Thai Grapheme-Based Speech Recognition

Paisarn Charoenpornsawat, Sanjika Hewavitharana, Tanja Schultz

Interactive Systems Laboratories, School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
{paisarn, sanjika, tanja}@cs.cmu.edu

Abstract

In this paper we present the results for building a grapheme-based speech recognition system for Thai. We experiment with different settings for the initial context independent system, different number of acoustic models and different contexts for the speech unit. In addition, we investigate the potential of an enhanced tree clustering method as a way of sharing parameters across models. We compare our system with two phoneme-based systems; one that uses a hand-crafted dictionary and another that uses an automatically generated dictionary. Experiment results show that the grapheme-based system with enhanced tree clustering outperforms the phoneme-based system using an automatically generated dictionary, and has comparable results to the phoneme-based system with the hand-crafted dictionary.

1 Introduction

Large vocabulary speech recognition systems traditionally use phonemes as sub-word units. This requires a pronunciation dictionary, which maps the orthographic representation of words into a sequence of phonemes. The generation of such a dictionary is both time consuming and expensive since it often requires linguistic knowledge of the target language. Several approaches to automatic dictionary generation have been introduced in the past with varying degrees of success (Besling, 1994; Black et al., 1998). Nevertheless, these methods still require post editing by a human expert or using another manually generated pronunciation dictionary.

As a solution to this problem, grapheme-based speech recognition (GBSR) has been proposed recently (Kanthak and Ney, 2002). Here, instead of phonemes, graphemes – orthographic representation of a word – are used as the sub word units. This makes the generation of the pronunciation dictionary a trivial task. GBSR systems have been successfully applied to several European languages (Killer et al., 2003). However, because of the generally looser relation of graphemes to pronunciation than phonemes, the use of context dependent modeling techniques and the sharing of parameters across different models are of central importance.

The variations in the pronunciation of phonemes in different contexts are usually handled by clustering the similar contexts together. In the traditional approach, decision trees are used to cluster polyphones – a phoneme in a specific context – together. Due to computational and memory constraints, individual trees are grown for each sub-state of each phoneme. This does not allow the sharing of parameters across polyphones with different center phonemes. Enhanced tree clustering (Yu and Schultz, 2003) lifts this constraint by growing trees which cover multiple phonemes.

In this paper we present our experiments on applying grapheme-based speech recognition for Thai language. We compare the performance of the grapheme-based system with two phoneme-based systems, one using a hand-crafter dictionary, and the other using an automatically generated diction-

Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, pages 17–20,
New York City, June 2006.

ary. In addition, we observe the effect of the enhanced tree clustering on the grapheme-based recognition system.

2 Grapheme-to-Phoneme Relation in Thai

In the grapheme-based approach, the pronunciation dictionary is constructed by splitting a word into its constituent letters. Previous experiments have shown that the quality of the grapheme-based recognizer is highly dependent on the nature of the grapheme-to-phoneme relation of a specific language (Killer, 2003). In this section we have a closer look at the grapheme-to-phoneme relation in Thai.

Thai, an alphabetical language, has 44 letters for 21 consonant sounds, 19 letters for 24 vowel sounds (9 short vowels, 9 long vowels and 6 diphthongs), 4 letters for tone markers (5 tones), few special letters, and numerals. There are some characteristics of Thai writing that can cause problems for GBSR:

- Some vowel letters can appear before, after, above or below a consonant letter. e.g. In the word "แมว" (/mae:w/), the vowel "แ" (/ae:/) appears before the consonant "ม" (/m/).
- Some vowel and consonant letters can be combined together to make a new vowel. e.g. In the word "มัว" /mua/, the vowel "ua" is composed of a vowel letter " ั " and a consonant letter "ว".
- Some vowels are represented by more than one vowel letter For example, the vowel /ae/ requires two vowel letters: "แ" and "ะ". To make a syllable, a consonant is inserted in between the two vowel letters. e.g. "และ" (/lae/). The consonant "ล" (/l/) is in the middle.
- In some syllables, vowels letters are not explicitly written. e.g. The word "ยก" (/yok/) consists of two consonant letter, "ย" (/y/) and "ก" (/k/). There is no letter to represent the vowel /o/.
- The special letter " ์ ", called Karan, is a deletion marker. If it appears above a consonant, that consonant will be ignored. Sometimes, it can also delete the immediately preceding consonant or the whole syllable.

To make the relationship between graphemes and phonemes in Thai as close as possible we apply two preprocess steps:

- Reordering of graphemes when a vowel comes before a consonant.
- Merging multiple letters representing a single phoneme into one symbol.

We use simple heuristic rules for this purpose; 10 rules for reordering and 15 for merging. In our initial experiments, reordering alone gave better results than reordering plus merging. Hence, we only used reordering rules for the rest of the experiments.

3 Thai Grapheme-Based Speech Recognition

In this section, we explain the details of our Thai GBSR system. We used the Thai GlobalPhone corpus (Suebvisai et.al., 2005) as our data set, which consists of read-speech in the news domain. The corpus contains 20 hours of recorded speech from 90 native Thai speakers consisting of 14k utterances. There are approximately 260k words covering a vocabulary of about 7,400 words. For testing we used 1,181 utterances from 8 different speakers. The rest was used for training. The language model was built on news articles and gave a trigram perplexity of 140 and an OOV-rate of 1.4% on the test set.

To start building the acoustic models for Thai, we first used a distribution that equally divided the number of frames among the graphemes. This was then trained for six iterations followed by writing the new labels. We repeated these steps six times. As can be seen in Table 1, the resulting system (Flat-Start) had poor performance. Hence we decided to bootstrap from a context independent acoustic model of an existing phoneme-based speech recognition (PBSR) systems.

3.1 Bootstrapping

We trained two grapheme-based systems by bootstrapping from the acoustic models of two different PBSR systems. The first system (Thai) was bootstrapped from a Thai PBSR system (Suebvisai et al., 2005) trained on the same corpus. The second system (Multilingual) was bootstrapped from the acoustic models trained on the multilingual GlobalPhone corpus (Schultz and Waibel, 1998) which shares acoustic models of similar sounds across multiple languages. In mapping phones to graphemes, when a grapheme can be mapped to

several different phones we selected the one which occurs more frequently.

Both systems were based on trigraphemes (+/-1) with 500 acoustic models. Training was identical to the Flat-Start system. Table 1 compares the word error rates (WER) of the three systems on the test set.

Flat-Start	Multilingual	Thai
37.2%	27.0 %	26.4 %

Table 1: Word error rates in % of GBSR systems with different bootstrapping techniques

Results show that the two bootstrapped systems have comparable results, while Thai system gives the lowest WER. For the rest of the experiments we used the system bootstrapped from the multilingual acoustic models.

3.2 Building Context Dependent Systems

For the context dependent systems, we trained two systems each with different polygrapheme units; one with trigrapheme (+/- 1), and another with quintgrapheme (+/-2).

The question set used in building the context dependent system was manually constructed by using the question set from the Thai PBSR system. Then we replaced every phoneme in the question set by the appropriate grapheme(s). In addition, we compared two different acoustic model sizes; 500 and 2000 acoustic models.

Table 2 shows the recognition results for the resulting GBSR systems.

Speech Unit	500 models	2000 models
Trigrapheme	26.0 %	26.0 %
Quintgrapheme	27.0 %	30.3 %

Table 2: Word error rates in % of GBSR systems using different speech units and the # of models.

The system with 500 acoustic models based on trigraphemes produced the best results. The higher WER for the quintgrapheme system might be due to the data sparseness.

3.3 Enhanced Tree Clustering (ETC)

Yu and Schultz (2003) introduced a tree clustering approach that allows the sharing of parameters across phonemes. In this enhanced tree clustering, a single decision tree is constructed for all sub-states of all phonemes. The clustering procedure starts with all the polyphones at the root of the tree. The decision tree can ask questions regarding the identity of the center phoneme and its neighboring phonemes, plus the sub-state identity (begin/middle/end). At each node, the question that yields the highest information gain is chosen and the tree is split. This process is repeated until the tree reaches a certain size. Enhanced tree clustering is well suited to implicitly capture the pronunciation variations in speech by allowing certain polyphones that are pronounced similarly to share the same set of parameters. Mimer et al. (2004) shows that this approach can successfully be applied to grapheme based speech recognition by building separate trees for each sub-state for consonants and vowels.

For the experiments on enhanced tree clustering, we used the same setting as the grapheme-based system. Instead of growing a single tree, we built six separate trees – one each for begin, middle and end sub-states of vowels and consonants. Apart from the question set used in the grapheme-based system, we added singleton questions, which ask about the identity of different graphemes in a certain context. To apply the decision tree algorithm, a semi-continuous recognition system was trained. Since the number of models that share the same codebook drastically increases, we increased the number of Gaussians per codebook. Two different values were tested; 500 (ETC-500) and 1500 (ETC-1500) Gaussians. Table 4 shows the recognition results on the test set, after applying enhanced tree clustering to the system based on trigraphemes (MUL-TRI).

	500 models	2000 models
MUL-TRI	26.0 %	26.0 %
ETC-500	16.9 %	18.0 %
ETC-1500	18.1 %	19.0 %

Table 3: Word error rate in % for the enhance tree clustering method

As can be seen from Table 3, the enhanced tree clustering has significant improvement over the best grapheme-based system. ETC-500 with relatively lesser number of parameters has outperformed ETC-1500 system. Performance decreases when we increase the number of leaf nodes in the tree, from 500 to 2000. A closer look at the cluster trees that used the enhanced clustering reveals that

50~100 models share parameters across different center graphemes.

4 Grapheme vs. Phoneme based SR

To evaluate our grapheme-based approach with the traditional phoneme-based approach, we compared the best GBSR system with two phoneme-based systems.

The first system (PB-Man) uses a manually created dictionary and is identical to (Suebvisai et al., 2005) except that we used triphones as the speech unit. The second system (PB-LTS) uses an automatically generated dictionary using letter-to-sound rules. To generate the dictionary in PB-LTS, we used the letter-to-sound rules in Festival (Black 1998) speech synthesis system trained with 20k words. We also applied the same reordering rules used in the GBSR system as described in section 2. Both the systems have 500 acoustic models based on triphones.

Table 4 gives the WER for the two systems, on the test set. Best results from GBSR systems are also reproduced here for the comparison.

Phoneme-based	
Using manual dictionary (PB-Man)	16.0 %
Using automatic dictionary (PB-LTS)	24.5%
Grapheme-based	
MUL-TRI	26.0 %
MUL-TRI with ETC (ETC-500)	16.9 %

Table 4: Word error rates in % of GBSR and PBSR systems

As expected, the manually generated dictionary gives the best performance. The performance between PB-LTS and grapheme based system are comparable. ETC-500 system has a significantly better performance than the automatically generated dictionary, and almost the same results as the phoneme-based baseline. This shows that grapheme-based speech recognition coupled with the enhanced tree clustering can be successfully applied to Thai speech recognition without the need for a manually generated dictionary.

5 Conclusions

In this paper we presented the results for applying grapheme-based speech recognition to Thai language. We experimented with different settings for the initial context independent system, different number of acoustic models and different contexts for the polygraphemes. We also tried the enhanced tree clustering method as a means of sharing parameters across models. The results show that the system with 500 acoustic models based on trigraphemes produce the best results. Additionally, the enhanced tree clustering significantly improves the recognition accuracy of the grapheme-based system. Our system outperformed a phoneme-based system that uses an automatically generated dictionary. These results are very promising since they show that the grapheme-based approach can be successfully used to generate speech recognition systems for new languages using little linguistic knowledge.

References

Stefan Besling. 1994. "Heuristical and Statistical Methods for Grapheme-to-Phoneme Conversion. In Proceedings of Konvens. Vienna, Austria.

Alan W. Black, Kevin Lenzo and Vincent Pagel. 1998. Issues in Building General Letter to Sound Rules. In *Proceedings of the ESCA Workshop on Speech Synthesis*, Australia.

Sebastian Kanthak and Hermann Ney. 2002. Context-dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition. In *Proceedings of the ICASSP*. Orlando, Florida.

Mirjam Killer, Sebastian Stüker, and Tanja Schultz. 2003. Grapheme Based Speech Recognition. In *Proceeding of the Eurospeech*. Geneva, Switzerland.

Borislava Mimer, Sebastian Stüker, and Tanja Schultz. 2004. Flexible Decision Trees for Grapheme Based Speech Recognition. In *Proceedings of the 15th Conference Elektronische Sprachsignalverarbeitung (ESSV)*, Cotbus, Germany, September.

Tanja Schultz and Alex Waibel. 1998. Development of Multi-lingual Acoustic Models in the GlobalPhone Project. In *Proceedings of the 1st Workshop on Text, Speech, and Dialogue (TSD)*, Brno, Czech Republic.

Sinaporn Suebvisai, Paisarn Charoenpornsawat, Alan Black and et.al. 2005 Thai Automatic Speech Recognition. Proceedings of ICASSP, Philadelphia, Pennsylvania.

Hua Yu and Tanja Schultz. 2003. Enhanced Tree Clustering with Single Pronunciation dictionary for Conversational Speech Recognition. In *Proceedings of the 8th Eurospeech*, Geneva, Switzerland.