Phonetically Distributed Continuous Speech Corpus for Thai Language

Chai Wutiwiwatchai[1], Patcharika Cotsomrong[2], Sinaporn Suebvisai[3], Supphanat Kanokphara[4]

Information Research and Development Unit
National Electronics and Computer Technology Center
112 Thailand Science Park , Paholyothin Rd., Klong 1
Klong Luang , Pathumthani 12120 Thailand
[1]chai@nectec.or.th, [2]aye@nectec.or.th, [3]sinaporn_s@notes.nectec.or.th, [4]supphanat_k@notes.nectec.or.th

Abstract

This paper proposes a work on phonetically balanced sentence (PB) and phonetically distributed sentence (PD) set, which are parts of the text prompt for speech recording in Large Vocabulary Continuous Speech Recognition (LVCSR) corpus for Thai language. Firstly, a protocol of Thai phonetic transcription and some essential rules of phonetic correction after grapheme-to-phoneme (G2P) process are described. An iterative procedure of PB and PD sentence selection is conducted in order to avoid tedious work of manual phone correction on all initial sentences. A standard text corpus, ORCHID, was chosen for the initial text. Analysis of several attributes such as the number of words, syllables, monophones and biphones, phone's distribution, etc., in both the PB and PD sets are reported. At the end, the final selected PB are partially compared to the American English TIMIT's PB set (MIT-450) and the Japanese ATR's 503 PB set.

1. Introduction

1.1. Thai LVCSR corpus project

Information Research and Development Unit (RD-I) at National Electronics and Computer Technology Center (NECTEC), Thailand, has initiated a project of the first Large Vocabulary Continuous Speech Recognition (LVCSR) corpus for Thai language since October 2000 (Wutiwiwatchai, 1999; Sornlertlamvanich & Thongprasirt 2001). The project is a collaboration among several Thai universities with NECTEC as the host center. The corpus aims for research and development of 5000-vocabulary dictation system. It consists of two sentence sets, a phonetically distributed (PD) set and a 5000-vocabulary coverage set. The PD sentence set must cover all phonemic units occur in the language. It is used for phone model initialization in the speech recognizer. The 5000-vocabulary coverage set contains 3 separated sets, training set (TR), development test set (DT), and evaluation test set (ET). Both sentence sets are carefully distributed for each speaker in recording session as shown in Table1. The speakers are requested to read assigned sentences in both clean and office environment. Full details of this corpus are presented in the proposal provided by Wutiwiwatchai (1999).

Group of Speakers	No. of Speakers	No. of Recorded Sentences			
		PD	TR	DT	ET
Group 1	100	20	80	-	-
Group 2	50	20	-	40	-
Group 3	50	20	-	-	40

Table 1: Distribution of sentences for recording

1.2. Thai language

Thai writing system is alphabetic-based. There is no space between words and no explicit rule of placing spaces between adjacent sentences. Word segmentation as well as sentence detection become crucial problems for natural language processing. Moreover, the definitions of both word and sentence are not unique especially when applied to different tasks. Thai syllable can be expressed in the form of $/C_i - V - (C_f) - T/$, where C_i denotes the initial consonant (either single consonant or consonant cluster), V denotes the vowel (either single vowel or diphthong), an optional C_f represents the final consonant, and T is the tone mark. More details of Thai phonetics defined in our work are presented in section 3.1.

1.3. Phonetically distributed sentence (PD) set

This paper focuses on the design of the corpus, selection procedure, and analysis of the final selected PD set. Normally, a set of phonetically balanced sentences (PB) is designed to minimally cover at most possible phonemic units occur in the language, in our case, phoneme pairs (or as called in this paper - biphones). The PD set does not only follow PB criterion, but the distribution of units in the set also corresponds to that of the original text. Due to its compactness and phone coverage property, speech utterances of PD set can be used for phone model initialization as well as for speaker adaptation problem.

The difficulties of Thai writing system as described previously enlarge the error of the grapheme-to-phoneme (G2P) module in which its performance strongly depends on word boundaries. ORCHID (Sornlertlamvanich, 1998), a sentence, word, and part-of-speech (POS) tagged corpus that is one of the standard corpus created for POS tagging research, was used as the initial text for PD selection. Section 2 describes an overall selection procedure beginning from the original text, ORCHID, to the final selected sentence sets. The implementation of the selection procedure including the protocol of Thai phonetic transcription, some essential rules of phonetic and the convergence of iterative selection are reported in section 3. Section 4 gives the analysis of the final PB and PD sets. The result is partially compared to the American English TIMIT's PB set (MIT-450) (Lamel, Kassel, & Seneff, 1986) and the Japanese ATR's 503 PB sentences (Iso, Watanabe & Kuwabara, 1988). Finally, section 5 summarizes our work.

2. Selection Procedure

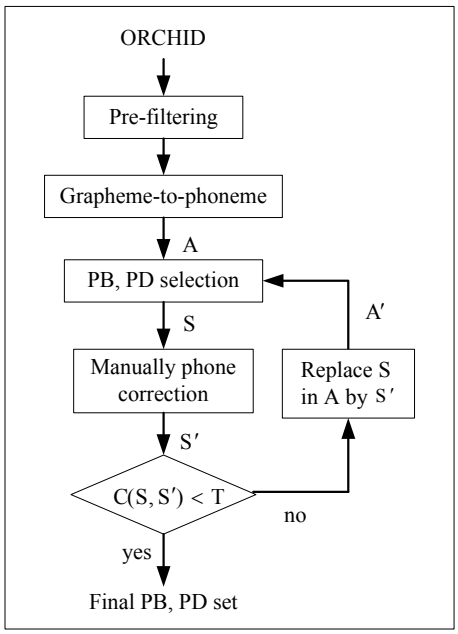

Figure 1: Overall PD selection procedure

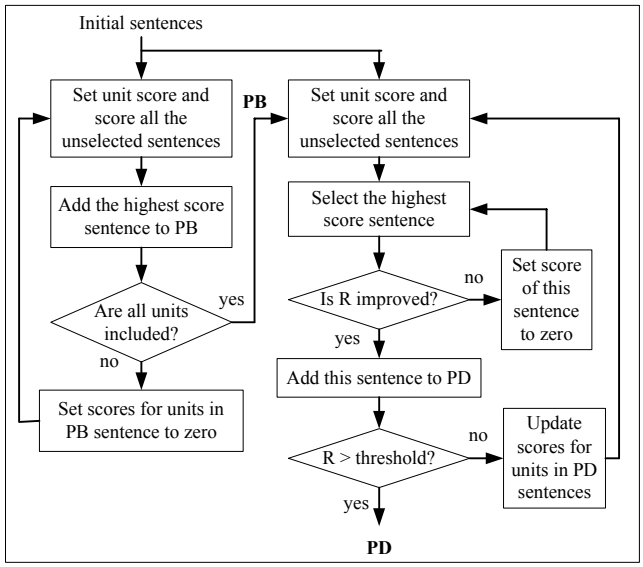

Figure 2: PB and PD selection procedure (Shen, et al, 1999)

Figure 1 illustrates an overall procedure of PD selection starting with manually-tagged ORCHID. After screening out the sentences containing English characters in the pre-filtering step, the remaining was processed by the newly improved G2P for Thai language (Tarsaku, Sornlertlamvanich & Thongprasirt, 2001) which has been suitably modified and adopted. Before selection, we got a total of 27,634 sentences denoted as A in Figure 1. The selection of minimal PB and PD sets were conducted with an approach proposed by Shen, et al (1999). To avoid the tiresome work of manual phone correction on all sentences in A, phone correction was performed only on the selected PD sentences. Let's define S = {PB, PD} as the selected sentence set (PB is always a subset of PD) and $S' = \{PB', PD'\}$ as the sentences after manual phone correction. $C(S, S')$, computed from S and S', is used for stopping criteria of iterative selection. S' becomes the final PB and PD set if $C(S, S')$ achieves a set threshold T. Otherwise, A will be replaced by A' (S is replaced by S') and will be an initial set of the next iteration. In our work $C(S, S')$ was the different between unique-biphone count of PB and PB' sets.

The sub-process of PB and PD selection is illustrated in Figure 2. 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 multiplied by the number of times it has been selected and included in PD set. The unit reduction score is again the reverse of unit frequency. R as described in the PD selection step is the degree to which the statistical distribution of the units in the selected sentence set is similar to that in the original text. The dot product between the distribution of units in original text and that in the PD set is a good measure for R. More details can be explored in Shen, et al (1999).

3. Implementation of Selection Procedure

3.1. Transcription protocol

In the selection procedure, a set of phones (phonetics) must be defined for G2P module and phone correction task. Table 2 overviews Thai phonetic system and Table 3 describes the phones used in each syllable in our work.

Consonant	Asp. stop Unasp. stop Voiced stop Nasal Fricative Semivowel Liquid	ph th kh ch p t k c b d m n ng f s h w j r l
Vowel	Long Short	i: e: x: v: q: a: u: o: @: i e x v q a u o @

Table 2: Thai phonetic system

C_i	Single: ph th ch kh p t c k h b d m n ng r l j w s f z Cluster: phr phl thr khr khl khw pr pl tr kr kl kw br bl fr fl dr	38 units
V	Single: i i: e e: x x: v v: q q: u u: @ @: Diphthong: ia i:a va v:a ua u:a	24 units
C_f	p^ t^ k^ m^ n^ ng^ w^ j^ f^ s^ ch^ l^	12 units
	Total	74 units

Table 3: Phonetic symbols used in our work

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/

(means "you") is always distorted to be /th-a-n-2/[1] (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/ (means "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/ (meant "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/ (means "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.

3.2. Convergence of sentence selection

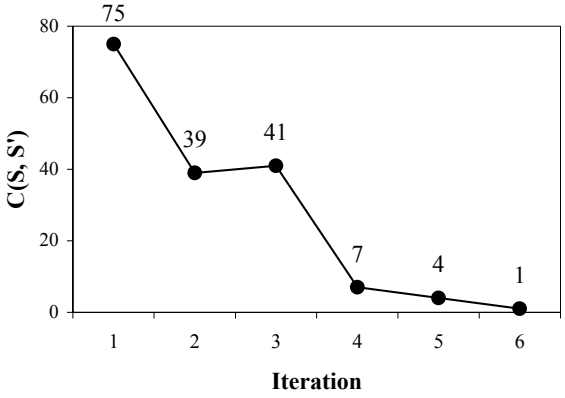

Figure 2: The value of $C(S, S')$ at each iteration

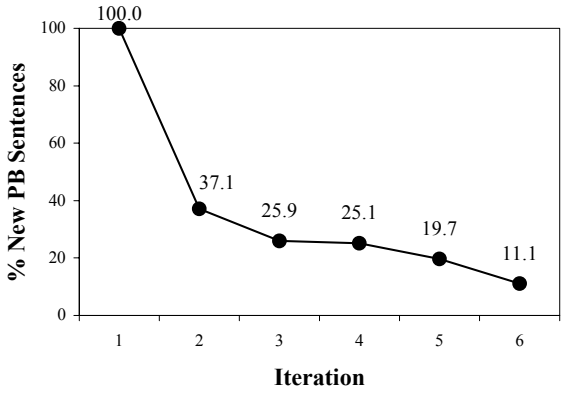

Figure 3: The number of new sentences occurred in PB set at each iteration

Manual correction of G2P's result is very wearisome. Instead of correction on the whole initial text, we performed an iterative selection as described in section 2. The difference between unique-biphone count of PB and that of PB' is defined for $C(S, S')$. By using '$C(S, S') < 3$' as our convergence criterion, the selection stopped at the 6th iteration. At this point, we obtained 802 PD sentences and 398 PB sentences containing 72 unique monophones and 1,628 unique biphones. Figure 2 shows the value of $C(S, S')$ and Figure 3 shows the percentage of new sentences that appeared in the PB set at each iteration.

4. Analysis of PB and PD Set

Some significant statistics of the final PB and PD sentence set compared to the original ORCHID are shown in Table 4. PB sentence sets of other languages have been created and used widely in speech research like in American English (MIT-450) (Lamel, Kassel, & Seneff, 1986) and Japanese (ATR-503) (Iso, Watanabe & Kuwabara, 1988). Our PB sentence set is compared to both standard PB sets in Table 5.

Attribute	ORCHID	PB	PD
No. of sentences	27,634	398	802
No. of words	352,113	3,980	8,833
No. of syllables	568,490	5,945	13,729
No. of phones	1,398,994	14,169	33,636
Avg. no. of words / sentence	12.7	10.0	11.0
Avg. no. of syllables / word	1.6	1.5	1.6
Avg. no. of phones / word	4.0	3.6	3.8

Table 4: Some attributes of the PB and PD set compared to the ORCHID

Attribute	Thai PB	MIT-450	ATR-503
No. sentences	398	450	503
PB unit	biphone	biphone	biphone, triphone
No. unique PB units	1,628	1,073	548

Table 5: Comparison among Thai PB set, American English MIT-450, and Japanese ATR-503

Table 5 shows that Thai PB set, MIT-450, and ATR-503 have some differences. ATR-503 sentence set considers the phone-class based triphones in addition to ordinary biphones. The phone-class triphones considered are in forms of /CVC/ and /VCV/ when one or more phones are the specific phone classes such as voiced and semivowel. Monophones defined in MIT-450 is exactly single phoneme, whereas consonant clusters and diphthongs in Thai PB set is defined as monophones. Furthermore, MIT-450 is designed for deeply research of acoustic-phonetic, hence more variations of phonetics for a unique phone are assigned.

A more essential attribute is the coverage of phone units to that exactly occur in the language. Our PD and PB set covers 72 of, overall, 74 monophones (97.3%). However, it can be noted that the monophones disappeared in the corpus are loan phones, which occur only in loan words from foreign language. According to the monophones defined in Table 3, there are 2,568 biphone combinations, in which only 1,792 biphones exist in Thai words. Our sentence set covers 1,628 biphones (90.9%), comparing to the ATR-503 which covers 548 of 625 context-dependent units (87.7%).

[1] A number behind each syllable denotes its tone mark (0-4).

Histograms of some common monophones and biphones, occur in the PD set, are shown in Figure 4 and Figure 5 respectively. A linguistic literature (Luksaneeyanawin, 1992) has presented a distribution of Thai vowel phones in a basis of frequency in different phone context. Although the histogram of vowels in our PD set, as shown in Figure 6, illustrates the cumulative count of phone occurrences regardless of its context, they are comparable because our selection approach tries to collect almost minimum number of sentences that cover the desired phonetic units. Since they correlate, the standard of the PD set can be assured.

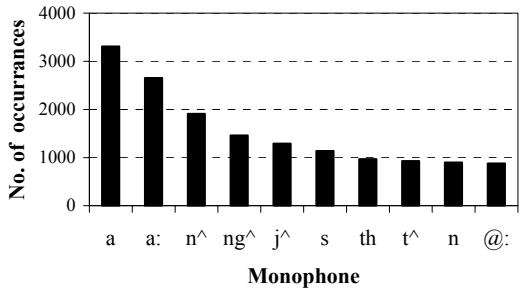

Figure 4: Histogram of some common monophones in PD set

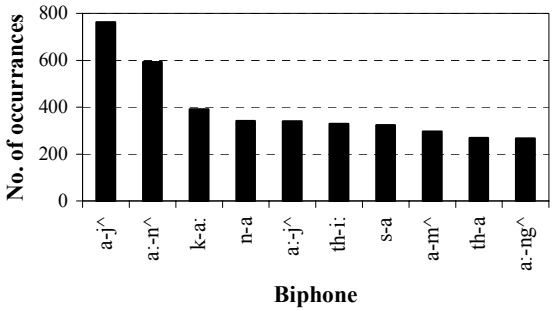

Figure 5: Histogram of some common biphones in PD set

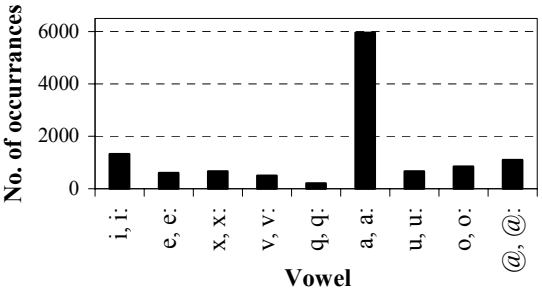

Figure 6: Histogram of single vowels in PD set

Lexical tone in Thai language is the most prominent and interesting characteristic. Since the G2P module also provides 5 tone markers (0-4) in the phonetic transcription, it would enable research on Thai continuous tone processing. In addition, the PD set covers 100% bitones (tone marks of syllable pair) and up to 91.2% tritones when computed on 10 tonal symbols (5 Thai tones, distinguishing between short and long vowels).

5. Summary

This paper describes a work of the first PB and PD sentence set in Thai, which is an important progress of Thai speech research. The PD set will be a part of text prompt used for recording in the first LVCSR corpus for Thai language. The PB and PD sets are analyzed in terms of several significant attributes e.g. number of context-dependent phones, coverage of the units compared to all units possibly occur in Thai language, etc. The partial comparison of the sets to the standard MIT-450 and ATR-503 PB sentence set is also presented. Although the coverage of biphones in our selection approach depends on how good the original text is, 90.9% biphone coverage with the distribution of some phones, which corresponds to the linguistic literature on Thai phones, can certain the standard of our corpus.

6. References

C. Wutiwiwatchai, 1999. NECTEC's Thai LVCSR corpus proposal. (in Thai)

V. Sornlertlamvanich, and R. Thongprasirt, 2001. Speech technology and corpus development in Thailand. *Proceedings of the Oriental COCOSDA Workshop*, pp. 44–47.

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

L. F. Lamel, R. H. Kassel, and S. Seneff, 1986. Speech database development: Design and analysis of the acoustic-phonetic corpus. *Proceedings of the DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, Palo Alto.

K. Iso, T. Watanabe, and H. Kuwabara, 1988. Design of Japanese sentence list for a speech database. *Proceedings of Spring Meeting – Acoustical Society of Japan*, pp. 89–90. (in Japanese)

P. Tarsaku, V. Sornlertlamvanich, and R. Thongprasirt, 2001. Thai Grapheme-to-Phoneme using probabilistic GLR parser. *Proceedings of European Conference on Speech Communication and Technology*, vol. 2, pp. 1057–1060.

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.

S. Luksaneeyanawin, 1992. Three-dimensional phonology: A historical implication. *Proceedings of the Third International Symposium on Language and Linguistics: Pan Asiatic Linguistics*, vol. 1, pp. 75–90.