Isarn Dialect Speech Synthesis using HMM with syllable-context features

Pongsathon Janyoi[1] and **Pusadee Seresangtakul**[2]

ABSTRACT

This paper describes the Isarn speech synthesis system, which is a regional dialect spoken in the Northeast of Thailand. In this study, we focus to improve the prosody generation of the system by using the additional context features. In order to develop the system, the speech parameters (Mel-ceptrum and fundamental frequencies of phoneme within different phonetic contexts) were modelled using Hidden Markov Models (HMM). Synthetic speech was generated by converting the input text into context-dependent phonemes. Speech parameters were generated from the trained HMM, according to the context-dependent phonemes, and were then synthesized through a speech vocoder. In this study, systems were trained using three different feature sets: basic contextual features, tonal, and syllable-context features. Objective and subjective tests were conducted to determine the performance of the proposed system. The results indicated that the addition of the syllable-context features significantly improved the naturalness of synthesized speech.

Keywords: Text-to-speech, speech synthesis, HMM, Isarn

1. INTRODUCTION

Text-to-speech (TTS) is the generation of synthesized human speech from unrestricted text. During the past few decade, several speech synthesis techniques have been proposed; such as formant synthesis [1] and di-phone concatenation [2]. However, these basic methods are unable to synthesize human-like speech, in that they employ only a small amount of speech data.

To generate high quality speech, unit selection [3] was proposed. This method generates syntactic speech by selecting and concatenating the corresponding speech units, which are stored as waveforms within the speech corpus. Its use of the original speech waveforms and digital signal processing smoothens the waveform at the concatenation point; which results in high-quality, natural speech. However, this method requires an extremely large speech corpus necessary to generate high-quality speech within an unrestricted text.

To overcome the limitations of the unit-selection method, HMM-based speech synthesis was subsequently proposed [4, 5]. It models speech parameters using a statistical model instead of the original speech waveform.

The performance of this method offers greater naturalness than that of the unit-selection approach, as the speech parameters of the speech units are smoothed and synthesized through a speech vocoder, rather than through the direct concatenation of the speech waveform. However, the use of speech parameters degrades the performance in terms of speech quality. Still, the HMM-based approach remains today's most popular method, due to its clear advantages over unit-selection, such as the smaller size of speech data [5–7] and the flexibility of voice characteristic conversion [8]. HMM-based speech synthesis systems are currently being utilized for several languages; including English [9], Chinese [10], and Thai [11].

Recently, Deep neural network (DNN) was successfully applied to speech synthesis [12, 13]. Its performance is higher than the HMM-based speech synthesis. However, the advantage of the HMM-based speech synthesis over the DNN-based is it has the lower computational cost at the synthesis time [12]. This is the main reason that we decide to apply the HMM-based speech synthesis in this work.

In HMM-based speech synthesis, contextual information is the most important factor controlling both the quality and the naturalness of the output speech, as the parameter sequence of speech units in continuous speech can vary, depending on the phonetic context. In practice, it is impossible to prepare the training data necessary to cover all possible context-dependent units. To solve this problem, a decision-tree-based context clustering [14] was applied to cluster the HMM state and share model parameters. Thus, we can generate the speech parameters of all possible phonetic contexts by tracking the decision-trees according to contextual information. The contextual information and question set used in clustering the HMM state therefore becomes important in achieving the optimal performance.

This paper proposes herein the development of an HMM-based speech synthesis system for the Isarn

Manuscript received on January 8, 2018 ; revised on June 8, 2018.

Final manuscript received on November 2, 2018.

[1,2] The authors are with the Department of Computer Science, Faculty of Science, Khon kaen University, 123 MitraphabRoad, Nai-Meuang, Meuang, Khon kaen, 40002 Thailand, E-mail: pongsathon@kkumail.com and pusadee@kku.ac.th

language, which is classified as a low, limited resource language. In previous work, we proposed a text analysis module for converting the input text to the corresponding linguistic specification [15]. We also proposed the Isarn HMM-based speech synthesis [16]. The system can generate the syntactic speech with the acceptable level. We investigated that both spectral and prosody parameter prediction is the main issue of the baseline system. This work emphasizes to improve the performance of prosody model of the baseline Isarn HMM-based speech synthesis by using the additional context features. We studied the influence of varied feature sets, and their influence upon tonal and syllable-context features. Tonal features were added to improve the modelling of the fundamental frequency contour (F_0) for Isarn, which is tonal language, and the syllable-context features were added to improve the performance of the duration model.

The rest of paper is organized as follows: Section 2 briefly introduces the Isarn Language, Section 3 describes the system architecture, Section 4 describes the implementation of Isarn HMM-based speech synthesis, and the experiments and results are presented in Section 5. Our conclusions and recommendations for future study are presented in the final section.

2. ISARN LANGUAGE

The Isarn language is dialect used in the Northeast of Thailand. In ancient times, Isarn text, using the Isarn Dharma alphabet, was depicted on palm leaves. Today's modernist is incapable of reading or writing the Isarn Dharma alphabet. To overcome this problem, the Thai alphabet, which is used throughout the daily life of Thai people, is used to represent Isarn text. The utterance unit of the Isarn language is a syllable, which is formed by the combination of the initial consonant or vowel, with or without the final consonant phonemes.

The Isarn syllable structure consists of C_i, V, C_f, and T; where C_i, V, C_f, and T represent the initial consonant, vowel, final consonant, and tonal marker, respectively; which are detailed below.

Consonants: There are 44 consonants in the Isarn language. The Initial and final consonants are represented by 20 and nine phonemes, respectively; as illustrated in Table 1. Double consonants do not appear in the Isarn language.

Vowels: There are 28 vowels, which are composed of 18 monophtongs and 6 diphthongs, as illustrated in Table 2.

Tones: Isarn is a tonal language, which means that a similar word, spoken in an incorrect tone will result in a different meaning. There are six different tones produced: mid (M), low (L), mid falling (MF), high falling (HF), high (H) and rising (R). Each tone has an individual pitch pattern contour. The average F_0 contours of the Isarn monosyllabic word, as uttered by a male native speaker, are illustrated in Figure 1.

Examples of the tonal effects upon Isarn pronunciation; for example, "k^{h}aː"; and the various possible linguistic meanings, are as such: M ค่า /k^{h}aː/ ("cost"), L ฆ่า /k^{h}àː/ ("kill"), MF ค้า /k^{h}âː/ ("trade"), HF คา /k^{h}ãː/ ("stick"), H ข่า /k^{h}áː/ ("galangal") and R ขา /k^{h}ăː/ ("leg").

Table 1: *Phonetic symbols of the Isarn consonants.*

Consonant	Phoneme		Consonant	Phoneme	
	Initial	Final		Initial	Final
ป	/p/	/p/	ม	/m/	/m/
ฏ,ต	/t/	/t/	น,ณ	/n/	/n/
จ	/c/	/t/	ง	/ŋ/	/ŋ/
ก	/k/	/k/	ร,ล	/l/	/n/
อ	/ʔ/	-	ฝ,ฟ	/f/	/p/
ผ,พ,ภ	/p^{h}/	/p/	ฉ,ช,ฌ,ซ,ศ,ษ,ส	/s/	/t/
ฐ,ฑ,ฒ,ถ,ท,ธ	/t^{h}/	/t/	ห,ฮ	/h/	-
ข,ฃ,ค,ค,ฆ	/k^{h}/	/k/	ว	/w/	/w/
บ	/b/	/p/	ย	/j/	/j/
ฎ,ด,ฑ	/d/	/t/	ญ	/ɲ/	/j/

Table 2: *Phonetic symbols of the Isarn vowels.*

Type	Short vowel		Long vowel	
	Grapheme	Phoneme	Grapheme	Phoneme
Monophthong	◌ะ	/a/	◌า	/aː/
	◌ิ	/i/	◌ี	/iː/
	◌ึ	/ɯ/	◌ื	/ɯː/
	◌ุ	/u/	◌ู	/uː/
	เ◌ะ	/e/	เ◌	/eː/
	แ◌ะ	/ɛ/	แ◌	/ɛː/
	โ◌ะ	/o/	โ◌	/oː/
	เ◌าะ	/ɔ/	◌อ	/ɔː/
	เ◌อะ	/ə/	เ◌อ	/əː/
Diphthong	เอียะ	/ia/	เ◌ย	/iːa/
	เออะ	/ɯa/	เออ	/ɯːa/
	ัวะ	/ua/	ัว	/uːa/

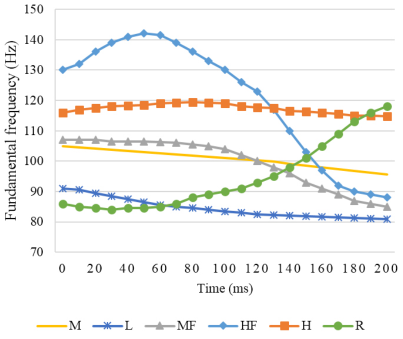

Fig.1: *Average F_0 contours of the six Isarn tones.*

3. ISARN TEXT-TO-SPEECH ARCHITECTURE

The architecture of the Isarn speech synthesis system consists of two modules: text analysis and speech synthesis as shown in Figure 2.

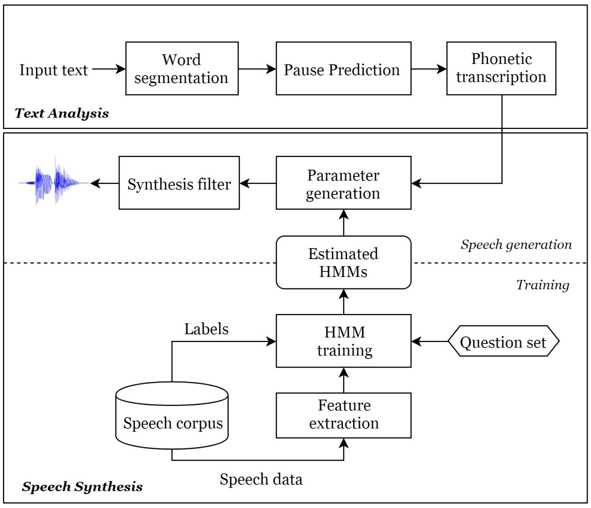

Fig.2: *Architecture of the Isarn text-to-speech system.*

3.1 Text Analysis

Text analysis is the process which converts the input text into linguistic specification, consisting of articulatory and prosodic features. As the front-end of the speech synthesis system, text analysis affects the intelligibility and naturalness of synthesized speech. Text analysis consists of three components: word segmentation, pause prediction, and phonetic transcription. Details of each are described in the following subsections.

3.1.1 Word Segmentation

Isarn is an un-segmented language. Its written style does not contain any symbols or spaces in which to inform the scope of the word boundaries. Several research efforts have found that invalid word boundaries may result in mispronunciation. Three main types of approaches have been proposed for successful word segmentation: rule-based, statistical-based, and hybrid approaches. Efforts in research proposing word segmentation for other languages [13–15], has suggested that the statistic-based approach achieves the best possible performance. However, the training model requires a fairly large text corpus. In this study, we employed the longest known matching algorithm for input segmentation, comprised of a dictionary containing 10,760 Isarn words.

3.1.2 Pause prediction

In speech synthesis, it is important to predict the prosody information necessary to produce natural sounding speech. One of important prosody essential to natural synthetic speech is the pause duration model. Pause duration prediction is the process of inserting silence into a long utterance, which has been found to improve the naturalness of syntactic speech [20]. For Thai, [21] suggests that a pause will necessarily occur after every eighth syllable. We adopted the rule of Thai pause prediction, and added the auxiliary rules to avoid adding pause in improper position.

3.1.3 Phonetic transcription

Phonetic transcription is the process of converting the input word into linguistic specification, consisting of the phoneme and tonal information. Therefore, this process proves the most challenging within the text analysis portion of the system, due to its direct influence on the intelligibility of the synthesis system. Before transcription; the input text, which consists of any number, symbol, or abbreviation of the Thai alphabet, must first convert these symbols into readable words. All unambiguous symbols and abbreviations are converted to a word by looking them up in the dictionary. We used general rules of pronunciation rules to convert any ambiguous symbols, as their pronunciation depends on their contextual factors. For example, the minus symbol (“-”) is pronounced as “**ลบ**” /lop/ (minus) when it occurs between two numbers in a mathematical expression, and as “**ฮอด**” /hɔ̂ːt/ (to) when it represents a period of time. The readable text produced is then converted into the linguistic specification, through a hybrid of the statistical model and rule-based approaches [15].

3.2 Speech synthesis

In the HMM-based Isarn speech synthesis system, herein the syntactic speech is generated according to linguistic specification, obtained from the text analysis module. The synthesis consists of two parts: training and synthesis, as illustrated in Figure 2.

3.2.1 Training Part

The training process aims to model each phone through two processes: feature extraction and HMM training. In the feature extraction stage, the speech parameters, which consist of the mel-cepstral coefficient and F_0, are extracted. The extracted parameters are then modelled by the HMM in the following steps, as shows in Figure 3.

The initial parameters of each monophone model are estimated by k-means clustering. The monophonic models are then re-estimated through an expectation–maximization (EM) algorithm [22]. In this step, the estimated models are used for synthesis; however, they typically produce unnatural speech. In order to improve the accuracy of the model, the context-dependent models are constructed according to the full-context label. The estimated parameters of

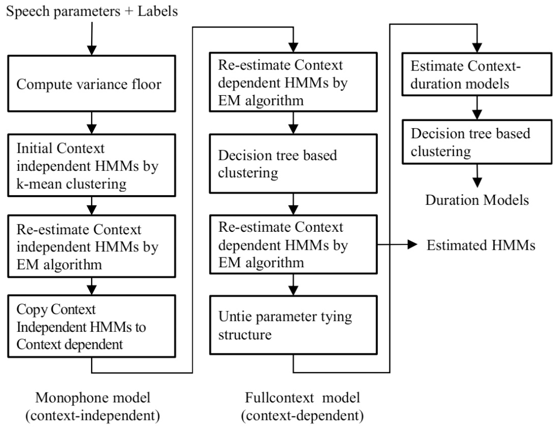
Speech parameters + Labels
Compute variance floor
Initial Context independent HMMs by k-mean clustering
Re-estimate Context independent HMMs by EM algorithm
Copy Context Independent HMMs to Context dependent
Re-estimate Context dependent HMMs by EM algorithm
Decision tree based clustering
Re-estimate Context dependent HMMs by EM algorithm
Untie parameter tying structure
Estimate Context-duration models
Decision tree based clustering
Duration Models
Estimated HMMs
Monophone model (context-independent)
Fullcontext model (context-dependent)

Fig.3: *The model training procedure.*

the monophonic model are used as the initial parameters of the context-dependent model. After which, the context-dependent models are re-estimated by the EM Algorithm.

However, the exponential increase in the context-dependent model has an influence on the amount of training data necessary to estimate each model. In practice, we cannot estimate an accurate model using sparse data. To overcome this problem, decision-tree-based context clustering was applied to share the parameters of similar models, having different context features. Each context-dependent model was then re-estimated using the cluster of the speech parameter. This work employs the conventional decision-tree-based context clustering technique available in the HTS Toolkit [23], using the minimum descriptive length criterion [24]. In the last iteration of re-estimation, the context-duration models were also estimated and clustered similarly to the Mel-cepstral and F_0. The training output produces three decision trees, which are used for tracking the Mel-cepstral, F_0, and duration. They are clustered independently, as each has a different influential context feature.

3.2.2 Speech generation

Speech generation is the process which generates syntactic speech. There are two steps involved: parameter generation and speech synthesis. To generate speech parameters, the linguistic specification, derived from the text analysis module, is converted into the context-dependent phoneme labels that correspond to the context features, which train the HMMs. After which, a sequence of context-dependent HMMs is concatenated. The state duration and speech parameters are then generated by tracking the respective decision tree. Lastly, the generated speech parameters are converted into speech waveforms through a speech vocoder.

4. ISARN TEXT-TO-SPEECH (TTS) CONSTRUCTION

This section describes the details of our system construction, including speech corpus construction, HMM topology and feature extraction, and contextual features.

4.1 Speech corpus construction

The quality of speech corpus is one of important factors in TTS construction, in which the number of phonemes within the speech corpus must be balanced. We therefore avoided the random selection sentence form text corpus by considering the number of samples of each phoneme. There are two steps involved in our corpus construction: phoneme balancing and voice recording. We preformed phoneme balancing in order to obtain clear speech in all phonemes. Firstly, we collected text from many sources, such as news, articles; however, the sentences in such sources proved to be deficient. Thus, we further collected text through the transcriptions of daily Isarn conversations. Input text was then converted into linguistic specification. We selected sentences that focused on phonemes, which lack samples. All sentences matching the phoneme criterion were selected. The selection process produced 4,400 sentences, which were recorded and included in the corpus. The Isarn speech corpus contains roughly six hours and 10 minutes of speech data.

In this study, speech samples were recorded by a native male speaker, in sound proof room, and saved in wave file (.wav) format with a single channel at a sample rate of 48 kHz, and a bit depth of 16 bits per sample. The recorded speech was down-sampled to 32 kHz, which did not deteriorate the quality of speech [25]. The recording process was divided into 20-minute sessions with a five-minute break between each, in order to help relax the speaker. Before the start of each new session, the speaker listened to the five recorded sentences from the previous session, in order to maintain his speaking style. The boundaries of the phonemes in the speech utterance were identified using the force alignment method, based on the mono-phonemic HMM. However, as the automatic method often identifies an incorrect boundary, the phoneme boundaries were manually refined by the linguist.

4.2 HMM topology and feature extraction

In this study, we deployed a seven-state HMM topology, in which “left-to-right no skip” and the first and the last states were not estimated parameters. The training data extracted the speech parameter through the use of a 25ms Hamming window and a 5ms frame shift. The WORLD speech vocoder [26] was used to extract spectral parameters, represented by the 60 order Mel-cepstral coefficients (including

Table 3: *Contextual features.*

Contextual Features	Description
Basic contextual features	
Phone	Before previous / previous / current / next / after next phoneme identities.
	Position of the current phoneme in the current syllable.
	Number of phoneme in the current syllable.
	Position of the current phoneme in the current word (forward/backward).
Syllable	Name of the vowel of the current syllable.
	Number of phonemes in the previous/current/next word.
	Position of the current syllable in the current word.
	Number of syllables in the previous/current/next word.
	Position of the current syllable in the current phrase.
	Position of the current syllable in the current phrase (forward/backward).
	Number of syllables in the current phrase.
	Position of the current syllable in the current word (forward/backward).
	Position of the current syllable (without tone) in the current word (forward/backward).
Word	Position of the current word in the current phrase.
	Number of words in the current phrase.
Phrase	Position of the current phrase in the current utterance.
	Number of phrases in the current utterance.
Tone features	
	Tone of before previous / previous / current / next / after next syllables.
	Tone of before previous / previous / current / next / after next phoneme.
Syllable-context features	
	Type of preivous / current / next syllable (dead/live).
	Tone of current and previous syllable are similar.
	Tone of current and next syllable are similar
	Syllable section in the current word (single, begin, middle, end).
	Syllable section in the current phrase (single, begin. Middle, end).

zeroth coefficients) and their dynamic features (delta and delta-delta) [5]. The excitation parameter was represented by log F_0, which was extracted through the RAPT approach [27].

4.3 Contextual features

Both the training and synthesis processes require contextual features. In the training process, they were used to construct the decision trees for sharing speech parameters. In the synthesis process, contextual features of each phoneme were used to select the appropriate speech parameters from the estimated HMMs. In this work, features were divided into three groups: basic contextual features, tonal features, and syllable-context features. The basic contextual features consist of the information at various levels; such as phoneme, syllable, word, phrase, and sentence. The tonal feature represents the tonal information of each syllable and the adjacent syllables. We also included syllable-context features, in order to improve the accuracy of the duration model. The syllable-context features contain the syllable type and syllable boundary types. Details of these contextual features are shown in Table 3.

Table 4: *Example Questions for Tree-Based Clustering.*

Type	Question	Description
Articulatory feature	QS “LL-Fricative” {f/*, s/*, h/*}	The 2nd left of current phoneme is fricative.
	QS “L-Nasal” {m/*, n/*, ng/*, y/*}	The left of current phoneme is nasal.
	QS “R-StopFinal” {p^/*, t^/*, k^/*}	The right of current phoneme is a final consonant with stop.
	QS “RR-LongDip Vowels” {iia/*, vva/*, uua/*}	The 2nd right of current phoneme is a diphthong vowel with a long sound.
	QS “L-Syllable Tone5” { *:5#* }	The right of current syllable has the rising tone (5 represents the rising tone).
Prosodic feature	QS “C-WrdNum Syls=2” {*-2^*}	The number of syllables in the current word equals two.
	QS “C-WrdSingle Syl” {*^Single+*,*^0+*}	The current word has one syllable.
	QS “R-WordCount Phone >=8<=12” {*;8/D:*, *;9/D:*, *;10/D:*, *;11/D:*, *;12/D:*}	The number of phonemes in the right word between 8 and 12.
	QS “C-PhraseNum Syls=8” {*=8+*}	The number of syllables in the current phrase equals eight.

In addition to the contextual features, a question set was also employed within the tree-based clustering of the training step. In this study, the question set was based on both articulatory and prosodic features. The articulatory feature considers phoneme identity and phoneme type, whereas the prosodic feature considers the number or position of the phoneme, syllable, and word. There are 1,195 and 396 questions within the articulatory features and prosodic features, respectively. Examples of the question sets for both articulatory and prosodic features are shown in Table 4.

5. SYSTEM EVALUATION

To evaluate the performance of the proposed system, we conducted both objective and subjective tests. The objective test was performed first, by comparing the generated speech parameters with the extracted speech parameters. Then, the results of objective tests were verified by the subjective test.

5.1 Objective test

In performing the objective test, we trained three systems, listed in Table 5, in order to investigate the influence of each feature set.

Table 5: *System for objective test.*

System	Detail
HMM-C	Basic contextual feature only
HMM-CT	Basic contextual feature + tone features
HMM-CTP	Basic contextual feature + tone features + syllable-context prosodic features

All systems were trained using 3,900 sentences, which were randomly selected from the speech corpus. The remaining 500 sentences became the test sentences in the objective test. Three measurement matrices: Mel-cepstral distortion (MCD), root means square error (RMSE), and voiced/unvoiced error rate (VER) were used to evaluate the accuracy of the speech parameter modeling.

The MCD was used to calculate the cepstral distance between the original Mel-cepstral and the predicted Mel-cepstral, which are defined as:

$$MCD = \frac{10}{ln10}\sqrt{\frac{2}{T}\sum_{t=0}^{T}\sum_{m=1}^{M}(c_{r,t}(m) - c_{p,t}(m))^2} \quad (1)$$

Where $c_{r,t}(m)$and $c_{p,t}(m)$ denote the m^{th} order Mel-cepstral at time t of the extracted cepstral of natural speech and predicted cepstral, respectively. M is the order of Mel cepstral, and T is the total number of frames.

The performance of F_0 modelling is evaluated in terms of the RMSE and VER. The RMSE were performed by considering only the frame in which the extracted F_0 for natural speech and the predicted F_0 value are voiced. Hence, the standard RMSE were modified as:

$$F_0RMSE = \sqrt{\frac{\sum_{t\in V}(f_r(t) - f_p(t))^2}{|V|}} \quad (2)$$

$$V = \{t : f_r(t) > 0 \wedge f_p(t) > 0, 0 < t < T\}$$

Where V is set of time indices that both the extracted F_0 for natural speech and the predicted F_0 value are voiced. $f_r(t)$, and $f_p(t)$ denote the extracted F_0 and the predicted F_0, respectively. T is the number of total frames.

The VER is calculated through:

$$F_0VER = 100(1 - \frac{|C|}{T}) \quad (3)$$

$$C = \{t : f_r(t) > 0 \wedge f_p(t) > 0 \vee f_r = 0 \wedge f_p(t) = 0\}$$

Where C denotes the set of time indices that both extracted and predicted F_0 values are voiced or unvoiced. $f_r(t)$ and $f_p(t)$ are the extracted F_0 and the predicted F_0, respectively.

Typically, The HMM-based TTS has two stages of speech parameters. The first stage determines the duration of each phoneme. The second stage generates the speech parameters corresponding to the determined duration. Therefore, the accuracy of the speech parameters depends on the duration model. To avoid the adverse effects of the duration model in this objective test, the test sentences were generated using the original phone duration. To evaluate the performance of the duration model, The RMSE was employed to assess the differences and similarities of the predicted and reference phone durations. The MCD, F_0 RMSE, and F_0 VER each feature set are presented in Table 6. Table 7 outlines the Duration RMSE of the system training with different feature sets.

Table 6: *MCD, F_0 RMSE and F_0 VER of the three systems training within the different feature sets.*

System	MCD (dB)	F_0 RMSE (Hz)	F_0 VER (%)
HMM-C	5.68±0.17	11.84±3.30	8.58±3.32
HMM-CT	5.67±0.16	9.07±2.86	8.51±3.22
HMM-CTP	5.67±0.16	8.86±2.79	8.64±3.37

Table 7: *DurationRMSE of the three systems training within the different feature sets.*

System	Duration RMSE (ms)	
	Phoneme	Syllable
HMM-C	26.58 ± 1.08	45.00 ± 2.16
HMM-CT	26.38 ± 1.06	44.11 ± 2.13
HMM-CTP	26.26 ± 1.07	43.68 ± 2.13

Based on the results listed in Table 6, the F_0 RMSE of the HMM-CT and HMM-CTP systems were lower than those of the HMM-C. This indicates that the inclusion of tonal features improves the prediction performance of the F_0 model without the distortion of other models. In the case of duration model, we found that the HMM-CTP achieved the best performance. Both phone and syllable duration RMSEs decreased when syllable-context features were added. This suggests that the addition of syllable-context features improves the prediction performance of the duration model.

Because we noticed that the results of the objective tests of all systems were not much different from one another, we also conducted the subjective test, in order to validate the results of the objective test.

5.2 The subjective test

The overall naturalness of the syntactic speech was represented through mean opinion score: MOS evaluation. The score of each utterance was rated on a scale of one to five, where (1) indicates extremely unnatural speech, and (5) indicates very natural speech.

Natural speech and vocoder speech were adopted as the baseline system for comparison with the syntactic speech of the proposed system. In this experiment, only the HMM-CT and HMM-CTP systems were included, because the F_0 RMSE of the HMM-CT and HMM-CTP systems were clearly lower than that of the HMM-C system. Table 8 summarizes the system comparison in the MOS test. The subjective test consisted of 15 sentences, from several different sources. In total, 60 utterances were derived from four sources. Because we anticipated that the listeners would grow fatigued listening to numerous utterances, and perhaps influence their opinion and expression of them, we divided the listening into two sessions, separated with a five minute break.

Table 8: *System comparisons in the MOS test.*

System	Detail
Natural	Natural speech.
Vocoder	Synthesize form the original speech parameters.
HMM-CT	Basic contextual feature + tonal features.
HMM-CTP	Basic contextual feature + tonal features + syllable-context prosodic features.

Twelve native listeners participated in the listening test. The order of utterances was also randomly arranged. The listeners were able to listen to each utterance at most two times, and were not permitted to return to the previous utterance. The results of the MOS test are given in Figure 4. Table 9 shows the mean and standard deviation of MOS for each system.

Table 9: *The detail statistics obtained by MOS test.*

System	Mean	SD
Natural	4.56	0.42
Vocoder	4.09	0.52
HMM-CT	3.51	0.80
HMM-CTP	3.71	0.68

The results of the MOS test indicated that the MOS of both the HMM-CT and HMM-CTP systems fell within acceptable, natural levels. We did notice that the MOS of the HMM-CTP system was higher than that of the HMM-CT system. This suggests that the syllable-context feature significantly improves speech naturalness. However, while this positive trend was demonstrated through the objective test, the gap of the MOS between natural speech and the HMM-CTP system is still wide.

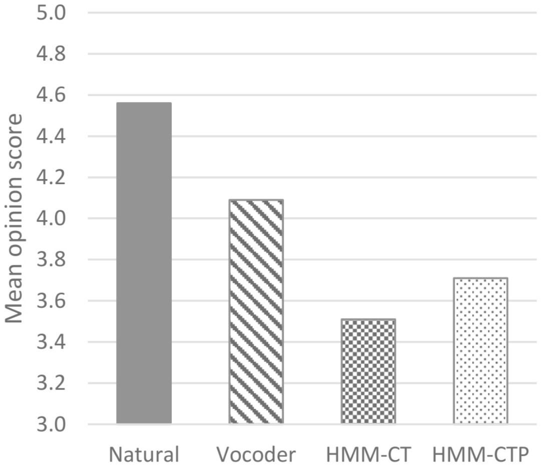

Fig.4: *The results of the MOS test.*

6. CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORKS

This research proposes an Isarn speech synthesis system based on the HMM. The speech data used for training consisted of 3,900 utterances, and the performance of the proposed system was evaluated through both objective and subjective tests. We trained the proposed system with three different feature sets: basic contextual, tonal, and syllable features. Both the objective and subjective tests indicated that the tonal features and syllable-context features are capable of improving the naturalness of the synthesized speech. The results suggest that the system is capable of generating natural, understandable speech; useful in the real word. However, problem areas do exist that degrade both the naturalness and intelligibility of the proposed system. Test results have shown that the MOS of the synthesized speech is still lower than the MOS of the natural and vocoder speech, thus indicating the limitations of the proposed system. We have assessed that the generation of both the spectral and prosodic parameters are the cause of the problem. In the case of prosodic features, it hard to model only phone/state level because it has the complex variation, which depends on the various factors such as coarticulation effect of Isarn tones in running speech. We hope to solve these problems in future works.

ACKNOWLEDGMENT

This work was supported through a Computer and Information Science Interdisciplinary Research Grant from the Department of Computer Science, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand.

References

[1] D. H. Klatt, " Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, vol. 67, no. 3, pp. 971-995, 1980.

[2] C. Hamon, E. Mouline, and F. Charpentier, " A diphone synthesis system based on time-domain prosodic modifications of speech," in *, 1989 International Conference on Acoustics, Speech, and Signal Processing, 1989. ICASSP-89*, 1989, pp. 238-241 vol.1.

[3] A. J. Hunt and A. W. Black, " Unit selection in a concatenative speech synthesis system using a large speech database," in *, 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings*, 1996, vol. 1, pp. 373-376 vol. 1.

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, " Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," *Proc EUROSPEECH*, vol. 5, pp. 2347-2350, 1999.

[5] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, " Speech parameter generation algorithms for HMM-based speech synthesis," *Acoust. Speech Signal Process. IEEE Int. Conf. On* , vol. 3, pp. 1315-1318, Jun. 2000.

[6] A. Gutkin, X. Gonzalvo, S. Breuer, and P. Taylor, " Quantized HMMs for low footprint text-to-speech synthesis," in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2010, pp. 837-840.

[7] B. T th and G. Nmeth, " Optimizing HMM Speech Synthesis for Low-Resource Devices," *J. Adv. Comput. Intell. Intell. Inform.* , vol. 16, no. 2, pp. 327-334, Mar. 2012.

[8] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, " A Style Control Technique for HMM-Based Expressive Speech Synthesis," *IEICE Trans.*, vol. 90-D, pp. 1406-1413, Sep. 2007.

[9] K. Tokuda, H. Zen, and A. W. Black, " An HMM-based speech synthesis system applied to English," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002*, 2002, pp. 227-230.

[10] Y. Qian, F. Soong, Y. Chen, and M. Chu, " An HMM-Based Mandarin Chinese Text-To-Speech System," in *Chinese Spoken Language Processing* , Q. Huo, B. Ma, E.-S. Chng, and H. Li, Eds. Springer Berlin Heidelberg, 2006, pp. 223-232.

[11] S. Chomphan and T. Kobayashi, " Implementation and evaluation of an HMM-based Thai speech synthesis system," in *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, 2007, pp. 2849-2852.

[12] H. Ze, A. Senior, and M. Schuster, " Statistical parametric speech synthesis using deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7962-7966.

[13] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, " On the training aspects of Deep Neural Network (DNN) for parametric TTS synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3829-3833.

[14] J. J. Odell, *The Use of Context in Large Vocabulary Speech Recognition.* 1995.

[15] P. Janyoi and P. Seresangtakul, " Isarn phoneme transcription using statistical model and transcription rule," vol. 59, pp. 337-345, 2014.

[16] P. Janyoi and P. Seresangtakul, " An Isarn dialect HMM-based text-to-speech system," in *2017 2nd International Conference on Information Technology (INCIT)*, 2017, pp. 1-6.

[17] C. Haruechaiyasak, S. Kongyoung, and M. Dailey, " A comparative study on Thai word segmentation approaches," in *2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, 2008, vol. 1, pp. 125-128.

[18] L. Du *et al.*, " Chinese word segmentation based on conditional random fields with character clustering," in *2016 International Conference on Asian Language Processing (IALP)*, 2016, pp. 258-261.

[19] T. P. Nguyen and A. C. Le, " A hybrid approach to Vietnamese word segmentation," in *2016 IEEE RIVF International Conference on Computing Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*, 2016, pp. 114-119.

[20] N. P. Narendra, K. S. Rao, K. Ghosh, R. R. Vempada, and S. Maity, " Development of syllable-based text to speech synthesis system in Bengali," *Int. J. Speech Technol.*, vol. 14, no. 3, p. 167, Sep. 2011.

[21] S. Luksaneeyanawin, " A Thai text-to-speech system," in *Proceedings of the Regional Workshop an Computer Processing of Asian Languages (CPAL)*, Asian Institute of Technology, 1989, pp. 305-315.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, " Maximum Likelihood from Incomplete Data via the EM Algorithm," 1976.

[23] H. Zen *et al.*, " Recent development of the HMM-based speech synthesis system (HTS)," 2009.

[24] K. Shinoda and T. Watanabe, " MDL-based context-dependent subword modeling for speech recognition," *Acoust. Sci. Technol.*, vol. 21, no. 2, pp. 79-86, Jan. 2001.

[25] A. Stan, J. Yamagishi, S. King, and M. Aylett, " The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech syn-

thesis system using a high sampling rate," *Speech Commun.* , vol. 53, no. 3, pp. 442-450, Mar. 2011.

[26] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Trans. Inf. Syst.*, vol. E99.D, no. 7, pp. 1877-1884, Jul. 2016.

[27] W. B. Kleijn and K. K. Pailwal, "A Robust Algorithm for Pitch Tracking (RAPT)," in *Elsevier*, 1995, pp. 495-518.

Pongsathon Janyoi received the B.Sc. degree and M.S. degree in Computer Science from Khon Kaen University, Khon Kaen, Thailand, in 2010 and 2015, respectively. Currently, he is a Ph.D. candidate in Natural Language and Speech Processing Laboratory, Department of Computer Science, Khon Kaen University. His current research interests include speech synthesis, automatic speech recognition and artificial intelligence.

Pusadee Seresangtakul received a B.Sc. degree in Physics from Khon Kaen University, Khon Kaen, Thailand in 1986, and M.S. degree in Computer Science from Chulalongkorn University, Bangkok, Thailand in 1991. She received Ph.D. from Graduated School of Engineering and Science, University of the Ryukyus, Japan in 2005. She is currently an assistant professor in the Department of Computer Science, Faculty of Science, Khon Kaen University. Her research interests include natural language and speech processing, machine learning and artificial intelligence.