An In-car Chinese Noise Corpus for Speech Recognition

Jue Hou[1,2], Yi Liu[1,3], Chao Zhang[1,2], Shilei Huang[3]
[1]Center for Speech and Language Technologies, Division of Technology Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, Beijing, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[3]Shenzhen Key Laboratory of Intelligent Media and Speech, Shenzhen, China
E-mail: houj@cslt.riit.tsinghua.edu.cn, eeyliu@tsinghua.edu.cn, zhangc@cslt.riit.tsinghua.edu.cn,
shilei.huang@imsl.org.cn

Abstract—**In this paper, we present an in-car Chinese noise corpus that can be used in simulating complicated car environment for robust speech recognition research and experiment. The corpus was collected in mainland China in 2009 and 2010. The corpus includes a diversity of car conditions including different car speed, open/close windows, weather conditions as well as environment conditions. Specially, the rumble strips are also taken into account due to the typical noise generated as the car is passing on. In order to use the corpus efficiently, we performed some acoustic signal analyses on those noise data, mainly focused on stationary properties and energy distribution in the frequency domain. We also performed ASR experiments using selected noise data from the corpus, by adding noise data to clean speech to simulate the in-car environment. The corpus is the first of its kind for in-car Chinese noise corpus, providing abundant and diversified samples for car noise speech recognition task.**

Keywords - in-car noise; Chinese speech database

I. INTRODUCTION

One of the most important applications of automatic speech recognition (ASR) technologies is controlling the car by speech without human hand operation, which can greatly improve the convenience and safety when driving [4]. There are many methods proposed for short phrase recognition or keyword spotting that can be used in such tasks. However, most state-of-the-art ASR systems fail to deal with the noise problem, thus the performance of ASR applications degrades dramatically when testing in a noisy environment, especially when real car environment for the noises in car is a mixture of several sources. Therefore the current speech recognition systems still need to be improved [6]. Thus, the noises appear during car driving should be carefully recorded and analyzed in order to tune a noise robust ASR system and finally get a higher accuracy.

There are two common ways to reduce the effect of additive noise. One is using noisy data in the training process, which means to add noise at different Source-to-Noise Ratios (SNRs) to existing clean speech data to make a simulation of real noisy data, and then using them to train acoustic models. The other methods are a series of proposed noise-reduction algorithms [2] such as using band pass filters, Cepstral Mean Subtraction (CMS), Parallel Model Combination (PMC), etc. Some of them are based on noise reduction that first samples the noises in real car environment to get the characteristics of the noise data, and then performs the noise reduction.

Since most of the ASR approaches are based on machine learning methods, researchers need appropriate noisy speech databases for experiments or evaluations. As far as we know, there is few noise corpus special designed for in-car environment, most of the current corpus are similar with AURORA [1][2]. Some of the researchers proposed a multimedia corpus which is recorded by several microphones with digital cameras simultaneously [4].

We established an in-car noise speech database on various conditions, using portable devices in car. The environment includes different kinds of roads (in the city or on expressway), at various speed and in different weather. There are 120 speakers in total, which are gender-balanced. The contents of the corpus are mainly composed of short phrase in Chinese and English, e.g. addresses, song names, and numbers. The total length of the database is about 50 hours, and can be used for acoustic model training and testing. The recording prompt includes syllable balanced sentences, short phrase, numbers, digit string, etc.

It is usually expensive to collect speech data in real environments; therefore it is very common to make artificial noisy speech by adding noise data to clean speech [1]. Although it is not an optimum solution since the real in-car environment is affected by the location of microphone and such complicated conditions, it is much more convenient for us to collect purely noise data instead of collecting speech data with noise.

However, the in-car environment is too complicated to be simulated by using artificial generated signals, for there are so many aspects which greatly influence the noise. Therefore, we designed a series of routine for collecting various kinds of noises separately in real car environment, so that the data can be added to clean speech for model training, noise analysis, or to study the common characteristics of real in-car noises. The total length of the noise database is about 50 minutes, but it can be further integrated with other existing speech corpus or the noisy speech database we recorded, thus to get mixed noise speech data in great length.

We use some of the common signal processing tools to perform a series analyses to show the detailed properties of recorded noises. What's more, we also perform some ASR experiments using several typical noise data in our corpus, therefore we can make a comparison among those noises.

The rest of this paper is organized as follows. Section 2 describes the design of our data collecting procedure. The equipments in our experiments (both hardware and software) are included, together with the detailed list of recording conditions. Section 3 is some analyses and comparisons on some typical kinds of noises. Part of the experimental results using the recorded data in ASR systems is shown in Section 4. Finally we conclude our work in Section 5.

II. Recording Conditions

A. Recording Configurations

As we were planning to record in-car noise corpus, we had to consider several typical circumstances. The database described in [5][7][8] defined a few environmental conditions, such as town traffic, rough road conditions, etc. There were quite many aspects, but still can be further extended to cover more common circumstances.

The in-car noises are mostly composed by the following four aspects. 1) The engine. The noise generated by the engine is usually louder if the car is accelerating or travelling at a higher speed than keeping a fixed lower speed. 2) The air conditioner. If the air conditioner is too close to the microphone, the wind from the air conditioner will certainly be caught by the recording devices. 3) The outside noise including rain drops on the windshield or other vehicles or pedestrians passing by. These kinds of noises will be able to get into the car if we do not close the windows tightly. 4) The audible rumbling which is transferred from the gears when the car is travelling on the rumble strips.

All of the factors we focused on are listed in Table 1. In our recording process, we selected several roads in Shenzhen city (which is located in Guangdong province in southern China), and driving the car at a relatively fixed speed. We called a ‘session’ to be a series composition of factors in Table 1, and kept recording for several minutes.

TABLE I. Various circumstances used in recording noise data

Factors	Value
Weather	Sunny, rainy
Road type	Suburb, downtown, expressway
Air conditioner	On, off
Window	Open, half open, slightly open, closed
Speed	High (80 km/h), low (30 km/h)
Rumble strips	Yes, no

B. Hardware and software used

The recording jobs were operated on Peugeot 307 for Guangzhou-Shenzhen expressway and BYD F3 or Buick GL8 for other roads. Besides the driver, there was one staff on board, sitting at the front passenger seat, with a laptop computer in hand. The wireless microphone with sound card produced by Sennheiser was linked to the laptop but not fastened. The software for sampling and quantizing was Cool Edit Pro v2.2.

C. File Format

We recorded several minutes for each session and there were altogether 62 files. All noise data were recorded in single channel, sampled at 16 kHz and quantized by 16 bit, and saved as wave format. The total length was about 49 minutes 35 seconds.

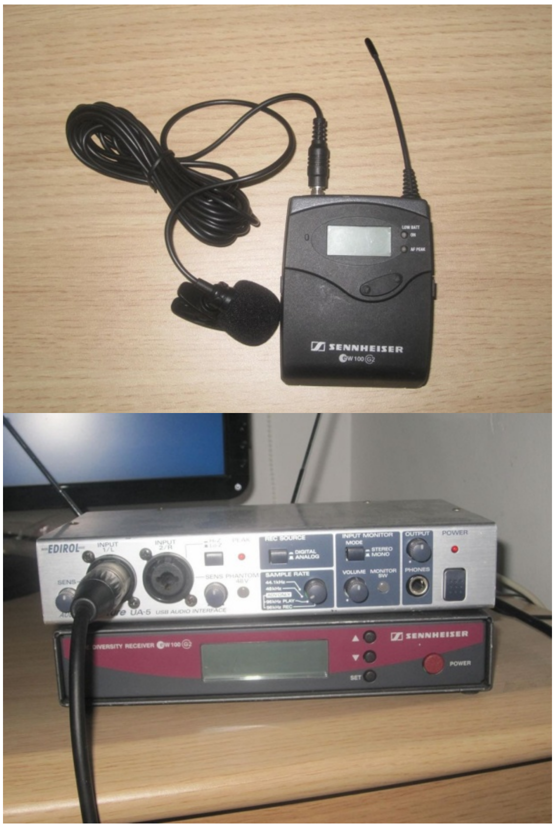

Figure 1. The wireless microphone used data collection.

III. Stationarity, Energy Distribution And Frequency Analysis

In order to illustrate the properties of the recorded noise data, further to find the most significant factor that influences the noise, we extracted the spectrogram and frequency percentage of each recorded noise file. Therefore we can make comparisons between those noise files.

A. Stationarity

When the car is traveling at a fixed speed on the road, it means the car is not accelerating or slowing down, then the noise generated by the engine should be fairly stationary [1]. Figure 2 shows the waveforms (on the top) and spectrograms (on the bottom) of two noise files recorded when the car was running in the suburb with windows closed and air-conditioner turned on. The only difference was that file 48.wav (on the left) was recorded when the car was on the rumble strips while 18.wav (on the right) was on plain road.

It is shown from Figure 2 that the noises were not stationary, but the energy distribution did not change much during the whole file. Also, the noise seemed to exhibit periodicity on rumble strips due to the fixed car speed and constant distances among the strips.

B. Energy distribution

On the energy distribution, it is shown in [3] that the car noises are mostly lowpass. We calculated the following 3 kinds of features: energy percentage in 0-100 Hz band, 0-200 Hz band and 90% energy's frequency. Here we discuss some of the factors that may influence the energy distribution: driving speed, air conditioner and outside environment with car windows.

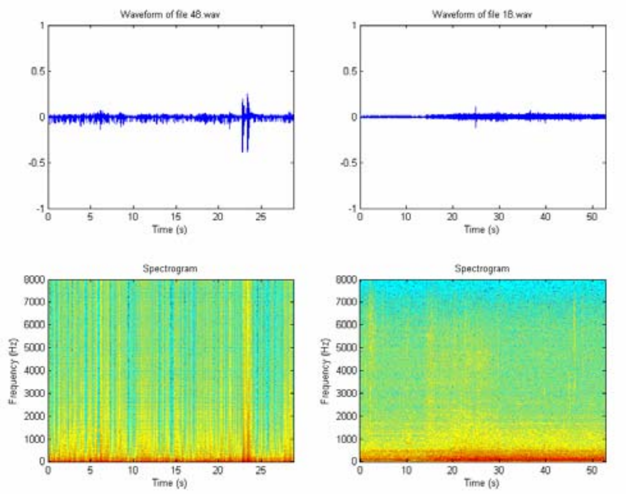

Figure 2. The waveforms and spectrograms of two noise files. The only recording condition in difference was whether the car was travelling on rumble strips or not.

1) *Driving speed*

Figure 3 shows the energy-frequency curve of two sessions, named as 20.wav and 31.wav. The windows were slightly opened (i.e. not totally closed) while the air conditioners were turned off. File 20.wav was recorded at 30 km/h but the other was 80 km/h. Both files were recorded in the suburb.

The results are as follows. Over 90% of the energy is kept in 0-200 Hz band, and about 50% of the energy is in 0-100 Hz band. The two curves in Figure 3 (in blue and brown) are very similar in shape. Therefore we can infer that the higher speed leads to an increase of energy in low-frequency band, but does not affect the energy distribution curve very much.

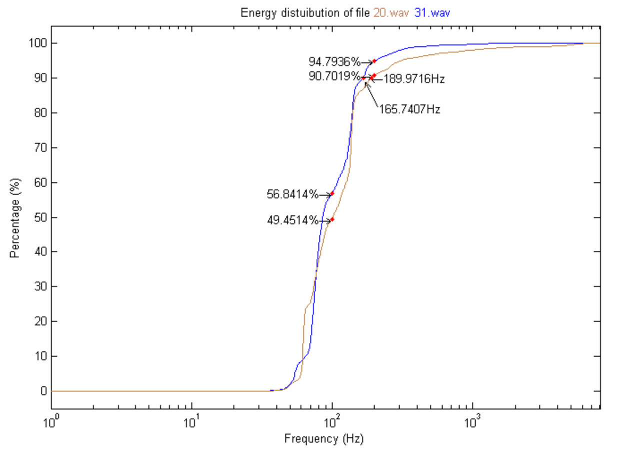

Figure 3. The energy distributions on frequency are slightly affected by the driving speed.

2) *Air conditioner*

Figure 4 illustrates how the energy contour is affected by the air conditioner. It can be seen from the figure that the two contours are almost the same (for file 5.wav and 9.wav, which are recorded in the downtown with windows slightly open). About 90% of the energy are kept in 0-100 Hz band, and over 96% are kept below 200 Hz.

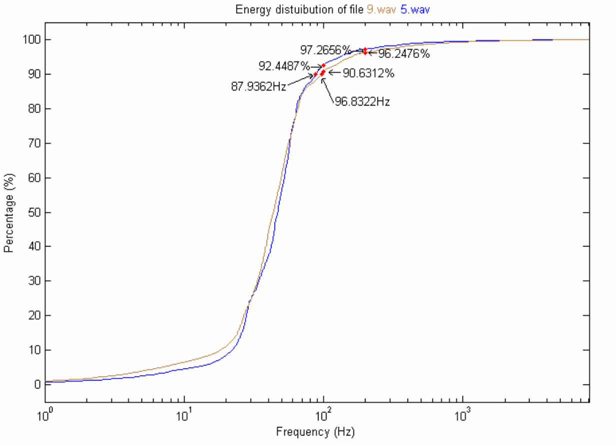

Figure 4. The energy distributions on frequency are almost the same regardless the status of air conditioner.

3) *Outside environment and windows*

We can compare the file 20.wav and 5.wav showed in Figure 3 and Figure 4. The rest conditions were same except the file 20.wav was in the suburb while 5.wav was in downtown. We can conclude from the contours that the low-frequency energy component mainly comes from the outside environment through the car windows, similar with [1].

Table 2 shows the relationship between energy and frequency of several noise files.

TABLE II. RELATIONSHIP BETWEEN ENERGY PERCENTAGE AND FREQUENCY

Windows	**Location**	**ID**	**0-100Hz (%)**	**0-200Hz (%)**	**90% Energy (Hz)**
Open	Downtown	5	92.4	97.3	87.9
	Suburb	20	49.5	90.7	190.0
Closed	Downtown	8	89.5	85.5	103.3
	Suburb	15	75.0	96.9	136.8

We can concluded from Table 2 that the noises in downtown have more low-frequency components than those in suburb, so that the window is slightly able to filter those components in downtown, which is mostly generated by the crowd people and other vehicles.

IV. EXPERIMENTAL RESULTS IN ASR

A. *Experimental Design*

We established an ASR system using selected speech data which were generated by mixing clean speech with recorded noise data at different SNRs.

The testing data set were 347 utterances in count, the contents of the testing set were all short phrases. The noise data chosen were recorded from downtown, suburb and the expressway. We also considered the status of the windows and the air conditioner.

The experimental design contained two parts. One was to compare the accuracy of the ASR system on a fixed SNR (say, 5dB) on different environments. The other was to compare different SNRs.

B. *Experimental Results*

1) *Accuracies on different conditions*

The detailed recognition accuracies under different circumstances are listed in Table 3. The word SO in the table is the abbreviation for "slightly open" for car windows.

TABLE III. ASR ACCURACIES ON DIFFERENT TESTING ENVIRONMENTS

Location	**Windows**	**Closed**	**SO**	**SO**
	Air conditioner	**Off**	**Off**	**On**
Suburb		95.39%	94.81%	95.39%
Downtown		94.81%	93.66%	93.95%
Expressway		95.68%	94.81%	95.39%

It can be concluded from Table 3 the accuracy on the expressway was higher than that in suburb, and lowest in downtown. In addition, the ASR systems performed better when the car window was closed.

The experimental results were as expected, for the energy distribution of human speech is quite far from the noise data, so that the speech were not affected much by those low-frequency-concentrated noises. As a result, the ASR systems performed best in clean environment with windows tightly closed.

2) *Accuracies on different SNRs*

The noise data chosen in this part belonged to the downtown sessions. We evaluated the ASR system's accuracy on four conditions, as shown in Figure 5.

It can be seen from Figure 5 that the accuracy is higher when the SNR is higher. In fact, SNR at 25 dB can be fairly regarded as 'clean', while 5 dB is quite noisy.

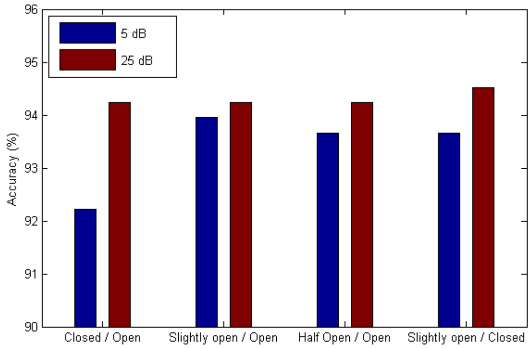

Figure 5. The accuracy of the ASR system on different SNRs. Higher SNR leads to higher accuracy.

V. SUMMARY

A noise-robust speech recognition system is an important part of the hands-free controlling system in car. As the noise environment is too complicated in real car environment, we established an in-car noise corpus which can be used to simulate the environment. The corpus contained several typical car conditions, such as car speed, weather, windows, air conditioner, and road types were considered. Also, we recorded a corpus of speech data in real car environment.

We conducted a series of analyses and experiments on those noise data, mainly focused on the stationarity and energy distribution. Many factors which had effects on those noises were carefully studied. We also performed several experiments on ASR by adding noise data to clean speech data at different SNRs to make comparisons. The results show that a clean outside environment and closed car windows lead to a better accuracy.

This corpus is the first of its kind in studying in-car noise environments, for its complexity and many aspects included. It presents enough samples of typical noises in car, and can be easily integrated with other speech corpus, e.g. the speech database we recorded in real car environment to get a mixed noise database which is useful for the development of robust speech recognition tasks.

ACKNOWLEDGMENT

This work was supported by the joint research grant of Natural Science Foundation of China (60975018), and Shenzhen Fundamental Research Grant (JC200903180702A).

REFERENCES

[1] David Pearce, Hans-Günter Hirsch, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions", ICSLP 2000 (6th International Conference on Spoken Language Processing), Beijing, China, 16-20, October 2000.

[2] J. C. Segura, A. de la Torre, M. C. Benitez, A. M. Peinado, "Model-Based Compensation of the Additive Noise for Continuous Speech Recognition. Experiments Using the Aurora II Database and Tasks", Eurospeech 2001.

[3] Firas Jabloun, A. Enis Cetin, Engin Erzin, "Teager Energy Based Feature Parameters for Speech Recognition in Car Noise", IEEE Signal Processing Letters, Vol. 6, No. 10, October 1999.

[4] Nobuo Kawaguchi, Kazuya Takeda, Fumitada Itakura, "Multimedia Corpus of In-Car Speech Communication", Journal of VLSI Signal Processing 36, 153–159, 2004.

[5] Asunción Moreno, Borge Lindberg, Christoph Draxler, Gael Richard, Khalid Choukry, Stephan Euler, Jeff Allen, "SPEECH DAT CAR. A Large Speech Database for Automotive Environments", Proc. II LREC, 2000.

[6] Satoshi Nakamura, Kazumasa Yamamoto, Kazuya Takeda, Shingo Kuroiwa, Norihide Kitaoka, Takeshi Yamada, Mitsunori Mizumachi, Takanobu Nishiura, Masakiyo Fujimoto, Akira Saso, Toshiki Endo, "Data Collection and Evaluation of Aurora-2 Japanese Corpus", ASRU 2003.

[7] Bjorn Schuller, Gerhard Rigoll, Michael Grimm, Kristian Kroschel, Tobias Moosmayr, Gunther Ruske, "Effects of In-Car Noise-Conditions on the Recognition of Emotion within Speech", DAGA 2007, Stuttgart, 305-306, 2007.

[8] R. F. Chen, C. F. Chan, H. C. So, Jonathan S. C. Lee, C. Y. Leung, "Speech Enhancement In Car Noise Environment Based On An Analysis-Synthesis Approach Using Harmonic Noise Model", ICASSP 2009, 4413-4416, 2009.