Thai Tagged Speech Corpus for Speech Synthesis

Chatchawarn Hansakunbuntheung, Virongrong Tesprasit and Virach Sornlertlamvanich
Information R&D Division, National Electronics and Computer Technology Center
112 Thailand Science Park, Phahon Yothin Road,
Klong 1, Klong Luang, Pathumthani 12120 Thailand
Email: {chatchawarnh,virong,virach}@nectec.or.th

Abstract

This paper presents a construction of a Thai speech corpus that aims to provide (1) a chunk of speech unit candidates for developing a unit selection speech synthesis system and (2) the linguistic tags and acoustic information for further development on Thai reading-style prosodic model of the system. In this case, the linguistic tags are composed of phoneme, tone marks, linguistic boundaries, POS, syllable position, voiced/unvoiced region, and tone/toneless region. And the acoustic information consists of energy values, pitch mark, F0, and segmental duration. This speech corpus contains 5,200 sentential utterances, which are selected to cover all Thai and foreign lent phones. Furthermore, the speech corpus extend its coverage to diphone, tri-phone, and, additionally, tri-tone combination. This coverage is used to obtain phone and tone variation in reading speech. The objective and subjective assessment are also operated to evaluate the coverage of speech units and linguistic patterns.

1 Introduction

From the past to the present, a large number of succeed in speech research are based on using speech corpus. Since the speech corpus can obtain variation of real phenomena of speech utterances, we are able to analyze the phenomena broadly and, also, establish general models for those phenomena. Likewise, many research on speech synthesis, which adept speech corpus, reveal great improvement on the quality and the naturalness of synthesized speech. The ways to adopt the speech corpus in speech synthesis field have been presented in many research.

To improve quality of the synthetic speech, a number of speech synthesis systems use the speech corpus themselves as a chunk of speech units for synthesizing, and select the suitable units to construct a target speech. The constructed speech could express the smoothness of spectral transition from one speech unit to the adjacent one since the spectral transition between the speech units can be gain from the speech corpus by using unit selection process. However, the quality of the synthetic speech depends on the quality of the speech units themselves, the coverage of variation of speech units and the number of similar speech units, which are used for the selection process. Thus the speech quality may be decreased if the speech corpus are improperly designed and recorded.

In addition, the speech corpus are also used as collection of prosodic variation for constructing prosodic models. The characteristic of prosodic variation are extracted and encoded as parameters in training process to establish a set of rules or a model for predicting the prosody characteristic. The proficient of a model is based on the variation of speech domain and the coverage of prosodic variation. Since the established model is based on the speech samples in the speech corpus, the corpus design can affect the derived model.

In Thai, there are a few of speech corpus for speech synthesis. Most of them were constructed as a part of speech synthesizing system, which generated speech from a set of static speech units.

This paper presents the construction of Thai speech corpus that aims to provide (1) a chuck of speech unit candidates for developing a unit selection speech synthesis system, and (2) the linguistic tags and acoustic information for further development on Thai reading-style prosodic model of the system. In this paper, we will describe details of corpus design including the corpus structure, text selection, speech recording, text tagging, linguistic labelling, acoustic information extraction, and corpus evaluation.

2 Corpus design

In the first step of corpus design, it is important to consider what the corpus would be used for future work on Thai speech synthesis. A text-to-speech system requires various kinds of information such as text variation, linguistic information, speech characteristics. Since Thai language is a tonal language, tonal information is also considered. These set of information can be retrieved from text and speech signal. The corpus is then considered to be composed of two subparts which are text part and speech part. Both parts are linked together by reference index.

2.1 Corpus Structure

The text part of the corpus was constructed in a hierarchical organization and written in a well-known XML standard format. In this part, text corpus were grammatically structured and tagged with additional linguistic information consisting of source profiles, paragraph tagging, sentence tagging, word tagging, part of speech, toneme and phoneme transcription (as shown in example in figure 1.) Each structured level consisting of corpora, text source, paragraph, sentence and word, was indexed for reference and linked with speech part. The data in this part are used for processing information such as text variation, sentence boundaries, word boundaries, and word pronunciation.

```
        <TAUTHOR>ผศ.ดร. เพียรศิริ วงศ์วิภานนท์และ อ.ดร. กิ่งกาญจน์ เทพกาญจนา </TAUTHOR>
        <EAUTHOR/>
        <TSOURCE>การประชุมทางวิชาการ ครั้งที่ 1, โครงการวิจัยและพัฒนาอิเล็กทรอนิกส์และคอมพิวเตอร์, ปีงบประมาณ 2531, เล่ม 1 </TSOURCE>
        <ESOURCE>The 1st Annual Conference, Electronics and Computer Research and Development Project, Fiscal Year 1988, Book 1 </ESOURCE>
        <TPUBLISHER>ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ, กระทรวงวิทยาศาสตร์ เทคโนโลยีและการพลังงาน </TPUBLISHER>
        <EPUBLISHER>National Electronics and Computer Technology Center, Ministry of Science, Technology and Energy </EPUBLISHER>
        <PAGE/>
        <YEAR>1989 </YEAR>
        <FILE/>
</PROFILE>
<CONTENTS>
<PRG ID=" 1">
        <SEN ID=" 1">
                <SENTEXT>โครงการวิเคราะห์ภาษาไทยเพื่องานแปลด้วยเครื่องคอมพิวเตอร์</SENTEXT>
                <WRD ID=" 1">
                        <WRDTEXT>โครงการ</WRDTEXT>
                        <POS>NCMN</POS>
                        <PHONETIC>khr-oo-ng^-0|k-aa-n^-0|</PHONETIC>
                </WRD>
                <WRD ID=" 2">
                        <WRDTEXT>วิเคราะห์</WRDTEXT>
                        <POS>VACT</POS>
                        <PHONETIC>w-i-z^-3|khr-@-z^-3|</PHONETIC>
                </WRD>
                <WRD ID=" 3">
                        <WRDTEXT>ภาษาไทย</WRDTEXT>
                        <POS>NPRP</POS>
                        <PHONETIC>ph-aa-z^-0|s-aa-z^-4|th-a-j^-0|</PHONETIC>
                </WRD>
                <WRD ID=" 4">
                        <WRDTEXT>เพื่อ</WRDTEXT>
                        <POS>RPRE</POS>
                        <PHONETIC>ph-vva-z^-2|</PHONETIC>
                </WRD>
```

Figure 1 An example of some part of the text-level corpus

In the speech part of the corpus, speech utterances were recorded sentence by sentence and linked with text by the indices. Each sentence has one phrase or more, and phrase boundaries were manually marked for study phrasing phenomena. Other tagged data include phone boundaries, syllable boundaries, syllable position in word and phrase, voice/unvoiced regions, tonal region, energy, and pitch mark.

2.2 Text selection

The whole sentence text was collected from Thai part-of-speech tagged corpus, named ORCHID [1]. The contents of ORCHID are based on Thai Junior Encyclopedias by Royal Command of His Majesty the King and the technical papers that appeared in the past six years of the proceedings of the National Electronics and Computer Technology Center annual conferences. The text corpus was marked in

three levels: paragraph, sentence, and word. Each word was given its part-of-speech tag. The whole text corpus has in total 568,316 words from 43,340 sentences.

In the step of sentence selection, firstly, undesired text were filtered out of the text corpus by the following constraints:

(a) each sentence must have the number of syllables greater than a selected threshold, and

(b) each sentence must contain Thai characters only.

Next, the filtered text corpus was parsed through the NECTEC's automatic grapheme-to-phoneme converter [2], which uses Probabilistic Generalized Left-to-Right Parser.

Afterward, we defined a set of tri-phones and tri-tones combinations as speech units for considering the unit coverage of the corpus. In Thai, the syllabic structure is (Ci)V(Cf), where Ci is initial consonant or initial consonant cluster, V is vowel or vowel cluster and Cf is final consonant of final consonant cluster. Since Thai is a tonal language, each syllable has a tone. There are 5 tones in Thai including mid, low, falling, high and falling tone. Considering Thai syllable structure and tones, a few sets of combinations are defined as follows.

(a) Sets of tri-phones combination
- Preceding phone + Current phone + Succeeding phone
  - $Cf_{-1} + Ci_0 + V_0$
  - $Ci_0 + V_0 + Cf_0$
  - $V_0 + Cf_0 + Ci_{+1}$

(b) Set of tri-vowels combination
- Preceding vowels + Current vowels + Succeeding vowels ($V_{-1} + V_0 + V_{+1}$)

(c) Set of tri-tones combination
- Preceding vowels + Current vowels + Succeeding vowels ($T_{-1} + T_0 + T_{+1}$)

Then, occurrence probabilities of the defined combination patterns were calculated from each sentence. Using equation (1) and (2), each sentence was given a score by their occurrence probabilities of the patterns and number of distinct patterns in the sentence.

$$Score_i = \sum_{j=1}^{N} f_{T_i} P_{Corpus}\left(T_j\right) \tag{1}$$

$$P_{Corpus}\left(T_j\right) = \frac{f_{Corpus,Tj}}{\sum_{k=1}^{N} f_{Corpus,T_k}} \tag{2}$$

Where f_{T_j} represents occurrence frequency of pattern T_j in the i-th sentence, $P_{Corpus}\left(T_j\right)$ represents occurrence probability of pattern T_j in the text corpus and $f_{Corpus,Tj}$ represents occurrence frequency of pattern T_j in the text corpus.

Using greedy algorithm, a minimum set of highest-score and pattern-fully-spanned sentences was selected. Finally, we got a phonetically balanced text corpus, which contains 439,401 diphones from 5,200 sentences.

2.3 Speaker selection

Since this corpus is used as a speech prototype for synthesis, then, some criteria were defined to select a proper speaker. The selection criteria for a speaker were such that the speaker must have clear articulation and standard Thai accent. Finally, a professional female speaker to read our sentences was chosen.

2.4 Recording conditions and tools

The recording environment is shown in figure 2. Each sentence was read and recorded under the following conditions:

- Recorded in silent room
- Recorded into a digital audio tape recorder (DAT Recorder) as 16-bits/sample raw data at sampling rate 44.1 kHz
- Monitoring waveform while recording
- Controlled by linguist and sound engineer
- At SNR > 36 dB

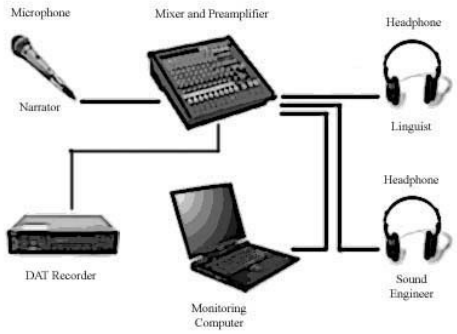

Figure 2 Recording environment

The speaker read each sentence once and was stopped for correcting if misreading occurred. In case of loan words like English words, the speaker read them as English-Thai accent.

2.5 Transcriptions

The set of Thai transcription [3] was designed to cover Thai phonemes and some foreign phonemes that occur in the initial and final position of syllable as shown in table 1 and table 2 respectively

Manner of Articulation		Foreign Phonemes
Bilabial +	Trill	br
	Lateral	bl
Alveolar +	Trill	dr
Fricative +	Trill	fr
	Lateral	fl

Table 1 Foreign phonemes occur in the initial position of Thai syllables

Manner of Articulation		Foreign Phonemes
Stop +	Fricative	ks, ts, ps
Fricative +	Stop	st, sk
	Fricative	fs, th, s
Nasal +	Fricative	ms, ns, ngs
	Affricate	nch
Lateral +	Fricative	ls, lf
	Affricate	lch
Glide +	Fricative	js, ws, jf, wf

Table 2 Foreign phonemes occur in the final position of Thai syllables

2.6 Linguistic labelling and acoustic information extraction

In the procedure of building speech corpus, speech segmentation and labelling are the most time-consumed comparing to the other procedures. Therefore, the automatic segmentation and labelling were developed to carry out some of these procedures.

To align phonetic transcription with speech utterances, our automatic speech segmentation and transcription alignment tool was applied in the first pass. The tool is based on our modified hidden Markov model using HTK [4]. The model uses the general hidden Markov phone model applied with our rule-based pronunciation variation technique [5]. The applied technique constructs alternative phone paths for network of hidden Markov models. Thus, the network can support variation in human's variation and the correctness of segmentation tool is improved.

Next, in the second pass, the phoneme transcriptions were manually checked and finely adjusted again by linguists. In addition, prosodic phrase boundaries were added at this step. In the stage of manual segmentation, the consistency across linguists is very important so that they were trained before starting labelling.

To mark syllable boundaries and syllable position, we developed a set of automatic tools to accomplish these tasks. With using phoneme alignment results, syllable boundaries were also marked automatically. In addition, syllable position in word and phrase were tagged automatically using prosodic phrase boundary information and word-tagged phonetic transcription.

In prosodic level, we selected Praat speech tools [6] to accomplish this work. The tool is developed by David J.M. Weenink and Paul P.G. Boersma at Institute of Phonetic Sciences (IFA). The tools is a system for doing phonetics by computer such as speech analysis, speech synthesis, and speech manipulation. In this case, we use the tool for marking voice/unvoiced region, and extracting pitch, F0 and energy values automatically. Since Tones in Thai is another significant role in Thai prosodic aspect, then tonal regions were marked additionally using our tonal region labelling tool. The labelling tool uses the voiced/unvoiced marking results from Praat Speech Tool and the marked syllable boundaries to analyze the tone/toneless region. The tonal region should be marked separately due to the fact that Thai tones always appear in voiced region, but not all voiced regions are presented as Thai tones. Additional prosodic symbols are also included to represent prosodic information as shown in table 3. Finally, energy curve were periodically calculated at 10 ms and stored in the speech corpus. The examples of speech labelling are shown in Figure 3 and 4.

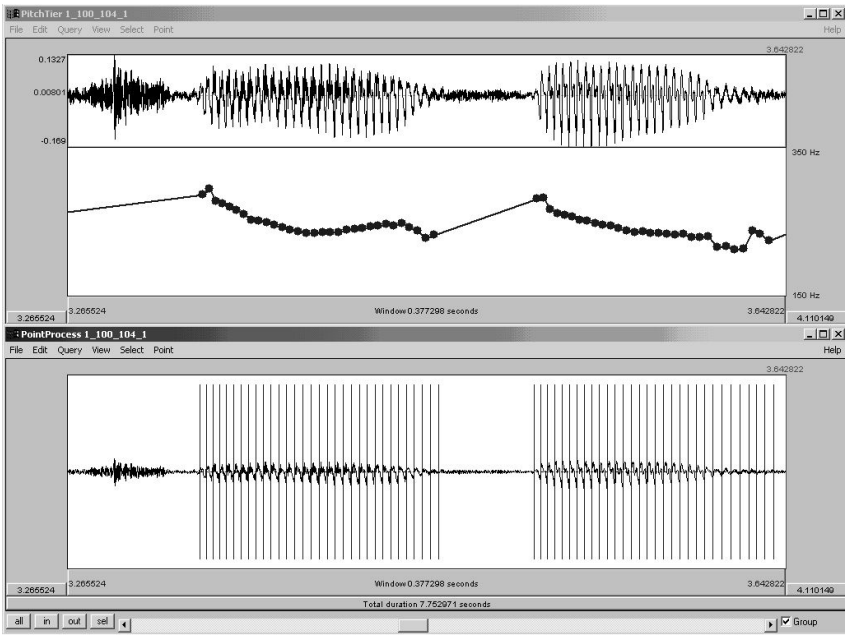

Figure 3 Example of acoustic information extraction

Prosodic information	Symbols
Tonal/Non-tonal region labels	T, NT
Voiced/Unvoiced region labels	V, UV

Table 3 Prosodic labels

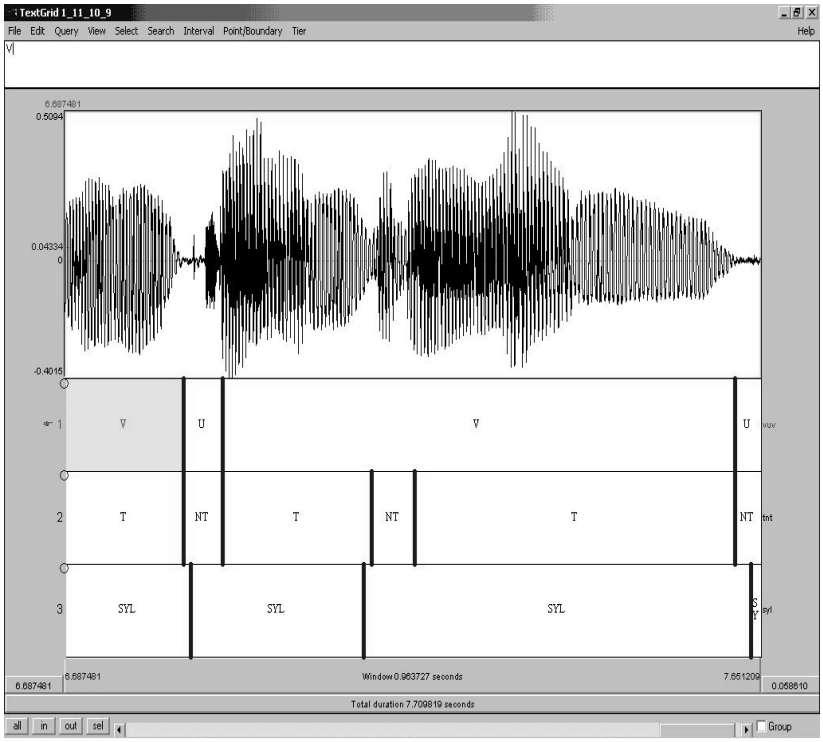

Figure 4 Examples of speech labels

3 Corpus evaluation

When the speech corpus has been constructed, evaluations were needed to determine how effective the speech corpus is. In this paper, objective and subjective assessments were performed as described below.

3.1 Objective assessments

In this measurement, the general statistical information and the coverage information were computed as shown in the table 4 and 5.

Number of sentences	5,200 sentences
Total duration	13.94 hours
Minimum sentence duration	1.53 seconds
Maximum sentence duration	51.44 seconds
Average sentence duration	9.7 seconds
Standard deviation of sentence duration	4.4 seconds

Table 4 General statistical information of the proposed corpus

Pattern type	Number of existing patterns in the corpus	% Coverage of the whole patterns
Phones (Thai only)	65	100
Phones (Thai and foreign)	89	100
Diphones (Thai only)	1,597	77.1
Diphones (Thai and foreign)	1,762	50.8
Tri-phones	10,781	39.6
Syllable (Ignore tones)	2,365	29.3
Tri-tones	180	87.8
Tri-vowels	7,934	50.9

Table 5 Coverage information of the proposed corpus

Considering the lowest level or phone level, the speech corpus spans over all the phone set. However, some unit types present small amount of coverage as occurred in tri-phones and syllable type. In case of syllable type, Luksaneeyanawin's works [7][8] recovered that total grammatical tonal syllables are 30,096 syllable patterns in Thai but only 26,928 tonal syllable patterns are admissible. In addition, only 5,912 tonal syllables (22% of the total admissible syllables) are used in the lexicons of Thai speakers. The study result showed that about 10% of the total tonal syllable are not acceptable and only about one fifth of the whole permitted syllables are used. Therefore, more study on unit distribution of the other cases to assess and improve the corpus is required.

3.2 Subjective assessment

Additional evaluation was presented, to compensate for the lack of information about unit distribution and to assess coverage of the proposed speech corpus in a real application. Therefore, unit selection speech synthesis system was applied here. To evaluate the quality of speech corpus, firstly, appropriate diphone and tone sequences from the corpus were selected to generate synthetic speech by the system. Then a simple concatenation technique without signal smoothing was used. Also, the missing diphones were listed by the system in case of some uncovered diphones were found.

Furthermore, the real-world text corpus, which collected from on-line newspapers in the Internet, was used for testing. The text categories cover various topics such as politics, business, economics, technologies, agriculture, arts, and sports. The testing text has statistic as shown in table 6 and the experimental results are shown in table 7.

Pattern type	Number of existing patterns in the test corpus
Words	57,404
Unique words	5,089
Diphones	172,551
Unique diphones	1,533

Table 6 General information of the testing text

Dipones	Number of missed diphones
Including tones	334
Excluding tones	261

Table 7 Subjective experimental result

The results show that the proposed corpus covered 1,272 of 1,533 tone-independent diphones or 82.8% of the diphones in real-usage text from on-line newspapers. In addition, the quality of synthesized speech is acceptable.

4 Conclusion

In this paper, we have presented the construction of Thai speech corpus for speech synthesis. This speech corpus consists of 5,200 sentential utterances that are tagged linguistic information and extracted acoustic information. These utterances cover all Thai phones and tones. However, it needs more speech utterances to cover at least all diphones, which obtain spectral transition between speech units. Considering the extracted acoustic information, we need further research to encode them for establishing the prosodic model effectively. In the stage of evaluation of speech unit coverage, the unit selection speech synthesis was applied. The synthesized speeches without prosodic adjustment are acceptable.

References

[1] Sornlertlamvanich, V., Charoenporn, T., and Isahara, H., 1997, *ORCHID: Thai Part-of-Speech Tagged Corpus,* National Electronics and Computer Technology Center Technical Report, pp. 5-19.

[2] Tarsaku, P., Sornlertlamvanich, V., and Thongpresirt, R., 2001, Thai Grapheme-to-Phoneme Using Probabilistic GLR Parser, *Eurospeech 2001*, vol. 2, pp. 1057-1060.

[3] Mittrapiyanuruk, P., Hansakunbuntheung, C., Tesprasit, V., and Sornlertlamvanich, V., 2000, Issues in Thai Text-to-Speech: The NECTEC Approach, *Proceedings of NECTEC 2000 Conference: ECTI Technologies with New Economy*, pp. 483-495.

[4] Young, S., Jansen, J., Odell, J., Ollasen, D., and Woodland, P., 1995, *The HTK Book (Version 3.0)*, Entropic Cambridge Research Laboratory, Cambridge, England.

[5] Kanokphara, S., Tesprasit, V., and Thongprasirt, R., 2003, Pronunciation Variation Speech Recognition without Dictionary Modification on Sparse Database, *International Conference of Acoustic and Speech Signal Processing 2003*, To be published.

[6] Luksaneeyanawin, S., 1992, Three-dimensional phonology: a historical implication, *Proceedings of the 3rd International Symposium on Language and Linguistics: Pan-Asiatic Linguistics*, Chulalongkorn University, pp. 75-90.

[7] Luksaneeyanawin, S., 1992, Hierarchy of Sound Distribution, *Proceedings of Conference to Commemorate the 30th Anniversary of the Faculty of Liberal Arts*, Thammasart University.

[8] Boersma, P., and Weenink, D., http://www.fon.hum.uva.nl/praat., Institute of Phonetic Sciences, University of Amsterdam, Amsterdam, Netherlands.