Buddhist Era). Thai native people use both pronunciations when read. We consistently assign every abbreviation to be read as its full pronunciation.

3) Nowadays, some word's pronunciations are not unique such as /ph-a-n-0/j-a:-0/ and /ph-a-n-0/r-a-3/j-a:-0/ (wife). Actually the correct pronunciations of these words have been defined officially, but many people still don't know which one is correct. In this case, we force every speaker to pronounce uniquely and correctly.

4) Orthographies of some loan words from foreign language are not defined officially. Spelling of these words usually deviate from their pronunciations especially in their tones. We have defined the orthographies and the pronunciations for our corpus specially. (C.Wutiwiwatchai et al. 2002)

Acknowledgements

The authors would like to thank Chai Wutiwiwatchai for his help in designing, and advice particularly. This research cannot be finished without Rachod Thongprosirt, who helps creating the collaboration with the universities and to set up the plan of this corpus. The authors wish to acknowledge the cooperation received from Montri Karnjanadecha (PSU) and Tanee Demeechai (MUT) in ORCHID-SPEECH CORPUS project, and NECTEC-ATR Thai speech corpus support received from SLT Dept. 2, ATR.

References

S. Luksaneeyanawin, 1993, Speech Computing and Speech Technology in Thailand, *Proceeding of the Symposium on Natural Language Proceeding in Thailand*, pp. 276-321.

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

J. L. Shen, H. M. Wang, R. Y. Lyu, and L. S. Lee, 1999, Automatic selection of phonetically distributed sentence sets for speaker adaptation with application to large vocabulary Mandarin speech recognition, *Journal of Computer Speech and Language*, vol. 13, no.1, pp. 7–98.

K. Sjölander, J. Beskow, 2000, Wavesurfer – An Open Source Speech Tool, *International Conference on Speech Processing (ICSLP)*, vol. 4, pp. 464-467.

P. Tarsaku, V. Sornlertlamvanich, R. Thongprasirt, 2001, Thai Grapheme-to-Phoneme using Probabilistic GLR Parser, *Eurospeech*, vol. 2, pp. 1057-1060.

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

R. Thongprasirt, V. Sornlertlamvanich, P. Cotsomrong, S. Subevisai, S. Kanokphara, 2002, Progress Report Corpus Development and Speech Technology in Thailand, International Coordinating Committee on Speech Database and Speech I/O system Assessment (*COCOSDA)*, pp. 300-306.

S. Kasuriya, T. Jitsuhiro, G. Kikui, and Y. Sagisaka, 2002, Thai Speech Recognition by Acoustic Models Mapped from Japanese, *Joint International Conference of SNLP-Oriental COCOSDA 2002*, pp. 211-216.

S. Kasuriya, V. Sornlertlamvanich, P. Cotsomrong, T. Jitsuhiro, G. Kikui, and Y. Sagisaka, 2003, Thai Speech Database for Speech Recognition, *International Coordinating Committee on Speech Databases and Speech I/O System Assessment*, pp. 105-111.