Speech Technology and Corpus Development in Thailand

Virach Sornlertlamvanich and Rachod Thongprasirt

Information Research and Development Division
National Electronics and Computer Technology Center
virach@nectec.or.th and rachod@nectec.or.th

ABSTRACT

This paper describes some recent activities on speech technology and corpus development in Thailand. Many speech corpus projects have been launched this year. The National Electronics and Computer Technology Center (NECTEC) recently provides a grant for two cooperative speech corpus projects to interested universities. The first project aims at developing a Thai speech corpus for the research on speaker-independent, large vocabulary, continuous speech recognition. The second project is to collect continuous speech of telephone number recorded through telephone line. Except for these projects, some NECTEC's internal corpus projects such as the corpus for speech synthesis development and the speech corpus for ATR Spoken Language Translation Research Laboratories are also under development. Besides the corpus development projects, many speech-technology researches are also discussed.

1. INTRODUCTION

This is an exciting time for speech technology in Thailand. Many speech corpora have been initiated. These corpora would help Thai researchers for advancing the speech-related technology in the future. Moreover, some private companies are interested in employing speech synthesizer to their systems.

This paper is divided into two sections. The first section describes recent activities in speech corpora, which includes the speech corpus for speech recognition, prosody tagging corpus, speech corpus for speaker identification (SID). The second section describes the speech technologies in Thailand over this past year.

2. SPEECH CORPUS

Probably the most prominent problem in Thai speech research is the lack of large speech corpus. Therefore, most speech-related researches in Thailand cannot develop into a full-fledged prototype or a commercial product. Fortunately, this year many speech corpus projects have been initiated. The following is the description of speech corpus projects by dividing them into three groups: the speech corpus for speech recognition, the speech corpus for speech synthesis and the speech corpus for speaker identification (SID).

2.1 SPEECH RECOGNITION CORPUS

Four speech corpora for speech recognition have been launched this year. Three of the corpus projects are initiated at NECTEC, one of which is subsidized by ATR in Japan, and the other two are granted to interested universities. Apart from NECTEC, Chulalongkorn's corpus project is also included. The following is details of these corpora.

2.1.1 NECTEC'S SPEECH CORPUS FOR SPEECH RECOGNITION

This project aims at collecting speech corpus for speaker independent, large vocabulary, continuous speech recognition. It is the cooperation project between NECTEC and universities. NECTEC provides funding and selects appropriate texts, and universities are going to record and label phonetic timing. Prince of Songkha and Mahanakorn University of Technology universities have been interested in joining this project so far. The project initially aims at recording at least 200 speakers with 100 males and 100 females. The sentences, extracted from articles, are divided into four sets: the phonetically balanced set (PB), the language-model training set (TR), the language-model development test set (DT) and the language-model evaluation test set (ET).

The PB set is designed not only to cover all possible combinations of Biphones but also to possess the same distribution as the text. The details of the construction the PB set can be found in [1], This PB set is used to train the acoustic model. The TR, DT and ET sets are designed to cover at least 5,000 words. The TR set is used to train additional acoustic and language models, and it has to contain at least 1,000 sentences. The DT set is used to test the speech recognition system at development time, and it has to contain at least 500 sentences. The ET set is used to evaluate the speech recognition system, and it must contain at least 500 sentences. All sentences in the TR, DT and ET must

have medium length and medium perplexity, which can be calculated from Bigram language model [2].

Two recording environments are needed: the quasi-quiet and the office rooms. A high-quality dynamic close-talk microphone is used in the quasi-quiet room and a medium-quality dynamic close-talk microphone and a medium-quality dynamic unidirectional microphone are used for recording in the office environments. The speech is first recorded on DAT and then transferred to PC at 16kHz with 16-bit linear quantization. These waveforms will be compressed with NIST Sphere with "shorten" compression type similar to that in TIMIT [3].

Each sentence in the PB, TR, DT and ET sets will have two transcription files: the word and the phonetic transcriptions. The transcription files for the PB set also contain the label of the starting time and the ending time of the words and phonemes, while for the other sets this timing label is omitted.

2.1.2 NECTEC'S SPEECH CORPUS FOR ATR

The ATR Spoken Language Translation Research Laboratories has subsidized NECTEC to create a Thai speech corpus to incorporate into its engine. The speech corpus is divided into three sets: the isolated-word speech database, the phonetically balanced sentence database and the script-scheduled speech database for hotel reservation task (HRT). The number of speakers is 40 (20 males and 20 females).

The isolated-word speech database contains 5,000 daily-used words, the phonetically balanced words, and Thai digits and some extra words such as credit-card type, money units, etc. Words in extra-word and phonetically balanced word sets are uttered by all speakers. The 5,000 daily-used word set is divided into five groups of 1,000 words. Forty speakers are divided into five groups of eight speakers, and each group is assigned to speak only one 1000-word group. The phonetically balanced sentences, spoken by all 40 speakers, are similar to the PB set described in the previous section. The script schedule task contains 50 conversations between a hotel clerk and a customer. Each conversation has been translated from English/Japanese text into Thai text. In this HRT set, each person are selected randomly to speak only five conversations.

2.1.3 NECTEC'S SPEECH CORPUS FOR TELEPHONE NUMBER RECOGNITION

This is another project that NECTEC grants funding to universities. The project aims at collecting Thai speech of telephone digit speaking over telephone channels. This project requires a large number of speakers (initially aims at 2,000 speakers). Each speaker has to speak 20 sets of telephone numbers designed to cover all possible combinations of digit pairs. Kasetsart university is probably going to join this project.

2.1.4 CHULALONGKORN'S SPEECH CORPUS

Besides NECTEC, Chulalongkorn University is also preparing for building speech corpus. Their initial aim is to cover about 1,000–2,000 words, and they have used tales as reference texts. Currently, they have started recording of a few people.

2.2 PROSODY CORPUS FOR SPEECH SYNTHESIS

This internal corpus project aims at tagging prosody information using extensible markup language (XML) for improving the quality of NECTEC's speech synthesizer. Sentences are selected from the Thai part-of-speech corpus, ORCHID [4]. The portion of selected sentences provides complete tri-phone and tri-tone combinations. Furthermore, these sentences must have at least seven syllables to avoid any short phrase. There are three main levels of tagging: sentence, word, and syllable. The prosody information is then tagged for each syllable. The prosody information is, for example, syllable duration, energy, pitch contour's parameters for each syllable is tagged after each syllable.

2.3 SPEECH CORPUS FOR SID

NECTEC already has a speech database of Thai isolated digits (0–9) via telephone for SID task. The database was collected from 50 speakers (25 males and 25 females). Each person was given a script containing 10 set of 10-isolated digit. Each 10-isolated digit set contains the digits from 0 to 9, but their positions in the set were randomized. In each session, each person calls to a voice modem acted as an answering machine and then uttered the given numeral scripts to be recorded by the voice modem. Each person had called and recorded his or her voice once a week for five consecutive weeks. There are two calling conditions: from the internal line and from the external line; the external call is a called a call outside office which has to be passed through PABX, while the internal call is the call inside office, which does not pass through any PABX.

3. RECENT SPEECH TECHNOLOGY IN THAILAND

Some of the speech technologies in Thailand have become mature enough to be commercialized especially speech synthesizers. For example, many companies are interested in incorporating speech synthesis into the their unified messaging systems (UMS). The following describes some recent speech-technology advancements occurring this year.

Like Chinese, Thai is a tonal language. Five tones are presented in Thai, which are the mid, the low, the falling, the high and the rising. Fig. 1 shows female's average F0 contour of five Thai tones produced in isolated.

Even though there are numerous researches on tone[5,6], tone research is still ongoing. Two new published works regarding tone will be described here. The first one is the studied of tone recognition of monosyllable [7]. In this work, the system is divided into three different modules: the F0 extraction, the F0 feature extraction, and the tone recognition. The F0 extraction draws F0 from monosyllable speech using modified short-term auto-correlation with center clipping method. The F0 feature extraction then determines the fitted parameters from the extracted F0. The fitting process is done by the polynomial regression function. The maximum a posteriori probability (MAP) is employed in the recognition process. The system was tested with 30 words containing all possible tones. The test was divided into two categories: the speaker-dependent and speaker independent. The best average performance F0 for the speaker-dependent test is 96.20%, and that of the speaker-independent test is 82.80%.

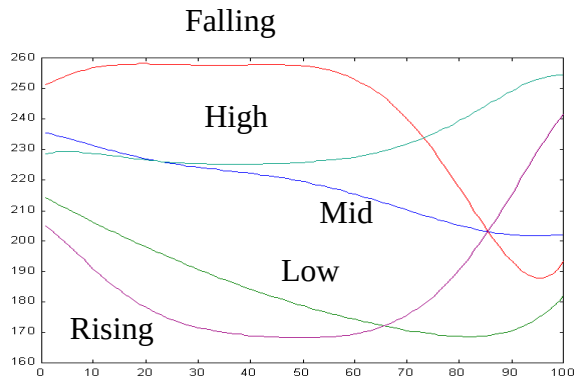

Figure 1: Female's average F0 contour of five Thai tones. Tone types are indicated.

The tone detection has also been added into recognition engines to improve recognition rate. In [8], the comparison of three different methods of adding tone detection into the hidden Markov model (HMM) was described. These methods are the joint detection method, the sequential detection method, and the linked detection method. In the joint detection method, the recognition is performed on the connected tonal syllables. In the sequential detection method, the base syllables (ignoring tone) are recognized first by using a HMM of connected syllables. The resulting syllable boundaries are then sent to the tone recognition. In the linked detection, the recognition of the base syllables and tones is done simultaneously. These systems were tested on 18 Thai sentences designed to be phonetically and tonally balanced and their distributions are similar to those in Thai language. The speeches have been recorded and then filtered to simulate telephone channels. It was found that the sequential detection and linked detection are comparable in performance, and they are superior to the sequential detection, but the linked detection requires less computational time than the sequential detection.

Various approaches have been proposed for Thai Grapheme-to-Phoneme (G2P) such as the dictionary-based, rule-based and statistical based approaches [9–12]. However, those techniques have some drawbacks such as the dictionary-based techniques [11] requires a large dictionary and cannot deal with an unknown word. Therefore, a new Thai G2P based on the probabilistic GLR has been proposed [13],[14]. This technique achieves 90.4% of word accuracy in ignoring vowel's length case and 72.87% of word accuracy in exact match case. This G2P is also adapted to use in new Thai soundex system. This system works by first changed the grapheme input into its corresponding phoneme by the G2P module. The resulting phoneme is then slightly modified into a soundex code. Because this soundex code is the authentic representation of an input phoneme, it is very efficient in grouping words of similar sound and separating words of different sound. This system had been tested on a name-searching task, and it was found that the system yields high precision and recall rates.

Kasuriya et al. [15] used the speech corpus for SID previously described to compare SID system based on dynamic time warping (DTW) and that based on multi-layer perceptron (MLP). First, they determined the identification rate for each digit by both methods, and they found that for all digits the DTW system performs better than the MLP system. Furthermore, they selected three most accurate digits for the DTW (5,2 and 0) and MLP (0, 2 and 1), and then concatenated them for testing the systems. It was again found that DTW system, which yields 97.8% identification rate, is better than the MLP system, which yields 96.30% identification rate.

Apart from the SID system, speaker verification systems based on discrete time warping (DTW) and Gaussian Mixture model (GMM) have been developed using parts of the SID engine [16]. Many scoring normalizations have been applied to those systems to determine system performance. These scoring normalizations are conventional scoring, cohort scoring, and Global speaker model (GSM). Moreover, a new scoring normalizations called Global anti speaker model (GASM) has been proposed and incorporated into those systems. The systems have been tested with Thai speech of the digit number "5", and it was found that systems based on GSM and GASM perform better than system based on cohort.

4. CONCLUSIONS

In this paper, we have discussed various speech corpora and some of the recent speech technology in Thailand. It is hope that when these corpora are available, speech technologies in Thailand would be dramatically growing. A growing interest in adopting speech synthesis program into commercial system would also persuade many people to speech technology fields.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Somchai Jitipunkul and his colleagues at Chulalongkorn University for information regarding speech technology and corpus development at Chulalongkorn univiersity.

REFERENCES

[1] J. L. Shen, H. M. Wang, R. Y. Lyu, and L.S. Lee, "Automatic Selection of Phonetically Distributed Sentence Sets for Speaker Adaptation with Application to Large Vocabulary Mandarin Speech Recognition," *Journal of Computer Speech and Language*, vol. 13, pp. 79–98, 1999.

[2] R. Rosenfield, "The CMU Statistical Language Modeling Toolkit and its use in the 1994 ARPA CSR Evalutaion," Carnegie Mellon University, 1994.

[3] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," NIST, 1993.

[4] V. Sornlertlamvanich, N. Takahashi, and H. Isahara, "Thai Part-of-speech Tagged Corpus: ORCHID," *Proceedings of first intenational workshop on east-asian language resources and evaluation*, pp. 131–138, 1998.

[5] Potisuk S., Harper M.P., and Gandour J., "Classification of Thai Tone Sequences in Syllable-Segmented Speech using the Analysis-by-Synthesis Method," *IEEE Trans. on Speech and Audio Processing*, vol. 7 pp. 95-102, Jan. 1999.

[6] Thubthong, N. "A Thai Tone Recognition System based on Phonemic Distinctive Features," Depeartment of Computer Engineering Faculty of Engineering, Chulalongkorn University.

[7] P. Charnvivit, S. Jitapunkul, V. Ahkuputra, E. Maneenoi, U. Thathong, and B. Thampanitchawong, "F0 Feature Extraction by Polynomial Regression Function for Mononsyllabic Thai Tone Recognition," to be published in Eurospeech 2001.

[8] T. Demeechai and Makelainen, "Recognition of Syllables in a Tone Language," *Speech Communication*, vol. 33, 2001, pp. 241–254.

[9] Chotimongkol, A. "Statistically Trained Orthographic to Sound Models for Thai," *Proceeding of the 6th International Conference on Spoken Language Processing*, pp. 533–554, Oct. 2000.

[10] Khamya, A. "SATTS: Syllable Analysis for Text-to-Speech System," *Proceeding of the Fourth Symposium on Natural Language Processing 2000 (SNLP'2000)*, pp. 336–340, May 2000.

[11] Luksaneeyanawin, S. "A Thai Text to Speech System," *Proceeding of the Conference of the Regional Workshops on Computer Processing of Asian Languages. Asian Institute of Technology*, pp. 305–315, 1990.

[12] Luksaneeyanawin, S. "Speech Computing and Technology in Thailand," *NLP in Thailand*, pp. 276–321, 1993.

[13] P. Tarsaku, V. Sornlertlamvanich, and R. Thongprasirt, "Thai Grapheme-to-Phoneme using Probabilistic GLR Approach," to be published in Eurospeech 2001.

[14] P. Tarsaku, V. Sornlertlamvanich, and R. Thongprasirt, "Grapheme-to-Phoneme for Thai," submitted to NLPRS 2001.

[15] S. Kasuriya, V. Achariyakulporn, C. Wutiwiwatchai, and C. Tanprasert, "Text-Dependent Speaker Identification via Telephone based on DTW and MLP," *Proceeding of the IASTED International Conference: Modelling, Identification, and Control*, 2001, pp. 190–194.

[16] C. Wutiwiwatchai, V. Achariyakulporn, and S. Kasuriya, "Improvement of Speaker Verification for Thai Language," to be published in Eurospeech 2001.