Performance Evaluation of Phonotactic and Contextual Onset-Rhyme Models for Speech Recognition of Thai Language

Somchai Jitapunkul[1]*, Ekkarit Maneenoi*[1]*, Visarut Ahkuputra*[1]*, and Sudaporn Luksaneeyanawin*[2]

[1]Digital Signal Processing Research Laboratory, Department of Electrical Engineering,
[2]Centre for Research in Speech and Language Processing, Department of Linguistics,
Chulalongkorn University, Bangkok, Thailand
somchai.j@chula.ac.th, ekkarit.m@student.chula.ac.th

Abstract

This paper proposed two acoustic modelings of the onset-rhyme for speech recognition. The two models are Phonotactic Onset-Rhyme Model (PORM) and Contextual Onset-Rhyme Model (CORM). The models comprise a pair of onset and rhyme units, which makes up a syllable. An onset comprises an initial consonant and its transition towards the following vowel. Together with the onset, the rhyme consists of a steady vowel portion and a final consonant. The experiments have been carried out to find the proper acoustic model, which can accurately model Thai sound and gives higher accuracy. Experimental results show that the onset-rhyme model excels the efficiency of the triphone for both PORM and CORM. The PORM achieves higher syllable accuracy than the CORM 2.74 %. Moreover the onset-rhyme models also give a more efficiency in term of system complexity compared to the triphone models.

1. Introduction

Presently, almost continuous speech recognition systems are based on the phones and their derivatives as acoustic units. The triphone has been the prominent acoustic unit in speech recognition for at least a decade. Due to the powerful acoustic modeling, triphone can model the most important coarticulatory effect from the neighboring phones.

The speech recognition system utilizing triphones as speech unit gives high performances, and it would probably work well with any languages. Although, the triphone seems to be a good speech unit for acoustic modeling, there are many disadvantages in applying this context-dependent phone unit. Since, triphone is a phone-derivative unit, it inherits some limitations of phone-based approaches namely the lack of an easy and efficient way for modeling long-term temporal dependencies [1]. The triphone unit spans an extremely short time interval. Consequently, integration of spectral and temporal dependencies is not easy.

For Thai language, researches on Thai speech recognition have been developed more than a decade. Various techniques were applied in the system [2]. Initially, Thai speech recognition systems were implemented based on word model, which has many limitations and disadvantages. The efficiency of speech recognition system using word model is still unsatisfactory. Presently, there is no Thai speech recognition system dealing with continuous speech. To implement a large vocabulary speech recognition system for continuous speech, subword unit should be exploited as an acoustic unit due to the drawbacks and limitations of word model.

The choice of speech unit is dependent on the language structure and the availability of sufficient training data. Since each language has its own attributes, choosing suitable speech units leads to effective utilization of the training data and a good performance of speech recognizer. With different language structure and special characteristics of Thai language, different recognition architecture has to be employed to perform much better in various aspects including accuracy and computation load.

A notion of onset-rhyme was proposed and applied to Thai continuous speech recognition [3, 4]. This acoustic unit outperforms the intra-syllable triphone in terms of accuracy and complexity. Actually, there are two types of onset-rhyme models, Phonotactic Onset-Rhyme Model (PORM) and Contextual Onset-Rhyme Model (CORM), but only CORM was employed to Thai continuous speech recognition [3, 4]. This paper will explore both two types of onset-rhyme models in order to compare which type will be more efficient and suitable for Thai speech recognition system.

This paper is organized as follows. In the next section, the details of the Thai spoken language are described. The onset-rhyme model is explained in details in section 3. In section 4, the experimental configuration and results are described. Finally, the results are analyzed and concluded in section 5.

2. Thai Spoken Language

This section is intended to provide the essential knowledge of Thai spoken language.

2.1. Thai Syllable Structure

Thai syllables are composed of three sound systems namely consonant, vowel, and tone [5]. The construction can be represented with the structure as illustrated in Figure 1. The combination of these sound units produces 26,928 syllables.

$$S \;=\; \overset{T}{c(c)V(V)(C)}$$

Figure 1: Thai syllable structure.

Where c is initial consonant, C is final consonant, V is vowel, and T is tone respectively.

2.2. Acoustical Properties of Thai Syllables

In the syllable structure, vowel and non-stop final consonant carry tone whereas an initial consonant does not perform this. Furthermore, the durations of final consonant following a short and a long vowel are statistically different. All statistical parameters of duration are shown in Table 1. Comparatively, The duration of final consonant following the short vowel or the weak vowel is longer than that of final consonant following the long vowel or the strong vowel. This relationship occurs only between the vowel and the final consonant. In contrast, the initial consonant is slightly affected by the duration of the vowel. Therefore, the final consonant is strongly influenced by the vowel duration.

3. Onset-Rhyme Acoustic Modeling

This section describes the details of onset-rhyme properties in both acoustical and phonological point of view.

3.1. Acoustical Properties

An onset consists of a releasing consonant and its transition towards the following vowels. Including a transitional portion between consonant and vowel, this onset model provides consonant-vowel combinations of Thai syllables. Along with the onset, a rhyme is composed of a vowel and an arresting consonant. This model contains a steady vowel portion plus the arresting consonant. The rhyme model provides vowel-consonant combinations of Thai syllables [3, 4]. The onset-rhyme segment is shown in Figure 2. Obviously, the onset covers the transition towards the vowel, which makes the onset precisely modeled.

According to the acoustic property of Thai syllable as described in section 2.2, in the syllable structure, the final consonant is strongly influenced by the vowel duration. The duration of final consonant following the short vowel or the weak vowel is longer than that of final consonant following the long vowel or the strong vowel as shown in Figure 3. This relationship occurs only between the vowel and the final consonant. In contrast, the initial consonant is slightly affected by the duration of the vowel. Hence, the vowel and the final consonant are tightly tied while an initial consonant is loosely tied with the vowel in the syllable. Consequently, decomposition of syllable into onset and rhyme is appropriate to Thai language.

Moreover, unlike other context-dependent phone units, the onset-rhyme is larger than triphone. The onset-rhyme is modeled with consonant including its transitional stage in onset and vowel along with final consonant in rhyme while the same phone units in triphone are differently modeled depended upon their contexts. The use of an acoustic unit with a longer duration facilitates exploitation of temporal and spectral variation simultaneously.

Table 1: Average duration of final consonant following short and long vowels

Final consonant	Duration final consonant (ms)		Final consonant	Duration final consonant (ms)	
	short	long		short	long
p	83.4	79.0	n	126.8	95.8
t	78.9	75.3	ng	120.6	96.3
k	80.1	77.2	j	111.5	100.7
m	123.5	98.6	w	108.3	99.2

3.2. Phonological Properties

Phonologically, the rhyme contains crucial prosodic information within the segment. The prosodic features, that the rhyme carried, are tone, stress, accent, and intonation [6]. For instance, the rhyme unit in Thai contains tone and stress information. Tones in Thai are also influenced by arresting consonant within the rhyme unit. [6] illustrated the use of tone information within the rhyme unit for tone recognition. The rhyme units give out better results than using the whole syllable or only within vowel segment. Both vowel and arresting consonant, making up a rhyme unit, have stored some prosodic information that is crucial for tone recognition.

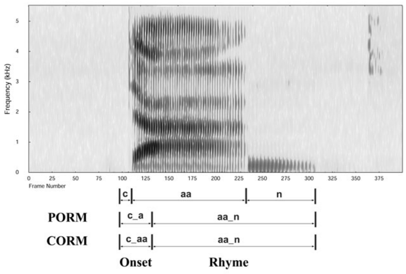

Figure 2: Onset-Rhyme segment

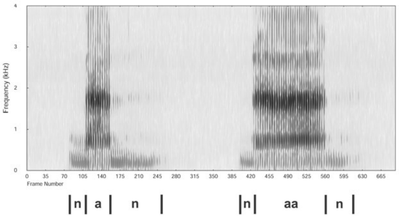

Figure 3: Relationship between vowel and final consonant duration.

3.3 Type of Onset-Rhyme

By considering the duration of the releasing consonant plus its transition preceding different vowel contexts, the onset-rhyme model is defined into two types: (1) Phonotactic Onset-Rhyme Model (PORM) and (2) Contextual Onset-Rhyme Model (CORM). These two models are generated from different combinations between releasing consonant and vowel. According to the duration of the releasing consonant and its transition, the phonotactic onset is differently created from the releasing consonant and its vowel context, even though the vowels are in the same pair. On the other hand, the contextual onset is similarly modeled from the releasing consonant on the same short-long vowel pairs. The number of phonotactic onset and contextual onset units are 792 and 297 units respectively while both models have the same 200 rhyme units as described in Table 2. This paper will explore both two types of onset-rhyme models in order to compare which type will be more efficient and suitable for Thai speech recognition.

3.3.1 Phonotactic Onset-Rhyme Model (PORM)

The onset units of the PORM are generated from combinations of releasing consonant in all possible vowel contexts. The phonotactic onset is differently created, according to the duration of the releasing consonant and its transition preceding the vowel, though the following vowel is in the same short-long pair. On the acoustic analyses, the patterns of formant transition of the same releasing consonant with different vowel contexts are similar except the transitional period as indicated in Table 3. Figure 4 shows the formant transitions of the releasing consonant [n] occurring in different contexts of vowel [i, ii, iia]. By considering the difference of releasing consonant plus transitional period of each vowel, onset units of PORM are individually modeled. The releasing consonant preceding the same short-long vowel pairs is differently modeled. For instance, the onset, consisting of the releasing consonant [n] and occurring in vowel [i, ii, iia], is separately modeled as [n_i, n_ii, n_iia]. According to their neighboring vowel, the onsets of PORM are thoroughly modeled.

Consequently, PORM will produce the most accurate onset units due to its completely contextual modeling.

3.3.2 Contextual Onset-Rhyme Model (CORM)

The contextual onset-rhyme models are proposed in this paper along with the phonotactic onset-rhyme models. Acoustic analyses conducted on Thai syllable showed similar pattern of formant transitions in particular cases. Apart from the duration of formant transition as indicated in Table 3, the formant patterns are similar in both short and long vowel context on the same releasing consonant. Figure 4 shows the formant transitions of the releasing consonant [n], occurring in three different contexts of vowel [i, ii, iia]. By disregarding the duration of formant transition, these onset units can be shared formant-transitional information. Therefore, combining similar onsets with short-long vowel pairs on the same releasing consonant substantially reduces the number of phonotactic onset units. The number of contextual onsets is reduced to 297 units compared with 792 units of the phonotactic onsets. The CORM gives a lower complexity in terms of search space than that of the PORM, which has the larger number of units.

Table 2: Numbers of speech units applying to Thai language

Unit	No. of units
Intra-syllable triphone	7,769
CORM	297O + 216R
PORM	792O + 216R

Table 3: Average duration of initial consonant preceding short and long vowels

Initial consonant	Duration of intial consonant (ms)		Initial consonant	Duration of initial consonant (ms)	
	short	long		short	long
p	66.40	74.73	m	76.69	83.31
t	67.85	72.78	n	73.26	78.58
k	66.40	73.62	ng	76.20	82.24
c	66.18	73.67	j	71.72	76.54
p^h	89.73	96.89	w	79.28	87.04
t^h	87.14	96.13	r	80.29	90.78
k^h	91.72	103.29	l	71.69	77.47
c^h	97.03	103.72	f	105.72	114.48
b	71.33	79.49	s	100.60	110.99
d	68.81	80.73	h	74.97	88.52
			?	60.63	65.50

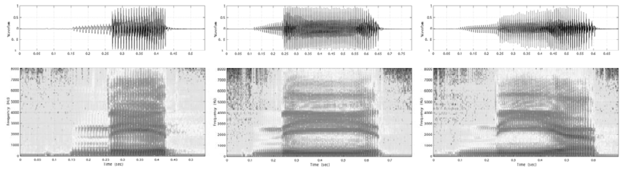
(a) /nit3/ (b) /niit2/ (c) /niiat2/

Figure 10: Spectrogram of: (a) /nit3/ (b) /niit2/ (c) /niiat2/

4. Experimental Results

All continuous speech recognition systems used in this paper are based on a standard LVCSR system, the HTK, developed by Cambridge University.

4.1. Speech Corpus

The speech corpus was recorded from 9 males and 11 females uttering in a reading style. The speech data were recorded with 16 bit resolution and 16 kHz sampling frequency in the office environment. A set of labeled training data, 661 utterances from each speaker, was used to initiate acoustic models. Another unlabeled utterances were used to train acoustic models after building them from labeled sentences. The numbers of unlabeled utterances, uttered by each speaker, is 1081. To evaluate the efficiency of the acoustic models, each speaker produces a set of 100 utterances consisting of 4,985 syllables. The read speech was selected due to less significant coarticulatory effects and pronunciation variations.

4.2. Acoustic Feature Extraction for Speech Recognition

The speech samples are passed through a signal processing routine. The 12-order MFCC and its temporal derivative are employed for speech parameterization [3, 4]. To compare the efficiency of the units, tone recognition is not implemented in this experiment.

4.3. Vocabulary and Language Modeling

Thai language is found to be relatively discrete to other western languages. Therefore, recognition of syllable sequence is more reasonable than word sequence in the utterance. The pronunciation dictionary is a syllable dictionary, consisting of 3,200 syllables excluding tones, which covers almost 33,000 vocabularies in Thai Royal Dictionary. In the recognition, two schemes, no language model and bigram language model, are applies. No language model is employed to evaluate actual performances of acoustic models. A bigram language mode is used to evaluate performance of the recognition system. The perplexity of this bigram language model is 42.78.

4.4. Construction of Acoustic Model

This section describes the implementation of two acoustic units. In this experiment the onset-rhyme was modeled and compared to the baseline triphone system. Construction of these acoustic models is described in the followings.

4.4.1. Intra-syllable Triphone

The phone model was used to create the triphone system. Primarily, a standard three state left-to-right topologies with no skip state was employed to initiate the acoustic model using a single Gaussian mixture from the labeled data. The standard Viterbi alignment process and Baum-Welch algorithm were applied to obtain the initial acoustic models. A set of initial acoustic models was trained with the embedded Baum-Welch algorithm. Then, the monophone model is used for creating the triphone system as the following procedures.

- Triphone construction: Monophone models were copied and tied their transition matrices. Then, triphones were initially trained using forward-backward algorithm. A tree-based clustering was applied to cluster data from a phonetic decision tree. Finally, state-tying, which merged any identical triphones, and re-training of state-tied triphones are the later processes.
- Mixture incrementing: A single Gaussian mixture would be split to attain the higher recognition performance. Finally, the splitting mixture models were re-estimated using forward-backward algorithm until the number of mixtures is sixteen components per state.

4.4.2. Onset-Rhyme Model

Initially, onset-rhyme labels were generated from manual labeling. The initial consonant segment and its transition towards the vowel, then, were converted to the onset segment. The rhyme segment is converted from the steady vowel portion and the optional final consonant. Since the rhyme consists of the entire vowel including the optional final consonant, this seems the rhyme comprises two connected phones. Therefore, the duration of rhyme is longer than that of phone unit. Then the number of HMM states used to model the rhyme is more than that of phone. A three state HMM was used for modeling onset containing the releasing consonant together with the transitional portion, whereas the rhyme consists of the steady vowel portion plus the arresting consonant is modeled with the six states of HMM. The onset-rhyme models were initiated and trained similar to monophone. Iteratively, re-estimation by four passes of Baum-Welch algorithm and increment of mixture components by a mixture-splitting procedure were employed to trained the models until the number of Gaussian mixture is increased to sixteen mixtures per state.

4.5. Performance Evaluation

The experiments were conducted on two schemes, recognition system using acoustic modeling only and recognition system using both acoustic and language modeling. Recognition results are shown and discussed in the following section.

4.5.1. Recognition Results of System Using Acoustic Model Only

In order to obtain an actual efficiency of acoustic model, language model should not be applied. Syllable recognition accuracy of triphone and onset-rhyme is shown in Table 4. Recognition results show accuracy of intra-syllable triphone, 34.39 %, less than the onset-rhyme. The PORM gives the highest accuracy at 46.79 %, and outperforms the CORM in terms of accuracy. However, with the number of units, the PORM has more complexity than CORM in terms of memory usage and search space.

Apart from the syllable accuracy, the recognition results of acoustic units of onset-rhyme were analyzed. The analysis of the recognition results in the acoustic level reveals the actual efficiency of the speech unit. Since the onset of PORM is completely modeled the initial consonant along with its transitional portion towards the vowel in every context, the PORM is more accurate than the CORM. Recognition results gave out the accuracy at 73.34 % of PORM onset better than 70.86% of CORM onset.

4.5.2. Recognition Results of System Using Acoustic Model and Language Model

Incorporating the language model can boost a performance of speech recognition system. The recognition results of the system using both acoustic and language model are shown in Table 4. Similar to the system using acoustic model only, the PORM attains the highest syllable accuracy at 76.71 %. The system using intra-syllable triphone performs 67.41 % accuracy, worse than the system utilizing onset-rhyme.

Like the previous section, the recognition results of acoustic units were also analyzed. These results were shown in Table 5. In this system, the accuracy of PORM onset still exceeds the CORM onset. Using the language model, the onset accuracy is increased around 10 %. In addition, the accuracy of the rhyme is substantially improved nearly 45-50 % when the language model is applied.

Table 4: Syllable recognition results

Speech Unit	Accuracy (AM)	Accuracy (AM+LM)
Intra-syllable triphone	34.39%	67.44%
CORM	44.05%	74.07%
PORM	46.79%	76.41%

Table 5: Acoustic unit recognition results

Speech Unit	Accuracy (AM)		Accuracy (AM+LM)	
	onset	rhyme	onset	rhyme
CORM	70.86%	53.37%	81.73%	79.02%
PORM	73.34%	54.34%	82.88%	79.09%

5. Conclusions

The onset-rhyme acoustic models are introduced in this paper for Thai languages. Two onset-rhyme models are proposed, Phonotactic Onset-Rhyme Model (PORM) and Contextual Onset-Rhyme Model (CORM). The models are applied to Thai in a continuous speech recognition system in order to illustrate their feasibility. The recognition results show major improvements over the triphome in many ways that indicates better performance of the models.

Since the onset of PORM is completely modeled the initial consonant along with its transitional portion towards the vowel in every context, the PORM is more accurate than the CORM. Recognition results gave out the accuracy at 74.07 % using the CORM and at 76.41 % using the PORM, respectively. However, PORM has more complexity than CORM in terms of memory usage and search space. Both PORM and CORM still excel triphone performance in terms of accuracy and complexity.

6. Acknowledgements

The authors would like to acknowledge TGIST and Government Research Grant in Research and Development Cooperative Project between EE Dept and Private Sector for partial financial support for the research to the second author.

7. References

[1] Ganapathiraju, A., “Syllable-Based large Vocabulary Continuous Speech Recognition”, *IEEE Trans. on Speech and Audio Processing*, 9(4): 358-366, 2001.

[2] Ahkuputra, V. et. al., “Comparison of Different Techniques on Thai Speech Recognition”. *Proc. IEEE Asia-Pacific Conf. on Circuits and Systems*, November 1998, 177 -180.

[3] Maneenoi, E. et. al., “Modeling of Onset-Rhyme for Speech Recognition of Thai Language”, *paper submitted in Eurospeech 2003,* Geneva, Switzerland.

[4] Maneenoi, E. et. al., “Acoustic Modeling of Onset-Rhyme for Thai Continuous Speech Recognition”, *Proc. 9th Australian Int. Conf. on Speech Science & Technology*, Melbourne, December 2002.

[5] Luksaneeyanawin, S., “A Three Dimensional Phonology: A Historical Implication”, *Proc. 3rd International Symposium on Language and Linguistics*, 75-90, 1992.

[6] Thubthong, N. et. al., “An Empirical Study for Constructing Thai Tone Models”. *Proc. SNLP-Oriental COCOSDA 2002*, 179-186.