Journal of VLSI Signal Processing 36, 153–159, 2004

Multimedia Corpus of In-Car Speech Communication

NOBUO KAWAGUCHI, KAZUYA TAKEDA AND FUMITADA ITAKURA
Center for Integrated Acoustic Information Research, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603 Japan

Received November 6, 2001; Revised July 11, 2002; Accepted July 22, 2002

Abstract. An ongoing project for constructing a multimedia corpus of dialogues under the driving condition is reported. More than 500 subjects have been enrolled in this corpus development and more than 2 gigabytes of signals have been collected during approximately 60 minutes of driving per subject. Twelve microphones and three video cameras are installed in a car to obtain audio and video data. In addition, five signals regarding car control and the location of the car provided by the Global Positioning System (GPS) are recorded. All signals are simultaneously recorded directly onto the hard disk of the PCs onboard the specially designed data collection vehicle (DCV). The in-car dialogues are initiated by a human operator, an automatic speech recognition (ASR) system and a wizard of OZ (WOZ) system so as to collect as many speech disfluencies as possible.

In addition to the details of data collection, in this paper, preliminary results on intermedia signal conversion are described as an example of the corpus-based in-car speech signal processing research.

1. Introduction

Providing a human-machine interface in a car is one of the most important applications of speech signal processing, where the conventional input/output methods are unsafe and inconvenient. To develop an advanced in-car speech interface, however, not only one but many of the real-world problems, such as noise robustness, distortion due to distant talking [1] and disfluency while driving, must be overcome.

In particular, the difficulty of in-car speech processing is characterized by its variety. The road and traffic conditions, the car condition and the driving movement of the driver change continuously and affect the driver's speech [2]. Therefore, a large corpus is indispensable in the study of in-car speech, not only for training acoustic models under various background noise conditions, [3, 4] but also to build a new model of the combined distortions of speech.

In order to keep pace with the ever-changing environment, it may be helpful to make use of various observed signals rather than to use the speech input signal alone. Therefore, to develop advanced speech processing for in-car application, we need a corpus (1) that covers a large variety of driving conditions, and (2) from which we can extract the conditions surrounding the driver. Constructing such an advanced in-car speech corpus is the goal of the project described in this paper.

In ongoing data collection, a specially built data collection vehicle (DCV) has been used for synchronous recording of seven-channel audio signals, three-channel video signals and vehicle-related signals. About 1 terabytes of data has been collected by recording three sessions of spoken dialogue in about 60 minute of driving for each of 500 drivers. Speech data for text read aloud has also been collected.

In the next section, we describe the DCV which was specially designed for the multichannel audio-visual data acquisition and storage. In Section 3, we present the details of the data collected and the methodology used for data collection. In Section 4, the spoken dialogue system that utilizes automatic speech recognition capability for evaluating the actual performance of the speech recognizer in an in-car environment is introduced. Finally, preliminary studies

Figure 1. Visual signal captured by the three cameras. (a) The driver's face (left upper), (b) the driver, the operator and the back view (right upper) and (c) front view (right bottom).

on intermedia data conversion will be described as an example of the corpus study, in Section 5.

2. Data Collection Vehicle

The DCV is a car specially designed for the collection of multimedia data. The vehicle is equipped with eight network-connected personal computers (PCs). Three PCs have a 16-channel analog-to-digital and digital-to-analog conversion port that can be used for recording and playing back data. The data can be digitized using 16-bit resolution and sampling frequencies up to 48 kHz. One of these three PCs can be used for recording audio signals from 16 microphones. The second PC can be used for audio playback on 16 loudspeakers. The third PC is used for recording five signals associated with the vehicle: the angle of the steering wheel, the status of the accelerator and brake pedals, the speed of the car and the engine speed. These vehicle-related data are recorded at a sampling frequency of 1 kHz in 2-byte resolution. In addition, location information obtained from the Global Positioning System (GPS) is also recorded at the sampling frequency of 1 Hz.

Three other PCs are used for recording video images (Fig. 1). The first camera captures the face of the driver. The second camera captures the conversation between the driver and the experiment navigator. The third camera captures the view through the windshield. These images are coded into MPEG1 format. The remaining two PCs are used for controlling the experiment. The multimedia data on all systems are recorded synchronously. The total amount of data is about 2 gigabytes for about a 60-minute drive during which three dialogue sessions are recorded. The recorded data is directly stored on the hard disks of the PCs in the car.

Figure 2 shows the arrangement of equipment in the DCV, including the PCs, a power generator with batteries, video controller, microphone amplifiers and speaker amplifiers. An alternator and a battery are installed for stabilizing the power supply. Figure 3 shows

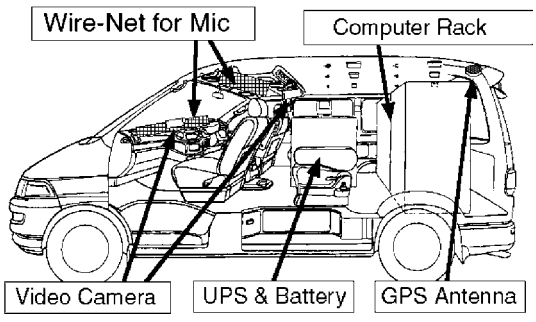

Figure 2. Configuration of DCV.

Table 1. Specifications of recording devices.

Type of data	Specifications
Sound input	16 ch, 16 bit, 16 kHz
Sound output	16 ch, 16 bit, 16 kHz
Video input	3 ch, MPEG1
Control signal	Status of accelerator and brake, angle of steering wheel Engine RPM, Speed: 16 bit, 1 kHz
Location	Differential Global Positioning System (DGPS)

Figure 3. The interior of the DCV. Wire nets are attached for various arrangement of microphones.

the interior of the DCV. Wire nets are attached to the ceiling of the car so that the microphones can be arranged in arbitrary positions. Figure 4 shows plots of the vehicle-related data such as the status of brake and accelerator pedals, the rotational speed of the engine motor, and the speed.

3. Speech Materials

The collected speech materials are listed in Table 2. The task domain of the dialogues is the restaurant guidance around the Nagoya University campus. In dialogues with a human operator and the WOZ system, we have prompted the driver to issue natural and varied utterances related to the task domain, by displaying a 'prompt panel'. On the panel, a keyword, such as *fast food*, *bank*, *Japanese food*, or *parking*, or a situation sentence, such as 'Today is an anniversary. Let's have a party.', 'I am so hungry. I need to eat!' or 'I am thirsty. I want a drink!', are displayed. In these modes, therefore, the driver takes the initiative in the dialogue. The

Table 2. Speech materials recorded in the experiment.

Item	Approx. time
Prompted dialogue	5 min
Natural dialogue	5 min
Dialogue with system	5 min
Dialogue with WOZ	5 min
Repeating phonetically balanced sentences (driving)	10 min
Reading phonetically balanced sentences (idling)	5 min

operator also navigates the driver to a predetermined destination while they are having a dialogue, in order to simulate the common function of a car-navigation system. All responses of the operator are given by synthetic speech in the WOZ mode. In addition, fully natural dialogues are also conducted between the driver and a distant operator via cellular phone. In such natural dialogues, the driver asks for the telephone number of a shop from the yellow pages information service. These natural dialogues are collected both when idling the engine and while driving the car.

All utterances have been phonetically transcribed and tagged with time codes. Tagging is performed separately for utterances by the driver and by the operator so that timing analysis of the utterances can be carried out. On average, there are 380 utterances and 2768 morphemes in the data for a driver.

In addition to the dialogues, speech of the text read aloud and isolated word utterances have also been collected. Each subject read 100 phonetically balanced sentences while idling the engine and 25 sentences while driving the car. A speech prompter is used to present the text while driving. The speech data of the read text is mainly used for training acoustic models. The set of isolated word utterances consists of digit strings and car control words.

4. Preliminary Results on Collected Data

Since dialogue between man and machine is one of our final goals, we are collecting man-machine dialogues using a prototype spoken dialogue system that has speech recognition capabilities. The task domain of the prototype system is restaurant information. Drivers can retrieve information and make a reservation at a restaurant near the campus by conversing with the system. The automatic speech recognition module of the system is based on a common dictation software platform known as Julius 3.1 [8].

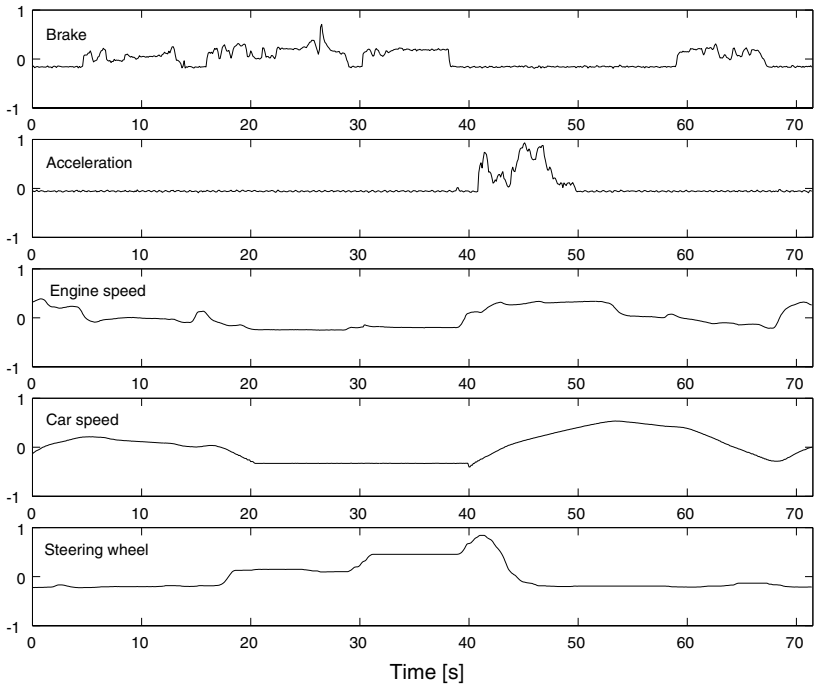

Figure 4. Vehicle-related signals. Brake and accelerator pedals, rotational speed of the engine motor and speed (from top to bottom).

A trigram language model with a 1500-word vocabulary is trained using about 10000 sentences. The main body of the training sentences is extracted from the human-human dialogue collected in the early stage of the experiment. The other sentences are generated from a finite state grammar that accepts permissible utterances in the task domain. State clustered triphone hidden Markov models consisting of 3000 states are used as acoustic models. The number of mixtures for each state is 16. The models are trained using 40,000 phonetically balanced sentences uttered by 200 speakers recorded in a soundproof room with a close-talking microphone [9]. The same microphone as in this recording is used for speech input in the prototype dialogue system. A preliminary evaluation of the speech recognition module of the system over 1,153 sentences of 48 speakers has given a word correctness of about 70% under real driving conditions in the city area, where the typical driving speed is less than 40 km/h.

The dialogue is controlled by transitions among 12 states, each of which corresponds to the database query results. When a set of particular conditions defined for a transition is satisfied, the predefined state transition occurs, invoking associated actions, i.e., generating speech responses. Examples of the dialogue states are listed in the Table 3. Up to today, 75% of the man-machine dialogues have been correctly completed by the system.

An example of a dialogue between the ASR system and a user is shown in Fig. 5.

5. Cross-Media Signal Conversion

Among the ongoing research projects concurrent to data collection, cross-media signal conversion makes use of the multimedia aspect of the corpus. In this

Table 3. Examples of the states defined for controlling dialogues.

ID	Function
1	Initial state.
2	More than three restaurants to announce
3	No restaurant to announce
4	Less than three restaurants to announce
5	The restaurant is decided
6	Error (the dialogue has too many turns.)
7	Error (the session is too long.)
8	Task completed

```
System: Please give me a condition for restaurant query.
User: Japanese restaurant that serves about 1000-yen food.
System: Japanese, 1000-yen. Is that right ?
User: Yes.
System: There are Mikado, Shogun and Kyoto near here.
System: Which restaurant do you prefer ?
User: I would like to go to Mikado.
System: Mikado. Is that right ?
User: Yes.
System: You can reserve a table for that restaurant.
System: Do you reserve ?
User: No.
System: Do you need a navigation to the restaurant ?
User: No.
```

Figure 5. An example of a dialogue between the ASR system and a subject.

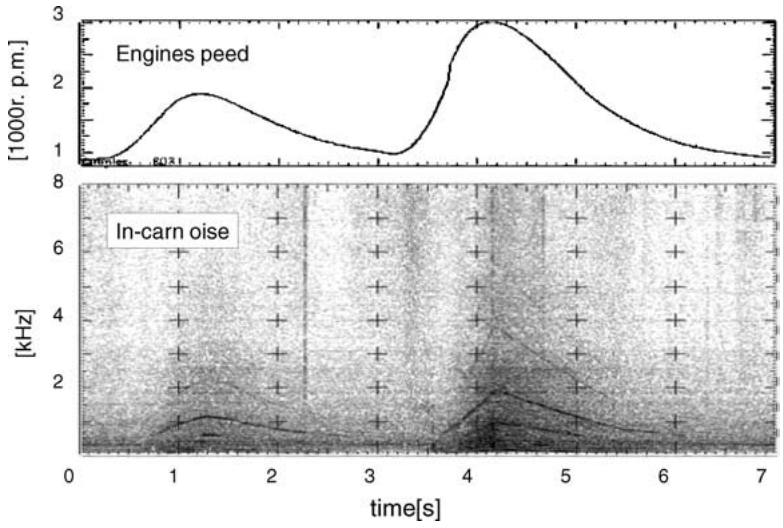

Figure 6. Engine speed and related in-car noise.

section, a typical example of the research effort, synthesizing in-car noise, is described.

Because the nonstationarity of the background noise is one of the most difficult problems in in-car speech processing, accurate prediction of the noise signal is expected to improve various speech processing performances. In this study, therefore, in-car noise is predicted using the driving signals, i.e., the car speed and the engine speed signals. As shown in Fig. 6, the engine speed governs the timbre of in-car noise by producing peaks in the spectrum. The power contour of the noise signal, on the other hand, is proportional to the car speed signal. Therefore, one can model the car noise signal as a superposition of the engine-speed-dependent and car-speed-dependent noise signals. Hereafter, these two noise signals are referred to as engine noise and road noise. Because both the engine-speed and car-speed signals are collected synchronously with the in-car noise signal, we can formulate the noise prediction problem as cross-media signal conversions, i.e., conversion from the engine-speed and car-speed signals to the in-car noise signal.

The engine-speed-dependent signal is synthesized by convoluting random white Gaussian noise with the frequency response of the given engine speed. The frequency response is calculated by adding some spectral peaks at the fundamental frequency, f_c, given as 16 times the combustion frequency, and their multiples

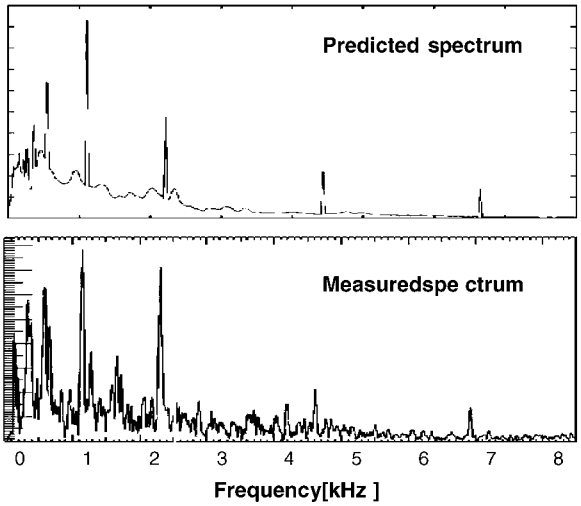

Figure 7. Spectrum of the synthesized (above) and the original (below) engine noise.

to the averaged spectral magnitude $\langle H(f)\rangle$, i.e.,

$$H(f) = \langle H(f)\rangle + A\delta(f - f_c) + \sum_{n=1}^{3} \frac{A}{n+1}\delta(f - 2nf_c) + \sum_{n=2}^{4} \frac{A}{n}\delta(f - 1/nf_c),$$

where $\delta(f - f_c)$ represents a line spectrum at frequency f_c ([Hz]). Therefore, the third term of the right hand side of the equation gives the harmonics of the fundamental frequency, whereas the fourth term gives the subharmonic components, which arise due to the nonlinearity of the vibration. The validity of the model can be seen in Fig. 7 where the predicted spectrum is seen to simulate the measured spectrum well.

The road noise, on the other hand, is generated by concatenating speed-dependent noise waveform prototypes. A noise waveform prototype is prepared for every 5 km/h speed, i.e., 5, 10, . . . , 85 km/h. For the generation of the noise signal, the prototypes are replaced every 2 seconds according to the car speed.

In a subjective test, the mean opinion score (MOS) of the generated noise signal was rated 3.1, whereas the recorded noise and white noise, whose amplitude envelope is proportional to the car speed, were rated 4.6 and 1.1, respectively. These results confirm the feasibility of cross-media signal conversion.

6. Summary

In this paper, we presented details of constructing a multimedia corpus of in-car speech communication. The corpus consists of synchronously recorded multichannel audio/video signals, driving signals and GPS output. The spoken dialogues of the driver were collected in various styles, i.e., human-human and human-machine, prompted and natural, for the restaurant guidance task domain. An ASR system was utilized for collecting human-machine dialogues.

To date, more than 500 subjects have been enrolled in data collection. Ongoing acoustic/language model training for robust speech recognition is being performed using the corpus. Furthermore, making full use of the multimedia characteristics of the corpus, a cross-media signal conversion study has begun.

The corpus will be available for various research purposes in early 2002, and is planned to be extended to English and other languages.

Acknowledgments

This research has been supported by a Grant-in-Aid for COE Research (No. 11CE2005).

References

1. J.C. Junqua and J.P. Haton, *Robustness in Automatic Speech Recognition*. Kluwer Academic Publishers, 1996.
2. D. Roy, "'Grounded' Speech Communication," in *Proc. of the International Conference on Spoken Language Processing, ICSLP 2000*, Beijin, 2000, pp. IV69–IV72
3. P. Gelin and J.C. Junqua, "Techniques for Robust Speech Recognition in the Car Environment," in *Proc. of European Conference Speech Communication and Technology, EUROSPEECH'99*, Budapest, 1999.
4. M.J. Hunt, "Some Experiences in In-Car Speech Recognition," in *Proc. of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, 1999, pp. 25–31
5. P. Geutner, L. Arevalo, and J. Breuninger, "VODIS—Voice-Operated Driver Information Systems: A Usability Study on Advanced Speech Technologies for Car Environments," in *Proc. of International Conference on Spoken Language Processing*, ICSLP2000, Beijin, 2000, pp. IV378–IV381.
6. A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, J. Allen, and Stephan Eule, "SpeechDat-Car: A Large Speech Database for Automotive Environments," in *Proc. of 2nd Int'l Conference on Language Resources and Evaluation*, Athens, LREC 2000.
7. N. Kawaguchi, S. Matsubara, H. Iwa, S. Kajita, K. Takeda, F. Itakura, and Y. Inagaki, "Construction of Speech Corpus in Moving Car Environment," in *Proc. of International Conference on Spoken Language Processing*, ICSLP2000, Beijin, 2000 pp. 362–365.

8. T. Kawahara, T. Kobayashi, K. Takeda, N. Minematsu, K. Itou, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano, "Japanese Dictation Toolkit: Plug-and-Play Framework for Speech Recognition R&D," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'99)*, 1999 pp. 393–396.

9. K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research, *J. Acoust. Soc. Jpn.(E)*, vol. 20, no. 3, 1999, pp. 199–206.