Overview of Recent Activities in East Asia: Speech Corpora and Assessment

ITAHASHI Shuichi[*1], *KUWABARA Hisao*[*2], *DAWA Idomuso*[*3], *LEE Yong-Ju*[*4],
Virach Sornlertlamvanich[*5], *WANG Hsiao-Chuan*[*6], *Wang Ren-Hua*[*7]

*1 University of Tsukuba, Tsukuba, Japan: itahashi@milab.is.tsukuba.ac.jp
*2 Teikyo University of Science and Technology, Uenohara, Japan: kuwabara@ntu.ac.jp
*3 Waseda University, Tokyo, Japan: dawa@shirai.info.waseda.ac.jp
*4 Wonkwang University, Iri, Korea: yjlee@wonkwang.ac.kr
*5 NECTEC, Bangkok, Thailand: virach@nectec.or.th
*6 National Tsing Hua University, Hsinchu, Taiwan: hcwang@ee.ntu.edu.tw
*7 University of Science and Technology of China, Hefei, China: rhw@ustc.edu.cn

ABSTRACT

This paper reports the recent activities on speech corpora and speech input/output systems assessment in East Asia. It includes the reports on Chinese, Japanese, Korean, Mongolian, Thai, and Chinese spoken Taiwan. Section 2 describes Chinese activities on performance assessment of speech recognition. Section 3 mentions Japanese effort on speech synthesizer performance evaluation and some recent speech related projects. Korean activities on speech corpora are described in Section 4. Mongolian dialectal speech corpus is introduced in Section 5. Section 6 deals with speech corpora and assessment in Taiwan. Finally, Section 7 outlines speech research and corpora in Thailand.

1. INTRODUCTION

This paper describes some recent activities on speech corpora and speech input/output systems assessment in East Asia [1]-[4].

East Asian languages exhibit a wide range of characteristics which are unlike European languages: (1) they embody considerable variety arising from different language families; (2) they use different orthographic system, such as Chinese characters, Korean syllabic alphabet, and Japanese Kana alphabet; (3) they employ various systems of romanization. It is quite natural to suppose that there would be ways of processing these languages which are different from and more suitable than those adapted to European ones.

Section 2 describes Chinese activities on performance assessment of speech recognition. Section 3 mentions Japanese effort on speech synthesizer performance evaluation and some recent speech related projects. Korean activities on speech corpora are described in Section 4. Mongolian dialectal speech corpus is introduced in Section 5. Section 6 deals with speech corpora and assessment in Taiwan. Finally, Section 7 outlines speech research and corpora in Thailand. The sections 2, 4, 6 are based on the references [5], [15] and [17], respectively.

2. NATIONAL PERFORMANCE ASSESSMENT OF SPEECH RECOGNITION SYSTEM FOR CHINESE

2.1 Outline

Under the support of National Hi-Tech Project 863 and National Foundation of Natural Science of China, the national assessment of speech recognition system for Chinese has been regularly carried out since 1991. The task and target for the assessment could be summarized as: 1) to establish assessment criterions for evaluating scientifically speech recognition algorithms and system performance for Chinese; 2) to promote the research and development of Chinese recognition system, and guide the research direction of speech recognition for Chinese; 3) to inspect the research progress and results on Chinese speech recognition in a science way, and afford objective reference to the decision-making branch of government; 4) to promote the academic exchange, improve and perfect available algorithms and system of speech recognition for Chinese; 5) to establish standard Chinese speech database and corpus, achieve resource sharing. Considering the development level of speech recognition in China, different testing tasks and methods are defined in different stages [5].

It can be said that, the research of Chinese speech recognition in China grows up from the speaker-independent, limited vocabulary, isolated syllables to speaker-independent, large vocabulary, continuous speech, especially, from the viewpoint of technology, the language models with a quite scale and good performance have been successfully applied instead of acoustic model only, that is an essential advance appeared in the past years.

This section mainly introduces 1998 national performance assessment of speech recognition systems, which not only reflects the state-of the-art of speech recognition for Chinese in the mainland of China, but also manifests a more scientific and perfect speech input performance assessment for Chinese. In the remainder of this paper, the guideline to the 1998 national performance assessment of speech recognition system is described in details. Furthermore the USTC97 –A read Putonghua corpus is also introduced as the basis for the assessment. How to count the errors in the continuous speech recognition for Chinese and how to compare the system performance are discussed. Finally the test results are showed there for reference [6].

2.2 Guideline to 1998 Assessment

The latest national performance assessment of speech recognition was held in April 1998 at Beijing. The guideline to 1998 assessment with all appendixes was announced half year before the assessment. Also a Putonghua corpus USTC97 was available on request.

2.2.1 Task and Method

The systems to be tested must be able to accept continuous speech in large vocabulary and speaker-independent as input, meanwhile output the recognition results in the form of Chinese characters and Pin Yin strings. The testing includes two parts:

A. System performance test. Test data are composed of 480 speech samples, which are digitized continuous sentences uttered by 12 speakers (6 males and 6 females), 40 sentences per speaker. Corresponding to each input sample the tested system should output a text file of recognition result including the character strings and Pin Yin strings.

B. Algorithm test in the acoustic level (excluding the language model). Test data are composed of 120 phrases composed of 4 syllables without any meaning. There are 6 speakers (3 males and 3 females), 20 phrases per speaker. For each input sample the tested system should output a text file of recognition result (syllable strings without tones) including 5 candidates

2.2.2 Corpus

The full text of 1993's and 1994's of "The People's Daily" and with around 6oooo words. All the testing sentences are selected from "The People's Daily" and cover different areas. The testing samples are collected in a similar way to the USTC97 from speakers whose native places cover 6 provinces and 3 large cities including Beijing, Tianjing, Shanghai, and Hebei, Shandong, Anhui, and so on.

2.2.3 Specifications of Evaluation

1. Character correct rates
2. Error rates with the wrong deletion, insertion and substitution
3 Sentence error rates (for A)
4 Syllable correct rates, counted separately for first to fifth candidates (for B)
5 Total time period for recognition of the testing materials and the robustness

2.3 Testing Results

Based on the least errors principle [8], the determination of recognition results is actually a search process of whole-path. Except to the large computation load, the length of route varies due to the deletions and insertions, so the search becomes much complicated. A recursive realization of non-fixed-length-path search algorithm was proposed to make the determination of recognition results automatic [8]; the details are omitted here.

2.5 Activities in Hong Kong

Putonghua corpus, HKU99, was developed at University of Hong Kong. The HKU99 corpus consists of a total of 74 speakers from 8 major dialect regions in mainland China. Each speaker read two scripts in Putonghua, one (189～307 sentences) unique to each speaker and another (213 sentences) shared by all the speakers [21].

3. SPEECH CORPORA AND ASSESSMENT IN JAPAN

3.1 Organizing the Creation and Utilization of Speech and Language Corpora

Owing to the need to prepare a systematic, common framework for collecting, creating, storing, distributing and sharing language and speech data in order to secure progress in future research, they established the Linguistic Resources Sharing Initiative (LRSI). Consisting of 23 members interested in speech and language corpora, the LRSI held a preparatory meeting and a symposium in 1994. However, there were serious concerns about the availability of financial support. Recently, GSK was established; GSK is the acronym for a Japanese phrase which means "Language Resources Sharing Organization". The initiating symposium was held in May 1999. This new effort is expected to effectively supersede LRSI.

3.2 JEIDA (Japan Electronic Industry Development Association)

Evaluating synthetic speech quality is an important part of promoting the research and development of speech synthesis. The Speech Input/Output Systems Expert Committee of JEIDA began reviewing the performance evaluation of speech synthesis-by-rule in fiscal 1992, and published a provisional version of "Guidelines for Speech Synthesizer Evaluation Methods" in fiscal 1994 [1].

Recently, a revised version was issued, with many examples of evaluation lists added, so that the guidelines can be directly applied to the performance evaluation of speech synthesizers [9]. Along with these guidelines, they prepared a "Commentary on the Guidelines for Speech Synthesis System Performance Evaluation Methods" to explain to users the background against which the guidelines were created [9].

They also issued "Standard of Symbols for Japanese Text-to-Speech Synthesizer" [11]. Its objective is to define the symbols that can be used in common for various applications and services utilizing Japanese text-to-speech synthesizers. It also aims to facilitate the system developers and the users and to extend utilization of speech synthesis technologies. Their English versions are in preparation.

3.3 Speech Related Projects in Japan

3.3.1 ATR-SLT (ATR Spoken Language Translation Research Laboratories)

ATR Interpreting Telephony Research Laboratories started its activities on speech research in 1986. ATR Interpreting Telecommunications Research Laboratories succeeded it in 1992. ATR Spoken Language Translation Research Laboratories started its work in March 2000.

The ATR database includes isolated words, phonetically balanced sentences, dialogues, spontaneous speech, and speech under noisy conditions. Most of these data have been manually segmented into phonemes and hand-labeled [12].

They also have a bilingual speech corpus of Japanese and English. It contains conversations between English and Japanese speakers through a human interpreter [13]. They are travel conversations between a tourist and receptionist of a hotel. This corpus is scheduled to be released in 6 CD-ROMs for research purposes.

Their latest corpus consists of words, sentences, and dialogues spoken by more than 3000 speakers from 47 cities in Japan. The words are selected from dictionaries which cover all phonemes currently used in Japanese, and the sentences are selected from the above Phonetically-Balanced 503 sentences. The dialogue data are designed for an appointment scheduling task in which two speakers collaboratively plan a visit to each other. In all, there are 31,589 words (13.8 h), 112,660 sentences (127.6 h), and 1,888 dialogues (53.4 h). This database is also expected to be released for research purposes [14].

3.3.2 RWC (Real World Computing) Project

In 1992 the Japanese government formulated a 10-year-program entitled the "Real World Computing Partnership" with a view to establishing new information processing systems. In relation to one of the research topics, the Real World Computing (RWC) project has produced speech and text corpora and made them available to the researchers [1].

The spoken dialogue corpus consists of dialogues between car dealers and customers, those between travel agents and their customers. Professional dealers and agents were employed to add reality in the conversations. To date, 60 samples of dialogues have been recorded, 48 of which were filed onto CD-ROMs including about 10 hours of speech waveforms with transcriptions and related-labeling. The dialogues are almost completely spontaneous but their acoustic quality is kept relatively high.

The read-speech corpus consists of read speech of broadcast manuscripts of financial news spoken by professional announcers. The read text has been selected from the manuscripts of actual news during recent 6 months provided by the NHK (Japan Broadcasting Corporation). The total amount of speech data is about 4 hours.

3.3.3 New Speech-Related Projects

Several new projects related to speech research have been launched recently.

The first is a new joint project initiated by two national organizations, the Kansai Branch of the Communications Research Laboratory (CRL) and the National Language Research Institute (NLRI), and sponsored by the Science and Technology Agency (STA). The title of the project is "Construction of 'Spoken Language Engineering' based on evaluation of the linguistic and paralinguistic structure of spoken language". Its main goal is to create a linguistic model of spoken language based on large-scale, spoken language corpora, which are to be developed in the project.

The second is a new project on Integrated Acoustic Information Research lead by Prof. F. Itakura of Nagoya University. It aims to integrate different aspects of human-sound relation by investigating: (1) how man can spatially localize sounds, (2) how to analyze and synthesize sounds, (3) how to recognize speech sounds, (4) how to understand spoken language by machine, and (5) how man perceives the acoustic environment by sound. In this project, academia and industry will work together intensively to develop and evaluate advanced sound systems. This project plans to create large-scale speech corpora and acoustic databases to be used in the project and also by the speech research community.

The budget allocated for each of the above two projects is one billion yen ($US 8 million) over 5 years.

Related Web Pages:

http://www.milab.is.tsukuba.ac.jp/corpus/corpora-e.html
http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html
http://www.milab.is.tsukuba.ac.jp/o-cocosda/
http://www.ntt-at.com/products/ivl/cont/
http://www.atr.co.jp/products_e/
http://www.jeida.or.jp/index-e.html
http://www.aiair.coe.nagoya-u.ac.jp/
http://tanaka-www.cs.titech.ac.jp/gsk/gsk-eng.htm
http://www.isip.msstate.edu/projects/jeida/

4. SPEECH CORPORA AND ASSESSMT IN KOREA

4.1 Korean COCOSDA and Its Activities

For efficient construction of speech corpora and its practical use, a study group was established a few years ago. It expanded into a discussion group of speech engineers and phoneticians to establish a new organization for systematic activities related to speech assessment and speech corpora in 1999. The organization is composed of working groups including speech recognition, speech synthesis and speech corpora and central coordination committee which manages them [15].

Until now Korean COCOSDA is characteristic of a study group but henceforth it is planned to coordinate related institutes to take part in it so as to establish recommendation with public confidence. Furthermore it is also planned to construct speech corpora available in common and to have speech corpora being used inside the institutes at present for common use. For such purposes, the experiences in other countries will be of help to Korean community.

4.2 International Construction of Korean Speech Corpora

Korean is used in South and North Korea, Korean self-governed districts in China and Korean residential districts in USA, Russia and Japan etc. And Korean used worldwide is the same in broad sense. But they are different in detail because it has been used by different norm respectively for half a century according to different ideology. So a lot of things are remained to be unified for integrated information processing. And climbing over ideological obstacles, many scientists tried to research to solve the problems. Since 1994, Hangul code, order of consonants and vowels, arrangement of computer keyboard and the related terms have been discussed by Korean scientist of South Korea, North Korea and China for 3 years. Participating organizations were Korean Language Information Society in South Korea, Chosun Computer Center, universities and research centers in North Korea, and Korean Language Information Society in Korean self-governed districts in China. And the results were published at International Conference on Computer Processing for Korean Language (ICCKL) held in Yanbian, China every year.

Additionally, in recent years, several subjects about spoken language were suggested in second phase and remained to be discussed. Collecting the spoken language data which is used in many different regions respectively will be useful not only for engineering usage but for study of phonetics. This plan was suggested in 1995 by Prof. Y-J Lee and the discussion started concretely in August 1999. Definite collecting target and method was discussed sufficiently by scientists in each region, but in the beginning how many changes and varieties Koran Languages have under regional or circumstantial differences will be seized. In other words, it is no doubt that the regional gap in the first target but it is expected that it will be expanded into the dialect. In the first step, speech scientists in South and North Korea and China will participate as the main group and thereafter scientists on other regions (Russia and Japan) etc.) will do. And they respectively are responsible for getting together investigators and speakers, recording and ensuring an estimate needed. For this purpose, it is planned to make representative liaison office in South and North Korea and China with investigators as central figure. In the beginning, it is planned to make triple liaison office in China and discuss by using internet. A working meeting will be opened three or four times in a year in International Conference on Computer Processing of Korean Language (ICCKL). And the data will be recorded in South and North Korea and Korean residential districts in China. Each raw data will be collected and arranged in South and North Korea and China. The raw data will be computerized and be converted into a database in each region if it is possible in view of technical and environmental conditions , but if it is impossible to do so, South Korea will do that. There will be many problems because of geographical and ideological differences between each region. Therefore Koran COCOSDA want to be a practical window to solve the problems.

4.3 Brief Introduction of Korean Text Corpora

In the early days, Korean text corpora were made to select a vocabulary entry for Koran language dictionary with universities as leaders but nowadays they are being constructed as basic materials for machine translation, information retrieval and NLP supported by the government.

4.3.1 Yonsei University Corpus

(http://lexeme.yonsei.ac.kr)

Yonsei University Corpus have been constructed to select raw materials for integrated Korean language dictionary since 1987. Center for Language and Information Development in Yonsei University is in charge of it and about 45 million words were collected. Header information consists of bibliographical information (including titles, authors, years and publishers) and decimal classification numbers. And the starting and ending point of the text is written and consist of SGML tag.

4.3.2 Korea University Corpus

Korea University Corpus was constructed in the size of 14 million words (KOREA-1 Corpus) with Institute of Korean Culture in Korea University as a leader in 1995. They are now expanding it to 100 million words. The norm of balance is written and spoken language, and the written language consists o pattern of materials (newspapers, periodicals and books) and contents (realistic and idealistic description).

4.3.3 KAIST Text Corpus

KAIST Corpus has been constructed to be used for developing various language processing tools in Korea Advanced Institute of Science and Technology since 1994. About 70 million words have been collected as a raw source, about 15 million words as a tagged corpus and about 4 thousand sentences as a corpus with grammatical information. Basically a user can extract the information needed by the connection with TDMS (Text and Dictionary Management System) in KAIST corpus. TDMS can manage the text based on SGML and electronic dictionary. The tag using TDMS is composed of file information, bibliographical information and input information; and classification information is added for classification of corpus and the balance.

The primary purpose of tagged corpus is to provide training data for construction of automatic POS tagger. And secondly, if the large tagged corpus is supported, it can be used in the linguistic study. Tagged corpus of 15 million words is constructed, though there are some errors during the process of construction in large scale.

5. A DATABASE FOR MONGOLIAN SPEECH CONSIDERING DIALECTAL CHARACTERISTICS

Mongolian, one of the most important spoken oriental languages, is used mainly in Mongolia, parts of China and Russia, and a couple of other neighboring countries. The speaking population is about 7.5 millions [16].

Because of its historical and geographical background, like some other languages, Mongolian has several dialectal variations in its linguistic, phonetic and graphic expressions. Although people speak the same language, it could happen sometimes that they cannot understand at all by listening or reading the different dialects. This becomes a serious problem in cultural, educational and economical developments in their countries. Furthermore, unlike the other major languages, the basis of linguistic and phonetic research has not been investigated well yet.

In our study, we basically classify the Mongolian dialects into three fundamental groups, i.e., Khalha(Mongolian), Chhar(Inner Mongolian, China), and Oirat(Kalmyk, A.R. USSA, Xin Jiang, China), which are used widely by large population, and collect and build the database in the department of information and computer science, Waseda University, from 1995. we have so far collected 10.024s speech and built several types of database.

Note that the meaning of a word (dialect), here has not only the difference of their spoken mode but also the difficulty of the language communication because of their graphic and phonetic systems.

They used the database for the acoustic analysis in dialectal speech, classifying and discrimination analysis of dialects , recognition of dialectal speech and converting process from spoken to written for Mongolian speech, etc. The result showed that the database is intuitive and effective. also it will be expected to use in linguistic, phonetic, dialectal, and educational research.

Generally, it is important to have a large scale database in research areas like phonetics, speech analysis, speech recognition, language translation, etc. Needless to say, such database should have the strategy in addition to its quantity and quality. They are going to continue collecting more data from various regions to expand the scale of the database.

6. SPEECH CORPORA AND ASSESSMENT IN TAIWAN

The research of Mandarin speech processing on Taiwan is in its turning point. Several systems have been developed and even commercialized. Now they continuously face the competition from research groups of foreign countries, especially those of US and European companies. It is the time for researchers in Taiwan to think about the future [17].

6.1 MAT Project

MAT(Mandarin speech across Taiwan) project started from 1995 and was sponsored by the National Science Council, Taiwan, ROC. This three-year project set its goal to generate a speech database of 5000 speakers. Besides, it aimed at setting the standard format for speech data collection [18]. Nine stations for speech data collection were set up around Taiwan. A speaker can input speech data by dialing up to one of the speech data collection stations, following the voice instruction, and speaking to the telephone handset.

The spoken materials are designed for speech model generation and speech recognition. The materials were extracted from two text corpora created in Academia Sinica [19]. They include 77,324 lexical entries and 5,353 sentences. Forty sets of speech materials were produced to generate 40 prompting sheets. The contents five subsets;

Sub-Database	Prompting Item No.	Speak Style	Description
MATDB-1	1 - 9	Spont.	short answer statements
MATDB-2	10 - 14	read	numbers pronounced in five different ways
MATDB-3	15 - 26	read	Mandarin syllables
MATDB-4	27 - 56	read	words of 2 to 4 syllables
MATDB-5	57 - 66	read	phonetically balanced sentences

In this three-year MAT project, more than 7000 speakers have input their speech data through telephone systems. Several speech databases were produced for different purposes.

MAT-800
This is a preliminary database containing speech data collected from 800 speakers (424 males and 376 females) in first year data collection. No screening mechanism is applied to discard those files with poor signal quality or bad recording conditions. This database is only for the self-testing and the evaluation of speech data collection systems. There is also carefully examined corpora **MAT-160** by 160 speakers and **MAT-2400** by 2400 speakers. **MAT-400** is a selected subset of MAT-2400 including 400.

MAT-2000
This is a product of the joint validation project conducted by Association of Computational Linguistics and Chinese Language Processing (ACLCLP) and Philips Innovation Center-Taipei (PICT) in 1999 [22]. Its source database is MAT-2400. Two versions are available now. The version of MAT-2000Edu is for education and research use only and the version of MAT-2000Com is for commercial use.

MAT-2500Ext

Sub-Databases	Prompting Item No.	Description
MATDB-6	67 - 68	Numbers
MATDB-7	69 - 72	Special terms
MATDB-8	73 – 82	Words of 2 to 4 syllables
MATDB-9	83 - 90	Sentences

MAT-2500Ext are the speech files of those extra materials provided by 2573 speakers. Since the speech materials are different from those previous MAT databases, it allows the use as testing data for speech recognition. A validation project of MAT-2500Ext is undergoing. This is a collaborated project by Computer & Communication Laboratories of ITRI and ACLCLP.

6.2 Assessment of Speech Recognition

Techniques

The assessment of speech recognition techniques is designed for promoting the research level and the use of common databases. The assessment program was conducted by the ACLCLP in 1998 and 1999.

6.2.1 First Assessment (June 1998 – April 1999)

Topics of Assessment

Three topics of speech recognition assessments are designed:

(1) Continuous Mandarin syllables recognition

Only base syllables are recognized. For each input utterance, the recognition result is a Pinyin transcript without tone indexes. MAT-160 is provided for training speech models, but the participants can also use other databases as they can get.

(2) Connected digits recognition

NUM-100 is provided for training the digit models.

(3) Isolated digit recognition

The recognition algorithm should be capable to reject non-digits. NUM-100 is also provided for this task.

Databases

All the speech data are recorded in 8 kHz sampling rate with 16 bits per sample.

(1) Databases for model training

MAT-160: Telephone speech of 160 speakers.

NUM-100: Microphone speech of 100 speakers.

(2) Databases for self testing

Test-500s: telephone speech. It contains the speech files of 200 words (50 single syllable words, 50 2-syllable words, 50 3-syllable words, and 50 4-syllable words) and 300 sentences.

Test-500c: telephone speech. It contains the speech files of 200 4-digit numbers and 300 7-digit numbers.

Test-500d: noisy speech. It contains the speech files of 400 isolated digit utterances and 100 non-digit (isolated syllable) utterances. Five levels of white noise signal, say SNR equal to 10 dB, 15 dB, 20 dB, 30 dB and no noise, are randomly added to the speech signal to generate noisy speech.

(3) Databases for final testing

Test-1000s, Test-1000c, and **Test-1000d** are for final testing. They are similar to the corresponding Test-500 databases except the size is doubled.

Assessment Procedure

The procedure of speech recognition assessment is designed in two phases.

Phase #1: Preliminary test

The training speech database (MAT-160 or NUM-100), test database (Test-500s, Test-500c, or Test-500d), and a reference text file (Reference-500s, Reference-500c, or Reference-500d). After six months to develop its speech recognition systems, each participant team submits the resulting file for the test database of 500 utterances.

Phase #2: Final test

When the format of resulting file is examined and proved to be correct, a test set of 1000 speech files (Test-1000s, Test-1000c, or Test-1000d) is handed to the participant team on a scheduled date. The resulting file is generated and returned immediately. The final examination is to evaluate the performance and calculate the recognition rate.

6.2.2 Second Assessment (September 1999 – August 2000)

Topics of Assessment

The topics of speech recognition assessment are similar to those in the first assessment except there are two categories for each topic; (A) constrained training data and (B) unconstrained training data. For the constrained training, only the provided training databases, MAT-160, MAT-400 and NUM-100, can be used.

Databases

The data sizes have been increased and the speech materials are changed.

Test-1500s: telephone speech. It contains the speech files of words and sentences.

Test-1500c: telephone speech. It contains the speech files of 1 to 7 digits numbers.

Test-1200d: noisy speech. It contains the speech files of 800 isolated digit utterances and 400 non-digit (isolated syllable) utterances. White noises are randomly added to the speech signal to generate noisy speech.

6.2.3 Test Results

The results show that the improvement of syllable recognition is obvious. For category B in second assessment, the syllable recognition rate was 88% (12% improvement). For isolated digit recognition, the accuracy dropped which came from the increase of test data size and variety of noise conditions.

6.3 Other Speech Corpus

The recently collected speech data are the microphone speech provided by 300 speakers. This is a project of the collaboration work by National Taiwan University, National Cheng Kung University and National Chiao Tung University. The goal of this data collection is to generate database for the development of dialogue systems. The produced database is named TCC-300. The verification is still undergoing.

7. SPEECH RESEARCH AND CORPORA IN THAILAND

Thai has its own unique characteristic. In writing Thai, there is no explicit marker for separating sentences, phrases and words [20]. Regarding the speech, Thai, like Chinese, is a tonal language. The tonal perception is important to identify the meaning of speech utterance. These unique characteristics need some suitable language specific algorithms to handle. In Thailand, the current research in speech technology can be divided into 4 major fields: (1) speech analysis (2) speech and speaker recognition and (3) speech synthesis. Most of the research topics in (1) have been done by the linguists focusing on the basic study of Thai phonetic and phonology. The study provides a significant background for the research in the others.

7.1 Research on Speech Recognition

The research in speech recognition became significant since early 80's. However, most research topics were dedicated to language study rather than system development. The research topics can be roughly classified into 3 groups. First is the investigation of isolated word-based recognition including isolated digits and polysyllabic words. Various recognition engines as well as speech features were continuously improved. It was reported that during the past 5 years the Digital Signal Processing laboratory, with the collaboration with Linguistics laboratory, Chulalongkorn University, had successfully developed an isolated word-based recognition system using the suitable speech features i.e. Mel-frequency cepstral coefficient (MFCC) with several well-known engines e.g. neural network with fuzzy techniques, and hidden markov model (HMM) [23]. Another research work was conducted in Computer Engineering department, King Mongkut Institute of Technology Ladkrabang, proposing the use of Karhunen-Loeve transformation in 1999.

Second is phoneme-based continuous speech recognition with limited vocabulary size. One of the reported approaches is to recognize isolated phonemes which are segmented from continuous speech. The research aims at finding optimal speech features and recognition engines. The other approach becomes a state-of-the-art recognition system. It is the utilization of acoustic and language model under an efficient search algorithm as successfully implemented in many other languages. There is only one report, in this year, on continuous speech recognition using finite state network as language model and continuous-HMM as acoustic model in recognizing Thai inquiry sentences in spoken dialogue system [24]. Currently, there is a project of Thai large vocabulary continuous speech recognition (LVCSR) in the National Electronics and Computer Technology Center (NECTEC), aiming at developing a speech dictation system.

Third is the research on some task specific systems e.g. Thai tone recognition, which has been deeply researched and continuously reported by Potisuk, S. and Thubthong, N., i.e. a syllable detection system on connected speech, and other prosodic researches on Thai. These related research topics are significantly necessary for speech synthesis and recognition.

7.2 Research on Speaker Recognition

The Digital Signal Processing laboratory, with the collaboration with Linguistics laboratory, Chulalongkorn University, has reported a text-dependent speaker identification system using MFCC and HMM in 1999 and a further progress on speaker verification system. Another significant engine of text-dependent speaker identification has been proposed by NECTEC. Current progressive system reported in this year uses either DTW or Gaussian mixer model (GMM) with MFCC speech feature [25].

7.3 Research on Speech Synthesis

The research in speech synthesis does not attract Thai researchers as much as the speech recognition. The first well-known text-to-speech synthesis system for Thai is developed by Luksaneeyanawin, S. and et. al. in 1992 [26]. This system is based on waveform concatenation of demisyllable unit. It includes text analysis which parses an input string into a syllable sequence then groups the syllable sequence to construct word and phrase sequences respectively. NECTEC later developed a system based on the demisyllable concatenation approach [27]. It differs from the former system in the text analysis, which is done in the top-down fashion, and the prosody generation is also considered as an important part of the system. Other research works are dedicated to some specific issues rather than the whole system.

In addition, there are some research works in letter-to-sound conversion using a statistical model and decision tree. Also, NECTEC has applied the probabilistic POS tagging model to solve the sentence segmentation problem. The model is implemented in the text analysis part of the NECTEC's Thai TTS system.

7.4 Speech Corpus

One serious problem on Thai speech technology development is the lack of large speech corpora. For the speech recognition task, especially the LVCSR, a large scale speech corpus is needed. NECTEC is now preparing to create a 5K vocabulary speech corpus including phonetically-balanced set. Moreover, for the speech synthesis task, a separated clean speech corpus is now designed for prosodic analysis for improving the naturalness in speech synthesizer.

8. CONCLUSION

This paper summarized some recent activities on speech corpora and speech input/output systems assessment in East Asia during 1999

We expect participation from Singapore and India next time.

9. REFERENCES

[1] S. Itahashi, ed. : Proc. First International Workshop on East Asian Language Resources and Evaluation (EALREW98), Tsukuba, Japan (1998).

[2] S. Itahashi, "Foreword to Special Issue on Speech Database/Assessment for Oriental Languages," JASJ, Vol.(E)20, No. 3, pp. 159-161 (1999).

[3] S. Itahashi, "On Recent Speech Corpora Activities in Japan," JASJ, Vol.(E)20, No. 3, pp. 163-169 (1999).

[4] L. S. Lee, ed. : Proc. Second International Workshop on East Asian Language Resources and Evaluation (EALREW99), Taipei, Taiwan (1999).

[5] R. H. Wang, "National Performance Assessment of Speech Recognition System for Chinese," Proc. EALREW99, Taipei, pp. 41-44 (1999).

[6] R. H. Wang, et al, "The assessment of continuous speech recognition for Chinese". Progress on Intelligent Computer Interface and Application, pp.157-162, 1997. (In Chinese)

[7] Yiqing Zu, “Sentence design for synthesis and speech recognition”, Proc. of 5th European Conference on Speech Communication and Technology, Vol. I2, pp. 743-746, 1997.

[8] Yongsheng Teng, “A study on some elementary problems in Chinese information processing”, Master thesis of USTC (1998).

[9] JEIDA, “Guidelines for speech synthesis system performance evaluation methods,” (in Japanese), JEIDA-G-24-2000,Japan Electronic Industry Development Association (2000).

[10] JEIDA, “Standard of symbols for Japanese Text-to-Speech synthesizer,” (in Japanese), JEIDA-62-2000, Japan Electronic Industry Development Association (2000).

[11] S. Itahashi, “Guidelines for Japanese Speech Synthesizer Evaluation,” Proc. LREC2000, Athens, pp. 655-660 (2000).

[12] T. Takezawa, T. Morimoto, and Y. Sagisaka, “Speech and Language Databases for Speech Translation Research in ATR,” Proc. First International Workshop on East Asian Language Resources and Evaluation (EALREW98), Tsukuba, Japan, pp. 148-155 (1998).

[13] T. Matsui, M. Naito, H. Singer, A. Nakamura, and Y. Sagisaka, “Japanese Spontaneous Speech Database with Wide Regional and Age Distribution,” Proc. Eurospeech99, Budapest, pp. 2251-2254 (1999).

[14] T. Takezawa, “Building a Bilingual Conversation Database for Speech translation Research,” Proc. EALREW99, Taipei, pp. 17-20 (1999).

[15] Y. J. Lee, “Some Activities of Korean COCOSDA and Korean Text Corpora,” Proc. EALREW99, Taipei, pp. 9-12 (1999).

[16] I. Dawa, S. Okawa, and K. Shirai, “Design and assessment of Mongolian dialectal speech input sysytem,” Proc. EALREW99, Taipei, pp. 45-48 (1999).

[17] H. C. Wang, “Speech Research Infra-Structure in Taiwan – From Database to Performance Assessment,” Proc. EALREW99, Taipei, pp. 53-56 (1999).

[18] H. C. Wang, “MAT – A project to collect Mandarin speech data through telephone networks,” Computational Linguistics and Chinese Language Processing, vol.2, no. 1, pp. 73-90 (1997).

[19] C. Y. Tseng, “A phonetically oriented speech database for Mandarin Chinese,” Proc. ICPhS95, Stockholm, pp. 326-329 (1995).

[20] T. Charoenporn, A. Chotimongkol, and V. Sornlertlamvanich, “Automatic romaniztion for Thai,” Proc. EALREW99, Taipei, pp. 137-140 (1999).

[21] Q. Huo, and B. Ma, “Training material considerations for task-independent subword modeling: Design and other possibilities,” Proc. EALREW99, Taipei, pp. 85-88 (1999).

[22] H. C. Wang, F. Seide, C. Y. Tseng, and L. S. Lee, “MAT2000 – Design, collection, and validation on a Mandarin 2000-speaker telephone speech database,” to appear in ICSLP2000, Beijing, 2000.

[23] S. Jitapankul, S. Luksaneeyanawin, V. Ahkuputra, E. Maneenoi, S. Kasuriya, and P. Amornkul, “Recent advances of Thai speech recognition in Thailand”, *Proceedings of the 1998 IEEE Asia-Pacific Conference on Circuits and Systems (APCCAS)*, pp. 173-176, 1998.

[24] W. Kasemsiri, C. Kimpan, and R. Kongkachandra, “Refined language modeling for Thai continuous speech recognition”, *Proceedings of the 4th Symposium on Natural Language Processing (SNLP)*, pp. 252-262, 2000.

[25] C. Tanprasert, and V. Achariyakulporn, “Comparative study of GMM, DTW, and ANN on Thai speaker identification system”, Forthcoming in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 2000.

[26] S. Luksaneeyanawin, et al., “A Thai text-to-speech system”, *Proceedings of 4th NECTEC Conference*, pp.65-78 (in Thai).

[27] P. Mittrapiyanuruk, C. Hansakunbuntheung, V. Tesprasit, V. Sornlertlamvanich, “Improving naturalness of Thai text-to-speech synthesis by prosodic rule”, Forthcoming in *International Conference on Spoken Language Processing (ICSLP)*, 2000.