Progress Report on Corpus Development and Speech Techonology in Thailand

Rachod Thongprasirt, Virach Sornlertlamvanich, Patcharikra Cotsomrong
Sinaporn Subevisai and Supphanat Kanokphara

Information Research and Development Division
National Electronics and Computer Technology Center
112 Paholyothin Rd., Klong 1, Klong Laung, Pathumthani, Thailand 12120
e-mail : {rachod, virach, aye, sinaporn, supphanat_k}@nectec.or.th

Abstract

Last year National Electronics and Computer Technology (NECTEC) launched a speech corpus project for building a large-vocabulary speaker independent, continuous speech-recognition system. It is a cooperation project between NECTEC and universities with NECTEC as a host center. This paper gives details of the corpus including the sentence selection, the sentence distribution method and interesting statistics of the corpus. Additionally, speech recognition researches at NECTEC are also given.

1 Introduction

One important mean of communication between men is speech. It would be nice if a computer can communicate to us via speech as well. Currently, the advance in speech technology has made it possible to do just that. For example, many speech recognition engines are on the market, but none is Thai language.

In order to build such a Thai speech-recognition system, modern technology requires a *large* Thai speech corpus. Even though there are some Thai speech corpora, they are too small to build a usable application. Fortunately, this problem will soon be overcome with the launch of NECTEC's speech corpus for speech recognition project described in this paper. This corpus project is built with cooperation between NECTEC and universities. After the end of the project, this corpus will be made available to public.

The outline of the paper is as follows. Section 2.1 describes an overview of the speech corpus. The reader should become familiar with this corpus quickly after reading this section. The text-selection process, the recording environments, the speaker's information, and the transcription files are explained in Section 2.2, 2.3, 2.4 and 2.5, respectively. In Section 2.6, the paper will discuss how to sentences to various institues. Next, in Section 2.7 interesting statistics of the corpus will be given.

2 Speech Corpus for Speech Recognition

This section describes in detail the speech corpus for speech recognition developed by NECTEC with cooperation from universities. It will explain the distribution of sentences to involved institues, the text-selection process, and the recording environments. In addition, examples of transcription and lexicon files are given. Finally, interesting statistics of the sentences in the corpus are also discussed.

2.1 Overview

Last year NECTEC launched a speech corpus project for building a large-vocabuary speaker-independent, continuous speech recognition system. This corpus aims at covering 5,000 vocabularies. One purpose of this project is to gain cooperation between various researchers from many institutes; therefore, it is intentionally set to be a cooperation project between NECTEC and universities. NECTEC

provides funding and texts to be recorded by universities. Prince of Songkha University and Mahanakorn University of Technology have joined the project. Each university is going to record from 100 speakers a piece, and NECTEC is going to record from 48 speakers. Hence, the total number of speakers is 248. An equal number of female and male speakers is required. Table 1 below shows the distribution of sentences to each institue.

Table 1. The distribution of sentences to each institute. PSU, MU, and NEC stand for Prince of Songkha University, Mahanakorn University of Technology, and NECTEC, respectively.

Institute	Num. of speakers	Number of sentences/speaker			
		PD	TR	DT	ET
PSU	100	20	80	-	-
MU#1	50	20	-	40	-
MU#2	50	20	-	-	40
NEC#1	24	20	80	-	-
NEC#2	12	20	-	40	-
NEC#3	12	20	-	-	40

The corpus is divided into 4 sets as shown in Table 1: the phonetically distributed set (PD), the training set (TR), the development test set (DT), and the evaluation test set (ET). The PD set is used to create the initial acoustic models. The TR set is used to train additional acoustic and language models. The DT and ET sets are used for testing the system at the development and evaluation times.

It can be seen from Table 1 that each speaker has to utter exactly 20 sentences from the PD set. A careful reader may notice that a speaker of the TR, DT, and ET sets are mutually exculsive; that is, a speaker of one of this set cannot utter the other sets. The last three groups belonging to NECTEC are a smaller version of the first three groups. They are used as a simulated version to determine in advance any problem that may occur during recording sessions.

2.2 Database Creation

The database creation is divided into three phase: text management, PD selection and the TR/DT/ET selection phases. The text-management process is shown in Figure 1. Firstly, sentences are automatically extracted and then manually selected from a reference text database; any problematic sentence is excluded at this phase. For the PD set, the text from ORCHID (Sornlertlamvanich et al. 1998) are used, for the other sets text from various places such as journal or magazine are used. Each selected sentence is then transformed into its equivalent non-verbal form by removing and/or changing any special symbols such as hypen, question mark and repeater symbol. In addtion, any foreign word is changed into its corresponding Thai word, and any parenthetical information is discarded. After all usable sentences have been selected, the texts are word-segmented automatically and then manually rechecked. After that the last process is to map from graphemes of every sentence into their equivalent phonemes. Again, this process is first done automatically and then manually rechecked.

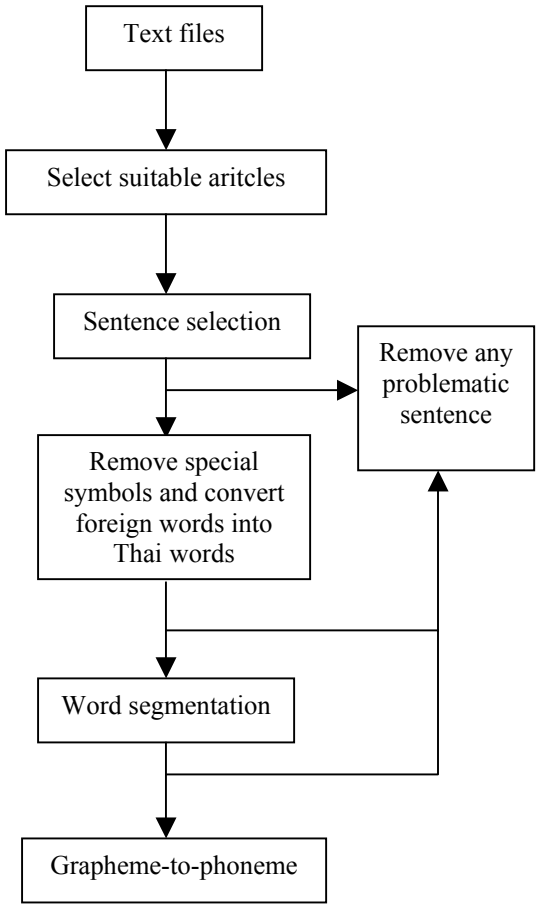

Figure 1. The text-management process.

In order to form the PD set, the list of all possible biphone pairs has been made. Theoretically, there are 2,568 possilble Biphone pairs (Luksaneeyanawin 1993). However, by eye inspection, some biphone pairs has never occurred, and it was found that the actual number of biphone pairs occur in the PD set is 1,605. Having known all possible biphone pairs, we first create a phonetically balanced set (PB) and then the PD set can be formed from the PD set as discussed in (Shen et al. 1999, Wutiwiwatchai et al. 2002). Figure 2 summaries all processes of this phase.

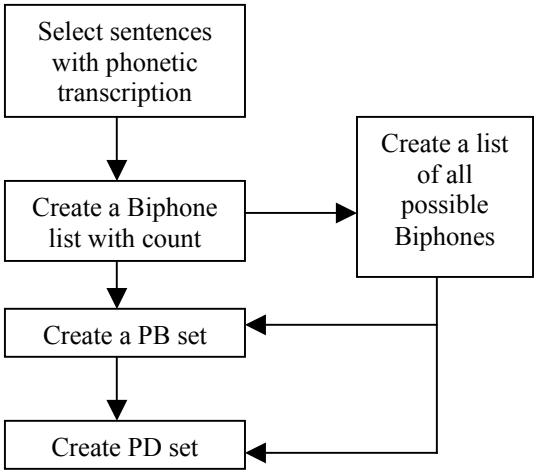

Figure 2. The PD selection process.

The constructions of the TR, DT, and ET are shown in Figure 3 below. The number of selected sentences is 180,266. Firstly, sentences are divided into two sets. The first set, about 90% of the text, is used to create a Bi-gram language model (Rosenfield 1994) and the most 5,000 frequent word list. Sentences in the second set, the rest 10%, are selected into the TR, DT and ET sets provided that it satisified all of the following constraints:

1. All words in the sentence are in the 5,000 word list.
2. The sentence has medium perplexity.
3. The sentence is not too long or too short.

All sentences are distributed into the TR, DT and ET with a constraints the TR set has to cover all 5,000 frequent words, and the smallest allowable number of sentences for the TR, DT and ET sets are 1,000, 500 and 500, respectively.

2.3 Recording Environments

Recording environments are divided into two: the clean speech (CS) and the office environments (OF1 and OF2). These two are separated by the presence of background noise such as air condition noise; that is, in the CS case there is no such presence, but in the OF1 and OF2 background noises can be audible. Furthermore, the signal to noise ratio of the CS must above 30dB, and those for the OF1 and OF2 must exceed 20dB. Microphones use in these two environments are different as well. For the CS, a high quality head set (Senheiser HMD-410 close-talk) is employed, while lower quality ones are used for the OF1 (a unidirectional microphone: SONY F-720 Dynamic microphone) and OF2 (a close-talk TELEX H-41 headset having a condenser microphone).

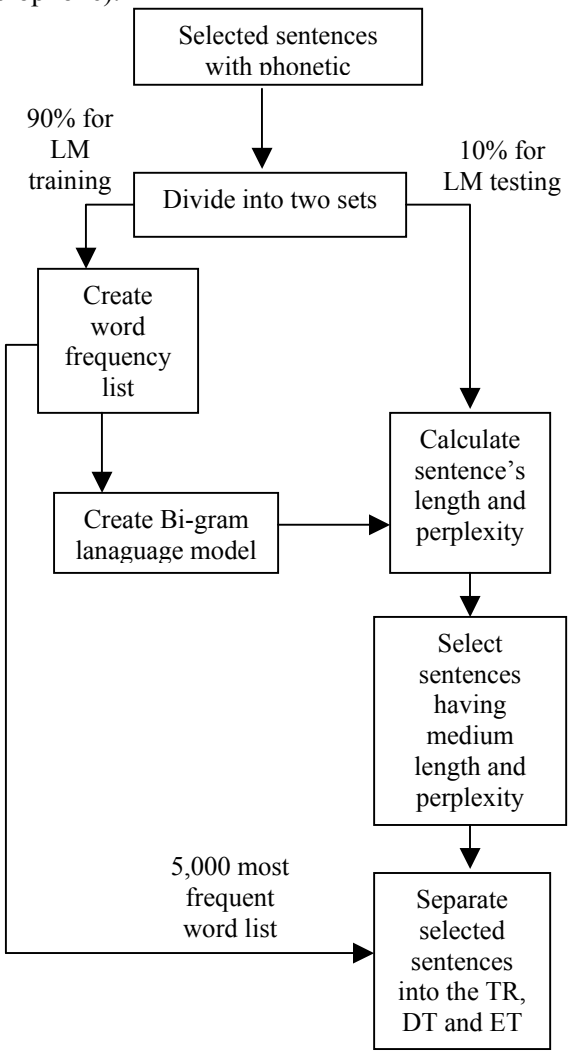

Figure 3: The selection of the TR, DT and ET sets.

Figure 4 shows the diagram of the instrument setup. Firstly, speeches are recording into a DAT tape with a sampling rate of 48 kHz and 16-bit quantizations. Recorded speeches were then played back into PC via an optical connection to a sound card. A waveform is then downsampled into 16 kHz (also use the prefilter to avoid aliasing). Note that a power supply is needed for the OF2 environment, since the TELEX headset has a condenser microphone.

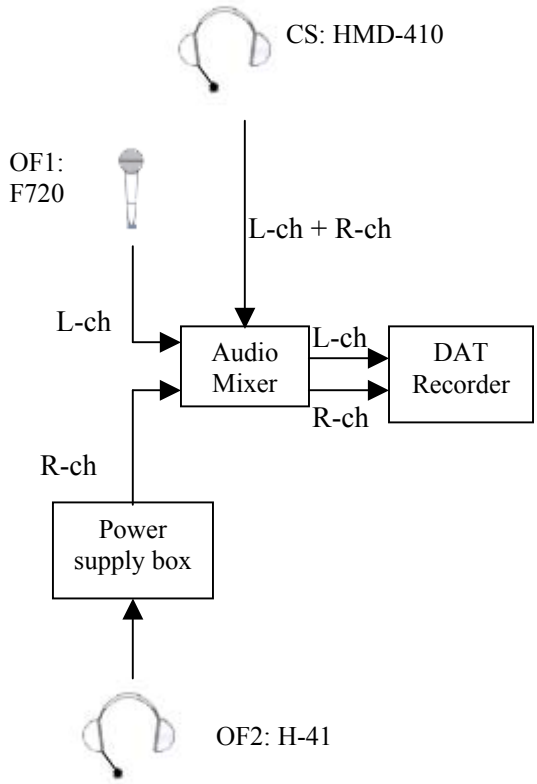

Figure 4. The setup of instrument.

2.4 Speaker information

As mentioned the number of speakers in the corpus is 248. Table 2 shows the age distribution of speakers in the corpus. However, this distribution can be relaxed; it can vary by as much as twenty percents. Most speakers are selected to be in their twenties—one of the age group which uses computer frequently. A speaker file information contains information such as: place id., speaker id., first name, last name, gender, age, height, weight and domicile, where place id. is P, M and N for PSU, MU and NECTEC groups.

Table 2. The age distribution of speakers in the corpus.

Age range (year)	No. speakers
0–18	50
18–25	100
25–35	74
> 35	24

2.5 Transcription and Lexicon files

Each sentence has its associated transcrition files. For the sentence in the PD set, those files are the word and phonetic transcriptions with time label. The format of those files are:
Word transcription:

<Start> <End> <Word><\n>

Phonetics transcription:

<Start> <End> <Phonetics> <Tone><\n>

where the <Start> and <End> denote the starting and ending samples of each word or phoneme, respectively, the <Tone> is a tone indicator taking a value of 0, 1, 2, 3 or 4, and <\n> is a newline character. Two special symbols are also used in these two files: /sp/ and /sil/ which indicate short pause and silent period, respectively.

For sentences in the TR, DT or ET set, they also posses the word and phonetic transcription files, but without time indicator. The formats of transcription files are the same as those in the PD set, only without the <Start> and <End> fields.

The Lexicon file contains information of 5,000 vocabularies of our corpus. The format of the information for each word is:

<Word> <Pos> <Pron> <\n>

where <Pos> is a comma-separated list of part-of-speech of word (denoted by <Word>). These Pos's were defined in (Sornlertlamvanich 1998). The field <Pron> is a sequence of phonetic transcriptions of that word.

2.6 Sentence distribution method

As depicted in Table 1, there are six different groups: PSU, MU#1, MU#2, NEC#1, NEC#2 and NEC#3. This section will explain how to distribute sentences into these sets.

One way to distribute sentences is to randomly assign them to each group. A problem of this method is that a sentence may be

assigned into only one or two groups instead of distributed equally among six groups. Here we present a method for distributing sentences. Although the method presented here cannot be claimed to be optimum, it is clearly better than the purely random method.

Since for the PD set each speaker has to utter 20 sentences and there are 248 speakers, the total number of uttered sentences is 4,960. Since there are 802 sentences in the PD set, each sentence will be uttered 6.18 times on average. Therefore, we set that a sentence will be uttered at least 6 times, and at most 7 times.

Because the total number of PD sentences for the PSU set is 2,000, the number of times a whole PD set can be read is 2 (2,000/802 = 2.49). By doing this to the other groups, one can easily conclude that a whole PD set can be read twice by the PSU group, once from the MU#1 group, once from the MU#2, and once from the NECTEC's groups (NEC#1, NEC#2 and NEC#3). Hence, one can assign two whole PD sets to be read by PSU group, one to the MU#1, one to the MU#2 and one to the NECTEC's groups. This accounts for 5 times a sentence can be read; therefore, each sentence can additionally be read by either once or two twices. Now the question is how to distribute these left-over sentences to those groups. The way we have adopted was to use the normalized dot product as a measure of how to add these sentences into each group. The normalized dot product between a vector $V = (v_1, v_2,\ldots, v_n)$ and $W = (w_1, w_2, \ldots, w_n)$ for an integer value n, is defined as

$$V \circ W \equiv \frac{\sum_{i=1}^{n} v_i w_i}{|V||W|}$$

where $|V|$ and $|W|$ are the $L2$-norms of V and W, respectively. The following is the details of algorithm.

Let

$$P = (p_1, p_2,\ldots, p_{1605}),$$

where p_i, i = 1, 2, …, 1,605, represents the number of occurrances of the i-th Biphone pair actually occur in PD set.

Also, let

$$S_j = (s_{j1}, s_{j2},\ldots s_{j802}),$$

where s_{ji} is the number of occurrances of i-th biphone pair of the j-th sentence and i = 1, 2, 3,…, 1,605.

For each group, let G_k be the vector of Biphone pair where k = 1, 2, 3 or 4. Here, k = 1 is for the PSU group, k = 2 for MU#1 group, k = 3 for MU#3 group, and k = 4 for all NECTEC's groups.

After adding two whole PD sets to PSU group, and one to MU#1, one to MU#2, one for NEC groups, it can easily see that

$$G_1 = 2P,$$

and

$$G_2 = G_3 = G_4 = P.$$

Hence, the dot products between the vector G_k, k = 1, 2, 3 and 4, and vector P are one.

Now, with k = 1, the PSU group, select the sentence which after adding it to the list of sentences in the PSU group results in the highest normalized dot product of G_1 and P. Mark sentence to indicate that it has been used once. Next, with k = 2, the MU#1 set, select the sentence which results in the highest dot product. Add that sentence to the MU#1 list. Then, check whether that sentence has been used before, if so delete the sentence from the list. Keep doing the same for k = 3 and 4, and then looping back to k = 1. Keep doing this until the overall number of sentences in each group is satisfied, i.e. 2,000 sentences for PSU group, 1,000 sentences for each of MU#1 and MU#2 groups, and 960 sentences for the NECTEC's groups.

Having performed the above step, we know how many times each sentence will be uttered in each group. The next step is to distribute these sentences into speakers of each group. Of course, one can easily assign each sentence to a speaker randomly, but the problem with this random method is that a speaker may end up reading the same phonetic sequence or even the same sentence many times. To correct this problem, the following scheme was proposed.

Let

$$H_{i,k} = (h_{1,i,k}, h_{2,i,k}, \ldots, h_{N,,i,k}),$$

where N = 1,605 and $h_{j,i,k}$ represents the number of j-th biphones of the i-th person in the k-th group. For each k-th group, assign a sentence from its poll that maximizes the normalized dot product between $H_{1,k}$ and P, and remove that sentence from the list if the sentence has been used more than its limit. Then do the same thing for the next person until the first sentence is assigned to every person in the group. Next, going back to the first person in the group and assign the second sentence to him/her. Keep

doing this until there is no more sentence left in the poll. Perform the same step for the other groups as well. Note that in the searching for the sentence to be added, we only search for sentence which is not already assigned to that person thus preventing the same sentence to be assigned to the same person more than once.

The same algorithm for assigning sentence to each person in each group is also used for the TR, DT and ET sets. In this case, in stead of using a vector of the number of biphones we form the vector of the number of words. The algorithm would be the same; therefore, will not be discussed any further.

2.7 Statistics of the PD set

This section gives interesting statistics from the corpus. Only the PD set statistics will be given, since the other sets are still in the development process.

The statistics of the number of biphones, syllables and words of sentences in the PD set are shown in Table 3. Those statistics are the minimum, maximum, mean and the standard deviation of the number of biphones, the number of syllables and the number of words.

Table 3. Various statistics of the number of Biphones, syllables and words for the sentence in the PD set.

	Number of Biphones	Number of Syllables	Number of Words
Min	9	7	1
Max	130	117	52
Mean	36.13	17.07	9.78
Std.	19.48	11.60	5.91

Table 4 shows the comparision between the theoretical biphones and the actual number of biphones occur in the corpus. Four categories of biphone pairs are listed: C_iV, VC_f, C_fC_i and VC_i, respectively, where C_i denotes an initial consonant or cluster consonant, V denotes a vowel, and C_f denotes a final consonant. As one can see, the number of actual Biphone pairs is about 37% lower than that of the theoretical one.

3 Speech Technology in Thailand

Apart from corpus development, speech recognition projects were developed at NECTEC. In this section, the discussions of these works will be covered briefly. For more details, interested readers are recommended to look upon references.

Table 4. The comparison between the theoretical number and the actual number of biphone pairs.

Biphone pairs	Num. of Theoretical Biphone pairs	Num. of actual Biphone pairs
C_iV	912	579
VC_f	288	152
C_fC_i	456	328
VC_i	912	556
Total	2,568	1,605

Speech transcription is a crucial part in developing corpus. However, the process takes a great deal of time and effort for manual transcription. To overcome this obstacle, we developed an automatic speech transcription (Tarsakul and Kanokphara 2002). The system works as follows. The input texts are sent to an automatic Grapheme-to-Phoneme (G2P) module (Tarsakul et al. 2001). Errors in transcription and dictionary generated from G2P are corrected automatically by our re-label training with automatic phonetic correction and short-pause insertion. By doing this, the transcription and dictionary for the corpus can be corrected during training. The experimental result shows an improvement of 3.69% over the training without phonetic correction whereas the result form the automatic segmentation is inferior to the that from hand label only by 0.59%.

Since there are many variations in Thai speaking style, some variations can greatly degrade recognition rate. A pronuciation variation approach to speech recognition system has been developed (Kanokphara et al 2002). The system is modified from the autmatic speech transcription described above. The system is designed to support those variations in the corpus. The phoneme variations used in this system is tree-based. With variation decision tree, large variations in pronuciations in training

can be implemented. However, training in this way, the acoustic models become weak to variations in testing data. The robustness of the model can be augmented by tying all models in the same variation group. This modified model is refered to as "pronunciation variation model".

Conclusions

In this paper, we have described a speech corpus for speech recognition. This corpus is going to be collected by NECTEC, PSU, and MU. The total number of speakers are 248. Details of interesting statistics from the corpus are also descriped. This corpus will be opened to public when it is finished, and it would be a great resouce for speech recogintion researchers. In addition, we have covered some of the works recently developed by NECTEC team.

Acknowledgement

The authors would like to thank Chai Wutiwiwatchai, who plays an important role in designing speech recognition corpus. The first author also gives thanks to all the supports of people at the RD-I division of NECTEC. Without their helps, I would be busy packing my stuffes instead of writing this paper.

References

Sornlertlamvanich, Virach, Naoto Takatoshi, and Hitoshi Isahara. Thai Part-of-Speech Tagged Corpus: ORCHID, *Proceedings of the first international workshop on east-asian language resources and evaluation*, pages 131–138, Tsukuba, Japan. 1998.

Luksaneeyawin, Sudaporn. 1993. Speech Computing and Speech Technology in Thailand. *Proceedings of the Symposium on Natural Language Processing*, pages 276–321, Bangkok, Thailand.

Shen, Jia-lin, Hsin-min Wang, Ren-yuan Lyu, and Lin-shan Lee. 1999. Automatic Selection of Phonetically Distributed Sentence Sets for Speaker Adaptation with Application to Large Vocabulary Mandarin Speech Recognition, *Computer Speech and Language*, 13:79–98.

Chai Wutiwiwatchai, Patcharika Cotsomrong, Sinaporn Suebvisai and Supphanat Kanongphara. Phonetically Distributed Continuous Speech Corpus for Thai Language, *LREC 2002*, to be published.

Rosenfield, R. The CMU statistical Language Modeling Toolkit and its use in the 1994 ARPA CSR Evaluation. Carnegie Mellon University. 1994.

Tarsaku, Pongthai, Sornlertlamvanich, Virach, and Thongprasirt, Rachod. Thai Grapheme-to-Phoneme using Probabilistic GLR Approach. *Proceedings of European Conference on Speech Communication and Technology*, vol. 2, pp. 1057–1060.

Tarsakul, Pongthai, and Kanokphara, Supphanat, A study of HMM-based automatic segmentation for Thai continuous speech recognition system. To be published in SNLP 2002.

Kanokphara, Supphanat, Tesprasit, Virongrong, and Thongprasirt, Rachod. Pronunciation Variation Approach to Speech Recognition System, submitted to ICSLP 2002.