A THAI SPEECH TRANSLATION SYSTEM FOR MEDICAL DIALOGS

Tanja Schultz, Dorcas Alexander, Alan W Black, Kay Peterson, Sinaporn Suebvisai, Alex Waibel
Language Technologies Institute, Carnegie Mellon University
E-mail: tanja@cs.cmu.edu

1. Introduction

In this paper we present our activities towards a Thai Speech-to-Speech translation system. We investigated in the design and implementation of a prototype system. For this purpose we carried out research on bootstrapping a Thai speech recognition system, developing a translation component, and building an initial Thai synthesis system using our existing tools.

2. Speech Recognition

The language adaptation techniques developed in our lab [5] enables us to rapidly bootstrap a speech recognition system in a new target language given very limited amount of training data. The Thailand's National Electronics and Technology Center gave us the permission to use their Thai speech data collected in the hotel reservation domain. They provided us with a 6 hours text and speech database recorded from native Thai speakers. We divided the data into three speaker disjoint sets, 34 speakers were used for training, 4 speakers for development, and another 4 speakers for evaluation. The provided transcriptions were manually pre-segmented and given in Thai script. We transformed the Thai script into a Roman script representation by concatenating the phoneme representation of the Thai word given in the pronunciation dictionary. The motivation for this romanization step was threefold: (1) it makes it easier for non-Thai researchers to work with the Roman representation like in the grammar development, (2) the romanized output basically provides the pronunciation which makes things easier for the speech synthesis component, and (3) our speech engine currently does not handle Thai characters.

In our first Thai speech engine we decided to disregard the tone information. Since tone is a distinctive feature in the Thai language, disregarding the tone increases the number of homographs. In order to limit this number, we distinguished those word candidates by adding a tag that represents the tone. The resulting dictionary consists of 734 words which cover the given 6-hours database.

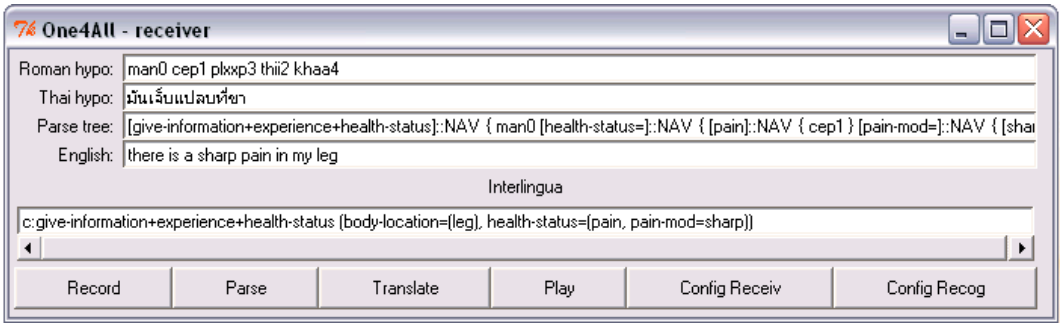
Building on our earlier studies which showed that multilingual seed models outperform monolingual ones [5], we applied phonemes taken from seven languages, namely Chinese, Croatian, French, German, Japanese, Spanish, and Turkish as seed models for the Thai phone set. Table 1 describes the performance of the Thai speech recognition component for different acoustic model sizes (context-independent vs. 500 and 1000 tri-phone models). The results indicate that a Thai speech recognition engine can be built by using the bootstrapping approach with a reasonable amount of speech data. Even the very initial system bootstrapped from multilingual seed models gives a performance above 80% word accuracy. The good performance might be an artifact from the very limited domain with a compact and closed vocabulary.

System	Dev Test	Eval Test
Context-Independent	85.62%	83.63%
Context-Dependent (500)	86.99%	84.44%
Context-Dependent (1000)	84.63%	82.71%

Table1: Word accuracy [%] in Thai language

3. Machine Translation

The Machine Translation (MT) component of our current Thai system is based on an interlingua called the Interchange Format (IF). The IF developed by CMU has been expanded and now encompasses concepts in both the travel and medical domains, as well as many general-use or cross-domain concepts in many different languages [4]. Interlingua-based MT has several advantages, namely: (1) it abstracts away from variations in syntax across languages, providing potentially deep analysis of meaning without relying on information pertinent only to one particular language pair, (2) modules for analysis and generation can be developed monolingually, with additional reference only to the second "language" of the interlingua, (3) the speaker can be given a paraphrase in his or her own language, which can help verify the accuracy of the analysis and be used to alert the listener to inaccurate translations, and (4) translation systems can be extended to new languages simply by hooking up new monolingual modules for analysis and/or generation, eliminating the need to develop a completely new system for each new language pair.
Thai has some particular characteristics which we addressed in IF and appear in the grammars as follows:



1) The use of a term to indicate the gender of the person:

```
Thai: zookhee kha1
Eng:  okay (ending)
s[acknowledge] (zookhee *[speaker=])
```

2) An affirmation that means more than simply "yes."

```
Thai: saap khrap
Eng:  know (ending)
s[affirm+knowledge](saap *[speaker=])
```

3) The separation from the main verb of terms for feasibility and other modalities.

```
Thai: rvv khun ca paj dooj thxksii
kyydaaj
Eng:  or you will go by taxi [can too]
s[give-information+feasibility+trip]
(*DISC-RHET [who=] ca paj
[locomotion=] [feasibility=])
```

4. Language Generation

For natural language generation from interlingua for Thai and English, we are currently investigating two options: a knowledge-based generation with the pseudo-unification based GenKit generator developed at CMU, which employs manually written semantic/syntactic grammars and lexicons, and a statistical generation operating on a training corpus of aligned interlingua and natural language correspondences. Performance tests as well as the amount and quality of training data will decide which approach will be pursued in the future.

5. Speech Synthesis

First, we built a limited domain Thai voice in the Festival Speech Synthesis System [1]. Limited Domain voices can achieve very high quality voice output [2], and can be easy to construct if the domain is constrained. Our initial voice targeted the Hotel Reservation domain and we constructed 235 sentence that covered the aspects of our immediate interest. Using the tools provided in FestVox [1], we recorded, auto-labeled, and built a synthetic voice.

In supporting any new language in synthesis, a number of language specific issues first had to be addressed. As with our other speech-to-speech translation projects we share the phoneme set between the recognizer and the synthesizer. The second important component is the lexicon. The pronunciation of Thai words from Thai script is not straightforward, but there is a stronger relationship between the orthography and pronunciation than in English. For this small set of initial words we constructed an explicit lexicon by hand with the output vocabulary of 522 words. The complete Thai limited domain voice uses unit selection concatenative synthesis. Unlike our other limited domain synthesizers, where they have a limited vocabulary, we tag each phone with syllable and tone information in selection making the result more fluent, and a little more general.

Building on our previous Thai work in pronunciation of Thai words [3], we have used the lexicon and statistically trained letter to sound rules to bootstrap the required word coverage. With a pronunciation model we can select suitable phonetically balanced text (both general and in-domain) from which we are able to record and build a more general voice.

6. Demonstration Prototype System

Our current version is a two-way speech-to-speech translation system between Thai and English for dialogs in the medical domain where the English speaker is a doctor and the Thai speaker is a patient. The translated speech input will be spoken using the built voice. At the moment, the coverage is very limited due to the simplicity of the used grammars. The figure shows the interface of our prototype system.

Acknowledgements

This work was partly funded by LASER-ACTD. The authors thank Thailand's National Electronics and Computer Technology Center for giving the permission to use their database and dictionary for this task.

References

[1] Black, A. and Lenzo, K. (2000) *"Building Voices in the Festival Speech Synthesis System"*, http://festvox.org

[2] Black, A. and Lenzo, K. (2000) *"Limited Domain Synthesis"*, ICSLP2000, Beijing, China.

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

[4] Lavie A. and Levin L. and Schultz T. and Langley C. and Han B., Tribble, A., Gates D., Wallace D. and Peterson K. (2001) *"Domain Portability in Speech-to-speech Translation"*, HLT, San Diego, March 2001.

[5] 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.