Syllable Structure Based Phonetic Units for Context-Dependent Continuous Thai Speech Recognition

Supphanat Kanokphara

Information R&D Division,
National Electronics and Computer Technology Center (NECTEC)
112 Paholyothin Rd., Klong 1, Klong Luang, Pathumthani 12120 Thailand
supphanat_k@notes.nectec.or.th

Abstract

Choice of the phonetic units speech recognizer is a factor greatly affecting the system performance. Phonetic units are normally defined according to the acoustic properties of a speech. Nevertheless, with the limit of training data, too delicate acoustic properties are ignored. Syllable structure is one of the properties usually ignored in English phonetic units due to a lot of possible onsets and codas. Some language like Chinese successfully gets the benefit from incorporating the syllable structure information in the phonetic units, as the language itself is naturally syllabic and has only small amount of subsegments (onsets, nuclei, and codas). Thai, as some point between English and Chinese, has larger subsegments than Chinese but not as much as English. The process of this paper can be classified into 2 main steps. First, prove that Thai phonetic units can be defined as a set of syllabic elements without any data sparseness problem. Second, demonstrate that syllable structure based phonetic units give better accuracy rate from integrating the syllable structure information and reduce a lot of number of triphone units because of left and right context constraint in the syllable structure.

1. Introduction

Hidden Markov Models (HMMs) become a common structure of acoustic models (AMs) since HMMs can normalize speech signal's time-variation and characterize speech signal statistically. These AMs are used to represent speeches or parts of speech in the optimal sense. Due to a large number of speeches in the real world, it is more practical to design AMs in the phonetic level. With the drawback effect, phonetic units representing parts of speech are not easily designed as a result of many variations in the phone. The variations in the phone are, for example, suprasegmental components, context-dependency, etc.

For a long period of human history, the phone variations have been studied and analyzed by many phoneticians. This phonetic knowledge is very valuable in AM design. Unfortunately, with the limit of speech database and computational speed, some useful information cannot be included in AM. Syllable structure is one of the information impractical to be included in English speech recognition system. Many researchers regard this problem and try to integrate this information in the system [1][2]. In order to illustrate this, consider "t" (in context "iy t er") in "beater", "beat Ernest" and "baby turned" are distinctive even it is the same triphone since the articulations are dissimilar in different position. One way to solve this problem is to model larger phonetic units. For example, "t" can be separated to "t-within-word", "t-final" and "t-initial", respectively. However, these extended phonetic units become worse due to the data sparseness problem as onsets and codas are the combination of many phones such as "pt" in "concept", and "pts" in "concepts". "t" in "concept" is "t-final" while "t" in "concepts" is not. This problem has been alleviated by phone-position-dependent tree-clustering technique [2].

Chinese language, on the other hand, has small number of syllabic elements (ignoring tone). This allows Chinese researchers designed the Initial/Final (IF) phonetic units [3]. In the aspect of position dependency, IF system is unsurprisingly efficient as Initials and Finals are in fixed position in syllable structure. IF system also simplifies syllable structure from (C)V(C) to only (Initial)-Final structure. The Extended IF (XIF) introduces more phonetic units called Zero-Initials. These Zero-Initials compacts the syllable structure from (Initial)-Final structure to Initial-Final structure. With XIF, two adjacent Finals cannot occur in the database and this drastically decreases the number of possible triphones.

Since NECTEC launched continuous Thai speech recognition project, Thai speech corpus has been carefully developed [4][5]. Thai phonetic units (ignoring tone) are designed according to the phonetic knowledge [6][7][8]. In the corpus, there are Initials and Zero-Initial like Chinese with the addition of Middles as we have larger number of nuclei and codas than Chinese. According to Thai phonetics, there is only a Zero-Initial, "z". Thai syllable structure is basically (C)CV(V)(C) with (C)CV(V)(C)(C) structure for some loan word. With Initial-Middle-Final (IMF) phonetic units, Thai syllable structure is reduced to Initial-Middle-(Final) structure in [4] and [5].

The paper is organized as follows. Section 2 outlines the experiment framework in this paper. In section 3, various types of phonetic units are tested and discussed in context-independent (CI) level. As an extension from section 3, section 4 extends the experiment to context-dependent (CD) level. Then, the result overviews are summarized in section 5.

2. Experimental Setup

Mainly speaking, the experiments in this paper can be divided into 2 main parts. First, Thai phonetic units are analyzed phonetically (acoustic properties of each speech). Second, complexity and accuracy of CD system are observed and discussed. As the first part cares only how well an AM can represent the phonetic unit, only CI experiment is enough to find the most appropriate Thai phonetic units. In the second

part, all across-word CD models (word transition with coarticulation) are constructed from the CI model counterpart [9]. Not only the word accuracy but also the computation complexity is concerned in this paper.

The database in these experiments is 390 Thai phonetic balance (PB) sentences. Thai PB sentences contain 1478 words. The average number of words per sentence is 10. The average number of phones per word is 3.6. 42 speakers (21 males and 21 females) are separated into 34 speakers (17 males and 17 females) for training and 8 speakers (4 males and 4 females) for testing. Speakers for training are required to read 376 from 390 sentences while speakers for testing are required to read 14 from 390 sentences. The read speech utterances (16 kHz sampling frequency with 16 bits quantization) are parameterized into 12 dimensional MFCC, energy, and their delta and acceleration (39 length front-end parameters). The AM topology is 5-state left-right models. The language model (LM) is trained from the training set using back-off bi-gram. All experiments are trained and tested by using HTK toolkit [10].

3. Phonetic Unit Analysis

Starting from the smallest phonetic units, Thai phonetic units are designed without cluster and final consonants. Cluster consonant is the sequence of two consonants, e.g., "kw" → "k w", "phl"→ "ph l". Long vowel is the concatenation of two vowels, e.g. "aa" → "a a", "vva" → "v va". These units are base units. There will be more phonetically modifications in order to display the improvement gained from adding these properties in the unit design. Table 1 shows the 35 base Thai phonetic units.

Consonants	k, kh, ng, c, ch, s, j, d, t, th, n, b, p, ph, f, m, r, l, w, h
Vowels	a, i, v, u, e, x, o, @, q, ia, va, ua
Special symbols	sil, sp

Table 1: Base Thai phonetic units (I).

3.1. Modifications

The modifications in this paper are short-long vowel, cluster consonant, final consonant and Zero or glottal stop, respectively.

3.1.1. Short-long vowel modification

Short and long vowels have the same acoustic properties with different time duration. With this modification, the Middle is changed from V(V) to nucleus. Phonetically, this step integrates duration information in the vowel units. Table 2 shows the 44 Thai phonetic units including long vowels. In this table, "aa", "ii", "vv", "uu", "ee", "xx", "oo", "@@", "qq", "iia" are added to the Table 1. "va" and "ua" are respectively changed to "vva" and "uua" because there is no "va" and "ua" sound in the corpus and they rarely occur in Thai language.

Consonants	k, kh, ng, c, ch, s, j, d, t, th, n, b, p, ph, f, m, r, l, w, h
Vowels	a, aa, i, ii, v, vv, u, uu, e, ee, x, xx, o, oo, @, @@, q, qq, ia, iia, vva, uua
Special symbols	sil, sp

Table 2: Thai phonetic units with short-long vowels (II).

Consonants	k, kh, ng, c, ch, s, j, d, t, th, n, b, p, ph, f, m, r, l, w, h
Cluster consonants	pr, pl, tr, kr, kl, kw, phr, phl, thr, khr, khl, khw, br, bl, fr, fl, dr, jf, ts
Vowels	a, aa, i, ii, v, vv, u, uu, e, ee, x, xx, o, oo, @, @@, q, qq, ia, iia, vva, uua
Special symbols	sil, sp

Table 3: Thai phonetic units with long vowels and cluster consonants (III).

3.1.2. Single-Cluster consonant modification

Similar to English, Thai also has cluster consonants in the speech. Fortunately, the finite number of cluster consonants allows the construction of cluster consonants as basic phonetic units. With this modification, the Initials and Finals are changed from (C)(C) to only one element. Phonetically, this step integrates Initial position dependency in Initials. Table 3 shows the 63 Thai phonetic units including long vowels and cluster consonants. In this table, cluster consonants are put into the Table 2.

3.1.3. Final consonant modification

Syllable position dependency is an important issue influencing system accuracy. With this modification and 3.1-3.2 modifications, Thai syllable forms (Initial)-Middle-(Final) structure. Phonetically, this step integrates syllable position dependency in IMF structure. Table 4 shows the 75 Thai phonetic units including long vowels, cluster and final consonants. In this table, "jf" and "ts" are removed from cluster consonants and final consonants are added to Table 3 and consonants are changed to initial consonants.

Initial consonants	k, kh, ng, c, ch, s, j, d, t, th, n, b, p, ph, f, m, r, l, w, h
Cluster consonants	pr, pl, tr, kr, kl, kw, phr, phl, thr, khr, khl, khw, br, bl, fr, fl, dr
Final consonants	k^, ng^, j^, t^, n^, p^, m^, w^, ch^, f^, l^, s^, jf^, ts^
Vowels	a, aa, i, ii, v, vv, u, uu, e, ee, x, xx, o, oo, @, @@, q, qq, ia, iia, vva, uua
Special symbols	sil, sp

Table 4: Thai phonetic units with long vowels, cluster and final consonants (IV).

3.1.4. Zero modification

Initial consonants	k, kh, ng, c, ch, s, j, d, t, th, n, b, p, ph, f, m, r, l, w, h, z
Cluster consonants	pr, pl, tr, kr, kl, kw, phr, phl, thr, khr, khl, khw, br, bl, fr, fl, dr
Final consonants	k^, ng^, j^, t^, n^, p^, m^, w^, ch^, f^, l^, s^, jf^, ts^, z^
Vowels	a, aa, i, ii, v, vv, u, uu, e, ee, x, xx, o, oo, @, @@, q, qq, ia, iia, vva, uua
Special symbols	sil, sp

Table 5: Thai subsegmental phonetic units (V).

Zero is a term defined in [3] in order to characterize Initial part of Zero-Initial syllable in Chinese phonetic units. Phonetically, this term can be replaced by glottal stop. Even though there is still no concrete conclusion about glottal stop characteristic as it fully depends on speech variation, the existence of glottal stop is confirmed by many phoneticians. With this modification and 3.1-3.3 modifications, Thai syllable forms Initial-Middle-Final structure. As Initial, Middle and Final is respectively equivalent to onset, nucleus and coda in phonetics, these phonetic units can be considered as subsegmental units. Table 5 shows the 77 subsegmental units in Thai. In this table, glottal stop "z" is added to initial consonants and "z^" is added to final consonants.

3.1.5. Finer initial glottal stop modification

Finer initial glottal stop, like simple glottal stop, still preserves the IMF structure with more information from the following vowel. There are 9 extended glottal stops in Thai, i.e., "_a", "_i", "_v", "_u", "_e", "_x", "_o", "_@" and "_q". Table 6 shows the 85 Thai subsegmental units with finer initial glottal stops. The idea of finer initial glottal stops is comparable to XIF set in [3].

Initial consonants	k, kh, ng, c, ch, s, j, d, t, th, n, b, p, ph, f, m, r, l, w, h, _a, _i, _v, _u, _e, _x, _o, _@, _q
Cluster consonants	pr, pl, tr, kr, kl, kw, phr, phl, thr, khr, khl, khw, br, bl, fr, fl, dr
Final consonants	k^, ng^, j^, t^, n^, p^, m^, w^, ch^, f^, l^, s^, jf^, ts^, z^
Vowels	a, aa, i, ii, v, vv, u, uu, e, ee, x, xx, o, oo, @, @@, q, qq, ia, iia, vva, uua
Special symbols	sil, sp

Table 6: Thai subsegmental phonetic units with finer initial glottal stops (VI).

3.1.6. Finer final glottal stop modification

Finer final glottal stops are the final consonants version of finer initial glottal stops. Table 7 shows the 93 Thai subsegmental phonetic units with finer glottal stops. In this table, "_a^", "_i^", "_v^", "_u^", "_e^", "_x^", "_o^", "_@^" and "_q^" are inserted to the Final consonants in Table 6.

3.2. Experimental results and discussion

Phonetic units in Thai have been developed according to the acoustic properties. The development starts from coarse to fine phonetic units. Coarse phonetic set yields lower in both accuracy rate and computational requirement while fine

Initial consonants	k, kh, ng, c, ch, s, j, d, t, th, n, b, p, ph, f, m, r, l, w, h, _a, _i, _v, _u, _e, _x, _o, _@, _q
Cluster consonants	pr, pl, tr, kr, kl, kw, phr, phl, thr, khr, khl, khw, br, bl, fr, fl, dr
Final consonants	k^, ng^, j^, t^, n^, p^, m^, w^, ch^, f^, l^, s^, jf^, ts^, "_a^", _i^, _v^, _u^, _e^, _x^, _o^, _@^, _q^
Vowels	a, aa, i, ii, v, vv, u, uu, e, ee, x, xx, o, oo, @, @@, q, qq, ia, iia, vva, uua
Special symbols	sil, sp

Table 7: Thai subsegmental phonetic units with finer glottal stops (VII).

phonetic set results in higher accuracy rate and consequently larger computational load. As evidenced in Table 8, with higher accuracy rate, it is unavoidably to design larger phonetic set for the system.

From Table 8, the most appropriate Thai phonetic set is V. Even though the VII provides highest word accuracy in the experiment, there is no significant change in word accuracy comparing with V. Therefore, VII is inappropriate for the system as it costs more number of units with only small improvement. Comparing I with V, V shows 9.38% higher accuracy rate, which is significantly large while the computation speed between these two phonetic sets are considerably not much different. IV and V are relatively the same as a little bit improvement comes with a few more units in the system. So V can be counted as the best phonetic set for CI Thai speech recognition.

The relationship between each modification reveals the effect of each phonetic property for acoustic modeling. By observing the data, every modification except V, VI and VII, indicates the significant change in accuracy rate. With these experimental results, the study of Thai phonetics can definitely help improving Thai speech recognition system. As the study of Thai glottal stop is still unclear, introducing glottal stop in phonetic units cannot reveal any outperformed result. Also, the extension of glottal stop in VI and VII, unlike Chinese, don't generate any good results. This means the language structure significantly affects the system performance and each language system should be designed upon to one's language.

Size of possible triphone is another factor concerned in this paper since CD takes part deeply in the acoustic modeling, as triphone can effectively relieve the coarticulation problem. From Table 8, the smallest number of triphone phonetic set is I. However, I phonetic set shows

Unit type	Number of units	Number of triphones	Accuracy rate (%)
I	34	35,937	43.16
II	44	74,088	45.21
III	63	167,662	48.05
IV	75	84,899	52.15
V	77	37,620	52.54
VI	85	40,260	52.73
VII	93	46,919	53.61

Table 8: The experimental results for various types of Thai phonetic units.

much lower accuracy rate than V while the number of triphone is relatively not too much different (37,620 for V and 35,937 for I). With this reason, V is also the best phonetic sets for CD Thai speech recognition.

4. Subsegmental CD Phonetic Units

From last section, segmental phonetic units are proved to be fruitful in both CI and CD speech recognition system. This section shows the implementation of subsegmental CD phonetic unit speech recognition system. Typically, triphone are too large and cannot be directly trained. Tree-based tied state clustering technique is usually employed to break this problem [11]. The algorithm starts from assigning a pool of states at the tree root. Then, split the states according to the phonetic questions until the stop criteria are found.

4.1. Phonetic questions

Phonetic questions are constructed from linguistic knowledge. It has been proved that addition of extra linguistically motivated questions will not degrade the performance, but add more chances of better group classification. [12]. Therefore, the question design is also crucial in the CD acoustic modeling. According to the questions, units producing certain types of articulatory effect are grouped together. The unit groups are designed in different size from large to small. Larger group is at the parent node while smaller group is at the child node. Thai phonetic question can also be grouped according to the place and manner of articulation for consonants and tongue position for vowels. For example, according to place of articulation, phonetic units can be classified as labial, alveolar, palatal, velar, and glottal. With Thai language specific characteristics, subsegmental phonetic units can also be grouped as follows.

- Short-long vowel group such as {“a”, “aa”}, {“ia”, “iia”}, etc.
- Single-cluster consonant such as {“b”, “bl”, “br”}, {“ph”, “phl”, “phr”}, etc.
- Initial-Final consonant such as {“t”, “ts^”, “t^”}, {“p”, “p^”}, etc.
- Glottal stop, i.e., {“z”, “z^”}.

Based on Thai linguistic knowledge, there are 116 question for base Thai phonetic units and 370 questions for subsegmental Thai phonetic units. As stated above, higher number of question means better phonetic unit classification. Predictably, higher-level linguistic structure units, like subsegmental units, should produce more phonetic questions. This is also another merit of subsegmental Thai phonetic set.

4.2. Experimental results and discussion

Subsegmental phonetic units are chosen as the standard unit set for CI acoustic modeling. This section also shows the efficiency of subsegmental phonetic units in CD acoustic modeling. For higher accurate AM, the experiment is done with multi-Gaussian CD system. Table 9 shows the experimental results. The best result is 4 mixtures CD AM (highest accuracy rate with appropriate number of Gaussians).

Number of mixtures	Number of Gaussians	Accuracy rate (%)
1	5,737	74.80
2	11,470	78.81
4	22,940	80.08
6	34,410	80.27

Table 9: The experimental results for CD subsegmental phonetic units.

5. Conclusions

From the study of phonetics and comparing with other languages, optimal phonetic units can be implemented. In this paper, the benefit of phonetic unit design is study in many aspects, i.e., word accuracy and computation complexity. According to the phonetic knowledge and experimental result, subsegmental phonetic units are proved to outperform in both conditions in CD speech recognition system.

In the future work, more information, e.g., tone, prosody, etc, will be integrated in the phonetic units in order to find better CD acoustic modeling.

6. Acknowledgements

Thank Virongrong Tesprasit for phonetic and phonological knowledge and discussion.

7. References

[1] E. Fosler-Lussier, S. Greenberg and N. Morgan, “Incorporating contextual phonetics into automatic speech recognition”, *Proc. ICPhS*, pp. 611-614, 1999.

[2] I. Shafran and M. Ostendorf, “Use of higher level linguistic structure in acoustic modeling for speech recognition”, *Proc. ICASSP*, vol. 2, pp. 1021-1024, 2000.

[3] J. Zhang, F. Zheng, J. Li, C. Luo and G. Zhang, “Improved context-dependent acoustic modeling for continuous Chinese speech recognition”, *Eurospeech*, vol. 3, pp 1617-1620, 2001.

[4] C. Wutiwiwatchai, P. Cotsomrong, S. Suebvisai and S. Kanokphara, “Phonetically distributed continuous speech corpus for Thai language”, *Proc. LREC*, vol. 3, pp. 869-872, 2002.

[5] R. Thongprasirt, V. Somlertlamvanich, P. Cotsomrong, S. Subevisai and S. Kanokphara, “Progress report on corpus development and speech technology in Thailand”, *Proc. SNLP-Oriental COCOSDA 2002*, pp. 300-306.

[6] L. Sudaporn, “Speech Computing and Speech Technology in Thailand”, *Proc. SNLP*, pp 276-321, 1993.

[7] V. Kanitthanan, “*Language and linguistics*”, Thammasat University Press, Thaland, 1990 (in Thai).

[8] N. Ranakiat, “*Theoretical and practical phonetics*”, Thammasat University Press, Thailand, 2000 (in Thai).

[9] A. sixtus and H. Ney, “Training of across-word phoneme models for large vocabulary continuous speech recognition”, *Proc. ICASSP*, pp. 849-852, 2002.

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

[11] S.J. Young, J.J. Odell, P.C. Woodland, “Tree-based state tying for high accuracy acoustic modeling”, *Proc. ARPA Human Language Technology Workshop*, Plainsboro, NJ, p. 405-410, Morgan Kaufmann, March 1994.

[12] J. J. Odell, “The use of context in large vocabulary speech recognition”, *PhD Thesis*, *Queens’ College*, 1995.