論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Development of a Thai Broadcast News Corpus and an LVCSR System
著者(和文)	ﾁｮﾝﾀｳｨｰｻﾀｰﾎﾟｰﾝ, 岩野 公司, 古井 貞熙
Authors(English)	Markpong Jongtaveesataporn, Chai Wutiwiwatchai, Koji Iwano, Sadaoki Furui
出典(和文)	, Vol. , No. 3-10-1, pp. 453-454
Citation(English)	日本音響学会2008年春季講演論文集, Vol. , No. 3-10-1, pp. 453-454
発行日 / Pub. date	2008, 3

Development of a Thai Broadcast News Corpus and an LVCSR System*

◎ Markpong Jongtaveesataporn[1)], Chai Wutiwiwatchai[2)], Koji Iwano[1)], Sadaoki Furui[1)]
1) Tokyo Institute of Technology 2) NECTEC

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 ready to operate for written text but not yet ready for natural spoken documents. This is especially true for resource deficient languages, such as Thai [1, 2]. In order to develop technology for Thai broadcast news multimedia processing, Tokyo Institute of Technology has initiated the construction of the first Thai broadcast news speech and language corpus. The preliminary targets of our corpus are collection of about 17 hours of television broadcast news speech and a text corpus transcribed from about 35 hours of television broadcast news.

2. Transcripts

Two types of transcripts are created: 1) Transcription and annotation (to which we refer as Speech Corpus) which is the exact reproduction of audio files under text form. The size of the Speech Corpus is 17 hours of broadcast news speech. 2) Transcription (to which we refer as Text Corpus) which is text spoken by announcers. This Text Corpus is derived from 35 hours of television broadcast news.

Only the portion of speech from professional announcers speaking in a studio is transcribed in the Speech Corpus. Speech from unknown announcers is not transcribed except for a few regular daily sections that contribute a considerable amount of data. It is made in XML format, developed manually by utilizing a software tool, called Transcriber [3]. The Transcriber tool applies a hierarchical layout to organize the annotation with the following elements: episode, section, turn, and segment. Moreover, it employs a background element to indicate the starting and ending point of the background noise, such as background music and sound from a news story. For each turn element, a number of attributes are tagged to describe its characteristics by speaker information (name and gender), mode (planned/spontaneous), and audio fidelity (high/medium/low).

The Text Corpus is also made in XML format by using the same software tool. However, no information for episode, section, and turn is identified since only text is needed. Speaking mode of planned or spontaneous is inserted into the text where the speaking mode changes.

2.1. Transcription Conventions

Transcription conventions are used as a guideline for the whole process as follows:

a. Sentence segmentation: There is no symbol for indicating sentence boundaries in written Thai text. A segment in the hierarchical layout is created on the basis of a sentence or a clause with the help of delimited breaths.

b. Word segmentation: There is no space or symbol between words in Thai text. This will make a lot of difficulties in the process of transcription and data checking. Therefore, word segmentation is performed by human transcribers when they make transcriptions.

c. Repeating word: In written Thai text, there is a Thai character representing a repetition of the preceding word. Transcript is made in the full text form instead of this character to avoid further converting task.

d. Abbreviation: No abbreviation is allowed, except for the case where it is spoken as the abbreviated form. In such a case, a sequence of letters is followed by a dot at the end, even though there are additional dots inside the sequence of letters in the typical writing form.

e. Abbreviation in English: In the Thai language, many English abbreviations are usually used especially in proper names. Each one is written in Thai characters and is considered as one word.

f. Number entity: A normalized form is used.

g. Bracketed tags: Special tags within brackets are used to describe some utterances and events, such as exclamations, and words/utterances expressing feelings and disfluencies.

3. Recording and transcription process

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, and 384Kbps. Only the left channel audio of video file was extracted and down-sampled at 16kHz with a resolution of 16bits, and encoded with the Microsoft PCM format.

News programs from a Thai TV broadcasting company were chosen to be collected. Three types of news programs, morning, noon, and evening news sessions, were recorded on weekdays. On weekends, only the evening news session was broadcasted and recorded. The recordings were made from February to April 2007. The total of 105 news episodes was recorded. Thirty-five evening news episodes were selected to make the Speech Corpus, and 70 news episodes covering morning, noon, and evening news reports were selected to make the Text Corpus.

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. Problems from transcribers were collected, and the revised manual was sent to a mailing group by the supervisor. Spellings of lexical entries were checked for consistency. The transcription and annotation set was checked again by 2 transcribers. A list of pronunciations was then created by a tool and revised by a human.

4. Thai Phonetics

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

Table 1: Thai phonetic symbols

C_i	Single: ph th ch kh p t c k h b d m n ng r l j w s f z Cluster: phr phl thr khr khl khw pr pl tr kr kl kw br bl fr fl dr
V	Single: a a: i i: e e: x x: o o: v v: q q: u u: @ @: Diphthong: ia i:a va v:a ua u:a
C_f	p^ t^ k^ m^ n^ ng^ w^ j^ f^ s^ ch^ l^

＊タイ語放送ニュースコーパスの作成とそれによる大語彙連続音声認識システムの構築
マッカポン・チョンタウィーサターポーン[1)]、チャイ・ウッティウィワッチャイ[2)]、岩野公司[1)]、古井貞熙[1)]
1) 東京工業大学 2) NECTEC

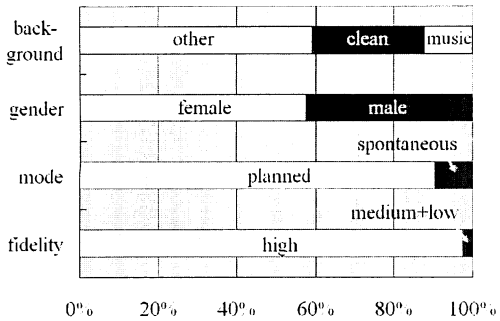

Figure 1: Time composition of the speech corpus

Table 2: Linguistic information of the Speech and Text Corpus

Attribute	Speech Corpus	Text Corpus
No. of sentences	11540	31863
No. of words	219k	587k
No. of unique words	10k	14k
No. of phonemes	839k	-

Table 3: WERs (%), PPs, and OOV rates for different F-conditions
(WERs in parenthesis excluded the result from unidentified speakers.)

Condition	Proportion		WER (%)						PP		OOV rate (%)	
					MLLR Adaptation							
					Male		Female					
	Time	#words	Male	Female	Cond.	Spk.	Cond.	Spk.	Male	Female	Male	Female
F0	28.10%	31892	44.4 (44.3)	40.8 (40.8)	39.5 (38.9)	38.8	37.5 (37.5)	37.5	244.5	263.0	1.2	0.7
F1	1.50%	2304	62.4 (62.4)	60.2 (60.2)	58.0 (58.0)	56.3	52.4 (52.4)	51.0	427.3	477.8	0.4	1.1
F3	11.50%	14676	82.2 (82.0)	72.4 (72.1)	64.6 (64.5)	65.5	55.0 (54.2)	51.5	302.4	251.0	0.9	1.0
F4	58.90%	64950	54.9 (53.9)	57.5 (57.2)	47.0 (45.8)	43.9	48.8 (47.9)	46.2	330.1	389.9	1.6	1.9
Overall	100%	113822	52.9 (52.4)	56.8 (56.5)	45.6 (44.8)	43.7	47.5 (46.8)	45.3	303.0	334.6	1.4	1.4

5. Evaluation

Figure 1 shows the composition of the speech database. The corpus consists of 4 male and 9 female speakers, and some unidentified speakers from 3 regular sections. Information of the Speech and Text Corpus is shown in Table 2.

We setup the first ASR system for the Thai broadcast news. A gender-dependent acoustic model of 3000-tied-state triphones with 8 Gaussian mixtures was trained from a newspaper read-speech corpus LOTUS [4] and a corpus collected by Furui Laboratory, by using HTK [5]. Tone information was not used. The amount of acoustic training data was 37 hours from 68 male and 68 female speakers.

To select a test set, a language model (LM) was trained from the transcribed text corpus by CMU SLM Toolkit [6]. The perplexity (PP) of each segment in the speech corpus was calculated against the LM. Segments were ranked by PP and 0.5% of highest and lowest ranked segments were excluded from the list. The number of words per segment was also utilized as a criterion for excluding segments from the list. Then, 3000 segments were randomly selected for each gender speech and were used for a test set. The test set was then partitioned according to the focus conditions employed in the DARPA HUB-4 Evaluation. Since the transcribed text corpus is rather small, transcriptions from segments that are not chosen to be included in the test set were added to the transcribed text corpus to train a new LM for the ASR system.

JULIUS [7] was used as a speech decoder. The resulting word error rates (WER), PPs, and OOV rates are shown in the Table 3. It also shows the WERs of the system obtained when the acoustic model was adapted by MLLR technique to each F-condition (indicated as Cond. column), using all the rest of F-condition speech data from all speakers. An experiment on speaker adaptation was also performed for known speakers, using 100 utterances of each speaker's speech data regardless of F-conditions. The resulting WER for speaker adapted systems is categorized into each F-condition in Table 3 (indicated as Spk. column).

6. Discussion

We have to note that the recording conditions between the training speech corpus and our broadcast news corpus are very different. With larger broadcast news speech corpus, an acoustic model trained from the same environment should yield better WER. Also, we realize that our transcription text corpus is rather small and the language model cannot predict 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 [2] for language modeling.

7. 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".

8. References

[1] C. Wutiwiwatchai and S. Furui, "Thai speech processing technology: a review," Speech Communication, 49(1), pp. 8-27, 2007.
[2] M. Jongtaveesataporn *et al.*, "Towards better language modeling for Thai LVCSR," Proc. Interspeech 2007, pp. 1553-1556, 2007.
[3] C. Barras *et al.*, "Transcriber: development and use of a tool for assisting speech corpora production," Speech Communication, 33(1-2), pp. 5-22, 2001.
[4] S. Kasuriya *et al.*, "Thai speech corpus for Thai speech recognition," Proc. COCOSDA 2003, pp. 54-61, 2003.
[5] http://htk.eng.cam.ac.uk
[6] http://www.speech.cs.cmu.edu/SLM_info.html
[7] http://julius.sourceforge.jp/en/julius.html