THAI AUTOMATIC SPEECH RECOGNITION

Sinaporn Suebvisai[1,2], Paisarn Charoenpornsawat[1], Alan Black[1,3], Monika Woszczyna[2], Tanja Schultz[1]
[1]Interactive Systems Laboratories, Carnegie Mellon University
[2]Multimodal Technologies Inc, Pittsburgh PA
[3]Cepstral LLC, Pittsburgh PA
E-mail: tanja@cs.cmu.edu

Abstract

We describe the development of a robust and flexible Thai Speech Recognizer as integrated into our English-Thai Speech-to-Speech translation system. We focus on the discussion of the rapid deployment of ASR for Thai under limited time and data resources, including rapid data collection issues, acoustic model bootstrap, and automatic generation of pronunciations. Issues relating to the translation and overall system will be reported elsewhere.

1. Introduction

This research was performed as part of the DARPA-Babylon program aimed at rapidly developing multilingual speech-to-speech translation capability in several languages. Building on our extensive background in ASR, language portability, and speech translation, our group has built Arabic-English and Thai-English Speech-to-Speech translation systems in less than 9 months per language [1]. This system has recently been used in an external DARPA evaluation involving medical scenarios between an American Doctor and a naïve monolingual Thai patient.

2. Thai Language Characteristics

With respect to speech recognition the Thai language bears challenging characteristics: (1) the usage of tones to discriminate meaning, which has an impact on the feature set used for acoustic modeling, (2) the relatively poor letter-to-sound relation, which makes the process of dictionary generation more challenging, and (3) the lack of word segmentation, which calls for automatic segmentation approaches to make n-gram language modeling feasible.

The Thai phoneme set (Table 1) consists of 21 consonantal phonemes, 17 consonantal cluster phonemes, and 24 vowels. Vowels are further divided into 9 short, 9 long vowels, and 6 diphthongs. Each vowel can carry one of 5 tones: low, mid, high, falling and rising. The syllable structure follows 4 patterns: CV, CCV, CVC_f and $CCVC_f$, with C, CC, C_f, V representing an initial consonant, a cluster consonant, a final consonant and a vowel respectively. Only 8 out of the 21 consonantal phonemes can be final consonants. The letter-to-sound relationship depends on context and position, for example the character "**ส**" in **สม** [soˇm] is pronounced as /s/ and in **รส** [rót] is pronounced as /t/. Moreover, there are many heterophone homographs in Thai such as "**แหน**" which can be pronounced [næˇ:] or [hæˇ:n] (see [2] for more details). Like Chinese and Japanese, Thai is written without any spaces between words. The correct segmentation of a sentence into words or phrases requires the full knowledge of the semantics of the sentence. For example, the word "**ตากลม**" can be segmented into "**ตา กลม**" (round eyes) and "**ตาก ลม**" (to expose wind) which are produced as [ta: | klom] and [tà:k | lom].

Initial consonants	p t c k ? ph th ch kh b d f s h m n ŋ w y r l
Cluster consonants	pr pl tr kr kl kw phr phl thr khr khl khw br bl fr fl dr
Final consonants	p t k m n ŋ w y
Short vowels	i ị u e ə o æ a ɔua ịa ia
Long vowels	i: ị: u: e: ə: o: æ: a: ɔ: ua: ịa: ia:
Tones	à a á â ǎ

Table 1: Thai phoneme set (IPA)

3. Language Data Acquisition

Here we describe our efforts in rapidly building up speech and text data resources for speech translation purposes.

3. 1 Speech Data

Hotel Reservation Data from NECTEC

For our early language adaptation experiments we received the permission from Thailand's National Electronics and Computer Technology Center (NECTEC) to use their Thai speech data collected in the hotel reservation domain. They provided us with a 6 hours text and speech database with high-quality 16kHz recordings from native Thai speakers. We used 34 speakers for training, 4 speakers for development, and 4 speakers for evaluation. NECTEC also provided manually pre-segmented transcriptions given in Thai script.

GlobalPhone style read newspaper articles

To create more general acoustic models we collected read speech data from native speakers based on the concepts of our multilingual data collection GlobalPhone [3]. The Thai speech data was recorded with a close-talk microphone in a push-to-talk scenario. Each speaker is prompted to read Thai

newspaper articles collected from the internet. More articles from the same newspapers were used to build statistical language models. Since the collection procedure does not require a time and cost consuming transcription process, the data could be recorded and were ready to use in less than a month. In total we collected 20 hours from 90 native Thai speakers in Bangkok, Thailand. The age of the 59 female and 31 male student speakers ranges between 18 and 25 years. Each speaker read on average 160 sentences, which corresponds to 20 minutes of speech. In sum, we recorded 14,039 sentences. Approximately 260,000 words were spoken, covering a vocabulary of about 7,400 words.

Babylon task-specific speech data

For the purpose of specializing our acoustic models and language models to the target task 'Medical dialogs', we additionally collected a small number of data from Thai native speakers. To avoid time consuming transcription we designed prompts which include word forms typically occurring in spontaneously spoken Thai speech and recorded 10 native speakers, out of which 8 were used for our adaptation experiments and 2 were used for testing.

Transcription Issues

The transcriptions of both, GlobalPhone and Babylon data were automatically segmented using a Thai word segmenter called 'Together' developed by Charoenpornsawat [4]. The segmenter requires a dictionary and provides various segmentation algorithms to automatically select suitable segmentations. Here we used a maximal matching algorithm including a method to handle context-independent segmentation ambiguities (see [5] for more details).

3.2 Text Data

For the development of MT components, bilingual data in the targeted domain are of major importance. We designed and used a very effective data collection procedure to provide such data, called *brainstorming*. In a brainstorming session up to three bilingual speaker are given a list of seed sentences in the source language and asked to create paraphrases from these sentences in the target language. Our experience shows that three people can spend up to 3 hours in a creative brainstorming session and process about 60 seed sentences. In 12 brainstorm sessions we collected about 3000 Thai sentences from 640 English seed sentences in the medical domain. The resulting bilingual corpus was successfully applied to enrich the Interlingua concepts and the statistical translators, as well as to train the Thai and English n-gram language models for Statistical MT and ASR.

Babylon task-specific text data

The Defense Language Institute (DLI) provided additional translations from monolingual collected English-English medical dialogs within the Babylon program. In total 2,507 pages of text were translated into Thai which adds up to 451,882 words. For building the task specific language models we used 350k words which had been available at the time of the experiments.

4. Automatic Speech Recognition

This section describes the rapid adaptation to Thai and the improvements made by considering the language's characteristics as described in section 2.

4.1 Automatic Pronunciation Generation

For a two-way speech translation system it is necessary to build a pronunciation dictionary that is needed in two components, the speech recognizer and the speech synthesizer. We share this dictionary between both components, which requires to also share a common phone set.

For high quality synthesis, we typically target correct pronunciations for at least 97% of all unique training words. Because the manual construction of new pronunciation dictionaries is too expensive we used novel techniques to make it more efficient. In previous work in Thai we had constructed a statistical letter-to-sound rule model for new words [6]. In this work we used a different phoneme set and also desired tonal information in each syllable. Using a more general lexicon construction method [7], we first manually transcribed the pronunciations of our base vocabulary, and from this data built a statistical letter-to-sound model. Using this model we predicted the pronunciation of additional words, and hand corrected errors. Iterating this method we quickly built a lexicon with pronunciations that covered our 8k word vocabulary. The final model achieves 56.84% word accuracy on a held out set of 621 unknown words.

4.2 Rapid Bootstrapping

The language adaptation techniques developed in our lab [8] enable us to rapidly bootstrap a speech recognizer in a new target language. Building on our earlier studies which showed that multilingual seed models outperform monolingual ones, we applied phonemes shared across seven languages (Chinese, Croatian, French, German, Japanese, Spanish, and Turkish) as seed models for the Thai phone set. In these first bootstrap experiments we used the data provided by NECTEC and disregarded the tone information. Since tone is a distinctive feature in the Thai language, this increases the number of homographs. In order to limit this number, we distinguished those word candidates by adding a tone tag. The resulting dictionary consists of 734 words which cover the given 6-hour database.

Table 2 describes the resulting performance for different acoustic model sizes indicating that a Thai speech recognizer can successfully bootstrapped with a reasonable amount of speech data. The good performance might be an artifact from the limited domain with a compact, closed vocabulary and low perplexity of the language model.

System	Dev Test	Eval Test
Context-Independent	14.4%	16.4%
Context-Dependent (500)	**13.0**%	**15.6**%
Context-Dependent (1000)	15.4%	17.3%

Table 2: WER for rapid bootstrap on HR corpus

Table 3 compares the performance of different acoustic models for bootstrapping a context independent system on the medical domain. We applied Thai models (Thai) build on GlobalPhone (see below), multilingual models (MM7) and English models (English). As expected, the Thai models work best, but the results also show that multilingual models outperform monolingual (English) ones.

System	Thai	MM7	English
Context-Independent	29.7%	32.5%	34.6%

Table 3: WER for rapid bootstrap on Babylon corpus

4.3 Phone set and Pronunciation Variation

After rapid bootstrapping we continued with building more generalized acoustic models using the GlobalPhone data and investigated the effect of enhancing the phone set and modeling pronunciation variants. Firstly, we investigated the impact of enhancing the baseline phone set by consonantal cluster phones. Rather than splitting up the 17 consonantal clusters into two separate phones (as in /kr/ composed of /k/ and /r/), we modeled the clusters as a single unit (/kr/). Secondly, we compare a single pronunciation dictionary with a multi-variant dictionary. To generate multiple pronunciation variants, we applied simple rules to handle the most common pronunciation variation effects when pronouncing words that include consonantal clusters. In this case, Thai speakers tend to omit /l/ and /r/, as for example, in the word "กลุ้ม" that should be pronounced [kluˆm], but many Thai people use [kuˆm].

System / #Acoustic Models	500	1000	2000
Baseline(#phones,single pron)	16.0%	15.2%	14.6%
Enhanced phone set	16.0%	-	14.4%
Pronunciation variants	15.6%	14.8%	**14.0**%

Table 4: WER on GP, Phone set and dictionary

The training was done on 80 speakers, 2 speakers were disregarded because of poor recording quality. For testing we used 1,181 utterances from 8 different speakers. The language model was built on news articles and gives a trigram perplexity of 140 and an OOV-rate of 1.4% on the test set using an 8k vocabulary. Table 4 gives the results for different model sizes. The modeling of pronunciation variants gives a significant improvement while the enhancement of the phone set does not seem to help.

All systems in Table 4 are based on quintphones, i.e. acoustic models consider two phonemes to the left and two to the right. An analysis on the Thai GlobalPhone data gave 65k triphones (±1), 184k quintphones (±2), and 242k septphones (±3). These numbers indicate a rather restricted phonology and correspond to the behavior of the Korean and Turkish parts of the multilingual GlobalPhone corpus. Korean is restricted mainly due to the segment length (our crossword polyphones reach one phoneme into the next word), and Turkish due to the small number of phonemes and the vowel harmony. We investigated a triphone scheme, but also a septphone scheme since the decision tree showed many questions reaching out to the ±2 neighborhood. With triphones we achieved 14.4%, with septphones 15.1%, so both could not outperform the quintphone system (14.0%).

4.4 Real-time Recognizer for Medical Dialogs

Since our target was to integrate a real-time Thai recognizer to our Thai-English speech translation system for medical dialogs, we adapted the general GP-based models to the medical domain using the medical data described in section 3. The experiments were performed on a fully continuous 3-state HMM system with 500 quintphone models using 32 Gaussians per state. The 13 mel cepstral coefficients, power, and the first and second derivatives had been reduced to 32 dimensions using LDA. For adapting the acoustic models, we used 2,433 utterances from 8 speakers of the Babylon data set. The test set consists of 322 utterances from two speakers. The trigram language model has a perplexity of 41.8 with an out-of-vocabulary rate of 0.48%.

In order to adapt the acoustic models to the medical domain using this very limited training material, we investigated 4 schemes. As a baseline we apply the acoustic models based on GlobalPhone training only (GP only), in the second scheme we use the Babylon training material to MLLR-adapt the GP models (GP+Bab MLLR). Thirdly, we joint the training material of both corpora, weighting the Babylon material by a factor of 2 (GP+Bab Mixed), the fourth scheme uses the GP models for initial alignments, but then completely retrain based on Babylon only material (Bab only). The third and the fourth scheme includes a re-clustering of the decision tree. Due to the limit of training data this results in only 378 models for the "Bab only" system. Table 5 shows the performance for these 4 adaptation schemes for the different phone sets and pronunciation variants as described in section 4.3.

System/Adaptation	GP only	GP+Bab MLLR	GP+Bab Mixed	Bab only
Baseline	24.6%	21.6%	20.6	21.5%
Enhanced phone set	23.1%	22.5%	-	22.6%
Pronunciation variants	23.7%	22.6%	**18.6**	19.6%

Table 5: WER on Babylon corpus

4.5 Tonal Features

Tones in Thai are more predictable and contain less information about the word identity compared to other tonal languages such as Chinese. For 8,112 distinct written forms in our dictionary, there were 7,733 distinct pronunciations with, and 7,272 distinct pronunciations without tone markup.

System / #Acoustic Models	GP	Babylon
Baseline (no tones)	16.0%	18.2%
Tone Tags	16.0%	19.1%
Tone Tags and Pitch Feature	16.2%	18.5%

Table 6: WER on GP+Bab, Tone Modeling

The experiments for the tonal features were produced on the same test set, but with a different recognizer: 600 triphone models, 12 mel cepstral coefficients with first and second derivatives reduced to 32 dimensions by removing higher second derivatives. The system is using the baseline dictionary and was trained using the 'GP+Bab Mixed' data from in the adaptation experiments.

For the second experiment tones were added to all vowels in the dictionary using statistical letter-to-sound models. Since the tones are also required for text-to-speech, they were then hand-corrected. The tones were used while building the cluster tree for the triphone model. While clustering different models can be assigned to the same phoneme if the tone differs and the gain in information is sufficiently large. The top three tone related questions used in this system were:

i:	current phoneme falling tone
a:	current phoneme falling tone
y	previous phoneme falling tone

For the final experiment, an additional pitch-feature (delta log pitch) was added and the system retrained. This resulted essentially in the same 'tone questions', but with a higher gain compared to the system without pitch feature. While this is an indication that the expected feature vector of a phoneme depends on the tone of the syllable, there seems to be little added discriminative value for ASR as seen in Table 6.

5. Conclusions

In this paper we described the development of a Thai speech recognizer using limited time and data resources. We successfully applied our rapid bootstrapping approach for initial acoustic models and our automatic dictionary generation scheme. Further experiments revealed that modeling of consonantal cluster phones do not show significant gains, while the introduction of pronunciation variants for words which include those cluster phones improved the performance. Using tonal representations for building the ASR cluster tree with or without a pitch-related feature does not seem to improve recognition performance. The described recognizer was integrated into our two-way English-Thai speech translation system and used in the external DARPA evaluation runs.

Acknowledgements

This work was partly funded by grants N66001-00-C-8007 and NBCHC030036 under the DARPA Babylon (CAST) and LASER-ACTD program: ``Mobile Speech-to-Speech Translation for Military Field Application." The opinions expressed in this paper do not necessarily reflect those of DARPA. The authors would like to thank DLI for providing the English to Thai translations. We also thank Virongrong Tesprasit who was a great help with her linguistic expertise of the Thai language. Furthermore, we are very grateful to Thailand's National Electronics and Computer Technology Center for giving the permission to use their database and dictionary for our bootstrapping experiments.

References

[1] Schultz, T., Alexander, D., Black, A., Peterson, K., Suebvisai, S., Waibel, A. (2004) *A Thai Speech Translation System For Medical Dialogs.* Proceedings of the Human Language Technologies (HLT), Boston, MA, May 2004.

[2] Tesprasit, V., Charoenpornsawat, P., and Sornlertlamvanich, V. (2003) "A Context-Sensitive Homograph Disambiguation in Thai Text-to-Speech Synthesis". In Proceedings of Human Language Technology Conference (HLT&NAACL 2003), Edmonton, Canada.

[3] Schultz, T., (2002) *GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University.* International Conference of Spoken Language Processing (ICSLP-2002), Denver, CO.

[4] Charoepornsawat, P. Together: Thai word segmentation program. [Online] http://www.thai.net/pcharoen/together.

[5] Meknavin, S., Charoenpornsawat, P., and Kijsirikul, B. (1997) "Feature-Based Thai Word Segmentation" In Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS 1997), Phuket, Thailand.

[6] Chotimongkol, A. and Black, A. (2000) "Statistically trained orthographic to sound models for Thai", ICSLP2000, Beijing, China

[7] Maskey. S, Black, A. and Tomokiyo, L. (2004) "Bootstrapping Phonetic Lexicons for New Languages" ICSLP2004, Jeju, Korea.

[8] Schultz, T. and Waibel, A. (2001) *"Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition"*, Speech Communication, Volume 35, Issue 1-2, pp. 31-51, August 2001.