Speech Information Technology & Industry Promotion Center in Korea: Activities and Directions

Yong-Ju Lee*, Bong-Wan Kim*, Yongnam Um*,

* Speech Information Technology & Industry Promotion Center
Wonkwang University, 344-2, Sinyong-dong, Iksan, Chonbuk, Korea
{yjlee, bwkim, umyongnam}@sitec.or.kr

Abstract

Speech technologies have developed substantially through the research and development of academia, industry and institute in Korea. Recently speech has featured as user interface in electronic products, computer, telecommunication, etc. As practical uses of speech technologies such as speech recognition and synthesis increase, difficulties arise in academia and industry with speech corpora and assessment. Thus Speech Information Technology & Industry Promotion Center (SITEC) was founded in 2001 under the auspices of the Ministry of Commerce, Industry and Energy to help solve the common difficulties in the field and to manage creation and distribution of speech resources in Korea. As the name indicates, the Center will work as the center for systematic national coordination of industries, schools, and associations. We will introduce the SITEC and present its current activities and future plans.

1. Introduction

Speech is the most efficient and effective means of communication between humans. Research on speech has been done in linguistics, phonetics, phonology, physiology, etc. as well as in speech engineering. Since 1970's technologies in computer, semi-conductor, and signal processing have progressed at a rapid speed, and much effort has been made to put the results into practical use. Recently speech has featured as a user interface in electronic products, computers, telecommunications, etc. We have experienced a significant growth in various speech technologies, and more and more products and services which use the technologies are being developed for everyday life. As practical uses of speech information technologies such as speech recognition and synthesis increase, difficulties arise in academic and industrial areas with respect to speech information technology, as with speech corpora for research and development and assessment for recognition and synthesis systems. Thus, Speech Information Technology & Industry Promotion Center (SITEC) was founded as a consortium in May 30, 2001, and Wonkwang University was selected as the host for the Center. We will introduce the SITEC and present its current activities and future plans.

2. Introduction to SITEC

Speech information technologies such as speech recognition, speech synthesis and speaker recognition/ verification stand out as one of the 10 most promising technologies in the 21st century. The world market for the products which use these technologies are expanding at a rapid speed. In Korea there are 20 or odd companies which provide these technologies and more than 100 companies which use them to develop their products. A further rapid growth is expected in the field of developing applications since speech information technologies can be applied to various industrial fields.

However, there are three issues for speech information technology industry in Korea: (1) industrial environments, (2) technology development, and (3) marketing. Current domestic industrial environments are characterized by small sizes of the companies, lack of cooperation among the companies, institutions, academia, and insufficient governmental policies. Current states of essential and primitive technology, number of specialized persons, environments for exchange of technology and industrial information, databases for common use, objective assessment of speech information technologies and evaluation of application performance, etc. are not sufficient for technology development. SITEC was founded in this context to foster speech information technology industry in Korea.

2.1. Its goal and activities

The ultimate goal of the Center is to boost speech technologies to the world level by helping solve the common difficulties in the field. The Center provides guidelines for objective performance assessment, hosts contests for performance assessment, participates in international activities for standardization, and encourages competition between companies through common benchmarking testing of systems.

It creates speech corpora and develops assessment technologies to support research and applications, and trains specialists. It aims at constructing infrastructures for information, technology and environment. Works for the three fields of infrastructure are detailed below.

2.1.1. Infrastructure for technology development

- Creation and distribution of speech copora

1. To predict the demands consistently and to produce and distribute speech corpora consistently depending on the stages of development of industrial technologies
2. To resolve technological and legal (copyright, etc) problems with production and distribution (sharing)
3. To encourage sharing of individually constructed speech corpora
4. Effective and stable acquisition or collection of domestic and foreign speech corpora through foreign cooperative institutions, overseas cooperative sites and domestic cooperative sites.

- Attainment and distribution of standards for evaluation of performance.

1. The Center encourages standardization of products and technologies in cooperation with the forum for standardization, and assist

partnership and joint development among the companies through standardizing the products and provides guidelines for objective performance assessment, hosts contests for performance assessment, participates in international activities for standardization, and encourage competition between companies through common benchmarking testing of systems.

2.1.2. Infrastructure for information

- Database on information about technologies Construction of database on overseas trends concerning speech information technologies through constant survey and compilation. It provides information about technologies and market trends through the webzine. It also constructs a database of papers on information about technologies.
- Database on the companies and persons specialized in the field

1. To construct database on information about the companies specialized in speech information technologies by fields
2. To construct database on persons with masters or higher degree working in the companies, schools or institutions

- On-line and off-line training about technologies

1. Development of the textbooks about advanced technologies
2. Development of educational contents based on the textbooks On-line training
3. To share the equipments for speech analysis, etc. in the Center

- Seminars and training sessions

1. Co-hosting of seminars and training sessions based on the textbooks about advanced technologies
2. To provide training sessions

- Newsletter Publishing

1. Regular publication of newsletters about the information on technology, market trends, industrial trends, etc.
2. Constant supply of the information through the webzine for industrial information

2.1.3. Infrastructure for industrial environment

- Assistance to joint researches among companies, schools and institutions

1. To enhance synergic effects among the companies by encouraging and supporting meetings among them
2. To contribute to the wide expansion of speech information technologies by supporting meetings among the companies
3. To encourage meetings among the companies, schools and institutions to find topics for joint researches and to exchange technologies, providing assistance to the researches

- Assistance to the academic activities of affiliated societies
- Assistance to the publicity of industrial information

1. To assist various events for the purpose of enhancing consumers' understanding
2. To assist SpeechTech Expo in collaboration with the association
3. To co-host the contest for speech information technologies with associations
4. To assist joint advertisements of the companies that are concerned with speech information technologies

- Assistance to participating in overseas expos and exhibitions

1. To distribute information on overseas expos and exhibitions to the concerned companies, encouraging them to participate, and describing reports on the events through the webzine

2.2. Resources and organization

Korean Speech Information Technology Industry Association and ten companies which include Samsung Electronics Co., Ltd. and LG Electronics Inc. are institutions or companies which subscribe to the Center.

The organization of our Center is shown in the chart. There are seven full time staff members, and there are committees, subcommittees and expert committees. Each committee consists of 30 or so professors and specialists with Ph.D.

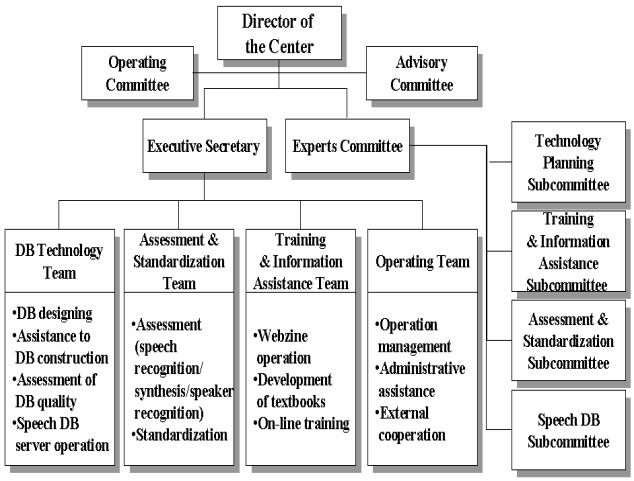

Figure 1: Organization of SITEC

3. Current activities and future plans

3.1. Affiliated sites for creation of speech corpora

The Center is cooperating with domestic and overseas affiliated sites which are shown in the map. The Center is collaborating with 11 domestic and 1 overseas cooperative sites. The Center also hopes for close relationship with overseas organizations like LDC, ELRA, and GSK.

Currently domestic affiliated sites are involved in collecting regional speech data, establishing regulations for labeling, and labeling speech data.

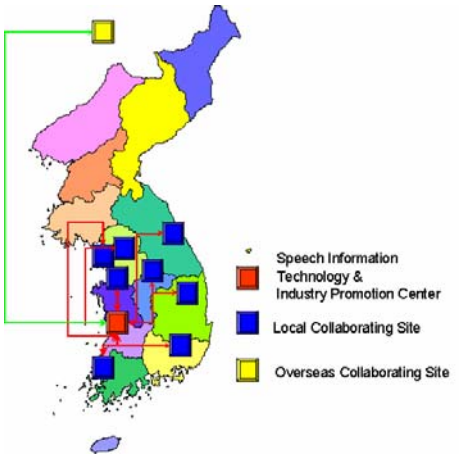

Figure 2: Affiliated sites

3.2. Creation of speech corpora

- Clean speech corpus

A speech corpus is being created for research for technological applications, based on 4,178 phonetically rich words containing various phonetic environments and syllable units. Data collection has been completed by regional affiliated sites for 500 speakers who are sampled according to sex, age, and region. The data is being labelled according to the conventions for labeling.

- Corpus for children's speech

A speech corpus is being created for pre-school and elementary school children. The word list consists of 1,233 words including 41 digits and command words, 340 four-digit numbers, 452 phonetically balanced words, 400 control command words. Each speaker reads 100 words. 500 children are currently being recorded.

- Corpus for read aloud sentences

A corpus for read aloud sentences collected through PC is being created for applications such as dictation. The reading list consists of 20,833 sentences which are made up of 10,000 words with high frequency extracted from a text corpus containing 43 million words, and it is structured for OOV (out of vocabulary) test, etc. as follows:

*5K set: the set of sentences composed of high frequency 5,000 words

*8K set: the set of sentences which are composed of high frequency 8,000 phonological words, excluding the sentences in the 5K set from 20,833 sentences

*10K set: the rest of the sentences excluding the sentences in the 5K and 8K sets from 20,833 sentences

Currently 400 speakers are being recorded. Each speaker reads 100 sentences.

- Corpus for prosody research and synthesis

A text was constructed by the following procedure:

10 million phonological phrases were extracted from a text corpus of 43 million phonological phrases (Choi 2001), considering genre (domain), and then 4,360 phonetically rich sentences were extracted by the greedy algorism.

The 4360 sentences were each recorded by one male professional announcer and one female professional announcer, and phonological and K-ToBI prosodic labeling is currently being done.

- Prototype for noise and speech database in the car noise environments

Recently speech recognition in car noise environments has been recognized. Thus demands have increased for corpora of speech in car noise environments. Data in various driving environments is being collected through 8 channels as a prototype to develop methods and procedures to create a corpus of speech in car noise environments.

- Encouraging to share existing corpora

Some proprietors have agreed to share their existing corpora. Thus those corpora are being assessed and supplemented. At present a corpus of telephone speech of 2,000 speakers, a corpus of 452 phonetically balanced words by 70 speakers for research, and a corpus of numbers by 500 speakers are being prepared for sharing.

3.3. Training and others

The Center is operating SITEC Academy for speech practioners and researchers on speech information technologies. It has offered two sessions for fundamentals of phonetics, and theory and practicum for phonemic and prosodic labeling. The fields of training in SITEC Academy will be expanded and the contents will be offered on-line. An expert group has been organized to establish guidelines for objective assessment of speech recognition and synthesis systems and selection of elementary technology in cooperation with a forum for standardization of speech information technology. The current states and difficulties in domestic speech information industry are being surveyed and analyzed in cooperation with Korean Speech Information Technology Industry Association, and Oriental COCOSDA 2001 (August 25, 2001) was sponsored by the Center.

3.4. Future plans

The Center will focus on the following works:

- To expand types and numbers of corpora consistently
- To establish a system of speech corpora distribution
- To construct a tera level of storage space for speech corpora
- To establish assessment guidelines for speech recognition and synthesis systems
- To construct a portal web site for speech information technologies
- To expand fields of training for SITEC Academy and on-line training
- To co-sponsor exhibitions for speech information technology products and invite the public for ideas on application with Korean Speech Information Technology Industry Association

4. Conclusion

In this paper we introduced SITEC, founded to support and promote industries using speech information technologies, and present its current activities and future

plans. The Center is working for boosting speech technologies to the world level by helping solve the common difficulties in the field. It provides objective evaluation for speech recognition and synthesis systems and manages creation and distribution of speech resources for research and development in Korea. The Center is currently creating 4 speech corpora including the clean speech corpus. It also encourages sharing of existing corpora among the academia, industry and institute.

5. References

Choi, Key-Sun. 2001. *KAIST Language Resources, 2001: Ministry of Science and Technology Core Softwares 1955-2000* (http://kibs.kaist.ac.kr). Ministry of Science and Technology.

Lee, Yong-Ju, et al. 2000. *A Study of Domestic and Overseas Trends of Speech Information Industry and Ways for its Development*. Ministry of Commerce, Industry and Energy.