TSynC-2: Thai Speech Synthesis Corpus Version 2

TSynC-2: คลังข้อมูลสำหรับการสังเคราะห์เสียงพูดภาษาไทยเวอร์ชั่น 2

Chai Wutiwiwatchai, Patcharika Chootrakool, Sittipong Saychum,
Nattanun Thatphithakkul, Anocha Rugchatjaroen, Ausdang Thangthai

Abstract - TSynC-2, the prototype of Thai speech synthesis corpus version 2, was designed and built on top of our 3-year experience in using TSynC-1 in the "Vaja" Thai speech synthesizer. The second version aims to resolve several problems occurred in the former one. Prompted sentences were selected from an extremely larger text corpus with a new selection criterion incorporating linguistic-phonetic knowledge. This could then drastically reduce the size of sentence set whilst enlarge the coverage of sound units. The process of adding sentences to cover rare or missing units was renewed to be more efficient. This second version was recorded from 2 carefully and scientifically selected speakers (male/femail). This article summarizes the overall concept and construction method of TSynC-2 with reports on data analysis, suggestion, and future plan.

Keyword - Speech corpus, Speech synthesis

บทคัดย่อ - ต้นแบบคลังข้อมูลเสียงสำหรับการสังเคราะห์เสียงพูดภาษาไทยเวอร์ชั่น 2 ได้ถูกออกแบบและสร้างขึ้นต่อจากคลังข้อมูลเสียงเวอร์ชั่น 1 ซึ่งใช้ในระบบสังเคราะห์เสียงพูดภาษาไทย "วาจา" มานานกว่า 3 ปี คลังข้อมูลเวอร์ชั่น 2 นี้แก้ปัญหาหลักที่เกิดขึ้นในเวอร์ชั่น 1 ได้แก่ ประโยคถูกคัดเลือกใหม่จากคลังข้อความขนาดใหญ่มากเมื่อเทียบกับเวอร์ชั่นแรกซึ่งจะช่วยลดขนาดของคลังและเพิ่มความหลากหลายของหน่วยเสียง ประโยคถูกคัดเลือกด้วยอัลกอริธึมใหม่ซึ่งใช้ฐานความรู้ทางสัทศาสตร์มาช่วยในการกำหนดหน่วยเสียงที่จำเป็นในคลังและตัดหน่วยเสียงที่ไม่จำเป็นออกไป ช่วยลดขนาดคลังได้มาก ในขั้นตอนการเตรียมประโยคยังเพิ่มเติมกระบวนการแต่งประโยคให้ครอบคลุมหน่วยเสียงที่เกิดขึ้นน้อยหรือหน่วยเสียงที่ขาดหายไป กระบวนการแต่งประโยคได้ถูกออกแบบอย่างมีประสิทธิภาพมากกว่าวิธีเดิมที่ใช้ในเวอร์ชั่น 1 คลังข้อมูลเสียงบันทึกจากผู้พูดจำนวน 2 คน (หญิง/ชาย) โดยมีกระบวนการคัดเลือกผู้พูดตามหลักวิชาการ บทความนี้สรุปแนวคิดและแนวทางในการสร้างคลังข้อมูลที่กล่าวมา รวมถึงวิเคราะห์ผลการสร้างคลังข้อมูล ข้อเสนอแนะและแผนการในอนาคต

คำสำคัญ - คลังข้อมูลเสียงพูด, การสังเคราะห์เสียงพูด

1. Introduction

Current state-of-the-art text-to-speech synthesis (TTS) is mainly based on unit concatenation. The algorithm simply picks up appropriate speech units in a speech corpus, concatenates them, and smoothens the speech signal in a post-processing step. The speech unit can range from a small size such as phonemes or phones, diphones, demisyllables, to a larger size such as syllables, and even words. The speech corpus contains specific utterances designed to cover all units, only one or more occurrences per unit. The more the number of unit candidates is, the better the quality of synthesized speech is obtained.

"VAJA", a Thai/English text-to-speech synthesizer developed by the National Electronics and Computer Technology Center (NECTEC) [1], employed a corpus-based unit-selection technique originated by [2]. In VAJA, three additional algorithms have been incorporated. The first algorithm is the design of tonal diphone, which explicitly labels tone in all diphones. Diphone selection hence considers the left tone, right tone, left phoneme and right phoneme simultaneously. Scanning over the speech corpus, the second algorithm segments an input phoneme sequence to be synthesized into subsequences based on left-to-right longest matching. Speech-unit candidates are then collected from the speech corpus for each phoneme subsequence and arranged in the search space. This method relies on the fact that no unit-concatenated speech is better than real human speech. The last algorithm solves the problem when a requested speech-unit is not in the speech corpus. In the case of VAJA where the basic unit is diphone, concatenation of half-phones replaces a missing diphone. As similar as other TTS systems, the unit-selection process in VAJA searches over unit candidates that match the desired unit in terms of phoneme context, tone, and phoneme duration. The desired phoneme duration is predicted by a linear regression estimator and TD-PSOLA is applied if the duration of the best selected-unit is still far from expectation [3].

A major problem found in the algorithm mentioned above is that the speech corpus needs to cover all possible tonal diphones. "TSynC-1", the first-version speech corpus used in VAJA [4], was thus as large as over 13 hours of speech. Nevertheless, it could not cover all tonal diphones appearing in Thai. The huge size of speech database requires higher computation and limits the system portability. To solve the problem, an approach to statistically eliminate redundant units was applied to reduce the corpus size [5]. Since in this method, diphones units were directly erased from speech utterances, more frequent signal discontinuity could be observed and the longest matching algorithm became less efficient.

An intensive design of Thai phoneme unit is an efficient way to systematically reduce the size of unit inventory and hence can help lowering the corpus size significantly. Instead of using the broad set of all tonal diphones, three assumptions regarding acoustic-phonetics of Thai are proved and used in a new compact design of diphone inventory. Wutiwiwatchai et al. [6] described the three assumptions with proofs and the way to incorporate the proven idea in the new corpus design. The paper explains an experiment simulating the overall corpus construction process using TSynC-1, which is summarily reviewed in this article. The idea has been applied to the construction process of TSynC-2. The overall procedure and analysis of the prototype TSynC-2 are given in this article.

2. Objective

The aim is to build a new version of Thai speech synthesis corpus, which is more efficient than its former version in terms of compactness and the variety of speakers.

3. Prototype Background and Status

Under a 3-year project on "Speech-to-Speech Translation" in NECTEC, one important research activity is to enhance the ability of "Vaja" Thai TTS. TSynC-1 is a phonetically-balanced speech database used in Vaja version 3 since 2003. Vaja has been gradually improved until version 5 in 2008. Most of improvements have been done by introducing prosody prediction modules [3] and better unit-selection algorithms [6], while the core speech database remains. It is known that major improvement of TTS can be obtained by enhancing the speech database and TSynC-1 has been found by experiences to be weak in many points. We have then developed the second version since 2007. Milestones of developing process are show in Table 1.

Table 1 Milestones of TSynC-2 construction

Dec 2006	Analyzing the problem of Vaja TTS in terms of its embeded speech database
Jan 2007	Studying and proposing ideas and procedure to construct TSynC-2
May 2007	Proving assumptions that will be used in compacting speech-unit inventory
Oct 2007	Build a very large text corpus to be used in sentence selection
Feb 2008	Transcribing pronunciations of words in the text corpus
Apr 2008	Selecting sentences from the text corpus and selecting appropriate speakers
Jun 2008	Handcrafting additional sentences to cover rare and missing units
Aug 2008	Recording speech
Sep 2008	Wrapping up the TSynC-2 corpus

4. Prototype Specification

The current TSynC-2 speech corpus contains

wav/	- All wave files recorded from 2 speakers (1 male/1 female)
wrd/	- Word transcription of each wave file
syl/	- Syllable transcription of each wave file
lab/	- Phoneme transcription of each wave file

Details and statistics of the corpus will be intensively explained in Section 5.

5. Design, Procedure and Results

The overall procedure of research and development on TSynC-2 is shown in Figure 1.

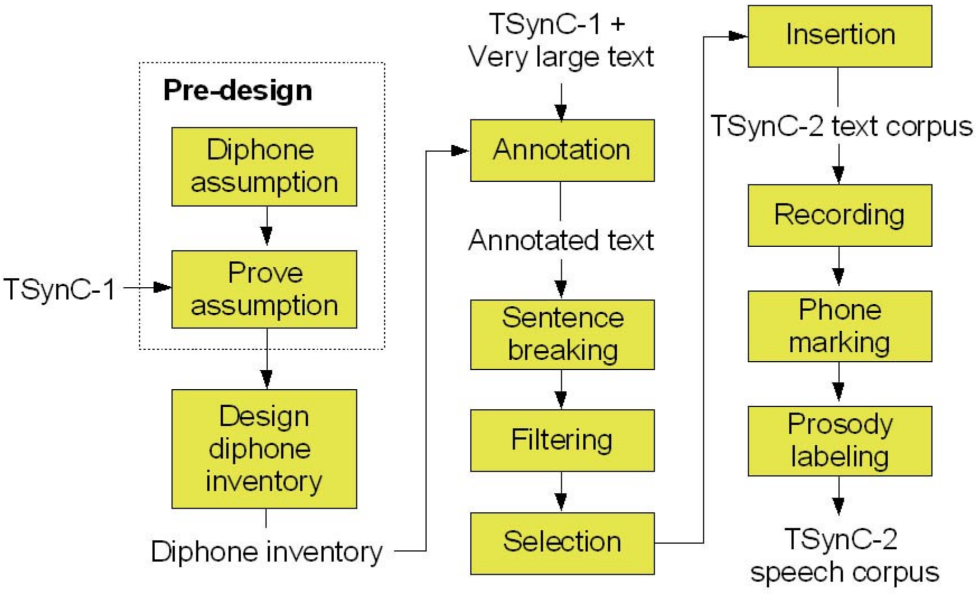


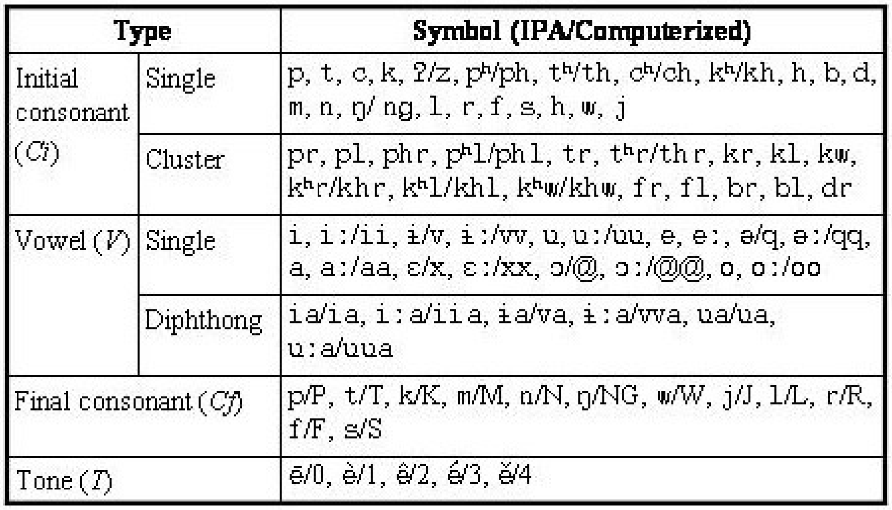
Figure 1 The overall procedure of TSynC-2 R&D.

The procedure can be explained in 4 major steps: designing a unit inventory, building a text corpus, creating TSynC-2 prompted sentences, and speech recording and annotation. Following subsections describe each step in detail.

5.1 Designing a unit inventory

This subsection spans over the first three blocks in the Figure 1. Constrainted by the current algorithm of Vaja, the fundamental unit used in the second designed was also "diphone", which spans from the middle of a phoneme to the middle of its succeding phoneme. The original idea in designing the unit inventory was simply to cover all possible diphones occurred in a language. Phonemes in Thai are presented in Table 2. The number of diphones is hence double of the number of phonemes in Table 2. Moreover, we also took Thai tones into account by tagging each phoneme with the tone level (0-4) of the syllable in which the phoneme located. This appeared to be an explosion of the number of distinguished units in the inventory. An enhanced design of the inventory rather requires basic knowledge as well as empirical observations on acoustic-phonetic characteristics. Based on works by Sproat [7], we explored the following three assumptions potentially expected to help reducing the size of inventory.

Table 2 Thai phonemes



Type		Symbol (IPA/Computerized)
Initial consonant (*Ci*)	Single	p, t, c, k, ʔ/z, pʰ/ph, tʰ/th, cʰ/ch, kʰ/kh, h, b, d, m, n, ŋ/ ng, l, r, f, s, h, w, j
	Cluster	pr, pl, ph r, pʰl/ph l, t r, tʰr/th r, kr, kl, kw, kʰr/kh r, kʰl/khl, kʰw/khw, f r, f l, br, bl, dr
Vowel (*V*)	Single	i, iː/ii, ɨ/v, ɨː/vv, u, uː/uu, e, eː, ə/q, əː/qq, a, aː/aa, ɛ/x, ɛː/xx, ɔ/@, ɔː/@@, o, oː/oo
	Diphthong	ia/ia, iːa/ii a, ɨa/va, ɨːa/vva, ua/ua, uːa/uua
Final consonant (*Cf*)		p/P, t/T, k/K, m/M, n/N, ŋ/NG, w/W, j/J, l/L, r/R, f/F, s/S
Tone (*T*)		ē/0, è/1, ê/2, é/3, ě/4

1) Coarticulation strength of phonemes

Our basic phoneme inventory relied on the Thai syllable structure / *Ci V (Cf) T* /, where *Ci*, *V*, *Cf*, and *T* denote an initial consonant, a vowel, an optional final consonant, and a tonal level, respectively. *Ci* can be a single consonant or a consonant cluster. *V* can be a single vowel or a diphthong. *T* is a symbol indicating the syllabic tone.

In our TTS system, a tonal diphone is a basic unit used in unit-selection. The tonal diphone is a speech chunk spanning from the center of a phoneme to the center of succeeding phoneme, with tonal levels marked for each phonemes. Our original speech corpus was simply designed to cover all possible tonal diphones. However, according to Sproat [7], it is obvious that some pairs of phoneme classes are less or no coarticulated, such as *stop-stop*, *stop-nasal*, *fricative-fricative*, etc. It is noted for our case that the basic unit includes consonant clusters, which comprise a stop and the /r/, /l/, or /w/. In such case, the unit can be in two classes depending on its context position. For example, /N-phr/ is included in the case *nasal-stop*.

2) Tone/Non-tonal phonemes

In the original design described in the previous subsection, every phoneme was marked with a syllabic tone no matter what type of the phoneme is. The design relied on an assumption that a syllabic tone may affect not only the vowel part, but also consonantal parts in a syllable. To our intuition, this is true if the consonantal part is voiced but is questionable otherwise. Therefore, this subsection carries out a subjective proof on the effect of tone over unvoiced consonants.

Unvoiced phone units in our design include *stop* initial-consonants /p/, /ph/, /t/, /th/, /k/, /kh/, /c/, /ch/, /z/, *fricative* initial consonants /f/, /s/, /h/, *stop* final consonants /P/, /T/, /K/, and *fricative* final consonants /F/, /S/. Since consonant clusters, appearing in the initial consonant part, always have a voiced unit /r/, /l/, or /w/ as an ingredient, they are potentially influenced by tones and not included in the unvoiced-unit set. A subjective test was set up for four cases:

- unvoiced *Cf* - voiced *Ci*
- voiced *Cf* - unvoiced *Ci*
- unvoiced *Cf* - unvoiced *Ci*
- unvoiced *Ci*

Couple samples of text were prepared in each case. For example, a word "SONTHI" /s o N 4 - th i 3/ was tested for the case "voiced *Cf* - unvoiced *Ci*", i.e. /N4-th3/. We then varied the tone of unvoiced unit /th/ from /th0/ to /th4/ and synthesize speech of the word for all variations. Finally, we ask subjects to listen and comment on the variations. As expected, all variations were considerably similar.

3) Vowel length

Although Thai explicitly distinguishes short and long vowels, it is possible to define two vowels that are qualitatively similar but quantitatively different as a single unit and convey the TD-PSOLA algorithm to modify the vowel length as needed. Sproat [7] showed that in Japanese, which has phonological vowel length opposition, only little difference in the F1/F2 space between long and short vowels was observed. In a listening test comparing the synthesis result of Japanese long vowels, using long vowel acoustic units as opposed to using short vowel acoustic units, subjects could not distinguish the two sets. He then collected only Japanese short vowels and lengthened their timing when they are phonologically long.

Thai can be considered in the same perspective by plotting the F1/F2 plane of all vowel units taken from the TSynC-1 speech corpus. Figure 2 illustrates the vowel space, where vowel symbols mark median values in Hz and boxes delimit their standard deviations. For sake of visualization and explanation, only some pairs of short/long vowels are shown in the Figure 2.

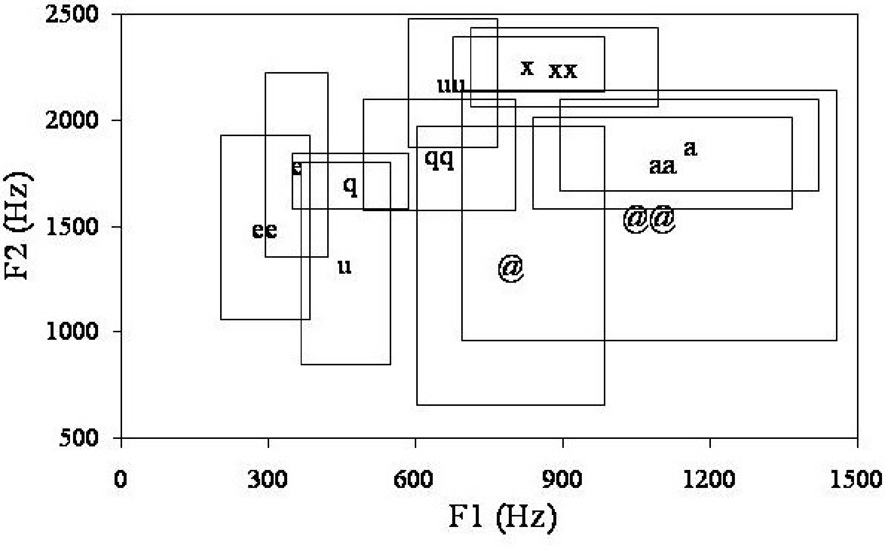


Figure 2 An F1/F2 plane of Thai vowels.

It is obvious that most of vowel pairs presented in the Figure 1 have significant spectral differences. Only few pairs of vowel such as /x/ and /xx/ can be designed using a unique symbol as their F1/F2 distributions are considerably overlapped. Therefore, we decide not to combine any short and long vowel in our unit inventory.

Applying the first assumption resulted 1,909 diphone units compared to 2,012 units in the former design of the TSynC-1. Tagging tonal levels to all diphones extremely increased the number of distinctive units. Employing the proof of the second assumption mentioned in the Section 2.2, only 5,823 out of 8,910 tonal-diphones found in the TSynC-1 were included in our new inventory.

5.2 Building a text corpus

The process of building text corpus, shown as the "Annotation" and "Sentence breaking" blocks in the Figure 1, includes

- Word segmentation
- Sentence breaking
- Phoneme transcription

Text used in building TSynC-2 was mainly from Newspapers. Word segmentation was done semi-automatically by applying SWATH, a NECTEC word segmentation tool, and then checking by humans. In total, there were approximately 13.5 million words. Sentence breaking was done manually. Phoneme transcription was then applied on the list of unique words collected from the word-segmented text and manual correction was also conducted finally. Table 3 summarizes statistics of the resulting text corpus.

Table 3 Statistics of text corpus.

No. of sentences	929,357
No. of words	13,535,098
No. of unique words	106,412

5.3 Creating prompted sentences

As shown in the Figure 1, the next step was to filter out sentences that were too short or too long, so that the final sentence set will be suitable for speakers to read. After filtering, the rest text corpus and its phoneme transcription were used to select a compact set of sentences that covered all diphone units appeared in the text corpus. Selection was performed by the following steps.

1) Calculating a sentence score of each sentence in the text using Equation 1 and 2,

$$Score_{Sent}=\sum(Score_{Diphone})/N_{Diphone} \qquad (1)$$

$$Score_{Diphone}=1/Freq_{Diphone} \qquad (2)$$

where $Score_{Sent}$ = Sentence score

$Score_{Diphone}$ = Diphone score

$N_{Diphone}$ = No. of diphones in the sentence

$Freq_{Diphone}$ = Frequency of the diphone in the whole text corpus

2) Selecting into the final set a sentence having the maximum sentence score
3) Setting zero to the frequency of all diphones already appeared in the selected sentence
4) Repeating steps 1 to 3 until all diphones appeared in the text corpus have been included

in the final sentence set.

To enhance the selected sentence set, three more steps were introduced. The first additional step tried to add more sentences having most frequently used triphones analyzed from the same text corpus. This is to ensure that the created corpus will provide long speech units for frequently used syllables or words. The process was done by sorting tonal triphones (triphones tagged with tones) by their frequencies and accounting whether top-rank triphones were already included in the selected sentences. As a result, there was no need to add more sentences since all top-rank triphones already existed in the sentence set selected previously.

The second additional step was to compare the diphone list accounted from the text corpus with another list of diphones possibly occurred in the Thai language. The latter list was prepared by linguists. Unfortunately, the list of all possible diphones in Thai was not so complete and it will take quite a long time to furnish. Analyzing on the current selected sentence set, the number of unique tonal diphones was significantly increased from that of the TSynC-1. We then decided to skip this additional step.

The final additional step was very important. Since the selection process was greedy, i.e. gradually add one sentence containing the maximum number of missing diphones whilst minimizing the number of exiting diphones, sentencs selected earlier were full of desired diphones while sentences selected later contains sparse number of desired diphones. A huge number of sentences then covered only a few desired diphones. We then audited these sentences by extracting only words containing such required diphones from the sentences and asked linguists to compose new compact set of sentences containing such extracted words. This process could dramatically reduce the number of sentences to be prompted to speakers in the recording step. Table 4 finalizes the TSynC-2 sentence set.

Table 4 Statistics of TSynC-2 text corpus.

No. of sentences	7,043
No. of syllables	105,604
No. of unique tonal diphones	17,977

For sake of recording convenience, we asked linguists to finalize the set by scanning over all sentences and writing down the exact orthographies of each sentence. This process is highly required since some words in the text are very hard to pronounced. Some may be able to pronounced in different ways. We needed to associate the orthography with its transcription given in the sentence selection step.

5.4 Speech recording and annotation

The first process of speech recording was to select appropriate speakers. There is actually no concreate solution how to select the best speaker to give recordings for TTS. According to literature reviews [8], we set up a subjective test to evaluate sounds of sample sentences given by candidated speakers. Five males and ten females read speeech were recorded in a controlled room. Four of them are announcer or have professional experience in anoucement with voice controllation, and eleven of them are resarechers. The speeches are given Mean Openion Scoring (MOS) on naturalness and clearness by 9 speech researchers and selected by maximum MOS.

Recording was performed in a control room using a microphone connected to PC laid outside the room. Speakers were prompted sentence by sentence on a display. Prompted sentences were shown synchronously on another display shown to a linguist controller sitting outside the room. The controller was able to correct any sentence by asking the speakers to repeat that sentence. Table 5 summarizes detailed conditions of recordings.

Table 5 Recording conditions.

Hardware	
Microphone	Plantronics monaural H-41
PC and sound card	Toshiba M600 (PM600 sound on board)
Approximate SNR	32.68 dB
Software	
Recording software	Adobe Audition 1.5
Wave file format	44 kHz, 16 bits, Mono, PCM Wav
Subject	
Speakers	Female: Ms. Thittha Saetachan Male: Mr. Roongchai Huayhongthong
Speaking style	Fluent speaking

After recording, speech utterances were tagged with phoneme boundary markers. In this first stage, an automatic force-alignment algorithm introduced in Hidden Markov Toolkit (HTK) [9] was applied. Manually adjusting the phoneme boundary will be performed in the next stage. Fundamental frequency (F0) was also extracted from the recordings using Praat [10]. The final spe ech corpus contains 9.2 hours of the male speaker and 9.5 hours of the female speakers. Average time used to record per person was within 1 week.

6. Discussion and Conclusion

TSynC-2 overcomes several problems found in TSynC-1. First, the size of speech corpus was reduced to approximately 66% then the minimum physical disc required to store the corpus is 130 MB with the ADPCM compression method. Second, the compact size of TSynC-2 prompted text allows us to repeat recording for more speakers within a short recording period. Third, the coverage of tonal diphones was however larger than that observed in the TSynC-1. This will provide a possibility to generate higher quality speech when combined in a TTS system. Forth, TSynC-2 contained both male and female speakers.

At present, the recording step was done. To gain benefit from TSynC-2, more works are required further, including phoneme boundary labeling, prosody tagging e.g. phrase breaks, stress, intonation, recording for more variety of speakers e.g. children and aged people. The TSynC-2 will be firstly incorporated in the newly developed Vaja 6.0, which is aimed to be more efficient than its previous versions.

References

[1] Mittrapiyanurak, P., Hansakunbuntheung, C., Tesprasit, V., Sornlertlamvanich, V., 2000. *Issues in Thai text-to-speech synthesis: the NECTEC approach*, NECTEC Annual Conference, Bangkok, pp. 483-495.

[2] Hunt, A., Black, A., 1996. *Unit selection in a concatenative speech synthesis system using a large speech database*, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 373–376.

[3] Rugchatjaroen, A., Thangthai, A., Saychum, S., Thatphithakkul, N., Wutiwiwatchai, C., 2007. *Prosody-based naturalness improvement in Thai unit-selection speech synthesis*,

ECTI International Conference, Chiangrai, vol. 2, pp. 1042-1045.

[4] Hansakunbuntheung, C., Tesprasit, V., Sornlertlamvanich, V., 2003. *Thai tagged speech corpus for speech synthesis*, International Conference on Speech Databases and Assessments (Oriental- COCOSDA), pp. 97-104.

[5] Hansakunbuntheung, C., Rugchatjaroen, A., Wutiwiwatchai, C., 2005. *Space reduction of speech corpus based on quality perception for unit selection speech synthesis*, International Symposium on Natural Language Processing (SNLP), pp. 127-132.

[6] Saychum, S., Rugchatjaroen, A., Thatphithakkul, N., Wutiwiwatchai, C., Thangthai, A., 2008. *Automatic duration weighting in Thai unit-selection speech synthesis*, ECTI-CON 2008.

[7] Sproat, R., 1998. *Multilingual text-to-speech synthesis, the Bell Labs approach*. Kluwer Academic Publishers, Norwell, MA.

[8] ผศ.บุญเกื้อ ควรหาเวช, คู่มือผลิตรายการวิทยุกระจายเสียง, 2540, pp. 203.

[9] The HTK book version 3.1, Cambridge University , December 2001, http://htk.eng.cam.ac.uk.

[10] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer", Online: http://www.praat.org, accessed on Dec 2007.