Thai Broadcast News Corpus Construction and Evaluation

Markpong Jongtaveesataporn*, Chai Wutiwiwatchai†, Koji Iwano*, Sadaoki Furui*

*Tokyo Institute of Technology
2-12-1, Ookayama, Meguro-ku, Tokyo 152-8552 Japan
{marky, iwano}@furui.cs.titech.ac.jp, furui@cs.titech.ac.jp

†National Electronics and Computer Technology Center
112 Thailand Science Park , Paholyothin Rd., Klong 1 Klong Luang , Pathumthani 12120 Thailand
chai@nectec.or.th

Abstract

Large speech and text corpora are crucial to the development of a state-of-the-art speech recognition system. This paper reports on the construction and evaluation of the first Thai broadcast news speech and text corpora. Specifications and conventions used in the transcription process are described in the paper. The speech corpus contains about 17 hours of speech data while the text corpus was transcribed from around 35 hours of television broadcast news. The characteristics of the corpus were analyzed and shown in the paper. The speech corpus was split according to the evaluation focus condition used in the DARPA Hub-4 evaluation. An 18k-word Thai speech recognition system was setup to test with this speech corpus as a preliminary experiment. Acoustic model adaptations were performed to improve the system performance. The best system yielded a word error rate of about 20% for clean and planned speech, and below 30% for the overall condition.

1. Introduction

Multimedia information has been getting more and more important especially on the internet in the past few years. Many broadcasting companies have started making their archives in digital format. Various services, such as indexing, are expected to be ready for processing multimedia information the same as what have been performed on written text data. Broadcast news transcription system is one of the applications to fulfill this desire. The Defense Advanced Research Projects Agency of the United States (DARPA) started research on automatic transcription of broadcast news in 1995 (Stern, 1997). Since then, this research theme has become attractive to many research groups. As the development of a state-of-the-art speech recognition system depends on large speech and text corpora, many collections of broadcast news corpora for various languages have already been created (Matsuoka et al., 1997; Federico et al., 2000; Graff, 2002; Wang, 2003).

However, that is not the case for resource deficient languages, such as Thai (Wutiwiwatchai and Furui, 2007). Until several years ago, research on Thai automatic speech recognition (ASR) was conducted on small vocabulary systems for a specific task. There were also some papers reporting about the development of large vocabulary continuous speech recognition (LVCSR) for the Thai language (Tarsaku and Kanokphara, 2002; Kanokphara et al., 2003; Suebvisai et al., 2005; Jongtaveesataporn et al., 2007). A few speech corpora were constructed by these research activities (Kasuriya et al., 2003a; Kasuriya et al., 2003b; Schultz, 2002). They are all read speech corpora recorded in a clean environment. Nevertheless, real-world speech data, such as broadcast news, always contains speech with background noise or spontaneity. Existing corpora cannot sufficiently represent real-world speech data. In order to develop technologies for Thai broadcast news multimedia processing, Tokyo Institute of Technology has initiated the construction of the first Thai broadcast news corpus. The preliminary target is a collection of about 17 hours of television broadcast news speech and a text corpus transcribed from about 35 hours of television broadcast news. Collaboration with National Electronics and Computer Technology Center (NECTEC) in Thailand was established and a joint effort was started to collect additional broadcast news speech with the target of another 50 hours.

This article is organized in the following way. Section 2 gives a basic background of the Thai language. Section 3 describes the structure of television broadcast news and conventions used in the transcription process. It then describes the specification of the corpus. Section 4 shows how the corpus was developed. The analysis of the resulting corpus is shown in Section 5. Section 6 describes how an LVCSR system for Thai broadcast news was setup and tested with the corpus. It also shows the experimental result and discussion on the result. Section 7 gives the conclusion of this paper.

2. Background of the Thai language

Thai is a tonal language. Text is written from left to right without sentence or word boundary markers. A space can be optionally inserted inside the text for aesthetic reasons. Since word definition is very ambiguous, word segmentation in Thai is not a trivial task.

The phonological system of Thai consists of 38 initial consonants, 18 vowels, 6 diphthongs, 12 final consonants, and 5 tones. Thai syllable structure is often described as /C_i V C_f/ where C_i, V, and C_f represent initial consonant, vowel, and final consonant respectively. Table 1 shows all consonants and vowels used in our system.

3. Corpus specification

There are three results from this project: a speech corpus, a text corpus, and a pronunciation dictionary. Firstly, this

C_i	Single: p p^h t t^h c c^h k k^h h b d m n ŋ r l j w s f ʔ Cluster: pr p^hr pl p^hl tr t^hr kr k^hr kl k^hl kw k^hw br bl fr fl dr
V	Single: a aː i iː ɨ ɨː u uː e eː æ æː o oː ɔ ɔː ə əː Diphthong: ia iːa ɨa ɨːa ua uːa
C_f	p t k m n ŋ w j f s c^h l

Table 1: Thai phonemes in IPA

section describes the structure of broadcast news and the conventions used in the transcription process. Then the description of each is presented.

3.1. Broadcast news structure

The hierarchical structure of television broadcast news programs can be depicted as in Figure 1. Each element is described as follows.

Episode: An episode refers to the recording of a particular broadcast news program at a time.

Section: A section represents a specific part in an episode, containing one specific news topic or reporting type. In other words, an episode is divided into many sections composed of untranscribed sections (e.g. TV commercial), fillers (e.g. greeting, headline news), and news reports.

Turn: A turn indicates a specific portion which contains speech from a single speaker.

Segment: A segment is a fragment of data, defined as a sentence of text, in a turn. Specifically, a segment itself contains transcribed speech data.

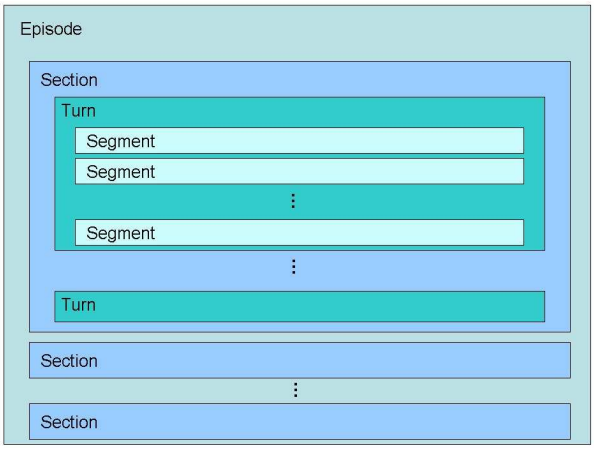

Figure 1: Broadcast news structure

3.2. Transcription conventions

Transcription conventions are used as a guideline for making the transcription as follows.

Sentence segmentation: A sentence refers to a segment in the structure of broadcast news. Normally, Thai text is written without sentence boundaries. Human transcribers were asked to create a segment on the basis of a simple sentence or a clause with the help of pauses in order not to make a segment too long.

Word segmentation: There is no word boundary symbol in Thai text. We realize that this will make a lot of difficulties in the process of transcription and data checking. Additionally, we think that existing morphological analyzers are not practical to be used in this situation because of errors caused by many proper names which are usually unknown words to the tools. Therefore, word segmentation is performed by human transcribers when they make transcriptions. As word definition in the Thai language is very ambiguous, and is not consistent among people, an instruction on some word segmentation patterns were given to transcribers. The patterns not included in the instruction set were left to the transcriber's decision.

Iteration marks: In written Thai text, there is a reduplication symbol used to represent a duplicate of the preceding word. Transcripts were made in the full text form instead of this character to avoid further conversion tasks at a later stage.

Thai acronyms: No acronyms were allowed, except for the case where an acronym was spoken. In such a case, a sequence of letters is followed by a dot character at the end, even though there are additional dot characters inside the sequence of letters in the typical written form.

English acronyms: An English acronym is usually used as a proper noun in Thai. Normally it can be written in either English or Thai character form. However, an English acronym in the transcript is transcribed in Thai character form.

Number entities: Numbers are transcribed in Thai character form.

Bracketed tags: Special tags within brackets are used to describe some speech events, and filled pauses. A list of these events was created and defined by bracketed Roman alphabet tags to be used in the transcription. Additionally, special tags are defined to describe 1) utterances with repairs or repetitions, 2) a foreign word, 3) utterances which cannot be clearly understood by transcribers, and 4) utterances whose signal may not be segmented correctly due to the sudden change of the speaker.

3.3. Speech corpus

The speech corpus is the collection of broadcast news recordings including corresponding structural descriptions and text transcriptions. For each turn, additional attributes are used to describe its characteristics by speaker information (name and gender) and mode (planned/spontaneous). In addition, the starting and ending of background noise, such as background music and sound from a news story, are indicated in transcriptions. We used a tool, called Transcriber (Barras et al., 2001), which utilizes the XML format to annotate the broadcast news structure.

At present, only the portion of speech derived from professional announcers speaking in a studio is transcribed. Speech from unknown announcers is not transcribed except for a few regular daily sections that contribute a considerable amount of data.

3.4. Text corpus

The text corpus is the collection of text transcribed from broadcast news recordings. It was transcribed exactly the same as the text associated with the speech corpus except that no structural information is annotated. Tags which describe the speaking mode as either planned or spontaneous were inserted.

3.5. Pronunciation dictionary

The pronunciation dictionary is made in simple tab-separated text format. A word with multiple pronunciations is entered with multiple records. Foreign words need special attention when their pronunciations are created. There are two speaking styles for the final consonant when a Thai speaks a foreign word, with original Thai pronunciation and with English-style pronunciation. For example, an English word "bus" may be pronounced as /bat/ (Thai-style) or /bas/ (English-style). This variation depends on many factors such as a speaker's age and education level, and situational context. Moreover, foreign words, especially proper names, can be pronounced in multiple ways. For instance, a surname such as, "Anderson", may be uttered as /ʔan deː san/ or /ʔæːn dəː san/ by announcers. This variation was examined in the transcribed corpus and multiple pronunciations for a foreign word were included in the dictionary.

4. Corpus development

4.1. Broadcast news collection

News programs from a public Thai TV broadcasting company were chosen to be collected. Three types of news programs: morning, noon, and evening news programs, were recorded. The recordings were made between February and April 2007. A total of 105 news episodes was recorded. Thirty-five evening news episodes (about 17 hours of the content to be transcribed) were selected to make the speech corpus, and 70 news episodes (about 35 hours of the content to be transcribed) covering morning, noon, and evening news reports were selected to make the text corpus.

The recording task was wholly done on a PC with an analog TV capture card that had an MPEG2 hardware encoder. The video was encoded in MPEG2 format and the audio was encoded in MPEG Layer II with a 48kHz sampling rate, stereo, at 384Kbps. Only the left channel audio of the video file was extracted and down-sampled at 16kHz with a resolution of 16bits, and encoded in the Microsoft PCM format.

4.2. Transcript development

There were 4 transcribers responsible for transcribing the speech corpus, and 7 transcribers responsible for transcribing the text corpus, guided by a supervisor. A transcribing manual was written, and demonstration and some training were performed before the work started. At the beginning of the work, problems received from transcribers were collected, and a revised manual was re-distributed to transcribers by the supervisor. When all transcriptions were completed, spellings of lexical entries were checked manually for consistency. After the lexical check was completed, the transcription and annotation of the speech corpus was checked again by 2 transcribers. A list of pronunciations was then created using a tool (Tarsaku et al., 2001) and revised by a human.

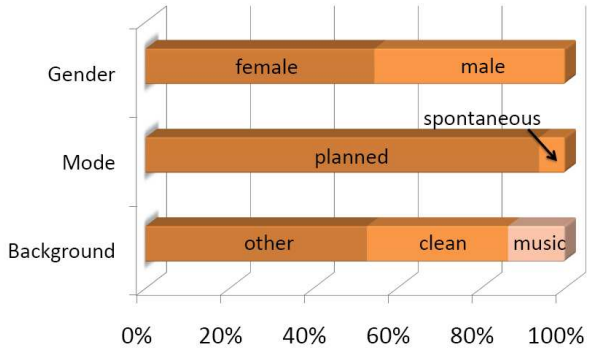

Figure 2: Time composition of the speech corpus

Attribute	Speech corpus	Text corpus
No. of sentence	13024	31816
No. of words	224k	573k
No. of unique words	10k	14k
No. of phonemes	899k	-

Table 2: Linguistic information on the speech and text broadcast news corpus

5. Corpus analysis

Some statistical information was analyzed from the corpora. The speech corpus contains 1613 sections which cover 4803 turns. Figure 2 shows the time composition of the speech corpus based on three features: gender, speaking mode, and background noise. It shows that the corpus contains more female speech than male speech. Around 54.6% of the corpus is spoken by female speakers while 45.4% of the corpus is spoken by male speakers. With regard to the number of speakers, the corpus consists of 8 female and 4 male speakers, and some unidentified speakers from 3 regular sections. The corpus is composed of 93.9% planned speech and 6.1% spontaneous speech. Regarding background noise, 33.7% of the material is clean speech. There exists 13.5% of speech with music noise and 52.8% of it with other noise. The analysis also shows that 0.2% of the corpus contains overlapped speech that is uttered by more than one speaker at a time.

Linguistic information of the speech and text corpus was investigated and shown in Table 2.

The speech corpus was partitioned according to the evaluation focus conditions (F-condition) employed in the DARPA Hub-4 evaluation (Stern, 1997). Information regarding the number of segments and the average length of segments categorized into each F-condition is shown in Table 3.

F-Condition	Number of segments		Average length (seconds)	
	Male	Female	Male	Female
F0	2409	1828	4.6	5.1
F1	254	254	2.1	2.4
F3	333	1580	4.1	4.6
F4	3175	3191	5.1	5.5

Table 3: Statistics for the speech corpus for each F-condition

F-Condition	Number of segments		Average length (seconds)	
	Male	Female	Male	Female
F0	626	407	4.7	5.3
F1	26	23	2.3	3.5
F3	70	337	5.1	5.0
F4	778	733	5.2	5.6

Table 4: Statistics for the test set for each F-condition

6. Speech recognition

6.1. Test set selection

It was difficult to establish a correct boundary for speech segments where the speaker's turn was changed abruptly. Since this kind of signal segmentation error may affect the recognition result, we excluded segments with this condition from the test set candidates. A language model was trained from the text corpus. The perplexity (PP) of each segment in the speech corpus was calculated against this language model. Segments were ranked by PP and 0.5% of the highest and lowest ranked segments were excluded from the list. Segments where the number of words is less than 4 were also removed from the list. There were 11778 segments left for test set selection. For each gender, 1500 speech segments were randomly selected to compose a test set. Table 4 shows the statistics of the test set separated by each F-condition.

6.2. Experimental setup

Since the speech corpus was rather small, it was used only for evaluation. We employed two newspaper read-speech corpora, LOTUS (Kasuriya et al., 2003b) and a corpus collected by Tokyo Institute of Technology, to train the acoustic model. All speech data sets available in the LOTUS corpus that were recorded with a dynamic close-talk microphone were used. The total amount of acoustic training data was 40.3 hours from 68 male and 68 female speakers. 25-dimensional feature vectors consisting of 12 MFCCs, their delta, and a delta energy were used to train gender-dependent acoustic models. Phones were represented as context-dependent, 3-state, left-to-right hidden markov models (HMM). The HMM states were clustered by a phonetic decision tree. The number of leaves was 1,000. Each state of the HMMs was modeled by 8 Gaussian mixtures. No special tone information was incorporated. The text corpus and transcriptions from the speech corpus that were not selected to be included into the test set were used as a training corpus for language modeling. The dictionary size was about 18k words. The TITech large vocabulary WFST speech recognition system (Dixon et al., 2007) was used as a speech decoder.

6.3. Experimental results

Some experiments were performed and the resulting word error rates (WER) are shown in Table 5. Perplexities and out-of-vocabulary (OOV) rates are shown in Table 6. Firstly, the acoustic models described in the previous subsection were used to decode the whole test set. The result was then classified into corresponding F-conditions and shown in the "No Adaptation" column. The overall WERs were at 36.5% and 41.1% for male and female speakers respectively. In detail, as expected, the system performed the best with the clean and planned speech (F0). With spontaneous speech (F1), the performance dropped even though it was clean speech. The performance became worse when the test speech was contaminated with noise. In the test set, the level of music noise (F3) appeared to be higher than the level of other noise (F4). This might be one reason why the WERs of F3 test set were higher than those of F4.

Next, the previous acoustic models were adapted using MLLR technique to each F-condition. The adaptation data for each F-condition were randomly selected from the rest of the speech corpus with the associated F-condition for any speakers. The amount of the adaptation data was 200 utterances. The results are shown in the "Condition" column under "MLLR Adaptation" in Table 5. As a result of MLLR adaptation, the overall WERs decreased 15.5% and 20.7% relatively for male and female speakers respectively.

Finally, speaker adaptation was performed to compare the result with F-condition adaptation. The amount of 200 utterances, used as the adaptation data, was randomly chosen from the speech corpus without considering F-condition. For a couple of speakers whose number of obtainable utterances was less than 200, all available data were used. The results were categorized into each F-condition and shown in the "Speaker" column in Table 5. This experiment was done only on the identified speakers' speech. In order to make the result comparable with other conditions, other experiments' WERs, not including results from unidentified speakers, are shown in parentheses. Except for the case of F3 male speech, the performance of speaker-adapted systems was slightly better than that of F-condition-adapted systems. Consequently, the overall WERs of 28.8% and 29.5% for male and female speakers were achieved.

6.4. Discussion

We would like to note that the recording conditions between the training speech corpus and our broadcast news corpus are very different. The training corpus consists of speech recorded in clean and office environments with a high-quality microphone. This broadcast news corpus was recorded with a TV capture card in the lossy compression format. The clean speech in the corpus was not really clean since there often appears noise from announcers or unknown sources such as the sound of paper flipping, or

F-condition	Time proportion	#words	WER(%)					
			No Adaptation		MLLR Adaptation			
			Male	Female	Male		Female	
					Condition	Speaker	Condition	Speaker
F0	35.3%	17160	27.2 (26.9)	24.9 (24.5)	23.3 (23.0)	21.6	22.2 (21.2)	20.5
F1	1.0%	629	48.8 (48.8)	37.1 (37.1)	43.3 (43.3)	41.6	29.5 (29.5)	28.6
F3	14.0%	7882	70.6 (70.6)	53.2 (53.1)	48.1 (48.1)	49.6	37.5 (36.4)	33.2
F4	49.7%	27542	39.9 (38.8)	44.1 (44.3)	34.5 (32.9)	31.4	36.1 (35.4)	32.9
Overall	100%	53213	36.5 (35.9)	41.1 (40.9)	30.9 (29.9)	28.8	32.6 (31.7)	29.5

Table 5: WERs (%) for different F-conditions with various acoustic models (WERs in parentheses: results obtained by excluding unidentified speakers, Condition: WERs after F-condition-dependent adaptation, Speaker: WERs after speaker adaptation)

F-condition	Perplexity		OOV rate (%)	
	Male	Female	Male	Female
F0	107.5	106.9	0.9	0.8
F1	126.4	100.1	0.9	0.6
F3	145.2	100.0	0.7	0.9
F4	141.6	157.6	1.5	1.9
Overall	126.9	125.6	1.2	1.3

Table 6: Test set perplexities and out-of-vocabulary rates for different F-conditions

bodies moving, etc. With a larger speech corpus, an acoustic model trained from the same environment should yield better WERs as we can observe from the F-condition based adaptation experimental result. At that time, the speaker adaptation system would give better WERs. Also, we think that our transcription text corpus is rather small and the language model could not predict some 2-grams and 3-grams properly. The use of newspaper text and some interpolation techniques needs to be applied to improve the language model performance. We are also planning to apply our automatic word segmentation techniques (Jongtaveesataporn et al., 2007) for language modeling.

7. Conclusion

This paper presented the construction of the first Thai broadcast news speech and text corpora. The speech corpus contains about 17 hours and the text corpus was transcribed from about 35 hours of television broadcast news. Specifications, transcription conventions, and recording and transcription processes were explained. The characteristics of the corpus were analyzed and described in the paper. The test set was then selected to test with the 18k-word Thai large vocabulary continuous speech recognition system. WERs for the baseline system were at 36.5% and 41.1% for male and female speakers respectively. The F-condition adaptation of acoustic models reduced WERs to 30.9% and 32.6% for male and female speakers respectively. Finally, the speaker adaptation yielded the best systems with WERs at 28.8% and 29.5% for identified male and female speakers respectively.

8. Acknowledgement

This work was supported in part by the 21st century COE program “Framework for Systematization and Application of Large-scale Knowledge Resources”. The speech corpus used for training the acoustic model was funded by the METI Project “Development of Fundamental Speech Recognition Technology”. The authors would like to thank Assoc. Prof. Somchai Thayarnyong for allowing us to use his computer for recording. We also would like to thank Jakrit Siriwathanakul for his help in setting up the recording environment.

9. References

C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. 2001. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1–2):5–22.

P. R. Dixon, D. A. Caseiro, T. Oonishi, and S. Furui. 2007. The TITech large vocabulary WFST speech recognition system. In *Proceedings of Automatic Speech Recognition and Understanding 2007*, pages 443–448.

M. Federico, D. Giordani, and P. Coletti. 2000. Development and evaluation of an Italian broadcast news corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC) 2000*.

D. Graff. 2002. An overview of broadcast news corpora. *Speech Communication*, 37(1–2):15–26.

M. Jongtaveesataporn, I. Thienlikit, C. Wutiwiwatchai, and S. Furui. 2007. Towards better language modeling for Thai LVCSR. In *Proceedings of Interspeech 2007*, pages 1553–1556.

S. Kanokphara, V. Tesprasit, and R. Thongprasirt. 2003. Pronunciation variation speech recognition without dictionary modification on sparse database. In *Proceedings*

of IEEE International conference on acoustics, speech, and signal processing (ICASSP) 2003, pages 764–767.

S. Kasuriya, V. Sornlertlamvanich, P. Cotsomrong, T. Jitsuhiro, G. Kikui, and Y. Sagisaka. 2003a. NECTEC-ATR Thai speech corpus. In *Proceedings of International Conference on Speech Databases and Assessments (Oriental-COCOSDA) 2003*.

S. Kasuriya, V. Sornlertlamvanich, P. Cotsomrong, S. Kanokphara, and N. Thatphithakkul. 2003b. Thai speech corpus for Thai speech recognition. In *Proceedings of International Conference on Speech Databases and Assessments (Oriental-COCOSDA) 2003*, pages 54–61, October.

T. Matsuoka, Y. Taguchi, K. Ohtsuki, S. Furui, and K. Shirai. 1997. Toward automatic recognition of Japanese broadcast news. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH) 1997*, volume 2, pages 915–918.

T. Schultz. 2002. GlobalPhone: a multilingual speech and text database developed at Karlruhe university. In *Proceedings of International Conference on Spoken Language Processing (ICSLP) 2002*.

R. M. Stern. 1997. Specification of the 1996 Hub 4 broadcast news evaluation. In *Proceedings of the 1997 DARPA Speech Recognition Workshop*.

S. Suebvisai, P. Charoenpornsawat, A. Black, M. Woszczyna, and T. Schultz. 2005. Thai automatic speech recognition. In *Proceedings of IEEE International conference on Acoustics, Speech, and signal processing (ICASSP) 2005*, pages 857–860.

P. Tarsaku and S. Kanokphara. 2002. A study of HMM-based automatic segmentation for Thai continuous speech recognition system. In *Proceedings of the Joint International Conference of SNLP-Oriental COCOSDA 2002*, pages 217–220.

P. Tarsaku, V. Sornlertlamvanich, and R. Thongprasirt. 2001. Thai grapheme-to-phoneme using probabilistic GLR parser. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH) 2001*, pages 1057–1060.

H. Wang. 2003. MATBN 2002: A Mandarin Chinese broadcast news corpus. In *Proceedings of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR) 2003*.

C. Wutiwiwatchai and S. Furui. 2007. Thai speech processing technology: A review. *Speech Communication*, 49(1):8–27.