Spontaneous Thai Speech Recognition

Monika Woszczyna[1,2], Paisarn Charoenpornsawat[2], Tanja Schultz[2]

[1]: Multimodal Technologies, Inc
[2]: Carnegie Mellon University
monika@cs.cmu.edu

Abstract

This paper expands previous work on Thai speech recognition, investigating pronunciation changes such as syllable and phoneme elisions as well as phoneme shifts in Thai spontaneous speech. We compare several approaches to model these effects in large vocabulary continuous speech recognition across multiple domains. This work includes experiments on two new speech databases that significantly alleviate the data sparseness problem of earlier publications. We found that given sufficient training data, a fully data driven approach using an allophone cluster tree yields the best results. Explicit modeling of pronunciation changes does not improve performance across domains.

Index Terms: Thai, speech recognition, spontaneous speech, pronunciation modeling, acoustic model sharing

1. Introduction

The underlying goal of the work in this paper was to build a Thai recognizer that works well for a variety of continuous speech tasks. During the initial analysis of the training data, we found several pronunciation changes that we need to address: syllable and phoneme elisions, as well as phoneme shifts due to fast, poorly enunciated speech and/or regional dialects.
General information about other issues in Thai speech recognition, such as letter-to-sound mapping, phoneme set selection, segmentation, and tonality can be found in earlier publications [1, 2, 3, 4, 5, 6, 7].

1.1. Syllable Elisions

The most striking pronunciation change in spontaneous Thai speech we observed in our data is the omission of entire syllables, mostly in frequent words. Table 1 shows some example words with their canonical and observed pronunciations.

Thai Word	Canonic	Observed
สวัสดี	**sa**waddee	waddee
คุณสมบัติ	khun**na**sombat	khunsombat
ปรกติ	prok**ka**ti	prokti
ประกอบธุรกิจ	prakypthu**ra**kit	prakypthukit

Table 1: Examples for syllable and phoneme deletions

Unfortunately, we did not observe enough examples to find patterns for these deletions and therefore decided to model them on a case by case base with explicit dictionary entries.

1.2. Phoneme Elisions

Phoneme elisions in Thai are most frequently observed in phoneme clusters such as /kr/ /khr/ /pl/ /phl/. In spontaneous or poorly enunciated speech, the second consonant can be slurred, muted or completely missing. There are several approaches to model this effect for speech recognition:

1. Model consonant clusters as individual phones. Since some consonant clusters are rare, this approach can lead to undertrained models [2,3].
2. Add pronunciation variants without the second consonant. This approach introduces additional homophones and contaminates the models of adjacent phones during training.
3. Add pronunciation variants that have zero-length phones for /r/ and /l/ in consonant clusters. These are available as context during clustering but do not map to speech frames.
4. When enough data is available, use allophone clusters with sufficiently large contexts. This will lead to context dependent models for the deleted phones that really cover the frames at the end and beginning of the adjacent phones.

1.3. Phoneme Shifts

In addition to the deletions described above we observed a number of phoneme shifts due to spontaneous speech and regional accents.

Canonic	Observed
/l/	/r/
/r/	/l/
long vowel	short vowel
short vowel	long vowel
/kw/	/f/
/khw/	/f/

Table 2: Examples for phoneme shifts

These substitutions occur much more frequently in certain phonetic contexts. We therefore expect the polyphone clustering to create separate models where necessary. To model shifts like /l/ ↔ /r/ more efficiently, we also tried building a modified allophone cluster tree that can share models between phonemes as described in [8].

2. Databases

We are using two new databases in addition to the GlobalPhone and Babylon databases used for prior research.

Domain	Training Hours	Testset Perplexity
AMI CallCenter Agent	19	64
AMI CallCenter Client	220	17
Broadcast News	12	212
GlobalPhone	20	169
Babylon	4	121

Table 3: Database Overview. Perplexity without transitions to/from out of vocabulary words.

2.1. Ami CallCenter

This data was recorded using close-speaking microphones on the agent-side of a live call center in Thailand. There are two kind of calls: client initiated calls (77 female speakers) and agent initiated calls (25 male speakers). The client initiated calls are more frequent and much simpler in structure. All data is manually transcribed. To reduce training time and limit the impact on the performance for other domains, we only used 25 speakers (~70 hours) of the client-initiated calls to train the acoustic models for all multi-domain recognizers in this paper. For training language models, single-domain, and domain-adapted recognizers, all available data was used.

2.2. Broadcast News

Our Broadcast News database was collected to obtain better polyphone coverage for spontaneous speech before the Ami CallCenter data was available. It consists of digitally compressed audio of 30 half-hour news shows that have been manually transcribed and cut into segments of 30-40 seconds. Each segment typically contains speech from a single speaker, but speaker changes are not labeled. A small number of anchor speakers dominate the database. High compression loss limits the usability of this data for the development of clean speech recognizers. This data was used for training multi-domain recognizers only.

2.3. GlobalPhone

The Thai GlobalPhone database is part of the multilingual GlobalPhone project. It contains read newspaper articles, recorded with a close-talk microphone in a push-to-talk scenario. The 20 hours of Thai data were collected from 90 students (59 female, 31 male) in Bangkok. Each speaker read on average 160 sentences. The average utterance length is around 5 seconds.

2.4. Babylon

This database consists of utterances from 'medical dialogs'. It was collected to improve the recognition performance for the Babylon speech-to-speech translation project. The language model data is created by rephrasing utterances from dialogs that are made-up, translated from English, or recorded in simulated 'medical interview' environments. The audio training data is generated by reading parts of the language model data.

3. Baseline System

All experiments reported in this paper were performed using the M*Modal Recognition Toolkit [9]. The initial version of the recognizer was trained on phoneme labels on GlobalPhone data provided by Carnegie Mellon University.

3.1. Segmentation and Language Model

Unlike English, the Thai language is written without spaces. For speech recognition, we need to introduce word-like units that:

- occur frequently enough for statistical language modeling
- provide sufficient discrimination between heterographs
- minimize the number of homophones
- minimize cross-word co-articulation

For the experiments in this paper we segmented the training text first using a segmenter from a previous project [10] and built a statistical language model on the output. For the final segmentation, we used this language model to find the segmentation that maximizes the probability of the segmented text given the model.

Domain dependent statistical language models (trigrams with Kneser-Ney backoffs) are computed on the final segmentation output and used for recognition. The domains are sufficiently dissimilar that interpolation with data from other domains did not improve test set perplexity or recognition performance.

3.2. Pronunciation Dictionary

We decided to generate Thai dictionaries using a bootstrap procedure: starting with a seed dictionary, we trained a statistical grapheme to phoneme tool based on a decision tree to predict the phonemes generated by graphemes in a particular context. This tool is used to generate more dictionary entries, which are then hand-corrected and added to the training set until the output of the statistical model is considered satisfactory to generate pronunciations for the remaining low-frequency words.

3.3. Acoustic and Phoneme Models

All experiments reported in this paper used the same MFCC front-end with a frame rate of 100 frames per second. We use a 3-state-skip topology for all phones, with a minimum phone duration of 2 frames. Zero-length phonemes were used for some experiments where noted. Triphones are clustered to 1,500, 3,000 and 6,000 context-dependent cross-word allophones as noted in the experiments. 'Pronunciations' in our system are a sequences of units that encode attributes such as phoneme identity, voicing, tone, position, domain, etc. These attributes are then used to select the topologies and mixtures for building the word-level HMMs.

For the experiments in this paper no tone models were used since earlier experiments [2,3] indicated that they do not improve ASR performance.

4. Experiments

4.1. Phoneme Shift / Elisions

4.1.1. Short Cluster Pronunciations

To better model consonant clusters in which the /l/ or /r/ are dropped, we first investigated introducing pronunciation variants without these phonemes during training and testing. These variants are preferred by the training alignment in a significant number of cases as seen in Table 4. A smaller number of short pronunciations indicates a lower speaking rate and a more carefully enunciated speech.

Domain	per cluster words	per total words
AMI CallCenter	33.6%	2.0%
GlobalPhone	30.4%	2.8%
Broadcast News	18.1%	1.5%
Babylon	14.4%	1.0%

Table 4: Percent of words using short pronunciation per number of words with consonant clusters and per total words.

We used the histogram of phone durations as a diagnostic tool to find problems with the modeling of individual phones: if a large number of occurrences are found for the minimum length enforced by the HMM topology, there usually is a problem in modeling this phone. If the length distribution is bi-modal, the phoneme might be used differently in different contexts. If we can find a rule to separate these (for instance different lengths of /l/ in clusters), it is often better to use two separate phones in the pronunciation dictionary.

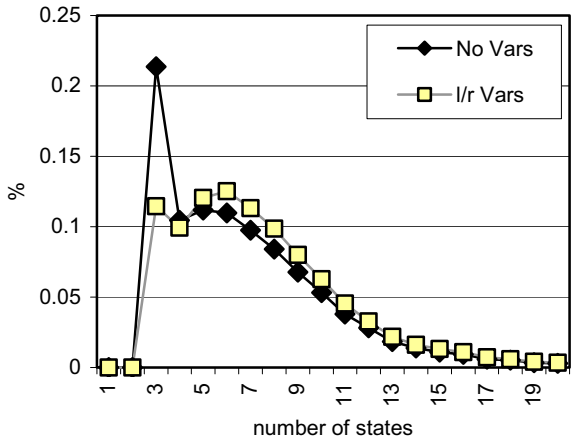

Figure 1: Duration distribution of phoneme /r/ with and without alternate pronunciations for phoneme clusters

Figure 1 shows that when aligning the training data with these variants, the number of occurrences of minimum-length /l/ and /r/ drops significantly. The minimum-length /l/ /r/ durations are therefore mostly due to words with phoneme clusters covered by the alternate pronunciations.

Domain	no variants	variants
CallCenter Agent	**57.52**	56.67
CallCenter Client	**78.68**	78.65
GlobalPhone	**78.92**	78.82
Babylon	**79.04**	78.75

Table 5: Word Accuracy using shorter variants without /l/ /r/ in phoneme clusters during training and testing

However, as can be seen in Table 5, training using pronunciations variants without /r/ /l/ for phoneme clusters slightly reduces the word accuracy in all cases. A closer investigation of the added errors shows that they are mostly due to contamination of neighboring phoneme models during training, and the introduction of homophones during testing.

4.1.2. Zero-length Phones

To avoid issues with phoneme contamination and homophones introduced by deleted phones, we introduced zero-length phones that affect the context in allophone clustering but do not map to any frames during training or recognition.

Domain	no variants	variants with zero-length phones
CallCenter Agent	**57.52**	57.29
CallCenter Client	78.68	**78.72**
GlobalPhone	**78.92**	78.54
Babylon	79.04	**79.49**

Table 6: Word Accuracy using zero-length phones

This resolves some problems with short pronunciations, but as shown in Table 6, it does not lead to a significant performance improvement over the baseline across all domains.

4.1.3. Model Sharing across Phones

Our baseline system grows a separate allophone cluster tree for every phone state. If there is a significant deviation between the normal phoneme and its realization in spontaneous speech, it might be more efficient to share data across phones [8]. Furthermore, if there is insufficient training data, we may be able to improve recognition performance by sharing data between different states of the same phone for some contexts.

Domain	no sharing	sharing
CallCenter Agent	57.52	**58.19**
CallCenter Client	78.68	**79.64**
GlobalPhone	**78.92**	77.97
Babylon	**79.04**	78.01

Table 7: Word Accuracy with/without sharing models across phonemes.

Table 7 shows that this approach does not work equally well for all domains in a multi-domain Thai system. We believe that most of the previously reported improvements from this technique are due to more efficient use of sparse data.

4.2. Domain Specific Acoustic Models

There are different ways to share models between domain dependent recognizers:

- Train completely separate models for each domain.
- Train a shared acoustic model over all domains. This is a robust baseline system for new domains with very little data (< 5 hours). If the acoustic environment differs significantly between domains, or if the pronunciation changes are more frequent in some (more spontaneous) domains, the resulting cross-domain models will have high variances.
- Introduce a domain attribute to decide which models to share across domains during allophone clustering.
- Train a shared model and adapt each allophone model to every domain using MLLR/MAP/ML (depending on the amount of available data per model).

The baseline acoustic model in this experiment has 1,500 allophone models. Taken together, the separate systems have 4 times as many parameters, and the adapted systems have almost that many. For the system with the domain attribute, we trained systems with 1,500, 3,000 and 6,000 allophones to illustrate the impact of the number of parameters on the word accuracy.

Domain	shared 1500	domain adapted 6000	separate 6000
CallCenter Agent	57.52	61.11	**61.69**
CallCenter Client	78.68	**80.97**	80.59
GlobalPhone	78.92	81.13	**81.35**
Babylon	**79.04**	78.35	75.85
Domain	domain attribute 1500	domain attribute 3000	domain attribute 6000
CallCenter Agent	58.09	59.87	61.04
CallCenter Client	79.37	79.67	80.63
GlobalPhone	79.74	80.30	80.87
Babylon	78.07	77.67	76.08

Table 8: Word Accuracy for different ways of sharing models across domains and different numbers of allophones.

5. Conclusions

Using pronunciation variations to model elisions in consonant clusters introduces homophones and fails to discriminate between acoustically different words. Systems that use zero-length phonemes to identify clusters perform better, but given enough training data do not consistently outperform systems that simply use wider contexts.

If the amount of training data for a domain exceeds 20 hours, an individually trained recognizer outperforms a recognizer with a joint acoustic model trained across multiple domains and little additional benefit is obtained from using cross-domain data for acoustic training. For domains that are close to each other with respect to acoustic environment, speaker characteristics, and polyphone coverage, as well as for domains with only small amounts of training data, a system trained on multiple domains provides best performance.

While we did not report numbers on combining the various techniques presented in this paper, exploratory experiments confirmed our expectation that the combination of algorithms designed to address spontaneous speech effects does not yield better performance.

6. Acknowledgements

The authors would like to thank Advanced Media, Inc (Japan) for funding for this research and for providing the Call Center database. Multimodal Technologies kindly allowed us to spend resources on research that went far beyond current business needs. This work builds in part on earlier research funded by grants N66001-00-C-8007 and NBCH030036 under the DARPA Babylon (CAST) and LASER-ACTD programs. The opinions expressed in this publication are the opinions of the authors and may not reflect those of the sponsors.

7. References

[1] Tanja Schultz, Alan W Black, Stephan Vogel and Monika Woszczcyna. *Flexible Speech Translation Systems*. IEEE Transactions on Audio, Speech, and Language Processing, Vol 14(2), March 2006.

[2] Sinaporn Suebvisai, Paisarn Charoenpornsawat, Alan Black, Monika Woszczyna and Tanja Schultz. *Thai Automatic Speech Recognition*. ICASSP 2005, Philadelphia, USA.

[3] Sawit Kasuriya, Supphanat Kanokphara, Nattanun Thatphithakkul, Patcharika Cotsomrong and Treepop Sunpethiniyom. *Context-independent Acoustic Models for Thai Speech Recognition*, ISCIT2004, Sapporo, Japan

[4] Tanja Schultz, Dorcas Alexander, Alan W Black, Kay Peterson, Sinaporn Suebvisai, and Alex Waibel. *A Thai Speech Translation System For Medical Dialogs.* HLT 2004, Boston, USA.

[5] Supphanat Kanokphara, Virongrong Tesprasit and Rachod Thongprasirt. *Pronunciation Variation Speech Recognition without Dictionary Modification on Sparse Database.* ICASSP 2003, Hong Kong, China.

[6] Supphanat Kanokphara. *Syllable Structure Based Phonetic Units for Context-Dependent Continuous Thai Speech Recognition*. Eurospeech 2003, Geneva, Switzerland.

[7] Paisarn Charoenpornsawat, Sanjika Hewaviharana and Tanja Schultz. *Thai Grapheme-Based Speech Recognition.* HLT-NAACL 2006, New York, USA.

[8] Hua Yu and Tanja Schultz. *Enhanced Tree Clustering with Single Pronunciation dictionary for Conversational Speech Recognition.* Eurospeech 2003, Geneva, Switzerland.

[9] Michael Finke, Jürgen Fritsch, Detlef Koll and Alex Waibel. *Modeling and Efficient Decoding of Large Vocabulary Conversational Speech.* Eurospeech 1999, Budapest, Hungary.

[10] Paisarn Charoenpornsawat, (2003) *SWATH: Thai Word Segmentation Program.* http://www.cs.cmu.edu/~paisarn/software.htm