Context-independent Acoustic Models for Thai Speech Recognition

Sawit Kasuriya, Supphanat Kanokphara, Nattanun Thatphithakkul,
Patcharika Cotsomrong, and Treepop Sunpethniyom
Speech Technology Section, Information Research and Development Division,
National Electronics and Computer Technology Center (NECTEC), Bangkok 12120 Thailand
Tel: +66-2564-6900, Fax: +66-2564-6873
E-mail: {sawitk, supphanat_k, nattanun_t, aye, treepop.sunpetchniyom}@notes.nectec.or.th

Abstract— **The main purpose of this paper is to investigate and construct the suitable acoustic models for Thai speech recognition. Many methods are presented in this paper including single phone modeling, base phone modeling, and tone modeling. The context-independent models and gender-independent models are often used in the acoustic modeling. Many modeling techniques are given from the linguistic-phonetic knowledge of Thai language. Experiments to evaluate the recognizer performance in word recognition are performed. The word utterances are used as training and testing data in our experiments. The experiments measure the word accuracy of each modeling methods when a number of mixtures are varied. The best accuracy is 87.03% from the single phone models without tone modeling (31 models) and 82.03% from the base phone models with tone modeling on vowels (171 models).**

Introduction

Many works on the speech recognition task have been done over the past decade. The large vocabulary continuous speech recognition (LVCSR) has been developed in many languages such as English, Chinese, Japanese, Korean, etc. Currently, their languages contain the huge data resources such as texts, speech utterances, and the parsing grammar. Many researchers have tried to investigate techniques of the global acoustic model, the acoustic model combination and context-independent model. All these techniques require large text and speech resources. The challenge is how to build the acoustic models from the limited data resources since some speech recognition applications do not required more than 90% of daily word coverage. In this paper, we propose many approaches for the acoustic model building in those conditions [1]. The rest of the paper is organized as follows. In section I, acoustic modeling is described. Section II describes speech database, Experiments and results are illustated in section III and section IV notes conclusions.

I. Acoustic Modeling

The first and foremost principle of the speech recognition that makes the system useful and powerful is the acoustic models. Acoustic modeling is a very important process because it directly effects the search speed, and accuracy. The design factors of acoustic modeling include the number of models which are suitable for the language coverage and the size of speech training database. The size of training database directly impacts the system performance. The number of acoustic models corresponds to linguistic knowledge of target language. Hence many acoustic modeling techniques are applied by following the linguistic knowledge. In this section, the basic of Thai phones and acoustic modeling techniques are described, which can be divided into two main techniques: tone modeling and no-tone modeling.

TABLE I
THAI PHONE SYMBOLS

Type of phone		Phone symbols
Initial consonants	Single	k, kh, ng, c, ch, s, j, d, t, th, n, b, p, ph, f, m, r, l, w, h, z
	Double	pr, pl, tr, kr, kl, kw, phr, phl, thr, khr, khl, khw, br, bl, fr, fl, dr
Vowels	Short	a, i, v, u, e, x, o, @, q
	Long	aa, ii, vv, uu, ee, xx, oo, @@, qq
	Diphthong	ia, iia, va, vva, ua, uua
Final consonants		k^, ng^, j^, t^, n^, p^, m^, w^, ch^, f^, l^, s^

A. Defining Thai Phone Models

The general forms of Thai syllables are $/C_{\mathrm{i}}V/$ and $/C_{\mathrm{i}}VC_{\mathrm{f}}/$ and the tone is marked onto each syllable. Five different tones in Thai are divided into two groups: (1) the static group–high, middle, and low tones, and (2) the dynamic group–rising and falling tones. Thai phonetic system has 21 single consonants, 12 double consonants, 24 vowels, and more than 5 double consonants that are used for pronouncing foreign words. A number of Thai phones are 74 (38+24+12) as shown in Table I. That also include 5 initial consonants (/br/, /bl/, /fr/, /fl/, /dr/) and 4 final consonants (/f^/, /s^/, /ch^/, /l^/) for foreign words. The character "^" is used to denote the difference between the initial consonants and the final consonants. More details are given in [2] and [3].

All of Thai phone symbols used in this paper are compatible with those used by other Thai researchers or linguists [2]. Next, the acoustic modeling method for Thai speech recognition is described.

B. Phone without Tone Modeling

The most basic acoustic modeling is done by ignoring tonal information. Some researchers use the tonal information as

TABLE II

31 ACOUSTIC MODELS

Type of phone	Phone symbols
Initial consonants (21 models)	k, kh, ng, c, ch, s, j, d, t, th, n, b, p, ph, f, m, r, l, w, h, z
Vowels (9 models)	a, i, v, u, e, x, o, @, q
Final consonants	k, ng, j, t, n, p, m, w, ch, f, l, s
Special symbols	sil

language modeling (meaning is used in syllable classification). Furthermore, the acoustic modeling without tonal information is very useful when the training database is small and does not cover all tone syllables. The scope of domain for some speech recognition applications does not require the tone modeling, e.g., voice command system in car environment, speech dialing, etc. In this paper, two methods are applied to build the phone models without tonal information: single phone models and base phone models.

1) Single Phone Models: Starting from the smallest phonetic units, Thai phonetic units are designed without double and final consonants. In this method, the double consonant is the sequence of two initial consonants, e.g., /kw/→/k/+/w/, /phl/→/ph/+/l/. The final consonant is substituted by the initial consonant symbol. And the long vowel is the concatenation of two short vowels, e.g. /aa/→/a/+/a/, because the short vowels and long vowels have the same acoustic properties with different time duration. In the same fashion, the diphthong are represented by two or three short vowels such as /vva/→/v/+/v/+/a/. These units are called the single units, because they are the basic phonetic units of Thai phonetic system. Table II shows the 30 single Thai phonetic units with a silence symbol of the acoustic models [4].

From Table II, the final consonants are used the same way as with the initial consonants (it does not use "^" to denote the difference). Therefore, the initial consonants are also used in the some way as the models for the final consonants. The number of single phone models is the smallest acoustic modeling in this paper.

2) Base Phone Models: From the previous subsection, the acoustic modeling will be phonetically modified further in order to display the improvement gained from adding these properties in the unit design. Hence, the long vowels, the diphthongs, the double consonants and the final consonants are added to the single phone models.

The single phone modeling presents the long vowel by integrating duration information of two short vowels. Diphthongs are the combination of vowels. The double consonants are the combination of two initial consonants. Generally, the combination of vowels or consonants causes change in the acoustic properties of phonetic units. Adding diphthongs or double consonants in the phonetic set can solve this problem. The initial consonants and final consonants are phones at the beginning and end of syllables. They have to use different acoustic models, because their acoustic properties are different. Introducing these phones to the acoustic models helps the system classify the phones easier. Table III shows the base phone modeling without tone.

TABLE III

75 ACOUSTIC MODELS

Type of phone		Phone symbols
Initial consonants (38 models)	Single	k, kh, ng, c, ch, s, j, d, t, th, n, b, p, ph, f, m, r, l, w, h, z
	Double	pr, pl, tr, kr, kl, kw, phr, phl, thr, khr, khl, khw, br, bl, fr, fl, dr
Vowels (24 models)	Short	a, i, v, u, e, x, o, @, q
	Long	aa, ii, vv, uu, ee, xx, oo, @@, qq
	Diphthongs	ia, iia, va, vva, ua, uua
Final consonants (12 models)		k^, ng^, j^, t^, n^, p^, m^, w^, ch^, f^, l^, s^
Special symbols		sil

TABLE IV

67 ACOUSTIC MODELS

Type of phone	Phone symbols
Initial consonants (21 models)	k, kh, ng, c, ch, s, j, d, t, th, n, b, p, ph, f, m, r, l, w, h, z
Short vowels (45 models)	a0, a1, a2, a3, a4, i0, i1, i2, i3, i4, v0, v1, v2, v3, v4, u0, u1, u2, u3, u4, e0, e1, e2, e3, e4, x0, x1, x2, x3, x4, o0, o1, o2, o3, o4, @0, @1, @2, @3, @4, q0, q1, q2, q3, q4
Final consonants	k, ng, j, t, n, p, m, w, ch, f, l, s
Special symbols	sil

C. Phone with Tone Modeling

Since Thai language is a tonal language, therefore the best acoustic modeling that is suitable for Thai speech recognition should include the tonal information. Therefore, the five tones are included in the acoustic models, and the modeling design question is how the tones should be included in those models. Many methods have been applied to the modeling, e.g., supra segmental, prosody modeling, syllable modeling, etc. In this approach, the tone modeling is focused on only vowels because the vowel period is the longest period when comparing with another phone. The effect of tone should be shown on the vowel period in each syllable [2].

1) Single Phone with Tone on Short Vowels: There are five tones in Thai language: mid, low, fall, high and rise. These tonal characteristics greatly affect the characteristic of vowels. In this paper, the digit numbers from 0 to 4 indicates mid, low, fall, high and rise, respectively. The single phone modification could include the tone information into the models and try to keep the concept of the single phone modeling (small phonetic units). Only the short vowels are represented the tonal and another phones are used the same condition with the single phone modeling. The number of these models with tonal is 67. These are shown in Table IV.

2) Single Phone with Tone on Vowels: This method appends the tonal information on all vowels, i.e., short vowels, long vowels and diphthongs, are separated with the digits 0 to 4. Therefore, the number of all vowel models is multiplied by five. They are increased from 24 to 120 models. The sum of acoustic models is 142 as shown in Table V.

TABLE V
142 ACOUSTIC MODELS

Type of phone	Phone symbols
Initial consonants (21 models)	k, kh, ng, c, ch, s, j, d, t, th, n, b, p, ph, f, m, r, l, w, h, z
Short vowels (45 models)	a0, a1, a2, a3, a4, i0, i1, i2, i3, i4, v0, v1, v2, v3, v4, u0, u1, u2, u3, u4, e0, e1, e2, e3, e4, x0, x1, x2, x3, x4, o0, o1, o2, o3, o4, @0, @1, @2, @3, @4, q0, q1, q2, q3, q4
Long vowels (45 models)	aa0, aa1, aa2, aa3, aa4, ii0, ii1, ii2, ii3, ii4, vv0, vv1, vv2, vv3, vv4, uu0, uu1, uu2, uu3, uu4, ee0, ee1, ee2, ee3, ee4, xx0, xx1, xx2, xx3, xx4, oo0, oo1, oo2, oo3, oo4, @@0, @@1, @@2, @@3, @@4, qq0, qq1, qq2, qq3, qq4
Diphthongs (30 models)	ia0, ia1, ia2, ia3, ia4, iia0, iia1, iia2, iia3, iia4, va0, va1, va2, va3, va4, vva0, vva1, vva2, vva3, vva4, ua0, ua1, ua2, ua3, ua4, uua0, uua1, uua2, uua3, uua4
Final consonants	k, ng, j, t, n, p, m, w, ch, f, l, s
Special symbols	sil

TABLE VI
96 ACOUSTIC MODELS

Type of phone	Phone symbols
Initial consonants (21 models)	k, kh, ng, c, ch, s, j, d, t, th, n, b, p, ph, f, m, r, l, w, h, z
Double consonants (17 models)	pr, pl, tr, kr, kl, kw, phr, phl, thr, khr, khl, khw, br, bl, fr, fl, dr
Short vowels (45 models)	a0, a1, a2, a3, a4, i0, i1, i2, i3, i4, v0, v1, v2, v3, v4, u0, u1, u2, u3, u4, e0, e1, e2, e3, e4, x0, x1, x2, x3, x4, o0, o1, o2, o3, o4, @0, @1, @2, @3, @4, q0, q1, q2, q3, q4
Final consonants (12 models)	k^, ng^, j^, t^, n^, p^, m^, w^, ch^, f^, l^, s^
Special symbols	sil

3) Base Phone with Tone on Short Vowels: As the counterpart methods from Table III and Table IV, only the short vowels of the base phone models are modified to have five tones and are used to represent the long vowels and diphthongs by two or more vowel concatenation. The number of base phone models with tone on short vowels is 96 as shown in Table VI.

4) Base Phone with Tone on Vowels: Combination with the consonant modeling method from Table III and the vowel modeling method from Table V, all vowels, which are short vowels, long vowels, and diphthongs, are attached with tones. The acoustic models are 171 as shown in Table VII. These models are the biggest models among all our approaches. All Thai phones use different symbols for each different phone position.

II. SPEECH DATABASE

The speech database used in this paper is a part of NECTEC-ATR Thai Speech Database called the isolated word set (DB1). It contains three subsets: 5,000 vocabularies subset (D0-D4), PB word subset (640 words) and extra word subset (D5). In this paper, the 5,000 vocabularies subset is used as a training data from 32 speakers (16 males and 16 females) and PB word subset is used for testing 10 speakers (5 males and 5 females) who are not in the training data.

TABLE VII
171 ACOUSTIC MODELS

Type of phone	Phone symbols
Initial consonants (21 models)	k, kh, ng, c, ch, s, j, d, t, th, n, b, p, ph, f, m, r, l, w, h, z
Double consonants (17 models)	pr, pl, tr, kr, kl, kw, phr, phl, thr, khr, khl, khw, br, bl, fr, fl, dr
Short vowels (45 models)	a0, a1, a2, a3, a4, i0, i1, i2, i3, i4, v0, v1, v2, v3, v4, u0, u1, u2, u3, u4, e0, e1, e2, e3, e4, x0, x1, x2, x3, x4, o0, o1, o2, o3, o4, @0, @1, @2, @3, @4, q0, q1, q2, q3, q4
Long vowels (45 models)	aa0, aa1, aa2, aa3, aa4, ii0, ii1, ii2, ii3, ii4, vv0, vv1, vv2, vv3, vv4, uu0, uu1, uu2, uu3, uu4, ee0, ee1, ee2, ee3, ee4, xx0, xx1, xx2, xx3, xx4, oo0, oo1, oo2, oo3, oo4, @@0, @@1, @@2, @@3, @@4, qq0, qq1, qq2, qq3, qq4
Diphthongs (30 models)	ia0, ia1, ia2, ia3, ia4, iia0, iia1, iia2, iia3, iia4, va0, va1, va2, va3, va4, vva0, vva1, vva2, vva3, vva4, ua0, ua1, ua2, ua3, ua4, uua0, uua1, uua2, uua3, uua4
Final consonants (12 models)	k^, ng^, j^, t^, n^, p^, m^, w^, ch^, f^, l^, s^
Special symbols	sil

The conditions of recording are as follows. All utterances were recorded in a quasi-quiet room. The qualities are around 20 dB. And only dynamic microphone (unidirectional microphone: SONY F720) is used in recording. A number of speakers are 20 males and 20 females (18 to 40 years old). All utterance is recorded in reading style and is middle and official dialect that is spoken in the middle area of Thailand. More details of this database are explained in [3].

III. EXPERIMENTS AND RESULTS

Investigations on the acoustic modeling methods for Thai speech recognition were carried out on two main techniques: tone modeling and no-tone modeling. In both experimental setups, we used HTK as a toolkit for building Hidden Markov Models (HMMs) from the same training data and evaluating with the same testing data. There are 39 dimensions (12 MFCC + log energy + 13 delta coefficients + 13 acceleration coefficients) used as the features. Only the number of acoustic modeling is different between each others.

A number of mixtures are varied to find out the best performance. In this section, we describe the approaches that are used and compare the results of six acoustic modeling methods as explained in the previous section. All modeling methods are monophone modeling or context-independent modeling. The implementation is divided into two experiments: acoustic modeling without tone experiment and acoustic modeling with tone experiment.

A. Acoustic Modeling without Tone Experiments

Comparison of the acoustic modeling methods between the single phone models and the base phone models indicates which phone modeling yields a better result in word recognition.

TABLE VIII

PHONE MODELING WITHOUT TONE MODEL RESULTS

Mixture	Word accuracy (%)	
	31 Models	75 Models
1	52.25	68.13
2	72.50	79.22
4	80.94	82.50
8	82.97	84.06
16	84.38	82.97
32	85.78	84.38
64	87.03	83.13
128	85.63	81.56

TABLE IX

SINGLE PHONE MODELING WITH TONE MODEL RESULTS

Mixture	Word accuracy (%)	
	67 Models	142 Models
1	54.38	65.48
2	70.47	72.50
4	77.97	77.03
8	77.50	76.41
16	78.91	77.66
32	78.28	75.00
64	76.72	71.41
128	54.38	65.48

From the results in Table VIII, the single phone models (31 models) have the highest word accuracy (87.03%) at 64 mixtures. On the other hand, the base phone models (75 models) reach to 84.38% of word accuracy at 32 mixtures. The single phone models yield higher word accuracy than the base phone models in the large size of mixtures (16 to 128).

B. Acoustic Modeling with Tone Experiments

There are two tone modeling experiments: the single phone model and the base phone model. The details of each modeling were described in section I.

1) Single Phone Model Experiment: The difference between 67 models and 142 models is tone modeling on only the short vowels or all vowels. The rest of Thai phones are using the same symbols of single phones.

The results from Table IX show that the 67 models gave higher word accuracy than 142 models at 16 mixtures. The 142 models gave the higher results at the small size of mixtures (1 and 2 mixtures). At a large size of mixture, the tone modeling on only the short vowels (67 models) gave the better results than the tone modeling on all vowels (142 models) in general case except at 128 mixtures.

2) Base Phone Model Experiment: This experiment is the same as the previous experiment for tone modeling methods. However, using the base phones to build the acoustic models makes the size of acoustic models larger than using the single phones.

From Table X, the results of the base phone with tone modeling on all vowels are mostly better than the base phone with tone modeling on only short vowels, but the improvement is insignificant. And at 16 mixtures, the word accuracy reaches the maximum for both modeling.

TABLE X

BASE PHONE MODELING WITH TONE MODEL RESULTS

Mixture	Word accuracy (%)	
	96 Models	171 Models
1	65.25	65.31
2	75.31	78.28
4	78.91	79.69
8	79.84	80.00
16	81.56	82.03
32	80.94	81.09
64	77.66	81.25
128	73.59	81.25

In this section, all results show the single phone models without tone modeling gave the best word accuracy (87.03%) at 64 mixtures. It is distinctively noticeable when comparing with the best result of tone modeling (82.03% from 171 models). The comparison between the single phone models and the base phone models can be indicated by two cases: no tone modeling and tone modeling. In the case of no tone modeling, the single phone models had the better result than the base phone models. On the other side, the base phone models gave the higher accuracy than the one in the tone modeling case.

IV. CONCLUSIONS

In this paper, we explore how the acoustic models can be builded for Thai speech recognition. The acoustic modeling can be presented in several methods: single phone modeling, base phone modeling, and tone modeling. Many design factors are considered for building the acoustic models to be suitable for recognizer including the size of acoustic models, the size of mixtures, accuracy rate, and process time. These factors are trade-off for each application system. For example, the voice command in car environment does not require the large size of command words and any words is not ambiguous in tonal. It is not necessary to use the acoustic models with tone modeling.

ACKNOWLEDGEMENTS

This research uses the partially part of NECTEC-ATR Thai speech database that is supported from SLT Dept. 2, ATR. The authors wish to thank their colleagues, Yoshinori Sagisaka, Genichiro Kikui, and Takatoshi Jitsuhiro for their advices.

REFERENCES

[1] Alex Waibel, Petra Geutner, Laura Mayfield Tomokiyo, Tanja Schultz an[d Monika Woszcayna, “Multilinguality in Speech and Spoken Language System”, Proceeding of IEEE, Vol. 88, No. 8, August 2000.

[2] S. Luksaneeyanawin, “Speech Computing and Speech Technology in Thailand”, Proceeding of the Symposium on Natural Language Proceeding in Thailand, pp. 276-321, 1993.

[3] S. Kasuriya, T. Jitsuhiro, G. Kikui, and Y. Sagisaka, “Thai Speech Recognition by Acoustic Models Mapped from Japanese”, Joint International Conference of SNLP-Oriental COCOSDA 2002, pp. 211-216, 2002.

[4] S. Kasuriya, V. Sornlertlamvanich, P. Cotsomrong, T. Jitsuhiro, G. Kikui, and Y. Sagisaka, “Thai Speech Database for Speech Recognition”, International Coordinating Committee on Speech Databases and Speech I/O System Assessment, pp. 105-111, 2003.