INTERSPEECH 2013

A Hybrid Language Model for Open-Vocabulary Thai LVCSR

Kwanchiva Thangthai, Ananlada Chotimongkol, Chai Wutiwiwatchai

National Electronics and Computer Technology Center, Thailand

{kwanchiva.thangthai, ananlada.chotimongkol, chai.wutiwiwatchai}@nectec.or.th

10.21437/Interspeech.2013-520

Abstract

This paper investigates the use of a hybrid language model for open-vocabulary Thai LVCSR. Thai text is written without word boundary markers and the definition of word unit is often ambiguous due to the presence of compound words. Hence, to build open-vocabulary LVCSR, a very large lexicon is required to also handle word unit ambiguity. Pseudo-morpheme (PM), a syllable-like sub-word unit specifically designed for Thai is considered to be a more well-defined unit. To overcome the problem of out-of-vocabulary words and to also reduce the size of the lexicon, a hybrid language model which combines word and sub-word units is proposed. Words and sub-words frequently found in several domains constitute open-vocabulary for general domain Thai LVCSR. To verify our scheme, we run recognition experiments on data from various tasks including broadcast news transcription, dictation and mobile speech-to-speech translation. Open-vocabulary Thai LVCSR using the hybrid language model obviously reduces the out-of-vocabulary problem. The proposed model having a much smaller lexicon size achieves a comparable recognition error rate to a baseline system using a full-word lexicon.

Index Terms: pseudo-morpheme, sub-lexical unit, open vocabulary LVCSR, hybrid language model

1. Introduction

One challenge in developing an open domain large vocabulary continuous speech recognition (LVCSR) system is lexical modeling, as the system has to handle unrestricted or open vocabulary. For any active language, new words, such as person names, place names and new technical terms, emerge every day. Even a very large text corpus would not be possible to cover all the lexicon in the language. Therefore, out-of-vocabulary (OOV) words are very likely to occur in speech recognition applications which operate in an open domain, for example, dictation, speech-to-speech translation and closed captioning. OOV is known to be a problem for LVCSR systems if there is no special treatment for OOV words. One OOV words usually translates into one recognition error or even more from erroneous context.

Sub-lexical unit is one commonly used technique when modeling OOV words in many languages [1, 2, 3, 4] as multiple sub-lexical units can be combined to form a new word which is not seen before in the training data. The design of a sub-lexical unit is crucial and depends on the characteristic of each language. *Morpheme*, a smallest meaningful unit in a language, becomes a natural choice for sub-word unit especially for highly inflected languages. In Thai, there is neither inflection nor derivative. Thus, another type of sub-word unit has to be considered. A letter in Thai writing system is a phonogram which roughly represents a phoneme or combination of phonemes. Therefore, each Thai word could be segmented into a set of syllable-like units. In this paper, we propose *pseudo-morpheme* (PM), a syllable-like unit in a written form, as a sub-lexical unit for Thai. According to Thai writing rules, PM is more deterministic when compared with word and has been shown to help mitigate the word segmentation problem [5]. Given a word or a string of text, PMs can be determined quite accurately with an automatic syllable segmentation tool. Therefore, PM is a good candidate for a sub-lexical unit for Thai LVCSR.

Nevertheless, in recognition, small units usually suffer from acoustic confusability and also cover shorter span of context in an n-gram language model. To resolve these problems, a hybrid language model which combines both unit types, PM and word, is proposed. A lexicon with mixed types of units has been used successfully, for example, in English and German LVCSR systems [6, 7]. The lexicon of our hybrid language model consists of frequent full-words and PMs from less frequent words. To verify our scheme, we run recognition experiments on data from various tasks including broadcast news transcription, dictation and mobile speech-to-speech translation. Several hybrid lexicons and language models were tested to investigate the effect of mixed unit types on the performance of LVCSR system. The reduction in OOV rate would verify whether the proposed hybrid lexicon of PMs and words could constitute an open vocabulary for Thai LVCSR while reduction in recognition error rate would be an ultimate goal for any LVCSR system.

This paper is organized as follows: section 2 describes the characteristics of Thai text. Section 3 explains pseudo-morpheme, a unit chosen as a sub-lexical unit for Thai. The proposed hybrid language model that combines both words and pseudo-morphemes is discussed in Section 4 while the recognition results obtained from the proposed models are presented and discussed in Section 5. We finally conclude our work and discuss future directions in Section 6.

2. Characteristics of Thai text

Thai is a non-segmented script language, i.e. there is no boundary marker between words while boundary markers on phrase and sentence levels are ambiguous. Furthermore, there is no capital letter to indicate the beginning of a new sentence or a proper noun. The definition of word unit is often ambiguous due to the presence of compound words. These characteristics become a challenge when processing Thai text.

The vocabulary growth of Thai text is illustrated by a type-token curve in Figure 1. This curve is plotted from 5 million words randomly selected from three text and speech corpora: BEST, LOTUS-BN and HIT-BTEC. The detail of each corpus is given in Section 4.1. To balance the amount of data from different corpora and domains, 500K words are selected from each of the 8 genres in BEST and 500K words each are selected from LOTUS-BN and HIT-BTEC. The total becomes 5 millions words from 10 different genres. We can see from the type-token curve that even with 5 million words, the vocabulary of Thai continues to grow. New names are the main cause of the vocabulary growth. When looking at a type/token ratio, the ratio for named-entity alone is 0.321 while the ratio for any types of words is 0.028. New words

  25 – 29 August 2013, Lyon, France

also arise from transliterated words and abbreviations. Since many named-entities are a compound word, this type of OOV words should be able to be modeled by sub-lexical units.

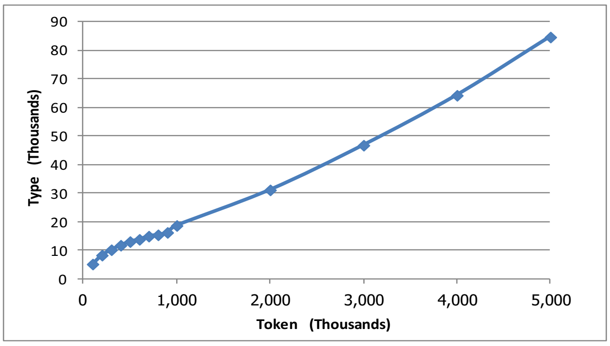

Figure 1: *Type-token curve of Thai text.*

3. Pseudo-morpheme

The design of a sub-lexical unit depends largely on the characteristic of each language. Since there is neither derivative nor inflection in Thai words, morpheme is not an appropriate sub-lexical unit for Thai. Nevertheless, as a letter in Thai is a phonogram which roughly represents a phoneme or combination of phonemes, Thai word could be segmented into a set of syllable-like units.

Basic Thai textual syllables can be represented in the form of {C_i, V, C_f, T} as shown in Figure 2, where C_i, V, C_f, and T denotes an initial consonant, a vowel, a final consonant, and a tone respectively. The word "ความมั่นคง" (stability) consists of three textual syllables: "ความ", "มั่น", and "คง". The second syllable "มั่น", pronounced as /mân/, has a basic syllable pattern with all four components {ม, ั, น, ่}. Each Thai letter and its corresponding phoneme (or combination of phonemes) in IPA are also illustrated in Figure 2. Some complex syllables may have multiple initial consonants, vowel forms or final consonants while some syllables may have components omitted. The first syllable "ความ", pronounced as /khwāam/, has two initial consonant 'คว' while the third syllable "คง", pronounced as /khōŋ/, has a vowel form omitted. Nevertheless, patterns of syllables can be defined and are known to be finite. Thai writing rules can be used to identify syllables and their components in a given text.

From the aforementioned characteristics, a suitable sub-lexical unit for Thai should be based on syllable. A *pseudo-morpheme* (PM) defined as a syllable-like unit in a written form is proposed [5]. Nonetheless, it should not be confused with a phonetic syllable obtained from word pronunciation as a PM may have multiple phonetic syllables. For example the fourth word "ปลัดกระทรวง" in Table 1, its first PM "ปลัด" corresponds to two phonetic syllables /pàʔ-làt/. PMs are separated by '-' in the second column. PM could also be considered as a textual syllable. Words in Table 1 are OOV words found in the test sets described in Section 5. The first and second words are person names while the third word "หัวหน้าชุด" (team leader) and the fourth word "ปลัดกระทรวง" (permanent secretary) are titles. The last word is a transliterated word from the word "GDP value". Named-entities and transliterated words are known to be the main causes of OOV as discussed in Section 2. By modeling these OOV words with a combination of PMs, they could be recognized correctly.

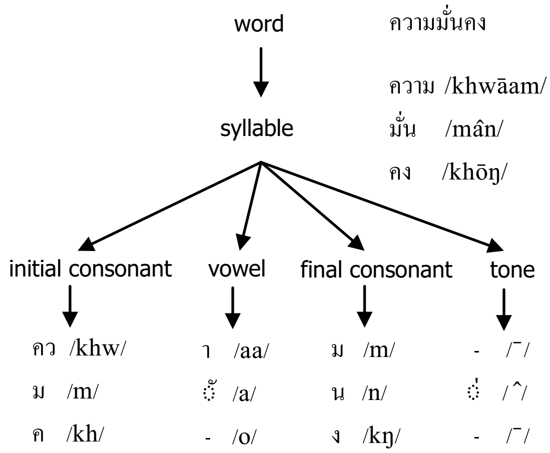

Figure 2: *Thai syllables and their components.*

Table 1. *Examples of words and their corresponding PMs and pronunciations.*

Word	PMs	Pronunciation
ธาดาธำรงเวช	ธา-ดา-ธำ-รง- เวช	/thāa/dāa/dām/rōŋ/wêet/
กิตติรัตน์	กิต-ติ-รัตน์	/kìt/tìʔ/rát/
หัวหน้าชุด	หัว-หน้า-ชุด	/hǔua/nâa/chút/
ปลัดกระทรวง	ปลัด-กระ-ทรวง	/pàʔ-làt/kràʔ/sūuŋ/
ค่าจีดีพี	ค่า-จี-ดี-พี	/khâa/cīi/dīi/phīi/

4. Hybrid language model

Even though both in-vocabulary words and out-of-vocabulary words can be reconstructed from PMs, using only PMs alone in a language model (LM) may not be suitable for the following reasons. A small lexical unit cover shorter span of context in an n-gram LM. Furthermore, in recognition, a small unit usually suffers from acoustic confusability. To resolve these problems, a hybrid language model which combines both unit types, PM and word, is proposed. We first explain the resources used for language model training in Section 4.1 then describe the training process in Section 4.2.

4.1. Language resources

Three large text and speech corpora BEST, LOTUS-BN and HIT-BTEC, are used for lexical modeling and language model training. As these corpora cover variety of domains, they are good resources for developing a hybrid language model for open-vocabulary LVCSR. Table 2 summarizes the amount of data from each corpus used in language model training.

BEST [8] is a large Thai text corpus developed under the BEST (Benchmark for Enhancing the Standard of Thai language processing) project. The corpus consists of articles from 4 different genres: academic article, encyclopedia, novel, and news. These articles were manually segmented into words by linguists. The BEST corpus contains approximately 5 million words and is available for research and academic purposes. After the first phase of the project, additional 2 million words from 4 genres, Buddhism, law, lecture and

Wikipedia, were collected and manually segmented. The total amount of available data is approximately 7 million words.

LOTUS-BN [9] is a Thai television broadcast news corpus which includes audio recordings of hour-long news programs and their transcriptions. There are 18 news topics in the corpus, for instance, politics, sport, and weather. The corpus contains approximately 100 hours of speech from 43 female speakers and 38 male speakers. Data in LOTUS-BN are divided into 3 sets: a training set (TR), a development test set (DT) and an evaluation test set (ET). There is no overlapping speaker among the TR, DT and ET sets. Only the TR set was used to train a hybrid language model.

HIT-BTEC, a parallel Thai-English corpus, is a Thai incorporated version of the HIT Olympic multilingual corpus (HIT) [10] and the Basic travel expression corpus (BTEC) [11]. This corpus was initially created as a resource for speech-to-speech translation research. The original HIT corpus was developed by Harbin Institute of Technology to include utterances from five domains related to Olympic Games, namely, traveling, dining, sports, traffic and business. The Thai version of BTEC and HIT was constructed under the Universal speech translation advanced research (USTAR) consortium. The total amount of data in the HIT-BTEC corpus is nearly 160,000 utterances.

Table 2. *Language resources for lexicon modeling and language model training.*

Corpus	Number of Tokens	Vocabulary size
BEST	7,818,410	110,334
LOTUS-BN (TR)	929,810	35,327
HIT-BTEC	1,745,680	20,307
All	10,493,900	143,622

4.2. Language model training

To train a language model, we start with creating a lexicon. The lexicon of our hybrid LM consists of frequent full-words and PMs from less frequent words. As all three corpora were manually segmented into words, a list of all distinct full-words and their frequency can be obtained straightforwardly. Words that occur more frequent than a threshold are added into the lexicon. The threshold is determined empirically from the experiments (See Section 5.2). We then look up the pronunciations of these words in LEXiTRON-Pro [12], a large pronunciation dictionary with over 130,000 of Thai words. For words that are not in the dictionary (about 40%), we generated their pronunciations using an automatic grapheme-to-phoneme conversion (G2P) [13]. For some words that their pronunciation cannot be obtained through the above process, for example, numbers and some English words, they are excluded from the lexicon.

From the list of full-word lexicon, all words in the training data that are not in the list are segmented into PMs. A syllable segmentation tool can be used to automatically segment a word or a string of text into PMs. A syllable segmentation tool which utilizes Thai writing rules and trigram statistics of syllables [14] is used in the paper. To remove redundant units, PMs that are also short words in the list of full-word lexicon are excluded. The pronunciation of each PM is obtained from G2P conversion. The PMs and their pronunciation are then added into the lexicon. Finally, we then train a trigram language model from the hybrid lexicon of both words and PMs using SRILM toolkit [15]. Chen and Goodman's modified Kneser-Ney discounting, which was found to be an optimal discounting technique from our preliminary experiment, is also applied.

5. Experiments

To verify our scheme, we run recognition experiments on data from various tasks including broadcast news transcription, dictation and speech-to-speech translation. The test data are described in Section 5.1 along with the experimental settings. The results from the recognition experiments are then discussed in Section 5.2.

5.1. Experimental settings

Our speech recognizer uses a context-dependent acoustic model with 3,000 tied-state triphones and 32 Gaussian mixtures. Each Hidden Markov Model (HMM) consists of 5 states. These HMMs are constructed from a 39-dimensional vector (12 MFCC, 1 zeroth-order cepstral coefficient, and their first and second derivatives) extracted from each frame of speech data. We trained an acoustic model using the Hidden Markov Model Toolkit (HTK) [16]. The acoustic model was initialized with phone-balanced read speech data from the LOTUS corpus [17]. To make the acoustic model more robust to noisy data, the model was retrained with multi-condition additive noises. Finally the model was retrained with 75 hours of LOTUS-BN training set (TR), which is the largest available open-vocabulary continuous speech data in Thai.

The test data are obtained from three different recognition tasks: broadcast news transcription (BN), dictation and mobile speech-to-speech translation (S2S). The *BN* test set consists of 2,200 utterances of 5 male and 5 female speakers taken from the evaluation test set (ET) of the LOTUS-BN corpus. The *Dictation* test set is a set of prepared speech of 300 utterances recorded by 3 speakers in office environment. The utterances cover 5 genres: newspaper, law, novel, social media and web board. The *S2S* test set is a field testing data of a speech-to-speech translation system in a sport and travel domain developed under the U-STAR project (http://www.ustar-consortium.com/). This data set consists of 2000 utterances recorded from mobile devices in real environment.

5.2. Experimental results

Several baseline and hybrid LMs were experimented to investigate the effect of mixed unit types on the performance of Thai LVCSR system. We use OOV rate (*OOV*) to measure the coverage of a given lexicon over a test set. OOV reported in this section is an effective OOV rate which account for the percentage of words that are not in the full-word lexicon and cannot be constructed from the PMs. Perplexity (*PP*) is used to measure the goodness of the proposed LM with respect to a test set. The overall effect of the hybrid LM on the LVCSR system is measured in terms of recognition error rate. Since mixed types of units, word and PM, are used in the LM, we compute the recognition error rate based on the smaller unit, PM. PM error rate (*PER*) is reported in this section.

The performances of the baseline LMs which use only full-words are reported in Table 3. B0 is a baseline LM which includes all full-words in the training data in its lexicon. B1 is an LM which includes only the words that occur more than 1 time in the training data. Similarly, B2 and B3 are LMs which includes only the words that occur more than 2 and 3 times in the training data respectively.

Table 3. *Recognition results using full-word LMs.*

Test Set	LM	B0	B1	B2	B3
	Vocab	*143,622*	*64,312*	*44,451*	*35,348*
BN	*PER (%)*	36.36	36.38	36.34	36.45
	PP	205.62	199.82	191.81	187.14
	OOV (%)	1.95	2.52	2.93	3.24
Dictation	*PER (%)*	33.00	32.93	32.82	32.94
	PP	186.94	186.96	180.61	174.67
	OOV (%)	2.26	2.41	2.74	3.08
S2S	*PER (%)*	49.57	49.60	49.45	49.52
	PP	121.90	137.98	133.19	131.04
	OOV (%)	0.79	1.04	1.44	1.65

Naturally, for all 3 test sets, the OOV rate is increased when a smaller lexicon is used. Nevertheless, the PER is not much increased or even slightly decreased in Dictation and S2S test set. Including low frequency words in the vocabulary may not be a good choice since noise words such as typos and word segmentation ambiguities may also be included. The increase in PP in both BN and Dictation test sets seems to indicate that the LM which includes low frequency words may suffer from a sparse data problem. Despite the use of 10-million-word training data, it is not sufficient to train a large vocabulary language model. From these reasons, B3 which uses approximately 35K words in its lexicon is an optimal full-word LM.

Table 4. *Recognition results using hybrid LMs.*

Test Set	LM	B3	H1	H2	H3
	Word	35K	20K	10K	5K
	PM	0K	15K	25K	30K
BN	*PER (%)*	36.45	**35.95**	36.45	36.98
	PP	187.14	183.06	161.60	301.39
	OOV (%)	3.24	0.48	0.40	0.37
Dictation	*PER (%)*	32.94	**32.38**	32.49	33.02
	PP	174.67	158.39	157.42	211.82
	OOV (%)	3.08	0.81	0.52	0.51
S2S	*PER (%)*	49.52	49.64	49.56	49.85
	PP	131.04	112.23	103.37	130.43
	OOV (%)	1.65	0.41	0.28	0.27

In the second experiment, we decide to keep the vocabulary size fixed to 35K entries but vary the amount of PMs and words in the hybrid LMs to examine the effect of mixed unit types on the performance of the LVCSR system. The amount of full-word and PM units used in the hybrid LMs (H1, H2 and H3) are shown in the second row and third row in Table 4 respectively. From the results we can clearly see that the hybrid LMs can reduce the OOV rates to less than 1% in all test sets (about 2% absolute reduction on average). The OOV rates of all hybrid LMs are even lower than the OOV rate of the baseline B0 using only 25% of the vocabulary size (35K vs. 140K). When more PMs are included in the lexicon, the OOV rate is reduced further. In terms of PER, the best results are obtained from H1 where 20K of words and 15K of PMs are used. The hybrid LM H1 can reduce PER by 0.50% absolute and 1.37% relative on the BN test set, and 0.56% absolute and 1.70% relative on the Dictation test set compared to the full-word LM B3. Moreover, the PERs of H1 is almost as good as or even better than the PERs of B0 in some test sets. However, when too many PMs are introduced as in the case of H3, PERs are increased. Too many small units seem add more confusability to the acoustic model. For a language model, too many small units make the trigram model covers shorter context span and cause PPs to increase.

Table 5. *Positive and negative effects on recognition results.*

Test Set	LM	B3	H1	H2	H3
	Word	35K	20K	10K	5K
	PM	0K	15K	25K	30K
BN	*OOV (%)*	3.24	0.48	0.40	0.37
	COOV (%)	-	26.61	25.82	25.36
	MIV (%)	38.70	36.03	36.27	36.59
Dictation	*OOV (%)*	3.08	0.81	0.52	0.51
	COOV (%)	-	31.66	31.66	26.81
	MIV (%)	39.61	38.32	38.54	38.36
S2S	*OOV (%)*	1.65	0.41	0.28	0.27
	COOV (%)	-	5.61	5.61	5.61
	MIV (%)	60.59	58.95	59.18	59.63

COOV is the percentage of correctly recognized OOVs while *MIV* is the percentage of misrecognized in-vocabulary words (negative effect). COOVs and MIVs are computed with respect to the 35K vocabulary of the baseline LM B3. By modeling OOV words with a combination of PMs in the proposed hybrid LMs, around 30% of OOV words can be correctly recognized in the BN and Dictation test set and around 6% in the S2S test set. Examples of correctly recognized OOV words, such as person names, titles and transliterated words, are shown in Table 1. Since MIV is not higher than the baseline, there is no negative effect from including PMs in the language mode.

6. Conclusions

We proposed a hybrid language model which combines word and sub-word units to alleviate the problem of OOV words in open-vocabulary Thai LVCSR. Pseudo-morpheme (PM), a syllable-like unit in a written form, is specifically designed as a sub-word units for Thai. The proposed hybrid lexicon can constitute an open vocabulary for Thai LVCSR as it can reduce the OOV rate to less than 1% in various recognition tasks including broadcast news transcription, dictation and mobile speech-to-speech translation. The OOV rates of the hybrid LM is even lower than the OOV rates of the full-word LM which uses all words in the training data while using only 25% of the vocabulary.

In terms of recognition performance, the best hybrid LM, which uses 20K of words and 15K of PMs, achieved the best recognition results on two recognition tasks, BN and Dictation. PER reduction is 0.50% absolute (1.37% relative) on the BN test set and 0.56% absolute (1.70% relative) on the Dictation test set compared with the full-word LM, which uses the same amount of lexicon entries, as the hybrid LM can correctly recognize about 30% of OOV words. On the S2S test set the recognition rate is comparable to the baseline. The vocabulary size of 35K is also practical as it is small enough to be used on mobile devices.

Although the hybrid LM can greatly reduce the number of OOV words, it can correctly recognized around 30% of them. To improve the recognition accuracy, the pronunciation variations of PMs should also be included. A variable-length n-gram language model [18] should also be considered to solve the problem of short context span of sub-word units.

7. References

[1] El-Desoky, A., Gollan, C., Rybach, D., Schlüter, R., and Ney, H., "Investigating the use of morphological decomposition and diacritization for improving Arabic LVCSR", in Proc. of Interspeech, Brighton, UK, Sep. 2009, pp. 2679 – 2682.

[2] Ordelman, R., Hessen, A. V., and Jong, F. D., "Compound decomposition in Dutch large vocabulary speech recognition", in Proc. of Eurospeech, Geneva, Switzerland, Sep. 2003, pp. 225 – 228.

[3] Rotovnik, T., Maučec, M. S., and Kačič, Z., "Large vocabulary continuous speech recognition of an inflected language using stems and endings", Speech Communication, 49(6):437-452, 2007.

[4] Kwon, O. W., and Park, J., "Korean large vocabulary continuous speech recognition with morpheme-based recognition units", Speech Communication, 39(3):287-300, 2003.

[5] Jongtaveesataporn, M., Thienlikit , I., Wutiwiwatchai, C., Furui, S., "Lexical Units for Thai LVCSR", Speech Communication, 51(4): 379-389, 2009.

[6] Rondeau, M. A., and Rose, R., "Developing a hybrid language model for open vocabulary automatic speech recognition in a lecture speech task", in Proc. of the International Conference on Information Science, Signal Processing and their Applications (ISSPA), Montreal, Canada, Jul. 2012, pp. 114-119.

[7] Shaik, M., Mousa, A. E. D., Schluter, R., and Ney, H., "Hybrid language models using mixed types of sub-lexical units for open vocabulary German LVCSR", in Proc. Interspeech, Florence, Italy, Aug. 2011, pp. 1411-1444.

[8] Kosawat, K., Boriboon, M., Chootrakool, P., Chotimongkol, A., Klaithin, S., Kongyoung, S., Kriengket, K., Phaholphinyo, S., Purodakananda,S., Thanakulwarapas, T., and Wutiwiwatchai, C., "BEST 2009: Thai Word Segmentation Software Contest", in Proc. of SNLP, Bangkok, Thailand, Oct. 2009, pp. 83-89.

[9] Chotimongkol, A., Saykhum, K., Chootrakool, P., Thatphithakkul, N., and Wutiwiwatchai, C., "LOTUS-BN: A Thai Broadcast News Corpus and Its Research Applications", in Proc. of Oriental-COCOSDA, Beijing, China, Aug. 2009.

[10] Yang, M., Jiang, H., Zhao, T., and Li, S., "Construct trilingual parallel corpus on demand", Lecture Notes in Computer Science: Chinese Spoken Language Processing, 4274:760-767, 2006.

[11] Kikui, G., Sumita, E., Takezawa, T., and Yamamoto, S., "Creating corpora for speech-to-speech translation", in Proc. of Eurospeech, Geneva, Switzerland, Sep. 2003, pp. 381-384..

[12] Chootrakool, P., Wutiwiwatchai, C., Kosawat, K., "A Large Pronunciation Dictionary For Thai Speech Processing", in Proc. of ASIALEX 2009 "Dictionary in Education", Bangkok, Thailand, Aug. 2009.

[13] Thangthai, A., Wutiwiwatchai, C., Rugchatjaroen, A., and Saichum, S., "A Learning Method for Thai Phonetization of English Words", Proc. of Interspeech, Antwerp, Belgium, Aug. 2007, pp. 1777-1780.

[14] Aroonmanakul, W., "Collocation and Thai Word Segmentation", in Proc. of SNLP-Oriental COCOSDA, Prachuapkirikhan, Thailand, 2002, pp. 68-75.

[15] Stolcke, A., "SRILM-an extensible language modeling toolkit", in Proc. of the International Conference on Spoken Language Processing (ICSLP), vol. 2, Denver, Colorado, USA, Sep. 2002, pp. 901 – 904.

[16] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[17] Kasuriya, S., Sornlertlamvanich, V., Cotsomrong, P., Kanokphara, S., and Thatphithakkul, N., "Thai Speech Corpus for Speech Recognition", In Proc. of Oriental-COCOSDA 2003, Singapore, Oct. 2003.

[18] Kneser, R., "Statistical language modeling using a variable context length", in Proc. of ICSLP 1996, Philadelphia, Pennsylvania, USA, Oct. 1996, pp. 494-497.