Thai Speech Recognition by Acoustic Models Mapped from Japanese

Sawit Kasuriya* Takatoshi Jitsuhiro* Genichiro Kikui* Yoshinori Sagisaka*,**

* Spoken Language Translation Research Laboratories,
Advanced Telecommunications Research Institute International,
2-2-2 Hikaridai Seika-cho, Soraku-gun,
Kyoto 619-0288, Japan
Tel.: +81-774-62-5403
Fax.: +81-774-95-1308(-9)

** GITI, Waseda University
29-7 Building 1-3-10 Nishi-Waseda Shinjuku-ku,
Tokyo 189-0051, Japan
Tel.: +81-3-5286-9853

e-mail : sawitk@slt.atr.co.jp, takatoshi.jitsuhiro@slt.atr.co.jp, genichiro.kikui@slt.atr.co.jp,
sagisaka@giti.waseda.ac.jp

Paper ID: SNLP COCOSDA-2002-NP0091

Keywords: Thai speech recognition, Isolated-word recognition, Acoustic models, Cross-language

Contact Author: Sawit Kasuriya

Abstract

This paper proposes a method of Thai speech recognition by using Japanese acoustic models. There was no Thai speech database for use in this study, and it takes a lot of time to correct a large number of speech database entries and label them. We also have only a small amount of Thai speech data, which was not alignment data. Therefore, we made initial models by mapping another language's acoustic models, and trained the initial models by using a small amount of data. We used phoneme mapping by the IPA table and made the initial Thai models from Japanese models. In our experiments, the initial models were trained by using isolated-word utterances by eleven Thai speakers. We evaluated the performance of isolated-word recognition with context-independent and context-dependent models.

Thai Speech Recognition by Acoustic Models Mapped from Japanese

Paper-ID: SNLP-O-COCOSDA-2002-NP0091

Abstract

This paper proposes a method of Thai speech recognition by using Japanese acoustic models. There was no Thai speech database for use in this study, and it takes a lot of time to correct a large number of speech database entries and label them. We also have only a small amount of Thai speech data, which was not alignment data. Therefore, we made initial models by mapping another language's acoustic models, and trained the initial models by using a small amount of data. We used phoneme mapping by the IPA table and made the initial Thai models from Japanese models. In our experiments, the initial models were trained by using isolated-word utterances by eleven Thai speakers. We evaluated the performance of isolated-word recognition with context-independent and context-dependent models.

1 Introduction

In a survey on Thai spech recognition, many reseachers were found to have focused on word-base recognition (Ahkuputra, 1997) and phoneme-base recognition (Ekkarit, 2000). Futhermore, research has never been done on continuous speech. The main problem was the need for a large Thai speech database. Therefore, National Electronics and Computer Technology Center, Thailand (NECTEC) were collecting and aligning a Thai speech database for speech recognition.

In this paper, the target language is Thai. We also had only a small amount of Thai speech data and none of the utterances were aligned. Recently, many researchers have focused on the question of how to build a large-vocabulary continuous speech recognition (LVCSR) system for a new target language using speech data from varied source languages. They have tried to build a multilingual speech recognition system (Tanja, 2001) (Uebler, 2001). Therefore, we used such a method and investigated Thai speech recognition. We made the initial Thai acoustic models by mapping from Japanese acoustic models. The Japanese acoustic models already exist. We used international phonetic alphabet (IPA) tables to map between Japanese and Thai phonemes. Our research was evaluated with Thai isolated-word recognition using Thai isolated-word utterances for training. They were spoken by eleven Thai speakers. In the experiments, we evaluated the initial models and trained models with four Thai speakers.

The organization of this paper is as follows. The second section introduces the making of Thai acoustic models by using Japanese acoustic models. Thai phonemes and mapping from Japanese phonemes to Thai phonemes are presented in this section. The next section describes the experiments. There are two experiments, the initial model experiment and the trained model experiment. And the last section contains our results.

2 Making Thai Acoustic Models by Using Japanese Acoustic Models

We made the Thai acoustic models by using Japanese acoustic models because we did not have a large enough Thai database to train Thai acoustic models directly and the Thai database did not have segment information. The Japanese acoustic models were trained from a very large database. This should make good initial models for the Thai acoustic models. Therefore, we converted the Japanese acoustic models to from the initial Thai acoustic models. After we got the initial models for the Thai acoustic models, they were trained with some Thai utterances using the Baum-Welch algorithm.

2.1 Thai phonemes

We will briefly describe Thai phonemes. The Thai language is a tonal language. Thai syllables are C_iV, C_iVC_f. There are about 30,000 syllables (Sudaporn, 1993). Each syllable has a tone and all tones have five several tones- a high tone, a middle tone, and a low tone in the static group, and a rise tone and a fall tone in the dynamic group.

The initial consonants (C_i) of the Thai language are 21 for single, twelve for double, and more than five initial double consonants for foreign languages. There are eight final consonants (C_f) and more than four for foreign languages. In Thai, there are nine short vowels, nine long vowels, and three short and three long double vowels. Thai vowels, single Thai consonants, and double Thai consonants are shown in Table B-2, B-3 and B-4 as a sequence in appendix B. All of these Thai phonemes refer to studies by Thai linguists (Sudaporn, 1993).

2.2 Mapping from Japanese phonemes to Thai phonemes

Japanese phonemes mainly have five vowels and 21 consonants for speech recognition, as defined by the Advanced Telecommunications Research Institute International (ATR). We have mapped the Japanese phonemes and Thai phonemes on the International Phonetic Alphabet (IPA) table, as shown in Figure 1 and Table 1.

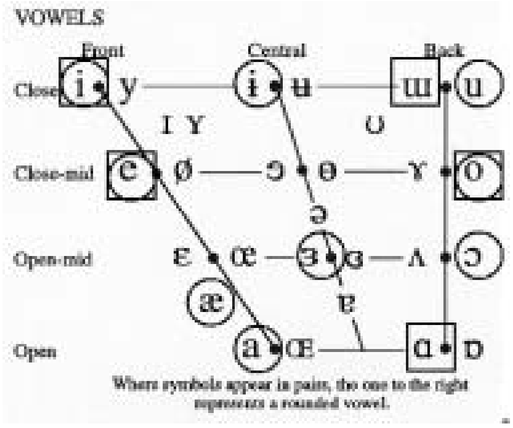
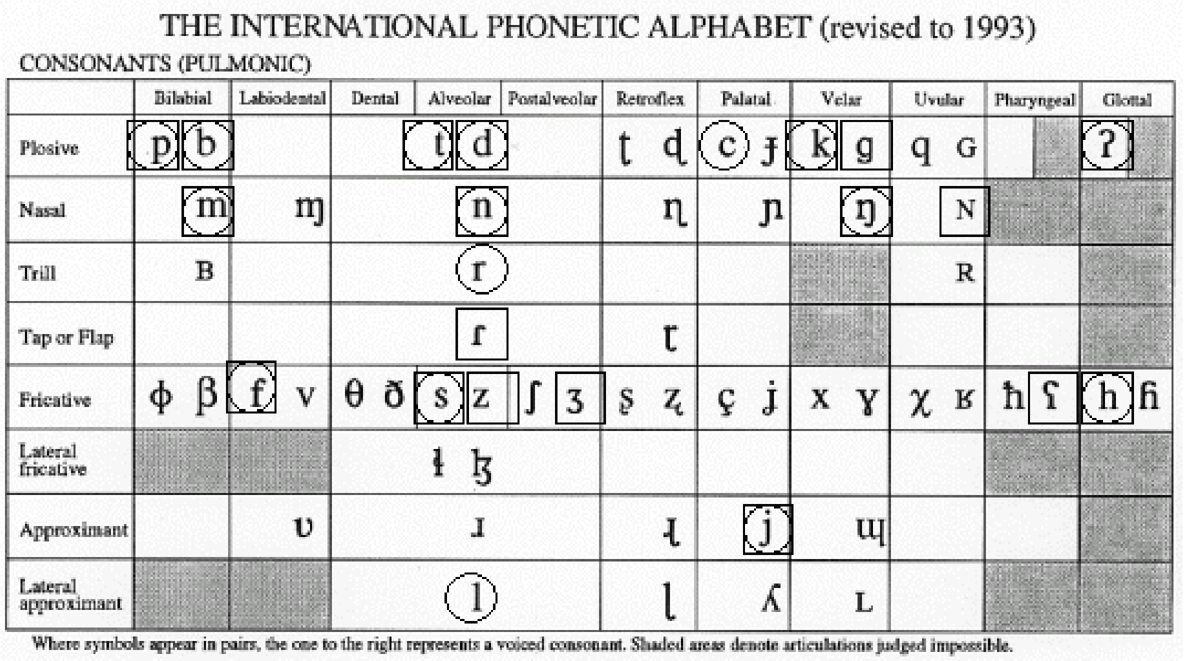


Figure 1. Japanese and Thai vowels on IPA

From Figure 1 and Table 1, the circles show the Thai phoneme positions, and the squares show the Japanese phoneme positions on the IPA table. Some Thai and Japanese phonemes have the same positions. That means a few phonemes should be similar to each other in occurrence and manner. It was easy to map from the Japanese phonemes to the Thai phonemes as shown in the highlight color of Table B-1. On the other hand, we tried to match Japanese phonemes with Thai phonemes by occurence, manner, and the acoustic of the phonemes as much as possible.

Table 1. Japanese and Thai consonants on IPA



THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

This mapping matched only single Thai and Japanese phonemes as shown in Table B-1. This conversion table did not consider the information from the Thai language tones. In the case of long vowels, double vowels, and double consonants, they were replaced with two single phonemes. For example, the long vowel of the Thai phoneme (/aː/) was defined by two Japanese vowels (/a/ + /a/). And the Thai double consonant (/kl/) was defined by the Japanese consonants (/k/ +/l/).

3 Experiments

In this paper, all experiments were performed using ATRSPREC and a Japanese database constructed by ATR. The Thai database was collected by NECTEC.

Context-independent and context-dependent acoustic models were used in these experiments.

3.1 Experimental conditions

All utterances used in these experiments were spoken by native Thai speakers using the Thai middle dialect.

3.1.1 Training

In these experiments, 12-order cepstral coefficients, 12-order delta-cepstral coefficients, and delta-powers were extracted for features.

The Japanese acoustic models were trained from 503 Japanese phonetic balanced sentences, spoken by 168 males and 232 females, with 19,948 utterances. Each phoneme was modeled with context-independent HMMs with 32 Guassian mixtures and 81 states, and context-dependent acoustic models (HMnet) (M. Ostendorf, 1997) with 5 Guassian mixtures and 1,403 states. These acoustic models were gender models. Only a silence model was trained to Thai speech data because the recording enviroment between the Japanese speech database and the Thai speech database was very different. After we converted acoustic models, we got Thai acoustic models with 93 states for context-independent models and with 1,935 states for context-dependent models. This included three states for silence models.

We converted 26 Japanense phonemes to 30 Thai phonemes using the conversion table shown in Table B-1. Thai acoustic models were trained by using 11,000 utterances from 5,000 Thai isolated-words of 11 Thai native speakers (6 females and 5 males). The training data required 3 hours and 53 minutes.

3.1.2 Evaluation

Two lexicons were used in our experiments: 50 isolated words and 250 isolated words. A male (M23) and a female (F03) were evaluated for the 50 isolated words. Two males (M05, M23) and two females (F03, F11) were evaluated for the 250 isolated words.

3.2 Initial model experiment

We evaluated the initial models that were converted from the Japanese acoustic models. Context-independent (CI) and context-dependent (CD) acoustic models were used in the experiments. The results indicated the performance of these initial models. As shown in Table 2, the context-dependent models exhibited a better performance than the context-independent models.

Table 2. Results of initial models

Lexicons	Speaker	Word Accuracy (%)	
		CI AM	**CD AM**
50 words	F03	58.00	58.00
	M23	50.00	72.00
	Average	**54.00**	**65.00**
250 words	F03	29.41	31.37
	F11	32.00	36.00
	M05	27.91	25.58
	M23	25.49	35.29
	Average	**28.70**	**32.06**

For some phonemes, these initial models can precisely segment utterances into phonemes. Therefore, these models can be trained to a certain extent.

3.3 Trained model experiment

The initial models were trained with 3 hours and 53 minutes of Thai utterances. Table 3 shows the results of this experiment.

These results indicate an improvement over the performance of the trained models. The improved performances of the context-independent models, compared with the previous experiments, were 17% and 30% for 50 and 250 isolated words. For example, the average word accuracy for the context-independent models with 50 isolated words was 54% when using the initial models and 71% when using the trained models, for a 17% increase. In the case of the context-dependent models with 50 and 250 isolated words, there was a 24% and 37% improvement. For example, the average word accuracy for the context-dependent models with 50 isolated words was 65% when using the initial models and 89% when using the trained models. That is a 24% increase.

Table 3. Results of trained models

Lexicons	Speaker	Word Accuracy (%)	
		CI AM	CD AM
50 words	F03	76.00	88.00
	M23	66.00	90.00
	Average	**71.00**	**89.00**
250 words	F03	62.75	68.63
	F11	64.00	84.00
	M05	58.14	65.12
	M23	49.02	58.82
	Average	**58.48**	**69.14**

When mapping the Japanese phonemes to Thai phonemes, it was difficult to map some of the Thai phonemes, such as “ɨ”, “ɜ”, “l”, and “r”. They are very different from Japanese phonemes. However, we tried to map some Japanese phonemes to them. Therefore, the results of the initial models were much lower than the results of models that were trained by a small amount of Thai speech data.

4 Conclusion

It takes a lot of time to correct a large amount of speech data and label them. We had only a small amount of Thai speech data, which was not alignment data. Therefore, we made initial models by mapping Japanese acoustic models, and trained the initial models by using a small amount of data. IPA tables were used for this mapping.

From our experiments, the context-dependent models of Japanese acoustic models that were converted to Thai acoustic models performed better than context-independent models in both experiments. Using Japanese acoustic models to convert Thai acoustic models was a good assumption as an indication in the first experiment.

The trained results exhibited improvements in the word accuracy rate under all conditions of the experiments, from about 17% to 40% when compared with the initial model experiments. The word accuracy rates of the 50 isolated words were better than those of the 250 isolated words, rising from about 13% to 33%.

References

Ahkuputra, V., Jitapunkul, S., Pornsukchandra, W., Luksaneeyanawin, S. 1997. “A Speaker-Independent Thai Polysyllabic Word Recognition Using Hidden Markov Model”, *Proceedings of the 1997 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, 593-599

Ekkarit Maneenoi, Somchai Jitapunkul, Visarut Ahkuputra, Umavasee Thathong, and Boonchai Thampanitchawong. 2000. “Thai monophthong recognition using continuous density hidden Markov model and LPC Cepstral coefficients”, The proceeding of the *6th International Conference on Spoken Language Processing*, IV: 620-623

Tanja Schultz, Alex Waibel. 2001. “Language-independent and language-adaptive acoustic modeling for speech recognition”, *Speech Communication*, 35(2001): 31-51

Ulla Uebler, “Multilingual speech recognition in seven languages”, *Speech Communication*, 35(2001): 53-69

Sudaporn Luksaneeyanawin, 1993. “Speech Computing and Speech Technology in Thailand”, *Proceeding of the Symposim on Natural Language Proceeding in Thailand*, 276-321

M. Ostendorf, H. Singer. 1997. “HMM topology design using maximum likelihood successive state splitting”, Computer Speech and Language, 11: 17-41

Appendix A. Thai Speech Database

The Thai speech database that is used at the Advanced Telecommunications Research International Institute (ATR) consists of three principle sets: the isolated-words set, phonetic balanced sentences, and hotel reservation transcription (HRT). The details of these sets are as follows:

(1) The isolated-word set contains 5,000 daily words, 640 phonetic balanced words, and 131 extra words for hotel reservation transcription.
(2) A 390 phonetic balanced sentence set.
(3) Each database has utterances for 50 dialogues of hotel reservation transcriptions

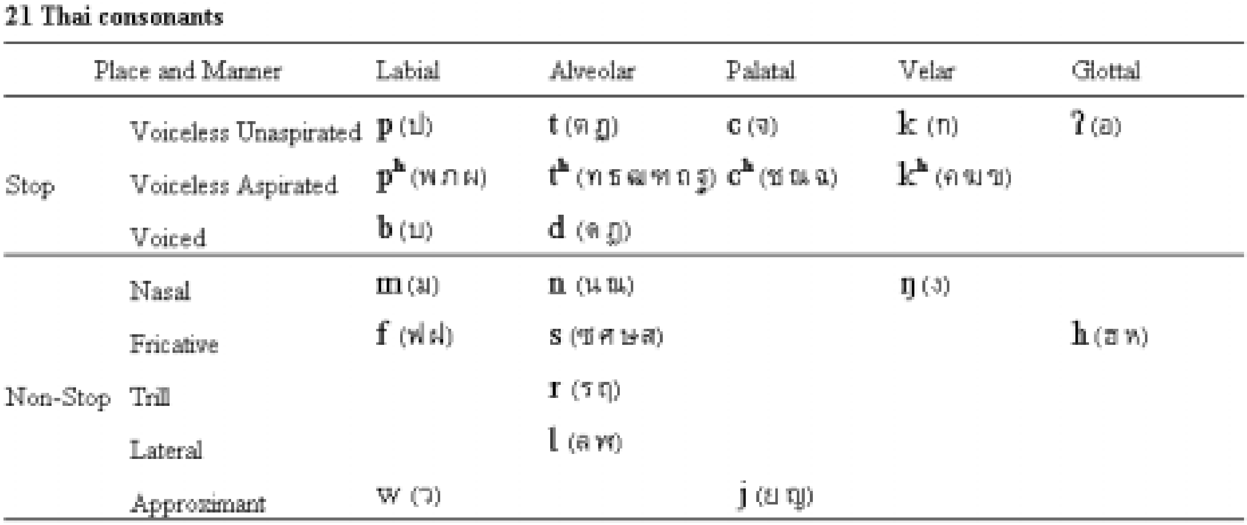
The reading speech data was spoken by 40 Thai native speakers (20 males and 20 females). They spoke in the middle dialect of the Thai language. All utterances were recorded in a quasi-quiet room.

Appendix B. Additional Tables

Table B-1. Conversion Table

Phone	TH	JA	Phone	TH	JA
Vowels	i	i	Consonants	t^h	t
	ɨ	i		c^h	tʃ
	u	ɯ		k^h	k
	e	e		b	b
	ɜ	e		d	d
	o	o		m	m
	æ	e		n	n
	a	ɑ		ŋ	ŋ
	ɔ	o		l	ɾ
Consonants	p	p		r	ɾ
	t	t		f	f
	c	tʃ		s	s
	k	k		h	h
	ʔ	ʔ		w	w
	p^h	p		j	j

Table B-2. Single Thai consonants



21 Thai consonants

	Place and Manner	Labial	Alveolar	Palatal	Velar	Glottal
Stop	Voiceless Unaspirated	p (ป)	t (ต ฏ)	c (จ)	k (ก)	ʔ (อ)
	Voiceless Aspirated	p^h (พ ภ ผ)	t^h (ท ธ ฒ ฑ ถ ฐ)	c^h (ช ฌ ฉ)	k^h (ค ฆ ข)	
	Voiced	b (บ)	d (ด ฎ)			
Non-Stop	Nasal	m (ม)	n (น ณ)		ŋ (ง)	
	Fricative	f (ฟ ฝ)	s (ซ ศ ษ ส)			h (ฮ ห)
	Trill		r (ร ฤ)			
	Lateral		l (ล ฬ)			
	Approximant	w (ว)		j (ย ญ)		

Table B-3. Double Thai consonants

12 double Thai consonants

Unaspirated Stop Set	pr (ปร)	tr (ตร)	kr (กร)
	pl (ปล)		kl (กล)
			kw (กว)
Aspirated Stop Set	p^hr (พร)	t^hr (ทร)	k^hr (คร)
	p^hl (พล)		k^hl (คล)
			k^hw (ขว)

Table B-4. Thai vowels

24 Thai vowels

Tongue Advancement / Tongue Height	Front	Central	Back
Close	i, iː (อิ, อี)	ɨ, ɨː (อึ, อือ)	u, uː (อุ, อู)
Mid	e, eː (เอะ, เอ)	ɜ, ɜː (เออะ, เออ)	o, oː (โอะ, โอ)
Open	æ, æː (แอะ, แอ)	a, aː (อะ, อา)	ɔ, ɔː (เอาะ, ออ)
Diphthongs	ia, iːa (เอียะ เอีย)	ɨa, ɨːa (เอือะ, อัวะ)	ua, uːa (อัวะ, อัว)