



An Overview of Speech Corpus Related Activities in Korea

- Focusing on SiTEC's activities -

Yong-Ju Lee

Speech Information Technology & Industry Promotion Center(**SITEC**),
KOREA

yjlee@wonkwang.ac.kr

Introduction



- **Speech technologies have developed substantially through the research and development of academia, industry and institute in Korea.**
- **As practical uses of speech technologies such as speech recognition and synthesis increase, difficulties arise in academia and industry with speech corpora and assessment.**
- **Speech Information Technology & Industry Promotion Center (SITEC)**
 - 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.
 - For national coordination of industry, academia, and associations

Introduction to SITEC (1)



■ Goal

- Creation and distribution of speech corpora
- Assessment methodologies of speech recognition and synthesis systems
- Collection and dissemination of information on industry and technology
- Short training & others

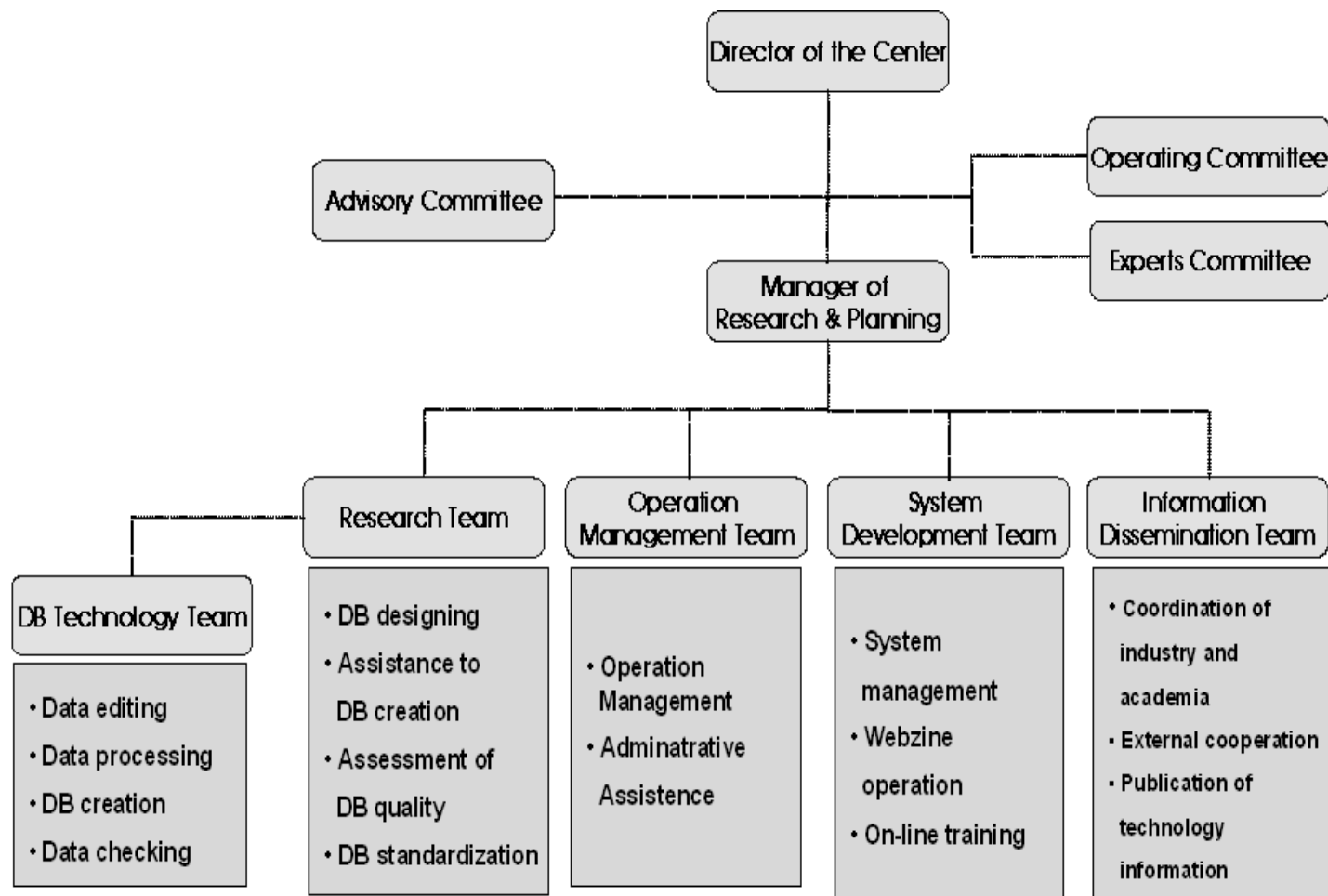
■ Operation

- Founded at Wonkwang University in May 2001
- Funded by a consortium of the government and 11 companies
- Approximately \$1,000,000 per year from 2001.5 to 2006.4 (5 years)
- Expected to be self-reliant after 5 years

Introduction to SITEC (2)



■ Organization



Introduction to SITEC (3)



■ Personnel

■ Full-time personnel

- Director & 10 persons

(4 doctorates, 4 masters, 1 bachelor and a secretary)

■ Part-time staff

- Processing staff of 10 persons
- For data collection, production, and testing



■ Participants

- Korean Speech Information Technology Industry Association
- 13 companies
 - Samsung Electronics Co., Ltd.
 - LG Electronics Inc.
 - Eoneo Inc.
 - Voiceware Co., Ltd.
 - Mediagen
 - D&M Technology Inc.
 - Zenersys Voice & Vision Technology
 - CoreVoice Inc.
 - Man to Machine Inc.
 - Extell Technology Corp.
 - Human Media Tech. Inc.
 - SL2
 - Hyundai Autonet

Committees



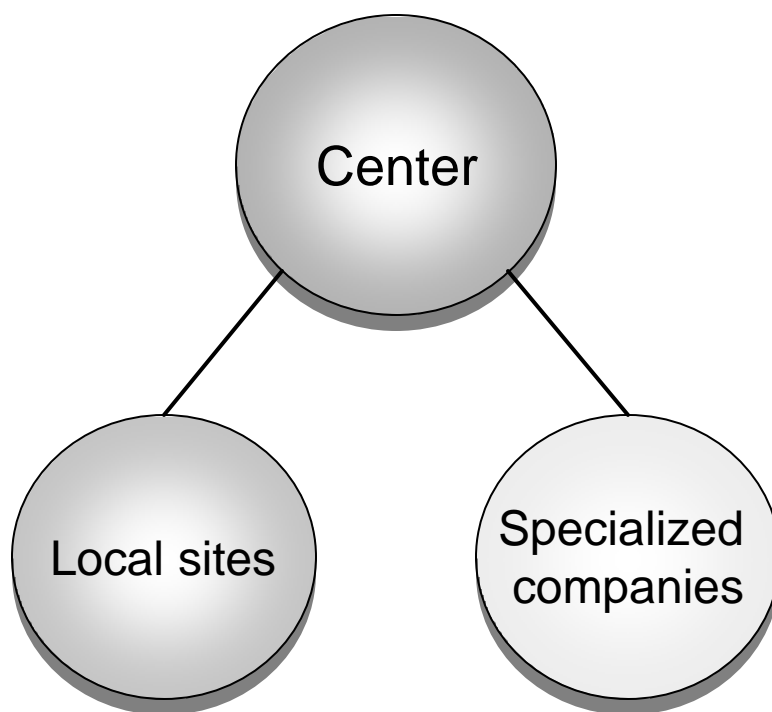
- **Advisory Committee**
- **Operating Committee**
- **Experts Committee**
 - SIG-Corpus (specification, assessment, reference)
 - SIG-Roadmap (planning)
 - SIG-Standard (standardization)
 - SIG-Tool (development and assessment of tools)
 - SIG-Edu (training courses development)
 - SIG-Info (Information: Survey & Analysis)
 - SIG-Assessment (assessment of speech recognition, synthesis, speaker verification)
 - SIG-Pub (publicity, conference, expo, etc.)

Creation of Speech Corpora



■ Corpora

- Local affiliated sites
- Recruiting and data processing companies
- In-house design, validation, collection & annotation

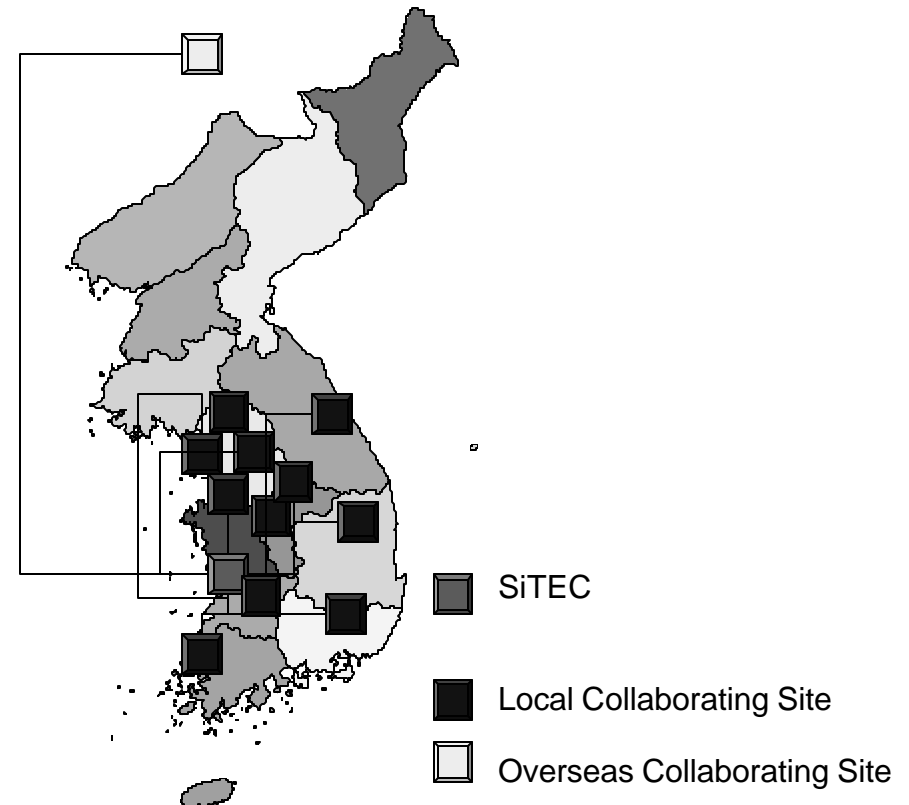


Affiliated Sites



■ Affiliated Sites

- Local sites
 - Total 12 university laboratories
- Local variants
- Expertise of experimental speech scientists
- Solidification of infrastructure for speech research





- **Speech Corpora for Car Applications**
- **Speech Corpora of Foreign Languages**
- **Speech Corpora for Basic Research**
- **Other Corpora**

Speech Corpora for Car Applications



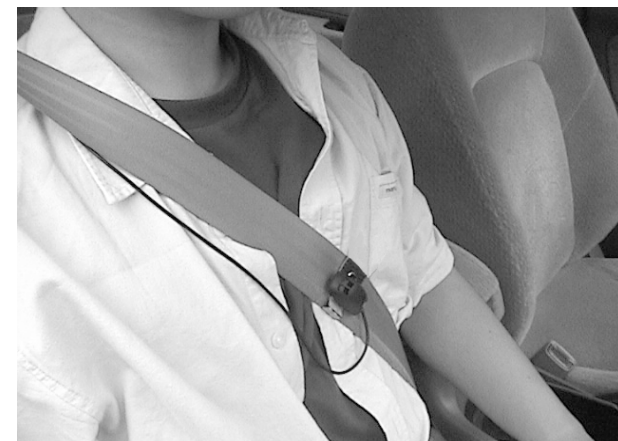
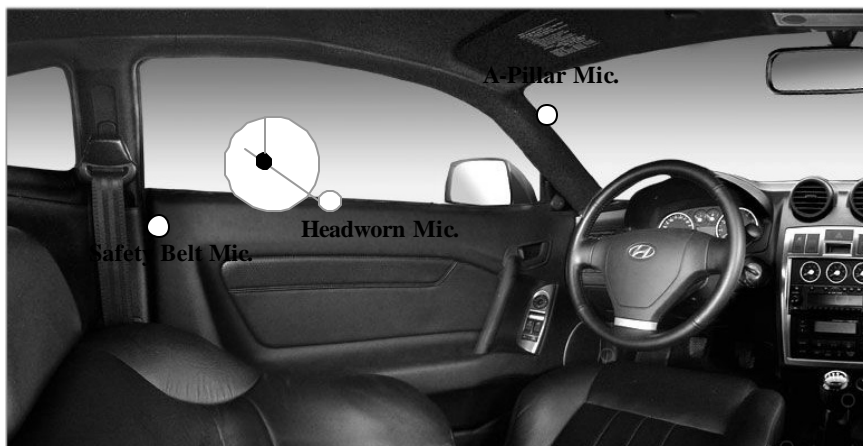
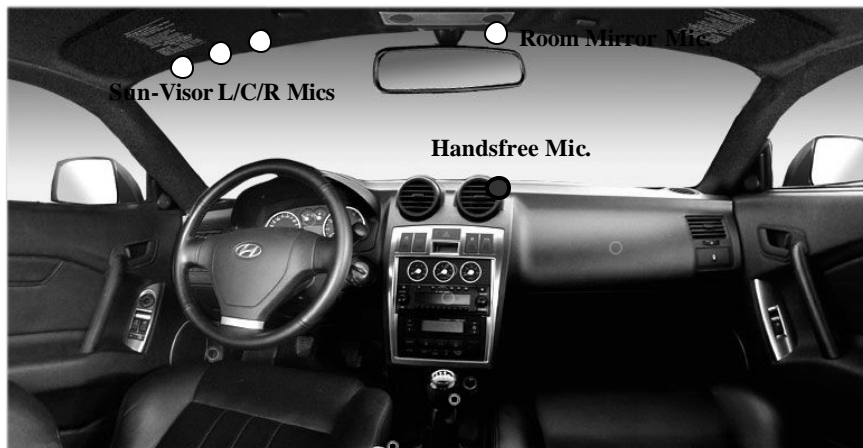
■ Car noise corpus

- 270 environments, 5 minutes per environment
- 7 channels + 1 hands free (simultaneous collection)

■ Large speech corpus of words in car environments

- 2,066 words
- 400 speakers (200 words per speaker)
- 8 channels (simultaneous recording)
- city & highway

■ Placement of microphones



Speech Corpora of Foreign Languages



■ Chinese speech corpus

- syllables, PBW words, four-digit numbers, etc.
- 300 speakers
- 110 tokens per speaker

■ English speech corpus

- 1,586 words
- 400 speakers
- 130 tokens per speaker
- native male and female adult speakers of English (recorded in America)

■ Spanish speech corpus

- 1,230 words + 5,670 sentences
- 300 speakers
- 130 tokens per speaker
- Hispanic speakers in Southwestern areas of America (recorded in America)



■ **Clean speech corpus of words**

- PRW 4,100 words
- 500 speakers, 417 words per speaker
- full prosody labeling

■ **Clean speech corpus of read sentences**

- 20,000 sentences
- 200 speakers
- 100 sentences per speaker

■ **Speech corpus of read sentences for prosodic synthesis**

- 2 professional speakers (one male and one female)
- 4,392 sentences per speaker
- full phonetic & K-ToBI prosody labeling



■ Corpus of read speech for dictation (I)

- 20,800 sentences
- 400 speakers
- 100 sentences per speaker

■ Corpus of read speech for dictation (II)

- 20,000 sentences
- 400 speakers
- 100 sentences per speaker

■ Corpus of numbers

- 25,000 unit numbers (2~3 syllables)
- 500 speakers
- 100 tokens per speaker
- simultaneous recording through 2 channels (PC & telephone)



■ Corpus of children's speech

- 1,283 tokens
- 500 speakers
- elementary school students (even distribution from 1st through 6th grades)
- 100 words per speaker

■ Corpus for embedded speech

- 4,199 tokens
- 300 speakers
- 107~108 tokens per speaker
- USB-DSP toolkit for collecting embedded speech

■ Corpus for microphone test

- Rerecording of PBW_SH DB(speech corpus) through HATS(Head And Torso Simulator) in a sound proof studio
- To compare the varying characteristics of different types of microphones (8 types of microphones)
- To compare the varying characteristics of different types of microphones depending on distance (5 cm, 10 cm, 20 cm, 50 cm, 100 cm)

Corpora of Other Organizations, Distributed by SITEC



- **Clean Speech Corpus of PBW 452 Words**
 - four-digit numbers, single digits, short sentences
 - 70 speakers, 452 words per speaker (one repetition)
- **Clean Speech Corpus of PBS 589 Sentences**
 - 20 speakers, 150 sentences per speaker
- **Digit Telephone Speech Corpus**
 - 2,000 four-digit numbers
- **PRW Telephone Speech Corpus**
 - PRW 3,800 words, 2,000 speakers
- **Trade-related Continuous Speech Corpus**
 - 3,008 sentences for trade consulting, 150 speakers
- **Speech Corpus of Digits in PC Environments**
 - telephone numbers, single digits, four-digit numbers, multi-digit numbers
 - 500 speakers

Speech Corpora under Creation(1)



| Title | Remarks |
|--|---|
| Speech corpus of foreign languages | For speech recognition <ul style="list-style-type: none">- English : PRW & short sentences- Chinese : words & sentences |
| Corpus for speech in car environments | Speech in real car environments <ul style="list-style-type: none">- Augmentation of content, conditions, & speakers |
| Corpus for speaker verification in car environments | Recording at varying times/periods |
| Multimodal speech corpus | Speech & lip movement images |
| Corpus for speech recognition evaluation | Comparable to AURORA 2.0 |

Speech Corpora under Creation(2)



| Title | Remarks |
|---|--|
| Corpus for welfare application | <ul style="list-style-type: none">- Prototype for speech of handicapped or elderly people |
| Corpus of speech in simulated environments | <ul style="list-style-type: none">- Car environments |
| Corpus of dialogue speech for recognition | <ul style="list-style-type: none">- Preliminary examination of domains, methods of collection & representation- Creation of prototype for examination |
| Corpus of speech of foreign languages by Korean speakers | <ul style="list-style-type: none">- Preliminary examination of demand and specification |
| Corpus of dialogue speech for synthesis | <ul style="list-style-type: none">- Preliminary examination of demand and specification |
| Corpus of high level noise | <ul style="list-style-type: none">- Preliminary examination of demand and specification |
| Corpus of emotional speech for synthesis | <ul style="list-style-type: none">- Preliminary examination of demand and specification |

Other Activities in Korea

in relation to Speech Corpus



■ Language & Speech Information Research Center

- Founded as an expansion of the research team on speech in Electronic and Telecommunication Research Institute (ETRI) in November 2001.
- Supported by the Ministry of Information and Communication for comprehensive research into speech information technologies.
- It is working for creating large speech corpora (several kinds of telephone speech corpora).
- <http://www.voice.etri.re.kr>

Conclusion



- **Introduction to SITEC**
- **Recent activities and next plans**
- **Other activities in Korea in relation to speech corpus**
- **Affiliated cooperation with organizations across the world, such as LDC, ELRA, and GSK, are also hoped for.**