Creation and Assessment of Korean Speech and Noise DB in Car Environment

Yong-Ju Lee*, Bong-Wan Kim, Young-Il Kim**,**
Dae-Lim Choi, Kwang-Hyun Lee**, Yongnam Um****

*Department of Electrical, Electronic and Information Engineering, Wonkwang University
344-2, Sinyong-dong, Iksan, Chonbuk, Korea
yjlee@wonkwang.ac.kr
**Speech Information Technology & Industry Promotion Center
344-2, Sinyong-dong, Iksan, Chonbuk, Korea
{bwkim, yikim, dlchoi, khlee, umyongnam}@sitec.or.kr

Abstract

Researches into robust recognition in noise environment, especially in car environment, are being carried out actively in speech community. In this paper we introduce three types of corpora that SITEC (Speech Information Technology & Industry Promotion Center) has created for research into speech recognition in car noise environment. The first is the recordings of 900 Korean native speakers, distributed according to gender, age, and region, who uttered command words in car environment. The second is the collection of mixed noise in 3 models of cars by while setting up various noise patterns which can be obtained with the car engine on or off, at different driving speed, and in different road conditions with windows open or closed. The third is the recording of simulated speech by HATS (Head and Torso Simulator) in car environment with the internal and external noise factors added. These three types of recordings were all made through synchronized 7 channels and a separate hands-free channel fixed in a car. The creation and specifications of these corpora will be reported on in detail.

1. Introduction

In general, the error rate of speech recognition in noise environment is higher than in laboratories or quiet offices. Especially it is hard to expect low error rate of recognition in car environment due to relatively high noise energy by various noise factors in driving environment. In order to improve speech recognition in noise environment, data collected in real noise environment is used in training instead of existing databases collected in studios or soundproof rooms, and various preprocessing techniques such as noise separation are used. Thus it is essential to collect speech data in real environment to improve speech recognition in noise environment.

In recent three years SITEC at Wonkwang University in Korea has been creating various speech corpora in real noise environment and distributing them to researchers and developers.

In this paper we will report on speech and noise corpora in car environment that SITEC has created. This paper is organized as follows. In Section 2 we introduce creation processes of the word speech DB recorded in driving environment. In Section 3 we introduce the noise DB collected in varying driving patterns and setting environment. In Section 3 we introduce the speech DB simulated in car environment using HATS (Head and Torso Simulator) instead of real speakers. Finally, in Section 5 we conclude by mentioning current distribution state of corpora in car environment and future plans for expanding.

2. Word Speech DB in Car Environment

In this section we introduce the speech DB which contains the recording of application command words read by speakers. This DB has been created by expanding 3 times in 3 years.

2.1. Data Collection

In Korea, researches into robust speech recognition is going on as in other countries. However, researchers have created small databases and used them for preprocessing algorithm and recognition experiments. There are many factors that can be controlled in designing specification such as type of cars, position and number of microphones, way of simultaneous recording, organization of prompt sheets, and definition and distribution of driving environment. Thus, considering this fact, 900 speakers were recorded for 3 years, reflecting the opinions of the specialists in academia, industry, and institutes who are interested in speech recognition in car environment. The corpus of word speech recorded in driving environment contains the recordings made in three types of cars on city streets and highways as in Table 1. Speakers read items based on command words provided by producers of related applications. Positions and models of 8 microphones are shown in Table 2 and their configurations are in Figure 1.

NVIRONMENT	DETAILS
Env. A (50%)	City Street (30~60km/h Speed)
Env. B (50%)	Highway (70~90km/h Speed)
Env. Common	Weather : Sunshine Road : Asphalt Window : Close Audio, Fan : Off Radio, Wipers : Off

Table 1 : Driving environment

CH.	POSITION	MICROPHONE MODEL
1	Head-worn	SHURE SM-10A
2	Upper A-Pillar	AKG C400-BL
3	Sun-Visor Left	AKG C400-BL
4	Sun-Visor Center	AKG C400-BL
5	Sun-Visor Right	AKG C400-BL
6	Room Mirror	AKG C400-BL
7	Safety Belt	AKG C400-BL
8	Center Front	DIGITEC ECM

Table 2 : Positions and models of microphones

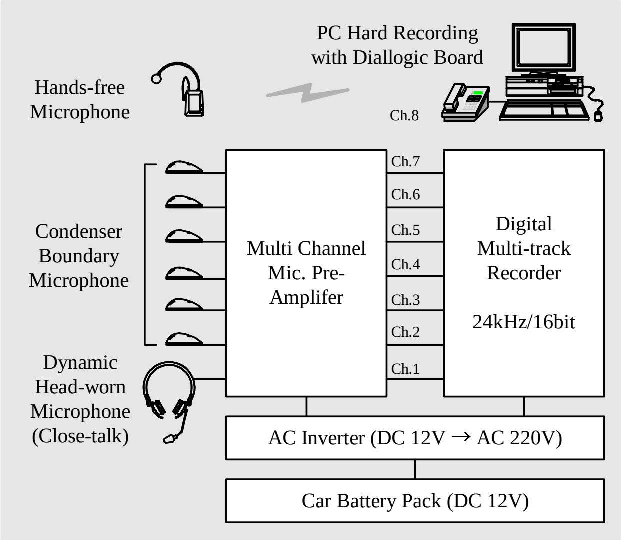


Figure 1 : The architecture of data collection system

2.2. Speakers and Prompting Items

Distribution of speakers by gender, age and region is shown in Table 3. Prompting sheets have 2,070 tokens that are composed of command words for communication systems like mobile phones and PDA and command words for car applications like car accessories, car audios, and navigation systems. The numbers of types of utterance items are shown in Table 4.

<table>
<tr><td rowspan="2">Gender</td><td colspan="10">Male</td><td colspan="10">Female</td></tr>
<tr><td colspan="10">50%</td><td colspan="10">50%</td></tr>
<tr><td rowspan="2">Age</td><td colspan="5">19~20</td><td colspan="5">21~29</td><td colspan="5">30~39</td><td colspan="5">40~50</td></tr>
<tr><td colspan="5">10%</td><td colspan="5">30%</td><td colspan="5">30%</td><td colspan="5">30%</td></tr>
<tr><td rowspan="2">Region</td><td colspan="4">SE</td><td colspan="4">YN</td><td colspan="4">HN</td><td colspan="4">CC</td><td colspan="4">GW</td></tr>
<tr><td colspan="4">40%</td><td colspan="4">20%</td><td colspan="4">20%</td><td colspan="4">10%</td><td colspan="4">10%</td></tr>
</table>

Table 3 : Distribution of speakers
(SE: Seoul, YN: Gyeongsang-do, HN: Jeolla-do
CC: Chungcheong-do, GW: Gangwon-do)

UTTERANCE ITEM	NUMBER
Digits	900
Dialing commands	63
Car audio commands	306
Control commands for car equipments	124
Navigation commands	57
PDA commands	121
Place names	446
Highway names	53

Table 4 : Utterance items

2.3. Segmentation and Transcription

Each token of speech mixed with noise in driving was segmented. Synchronized segmentation is necessary for the speech signals recorded simultaneously from 7 channels because they have time difference according to their distance from the sound source (mouth). In consideration of this, speech signals from head-worn microphone as reference channel were segmented so that 500 ms span was obtained before and after speech signals and time information about end point for reference channel was stored. Based on this time information, synchronized automatic segmentation for 7 channels was performed. For the channel of hands-free microphone segmentation was performed independently.

On the other hand, orthographic and phonetic transcription was done only for the channels of head-worn microphone (reference channel) and hands-free microphone. And noise transcription was done in addition for the reference channel, using the symbols shown in Table 5. However, in the case of the channel of hands-free microphone, only audible speech corresponding to the prompting item was transcribed in order to represent received pronunciation distortion depending on the transmission quality of the wireless network. Inaudible items were not transcribed and marked as “[inaudible]”.

Transcriptions for each speech data are stored as text file. The first level in the transcription file is for phonetic transcription, the second for orthographic transcription, and the third for noise transcription. An example is shown in Figure 2. A transcriber as in Figure 3 was developed by SITEC to prevent the worker’s error and enhance his consistency in his transcription. The transcriber program performs the following functions:

- to prevent error in using meta symbols
- to prevent pronunciation that cannot occur from being entered
- to check out incorrect orthographic word boundaries
- to report segmentation error
- to report clipping error automatically

SYMBOL	NOISE TYPE
/cl	Normal driving noise without other special noise
/ln	Blinker's clicking sound
/cw	Noise of other cars
/rn	Jolting noise
/on	Other noise

Table 5 : Noise symbol types of head-worn channel

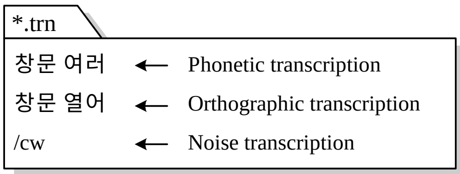
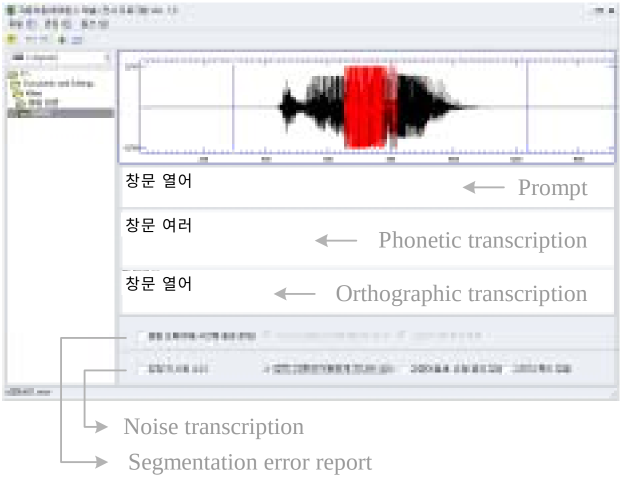

*.trn
창문 여러
창문 열어
/cw

Figure 2: Example of transcription file



창문 열어

창문 여러

창문 열어

Figure 3: Transcription program for car environment DB

3. Noise DB in Car Environment

This is the DB that reflects car noise environment. Three types of cars at 1,500cc, 2,000cc, and 2,500cc were used. Driving noise was collected for each type of car in the same 90 types of noise environment. Noise signals were recorded simultaneously through 8 microphones as for the speech DB in car environment. The most stable signal of 90-second span that does not have distortion or other noise was extracted from 5-minute signal data from the channel of head-worn microphone (reference channel), and synchronized segmentation was performed for 7 channels except the channel of hands-free microphone. On the other hand, separate segmentation was performed for hands-free channel because recording was made separately for the channel, but segmentation was made so that the span of signal as close as that in the reference channel should be obtained. 90 types of noise environment that were applied commonly to the three types of cars were composed by combining the factors shown in Table 6.

oad State	Asphalt / Cement / etc.
ngine State	On / Off
riving Speed	Stop / 30 / 50 / 80 / 100km/h
riving Environment	City Street / Country Road
indow State	Open / Close
ir Conditioner State	On / Off
ir Conditioner Aimming	Front Aim / Under Aim
udio State	On / Off
ic Genre	Ballade / Rock&Roll
Audio Volume	46dB / 68dB SPL_{ave}

Table 6: Factors of driving environment

4. Simulated Speech DB in Car Environment

Characteristics of received speech signals and noise patterns can be different due to internal and external noise factors in parking and driving environment, and acoustic models can have various patterns depending on individual speakers (drivers). In consideration of this, re-recording was made while 360 speakers from the PRW (phonetically rich words) corpus of 500 speakers created earlier by SITEC were being replayed through HATS (Head and Torso Simulator) as in Figure 4.

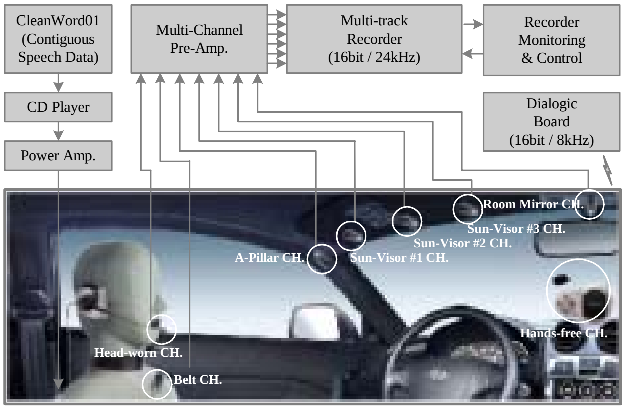


Figure 4 : Data collection system through HATS

320 speakers replayed and recorded in stable driving environment were used for creating training data, and 40 speakers replayed and recorded in 5 types of environment composed by variable factors occurring in actual driving environment were used for creating test data. Recording was made for training data at two types of speed (low and high speed), using cars at 2500cc, and for test data 5 types of environment were composed. Using information about the speakers, even distribution was made for replayed data of 360 speakers by gender, age, region and reading set. Recording channels and processing and transcription procedures are identical to those for the speech DB in car environment. Configurations for recordings for training and test data are shown in Table 7.

TRAINING DATA		
Data Env.	Drive Speed	Window / Fan
Env. 1	City Street, 40~60km/h	Close / Off
Env. 2	Highway, 70~90km/h	Close / Off
TEST DATA		
Data Env.	Drive Speed	Window / Fan
Env. A	City Street, Random Speed	Close / Off
Env. B	City Street, Random Speed	Close / On
Env. C	City Street, Random Speed	100% Open / Off
Env. D	Highway Driving	Close / On
Env. E	Highway Driving	30% Open / Off

Table 7: Configurations for speech recordings

5. Conclusion and Future Plan

In this paper we have reported on creation processes and characteristics of speech and noise corpora in car environment. At present the word speech DB of 900 speakers, the noise DB that contains noise in driving environment, and the simulated DB in car environment have been completed. They are being distributed constantly to companies and institutes. SITEC will supplement and expand more DB consistently based on the completed speech DB in car environment. SITEC plans to design specifications and procedures in cooperation with the specialist groups in academia, industry, and institutes.

References

Yong-Ju Lee, Bong-Wan Kim, Yongnam Um (2002). Speech Information Technology & Industry Promotion Center in Korea: Activities and Directions. In Proceedings of LREC 2002 (pp. 1851--1854). Paris: ELRA.

Yong-Ju Lee, Bong-Wan Kim, Yongnam Um (2001). Speech Information Technoloty & Industry Promotion Center: Introduction and Future Directions. In Proceedings of the Oriental COCOSDA Workshop 2001 (pp. 3--6). Daejon, Korea.

Speech Information Technology & Industry Promotion Center (http://www.sitec.or.kr)