PAPER *Special Section on Corpus-Based Speech Technologies*

Construction and Evaluation of a Large In-Car Speech Corpus

Kazuya TAKEDA[†a)], ***Member***, **Hiroshi FUJIMURA**[†], ***Nonmember***, **Katsunobu ITOU**[†], **Nobuo KAWAGUCHI**[††], **Shigeki MATSUBARA**[††], ***Members, and*** **Fumitada ITAKURA**[†††], ***Fellow***

SUMMARY In this paper, we discuss the construction of a large in-car spoken dialogue corpus and the result of its analysis. We have developed a system specially built into a Data Collection Vehicle (DCV) which supports the synchronous recording of multichannel audio data from 16 microphones that can be placed in flexible positions, multichannel video data from 3 cameras, and vehicle related data. Multimedia data has been collected for three sessions of spoken dialogue with different modes of navigation, during approximately a 60 minute drive by each of 800 subjects. We have characterized the collected dialogues across the three sessions. Some characteristics such as sentence complexity and SNR are found to differ significantly among the sessions. Linear regression analysis results also clarify the relative importance of various corpus characteristics.

key words: *speech corpus, in-car speech recognition, perplexity, SNR*

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 [1]. To develop an advanced in-car speech interface, however, not only one but many real-world problems, such as noise robustness, distortion due to distant talking [2] and disfluency while driving, must be overcome [3], [4].

In particular, the difficulty of in-car speech processing is characterized by its variety. Road and traffic conditions, the car's condition and the movements of the driver change continuously and affect the driver's speech. From the interface viewpoint, the difference in the navigator, e.g., a human operator or an automatic speech recognition system, may also cause variability [5], [6]. Therefore, a large corpus is indispensable in the study of in-car speech, not only for training acoustic models under various background noise conditions but also for building a new model of the combined distortions of speech [7]–[10].

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 applications, we need a corpus 1) that encompasses a wide 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.

For 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 from each of 800 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 Sect. 3, we present the details of the data collection scenario. In Sect. 4, the basic statistics of the collected corpus are summarized. Although the characteristics of the collected data are calculated for three different navigation modes directed by, i.e., a Human Navigator (HN), a Wizard of OZ (WOZ) and an Automatic Speech Recognition system (ASR), the aim of this section is not to find a consistent model of their difference*.

In Sect. 5, we show the results of speech recognition experiments using language and acoustic models trained for three dialogue modes. We also show the results of linear regression analyses between word accuracy and characteristics of the utterances for the three dialogue modes.

2. Data Collection Vehicle

The Data Collection Vehicle (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 with 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 RPM.

Manuscript received July 6, 2004.
Manuscript revised September 11, 2004.

[†] The authors are with the Graduate School of Information Science, Nagoya University, Nagoya-shi, 464–8603 Japan.
[††] The authors are with the Information Technology Center, Nagoya University, Nagoya-shi, 464–8601 Japan.
[†††] The author is with Meijo University, Nagoya-shi, 468–8502 Japan.
a) E-mail: takeda@is.nagoya-u.ac.jp

DOI: 10.1093/ietisy/e88–d.3.553

* In [5], experimental results comparing the acoustic and linguistic features of the user's dialogue utterances in different dialogue mode are analysed through simulated driving experiments.

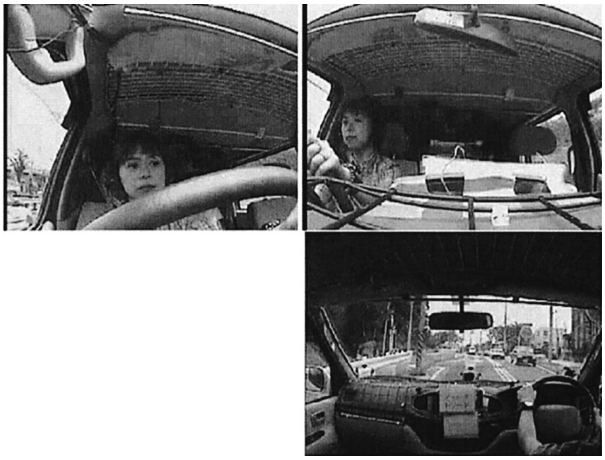


Fig. 1 Visual signals captured by the three cameras. (a) the driver's face (left upper), (b) the driver and the back view (right upper) and (c) front view (right bottom).

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

These vehicle-related data (Table 1) are recorded at a sampling frequency of 1 kHz with 2-byte resolution. In addition, location information obtained from the Global Positioning System (GPS) is also recorded at a 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 back view of dialogue between the driver and the experimental navigator. The third camera captures the front view through the windshield. These images are encoded in MPEG1 format. The two remaining PCs are used for controlling the experiment. The multimedia data from all systems are recorded synchronously. The total amount of data is approximately 2 gigabytes for approximately a 60-minute drive during which three dialogue sessions are recorded. The recorded data is stored directly 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. Wire nets are attached to the ceiling of the car so that the microphones can be arranged in arbitrary positions. Figure 3 shows the positions of the microphones used for the data collection in the DCV. The average SNR at each microphone position is listed in Table 2.

Figure 4 shows plots of the vehicle-related data such as the status of brake and accelerator pedals, the RPM of the engine motor, and the vehicle speed.

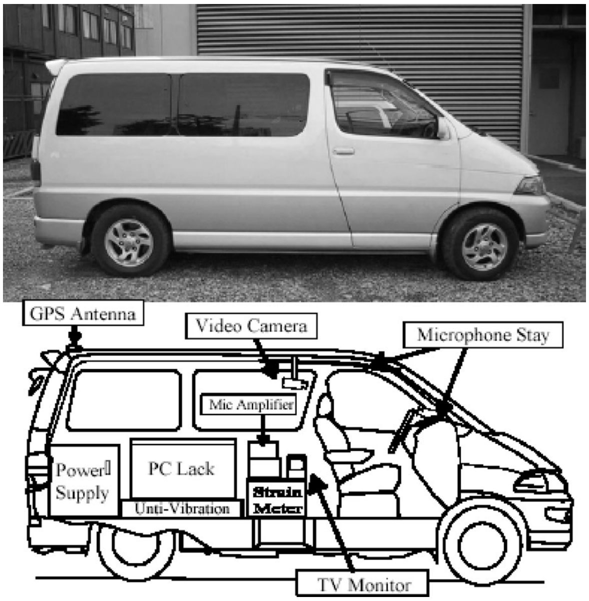


Fig. 2 Configuration of DCV.

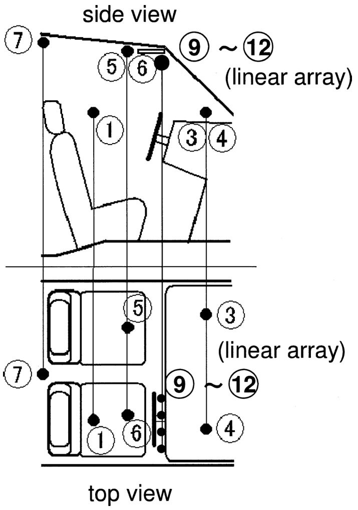


Fig. 3 Microphone positions for the data collection.

3. Speech Material and Collection Scenario

We have carried out our extensive data collection from 1999 through 2001 including over 800 subjects under both driving and idling conditions. The collected data types are shown in Table 3. In particular, during the first year, we collected the following data from 212 subjects: (1) pseudo information retrieval dialogue between a subject and the human navigator, (2) phonetically balanced sentences, (3) isolated words, and (4) digit strings.

In the 2000-2001 collection, however, we have included two more dialogue modes such that each subject has completed a dialogue with three different kinds of interface systems. The first system is a human navigator (HN), who

Table 2 Microphone positions and SNR at those positions.

#	position	ave. SNR (dB)	
		idling	driving
1	driver headset	-	-
2	navigator headset	-	-
3	on the navigator dashboard	8.03	5.25
4	on the driver dashboard	12.25	7.87
5	on the navigator visor	10.32	7.40
6	on the driver visor	13.74	11.52
7	on the center ceiling, front seats	10.29	8.23
8	on the center ceiling, back seats	-	-
9-12	linear array, on the driver visor	-	-

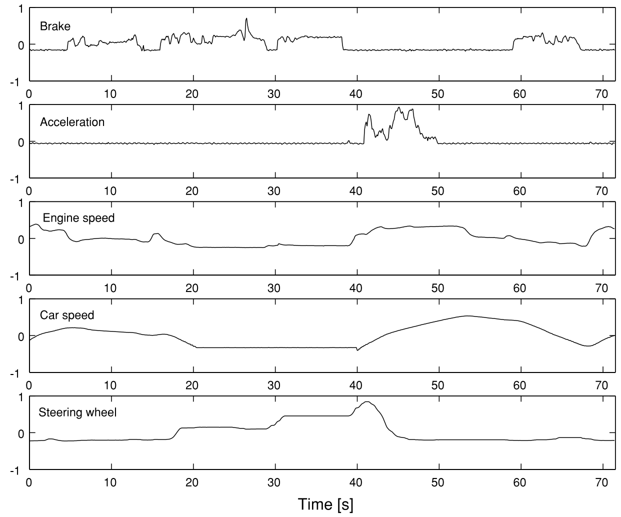


Fig. 4 Vehicle-related signals. Brake and accelerator pedals, the RPM of the engine motor, carspeed and steering wheel (from top to bottom).

Table 3 Speech materials recorded in the experiment.

1999 collection	212 subj.
Spoken dialogue with human navigator	11 min
Phonetically balanced sentences	
(Idling)	50 sent.
(Driving)	25 sent.
Isolated words	30 words
Digit Strings	4 digit * 20
2000-2001 collection	600 subj.
Spoken dialog with human navigator	5 min.
Spoken dialog with WOZ system	5 min.
Spoken dialog with ASR system	5 min.
Phonetically balanced sentences	
(Idling)	50 sent.
(Driving)	25 sent.
Isolated words	30 words
Digit Strings	4 digit * 20

sits on the back seat and converses naturally. The second one is a wizard of Oz (WOZ) type system. The final one is an automatic dialog set-up based on automatic speech recognition (ASR).

3.1 Multimode Dialogue Data Collection

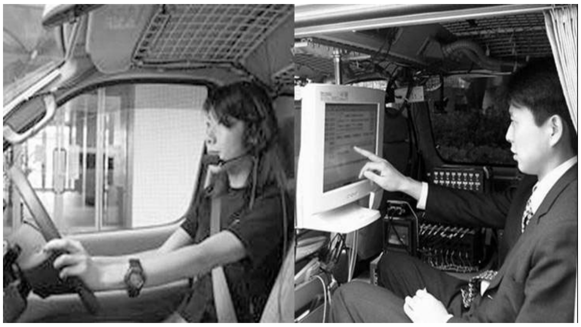


Fig. 5 A sample scene of dialogue recording using WOZ.

The primary objective of the dialogue speech collection is to record the three different modes of dialogue mentioned earlier. It is important to note that the task domain is the information retrieval task for all three modes.

To simplify the dialogue recording process, the navigator prompts each task by using several levels of a task description panel to initiate the spontaneous speech. There are a number of task description panels associated with our task domains. A sample set from the task description panels is as follows: 'Fast food', 'Hungry', 'Hot summer, thirsty', 'No money', and 'You just returned from abroad'.

All of our recorded dialogues are transcribed into text in compliance with a set of criteria established for the Corpus of Spontaneous Japanese (CSJ) [12]. We have collected more than 187 hours of speech data corresponding to approximately one million morpheme dialogue units.

3.1.1 Dialogue with Human Navigator (HN)

Navigators are trained in advance and have extensive information for the tasks involved. However, in order to avoid a dialogue divergence, some restrictions are placed on the way they can speak.

3.1.2 Dialogue with Wizard of OZ System (WOZ)

The WOZ mode is a spoken dialogue platform which involves a touch-panel input for the human navigator and a speech synthesizer output. The system has a considerable list of shops and restaurants along the route and the navigator uses the system to search for and select the most suitable answer for subjects' spoken requests (Fig. 5).

3.1.3 Dialogue with Spoken Dialogue System (ASR)

The dialogue system called "Logger" performs a slot-filling dialogue for the restaurant retrieval task. The vocabulary size of the task is 1,500 and the utterances are modeled using bigram language model. The close-talking microphone is used for the speech input in order to use Continuous Speech Recognition Consortium (CSRC) standard triphone HMM and a Julius decoder as an acoustic model and the LVCSR engine [11], respectively.

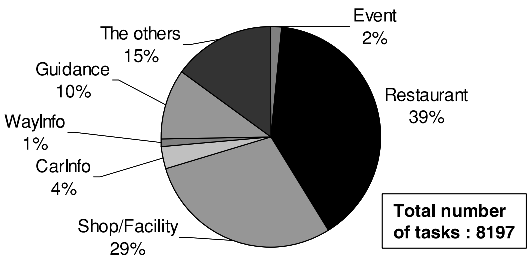


Fig. 6 Distribution of the task domain of all dialogue sentences.

3.1.4 Task Domains of HN and WOZ Sessions

We have categorized the dialogue sessions recorded through HN and WOZ modes into several task domains. In Fig. 6, we show a breakdown of the major task domains. It is easy to see that approximately forty percent of the tasks are related to restaurant information retrieval.

3.2 Phonetically Balanced Sentences

In addition to the dialogue speech, each subject has read 50 phonetically balanced sentences in the car while the vehicle was idling, and subsequently drivers have also spoken 25 sentences while driving the car. While idling, subjects have used a printed text posted on the dashboard to read a set of phonetically balanced sentences. While driving, we have employed a slightly different procedure for safety reasons. In this case, subjects are prompted for each phonetically balanced sentence from a head-set utilizing a specially developed waveform playback software. The phonetically balanced sentences are mainly used for acoustic model construction.

4. Corpus Evaluation

In this section, the basic characteristics of the collected data are given. Although statistics are calculated for each of the three dialogue sessions, the aim is not presenting a particular model of the difference among the sessions[†].

4.1 Corpus Size

The characteristics of the corpus used in this analysis are summarized in Table 4 for each dialogue session. We use 32,000 to 40,000 speech units uttered by 435 speakers for the analysis. A “speech unit” is a segment of speech that is separated by a silence of longer than 200 ms, therefore, in this corpus, most “utterances” consist of a “speech unit”. The boundaries between silences and utterances are determined manually. The average length of a drivers’ speech unit in the HN session is longer than the other two sessions in both duration, 2.26 sec., and number of morphemes, 8.72.

Table 4 Corpus statistics for 435 speakers.

(a) size

		HN		WOZ		ASR	
Total(sec)		101430		73116		80978	
Driver	Operator	0.40	0.60	0.39	0.61	0.21	0.79
Speech Unit		40560		32883		40149	
Driver	Operator	0.44	0.56	0.42	0.58	0.40	0.60
duration/unit(sec)		2.26		2.05		1.07	
Morph.		353875		195513		262354	
Driver	Operator	0.34	0.66	0.44	0.56	0.19	0.81
Morph./Unit		8.72		5.95		6.53	

(b) complexity (entropy)

bi-gram	trigram	18.1	7.7	14.1	7.1	9.1	6.6
vocabulary size		5001		3216		1839	

(c) Acoustic characteristics

		close	visor	close	visor	close	visor
SNR[dB]		23.0	10.6	24.0	11.3	26.0	12.9

(d) speaking rate (mora length)

SPR[msec/mora]		144.3		143.5		149.1	

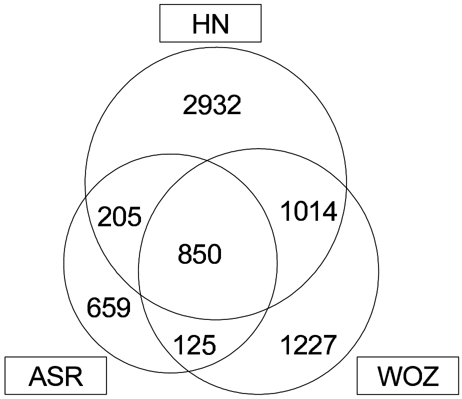


Fig. 7 Overlap between the vocabularies in difference sessions.

Since the dialogues consist of question-and-answer pairs, the ratio between the driver and navigator (system) utterances are not significantly different among sessions, i.e., about 40-45% of the utterances are made by the driver.

4.2 Complexity of the Utterance

As seen from the table, the vocabulary size used in the ASR session is less than half of that in the HN session and that for the WOZ session is in-between the two sessions. The overlap between the vocabularies in the different sessions is shown in Fig. 7. The vocabulary of the ASR session can be regarded as a subset of the HN session. The additional vocabulary of the HN session mainly consists of the names of restaurants and out-of-task words such as “business trip”, “birthday party”, “hungry”.

The bigram and trigram entropies of the utterances are almost proportional to the session vocabulary sizes, i.e., 18.1, 14.1 and 9.1 for HN, WOZ and ASR, respectively[††].

[†] Results of the comparative analysis of dialogues with different navigators has been reported in [5]. In [5], utterances under simulated driving conditions are analyzed, whereas utterances under real driving conditions are collected in this corpus.

[††] The causes of the characteristic differences between the human-human and human-machine dialogues are discussed in [5].

4.3 Acoustic Condition

The SNR condition of each utterance is estimated from the speech signal by fitting a two-mixture Gaussian distribution to the histogram of log-power values that is calculated in a frame-by-frame manner over an utterance. After finding two Gaussian distributions, the difference between the two mean values is used as the SNR of the utterance. Therefore, the estimated SNR is not accurate if the SNR of the original signal is negative [13]. The SNR of the utterances in the ASR session is better than those of the HN and the WOZ sessions by approximately 2 dB. Since the driving conditions, and therefore noise conditions, are designed to be the same across all three sessions, the driver speaks in a louder voice in the ASR session.

4.4 Speaking Rate

We define the speaking rate by using the average mora length. It is calculated using the result of forced alignment of the reference monophone label. The speaking rate of the utterances in the ASR session is slower than that in the HN and WOZ sessions, by approximately 5 msec/mora.

5. Speech Recognition Experiments

In this section, we discuss the characteristics of the corpus in terms of speech recognition accuracy.

5.1 Experimental Setup

Language models were constructed for each of the three sessions. The size of the text data for language model training is shown in Table 5. A training text was tagged by the morphological analyzer ChaSen [14], and then manually corrected. The morphological label includes part-of-speech information. Forward word bigram and backward word trigram models without any cutoffs are trained. An "open language model" is trained for each speaker, for which the utterances made by the speaker were not used, whereas a "closed language model" is trained using all utterances.

Two different acoustic models, i.e., the "close-talking" and "visor" models, were trained on the basis of the speech captured through a close-talking microphone (#1 in Fig. 3) and s distant microphone (#6 in Fig. 3), respectively. The number of speakers used for the acoustic model training is listed in Table 6. Excluding short utterances of less than ten syllables, (most of them are yes/no answers), 22.4 hours of speech signals were used for the experiment. The breakdown of the data is: 11,746 utterance units, 11.4 hours for the HN sessions; 8,550 utterance units, 7.84 hours for the WOZ sessions; and 4,878 utterance units, 3.10 hours for the ASR sessions.

In order to perform an open speaker set experiment for acoustic models, we divided all speakers into five groups and trained five different HMMs using the utterances of the four speaker groups. The feature parameters for the HMM acoustic model were 12MFCC, 12ΔMFCC and Δ log power. Although the original signal is sampled at 16 kHz, the bandwidth was limited to the range from 250 Hz to 8000 Hz. The basic structure of the HMM is three-state continuous density triphones that share 2000 states with 32 Gaussian mixture components. All triphones have a simple left-to-right topology except for the short pause which has a transition from start state to final state. Julius was used as the decoder.

Table 5 Training sentences for language models used for recognition experiments.

	HN	WOZ	ASR
number of subjects	535	586	575
male	342	337	368
female	193	209	207
# of speech units	22240	19044	21289
# of morpheme	149213	117250	66612
vocabulary size	5532	3694	2083
# of bi-grams	35095	22277	9850
# of trigrams	67972	44322	18403

Table 6 Training sentences for acoustic models used for recognition experiments.

	HN	WOZ	ASR
number of subjects	534	527	512
male	341	338	326
female	193	189	186

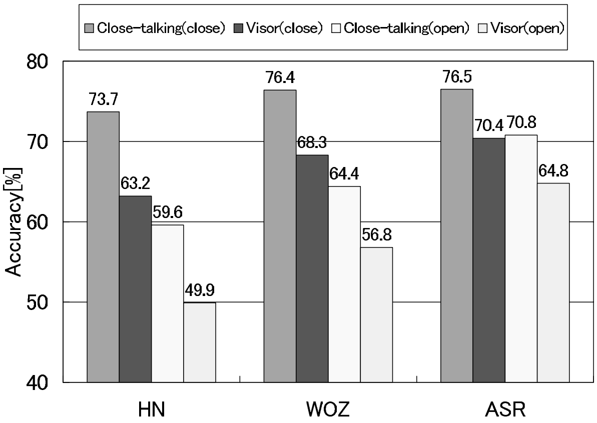


Fig. 8 Recognition performance for the utterances in three sessions. "open" represents the results using both the open language model and the open acoustic model, whereas "close" represents the results using both the close language model and the close acoustic model.

5.2 Performance Comparison over Sessions

The recognition performances for the utterances collected in three sessions are shown in Fig. 8. The performance is worst for the HN session utterances and best for the ASR session. In the visor microphone case, in particular, the performance difference between HN and ASR becomes approximately 15%. The difference in recognition accuracy among dialogue modes is remarkable under noisy conditions.

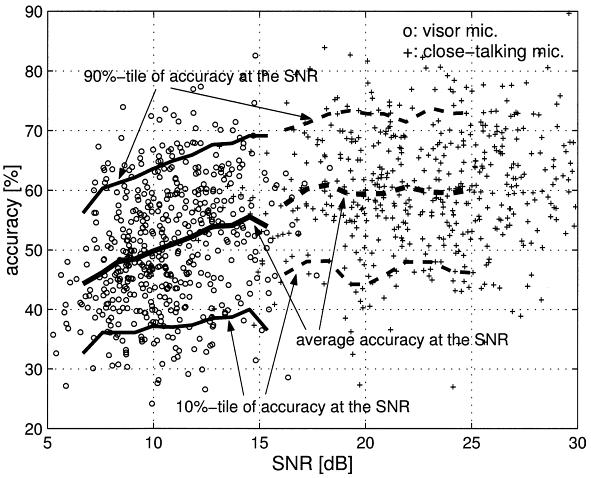


Fig. 9 Word accuracy score and SNR value averaged over each speaker for the utterances in the HN (human navigator) session. Thick and thin lines indicate, 10%-tile, average and 90%-tile accuracies at the SNRs. Broken and solid lines correspond to close-talking microphone and visor microphone results, respectively.

5.3 Performance Comparison over SNR, Entropy and Speaking Rate

In order to show the variabilities contained in the corpus, we have calculated the distributions of the recognition accuracy against the various characteristics of the speaker, i.e., SNR, entropy of the utterance, average mora length, and its standard deviation, for the utterances in the HN (human navigator) session.

In Fig. 9, the average word accuracy and SNR are plotted for all drivers. When the close-talking microphone is used, the dependency of the recognition accuracy on SNR is not so high, i.e., it is approximately 0.3%/dB. Cross speaker variability seems to be much larger. For the visor microphone speech, however, the dependency becomes higher, and reaches approximately 1.25%/dB[†].

In Fig. 10 the average word accuracy and entropy are plotted for all drivers. The average entropy of the speaker ranges from 3 to 6 bits. (The entropy of each utterance is calculated as the crossentropy against the trigram language model, then averaged for a speaker.) There is no difference between close-talking and visor microphones in terms of the dependency of the recognition accuracy on the entropy. It can be seen that the relationship between the entropy and the recognition accuracy is inversely linear and highly correlated.

In Fig. 11, the average word accuracy and mora length are plotted for all drivers. Unlike those reported for the monologue corpus [15], the effect of the speaking rate on the accuracy of dialogue speech recognition was small. As shown in Fig. 12, on the other hand, the average word accuracy depends more heavily on the variability (standard deviation) of mora length of a speaker.

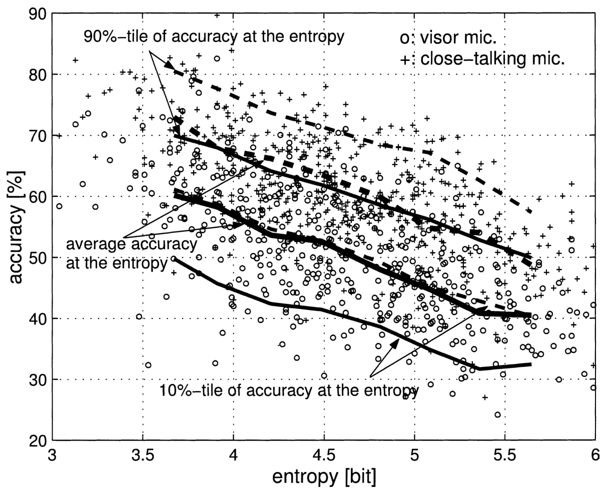


Fig. 10 Word accuracy score and entropy value averaged over each speaker for the utterances in the HN (human navigator) session. Thick and thin lines indicate, 10%-tile, average and 90%-tile accuracies at the entropies. Broken and solid lines correspond to close-talking microphone and visor microphone results, respectively.

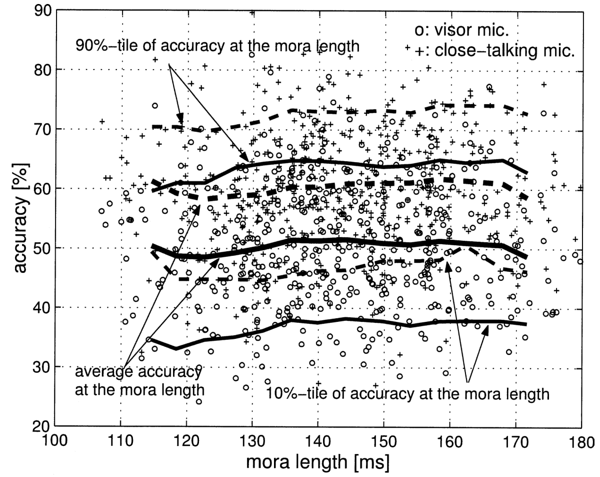


Fig. 11 Word accuracy score and mora length averaged over each speaker for the utterances in the HN (human navigator) session. Thick and thin lines indicate, 10%-tile, average and 90%-tile accuracies at the mora length. Broken and solid lines correspond to close-talking microphone and visor microphone results, respectively.

5.4 Regression Analysis

In Fig. 13, the correlation coefficients between word accuracy and entropy, SNR, mora length and its deviation are plotted. The highest linear correlation is found between accuracy and entropy. Particularly for the ASR session, the correlation coefficient of approximately 0.77 was obtained. The correlation between word accuracy and SNR is high in visor microphone speech cases where the correlation coefficient was approximately 0.25 to 0.28.

[†] It is a generally known fact and is not found by the authors that performance improvement due to the SNR improvement saturates under the high SNR condition.

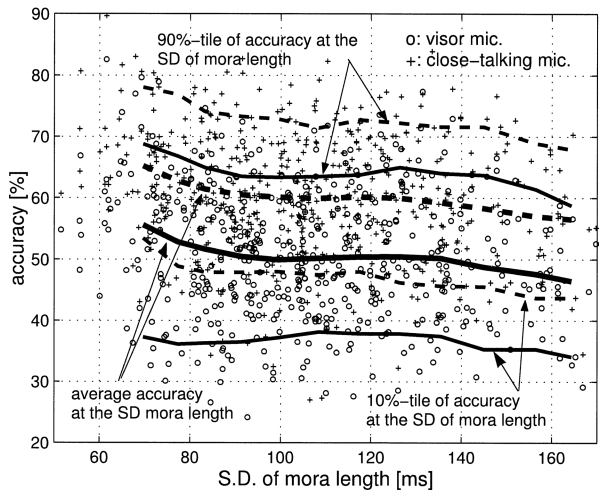


Fig. 12 Word accuracy score and standard deviation (S.D.) of the mora length distribution over each speaker for the utterances in the HN (human navigator) session. Thick and thin lines indicates 10%-tile, average and 90%-tile accuracies at the S.D. of mora length. Broken and solid lines correspond to the close-talking microphone and visor microphone results, respectively.

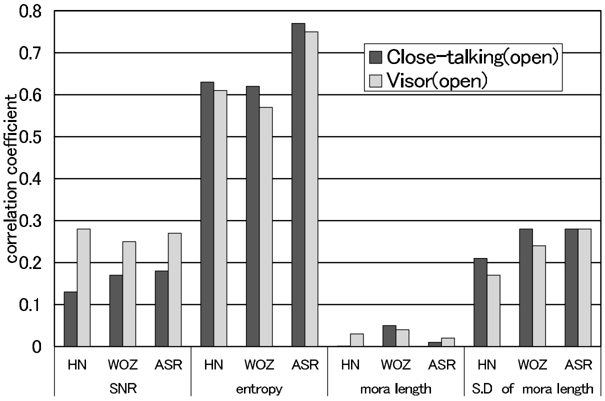


Fig. 13 Correlations analysis results between recognition accuracy and various corpus characteristics.

6. Discussion

In Figs. 9-12, 80% distribution ranges of accuracy are also given. A thick line shows the average accuracy and thin lines indicate the lower and upper bounds of the 80% of speakers. Although each point spreads over a wide range of accuracy in the original scatter plot, the averaged value shows a clearer relationship between the accuracy and factors, i.e., SNR, entropy, mora length and its standard deviation.

In Table 7, the average correlation calculated along the thick line and 80% distribution ranges are listed for each factor. The entropy, SNR (visor microphone case only) and the standard deviation of mora length have a high correlation with averaged accuracy, in this order. It can also be seen that the distribution range of the accuracies can be limited to 20% (for the entropy case) or 25% (other cases) by disregarding 20% outliers.

On the basis of the above results, some important facts concerning speech variability under real driving conditions have been found in the corpus.

Table 7 Correlation between word accuracy and corpus characteristics. 80% distribution ranges are calculated by averaging the distance between the two thin lines, i.e. 10%-tile and 90%-tile values, plotted in Fig. 9 to Fig. 12. The average correlation shows the correlation between the characteristic values, i.e., SNR, entropy, speaking rate and its deviation, and the "averaged" accuracy at that region, that is given by the thick lines in Fig. 9 to Fig. 12. All values are calculated for human navigator sessions.

	SNR	entropy	mora len.	S.D. of mora len.
for the visor mic.				
80% distribution range (%)	27.5	21.4	26.8	27.3
average correlation	0.97	−0.99	0.31	−0.89
for the close-talking mic.				
80% distribution range (%)	25.9	20.0	25.6	25.4
average correlation	0.33	−0.99	0.36	−0.93

- Even under the same condition, the distribution of the accuracy had a range up to 40%. 20% outliers contribute 13 to 20% of the 40% range.
- The SNR distributes in a range from 5 to 15 dB at the visor position, however, the correlation between SNR and accuracy was low, i.e., less than 0.3.
- A linear relation between accuracy and entropy was found.
- Unlike the monologue corpus results, the dependency on the speaking rate was small in the accuracy of dialogue speech recognition. The variation of the speaking rate has correlation with the accuracy.

From the fact that such important results can only be found by very large scale experiments, we can conclude that the in-car speech corpus will play a vital role in further speech research.

7. Conclusion

In this paper, we have presented a brief description of a large corpus of in-car speech communication. The corpus consists of synchronously recorded multichannel audio/video signals, driving signals, and a differential GPS reading. For a restaurant information query task domain speech dialogues were collected from over 800 drivers in three different modes, namely, human-human, WOZ and human-machine. In addition, we have experimented with an ASR system for collecting human-machine dialogues. Every spoken dialogue is transcribed with precise time stamp.

This paper also reported the characteristics of the collected conversational utterances. Four hundred and thirty five drivers' utterances for three different modes are characterized by vocabulary size, average perplexity of the sentences, SNR and speaking rate. Furthermore, through large-scale speech recognition experiments, several important results have been found. 1) The correlation between SNR and accuracy is low; 2) there is a linear relationship between accuracy and entropy; and 3) the dependency on the speaking rate was small in the accuracy.

Through these discussions, the effectiveness of the corpus in real-world speech recognition research was clarified. Currently, the corpus is being used for various speech related research including distributed microphone approach for in-car robust speech recognition [16], a statistical Spectral Subtraction method [17], corpus based dialogue controlling [18] and the driver verification [19].

Acknowledgements

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

References

[1] H. Abut, J.H.L. Hansen, and K. Takeda, eds., DSP for In-Vehicle and Mobile Systems, Kluwer Publishers, USA, 2004.

[2] J.C. Junqua and J.P. Haton, Robustness in automatic speech recognition, Kluwer Academic Publishers, 1996.

[3] P. Gelin and J.C. Junqua, "Techniques for robust speech recognition in the car environment," Proc. European Conference Speech Communication and Technology (EUROSPEECH '99), pp.2483–2486, Budapest, 1999.

[4] M.J. Hunt "Some experiences in in-car speech recognition," Proc. workshop on Robust Methods for Speech Recognition in Adverse Conditions, pp.25–31, 1999.

[5] T. Ito, M. Kai, Y. Iwamoto, M. Mizutani, H. Yuasa, T. Konishi, and Y. Itoh, "Comparison of linguistic and acoustic features caused by different dialogue situations in a Landmark-input Task," IPSJ Trans., vol.43, no.7, pp.2118–2129, July 2002.

[6] C. MacDermid "Features of naive callers' dialogues with a simulated speech understanding and dialogue system," Proc. 3rd European Conference on Speech Communication and Technology (EUROSPEECH 93), pp.955–958, Berlin, 1993.

[7] P. Geutner, L. Arevalo, and J. Breuninger, "VODIS — Voice-operated driver information systems: A usability study on advanced speech technologies for car environments," Proc. International Conference on Spoken Language Processing (ICSLP2000), pp.IV378–IV381, Beijing, 2000.

[8] J.H.L. Hansen, J. Plucienkowski, S. Gallant, B. Pellom, and W. Ward, "CU-Move: Robust speech processing for in-vehicle speech systems," Proc. International Conference on Spoken Language Processing (ICSLP2000), pp.I527–I530, Beijing, 2000.

[9] N. Kawaguchi, S. Matsubara, H. Iwa, S. Kajita, K. Takeda, F. Itakura, and Y. Inagaki, "Construction of speech corpus in moving car environment," Proc. International Conference on Spoken Language Processing (ICSLP2000), pp.362–365, Beijing, 2000.

[10] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, J. Allen, and S. Eule, "SpeechDat-Car: A large speech database for automotive environments," Proc. 2nd International Conference on Language Resources and Evaluation (LREC 2000), pp.373–378, Athens, 2000.

[11] A. Lee, T. Kawahara, K. Takeda, M. Mimura, A. Yamada, A. Ito, K. Itou, and K. Shikano, "Continuous speech recognition consortium — An open repository for CSR tools and models," Proc. International Conference on Language Resources and Evaluation (LREC2002), pp.1438–1441, LasPalmas, 2002.

[12] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003), Tokyo, (CDROM Proceedings), 2003.

[13] T.H Dat, K. Takeda, and F. Itakura, "Robust SNR estimation of noisy speech based on Gaussian mixtures modeling on log-power domain," COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction (Robust2004), Norwich, (CDROM Proceedings), 2004.

[14] Y. Matsumoto, A. Kitauchi, T. Yamashita, and Y. Hirano, "Japanese morphological analysis system ChaSen version 2.0 manual," NAIST Technical Report, NAIST-IS-TR99009, April 1999.

[15] T. Shinozaki and S. Furui, "Analysis on individual differences in automatic transcription of spontaneous presentations," Proc. International Conference on Acoustic Speech and Signal Processing (ICASSP2002), Orlando, vol.1, pp.729–732, 2002.

[16] H. Banno, T. Shinde, K. Takeda, and F. Itakura, "In-car speech recognition using distributed microphones — Adapting to automatically detected driving conditions," Proc. International Conference on Acoustic Speech and Signal Processing (ICASSP2003), pp.I324–I327, Hong Kong, 2003.

[17] T. Dat, K. Takeda, and F. Itakura, "Speech enhancement based on magnitude estimation using the Gamma prior," Proc. International Conference on Spoken Language Processing (INTERSPEECH 2004 - ICSLP), Jeju, 2004.

[18] H. Murao, N. Kawaguchi, S. Matsubara, Y. Yamaguchi, K. Takeda, and Y. Inagaki, "Example-based spoken dialogue system with online example augmentation," Proc. International Conference on Spoken Language Processing (INTERSPEECH 2004 - ICSLP), Jeju, 2004.

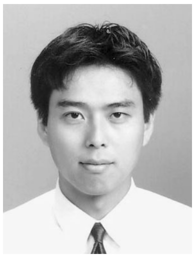
[19] K. Igarashi, C. Miyajima, K. Itou, K. Takeda, F. Itakura, and H. Abut, "Biometric identification using driving behavioral signals," Proc. International Conference on Multimedia and Expo (ICME2004), TP1-2(5), Taipei, 2004.

Kazuya Takeda received the B.S. degree, the M.S. degree, and the Dr. of Engineering degree from Nagoya University, in 1983, 1985, and 1994 respectively. In 1986, he joined ATR (Advanced Telecommunication Research Laboratories), where he involved in the two major projects of speech database construction and speech synthesis system development. In 1989, he moved to KDD R & D Laboratories and participated in a project for constructing voice-activated telephone extension system. He has joind Graduate School of Nagoya University in 1995. Since 2003, he is a professor at Graduate School of Infomation Science at Nagoya University. He is a member of the IEEE and the ASJ.

Hiroshi Fujimura received the B.E. degree from Nagoya University in 2003. He has been studying in Graduate School of Information Science of the Nagoya University. His research interest is speech recognition. He is a member of the ASJ.

Katsunobu Itou received the B.E., M.E. and Ph.D degress in computer science from Tokyo Institute of Technology in 1988, 1990 and 1993 respectively. From 2003, he has been an associate professor at Graduate School of Information Science of the Nagoya University. His research interest is spoken language processing. He is a member of the IPSJ and ASJ.

Nobuo Kawaguchi received the B.S. degree, the M.S. degree, and the Dr. of Engineering degree from Nagoya University, in 1990, 1992, and 1997 respectively. He was an Assistant, a Lecturer, and an Associate Professor at the Nagoya University. Since 2002, he has been an Associate Professor of Information Technology Center, Nagoya University. His research interests include mobile computing, mobile agent technology, and multi-modal user interface. He is a member of the IEEE, the ACM, the IPSJ, the ASJ, the JSSST, and the JSAI.



Shigeki Matsubara received the B.E. degree in electrical and computer engineering from the Nagoya Institute of Technology, in 1993, and the M.E. degree and the Dr. of Engineering degree in information engineering from Nagoya University, in 1995, and 1998, respectively. He was a Research Fellow of the JSPS from 1996 to 1998, and a Research Associate from 1998 to 2002 at the Faculty of Language and Culture, Nagoya University. Since 2002, he has been an Associate Professor of the Information Technology Center, Nagoya University. His research interests include natural language processing, spoken language processing, and digital library. He is a member of the ACM, the IPSJ, the JSAI, the NLP, and the JAIS.

Fumitada Itakura was born in Toyokawa near to Nagoya, in 1940. He earned undergraduate and graduate degrees at Nagoya University. In 1968, he joined NTT's Electrical Communication Laboratory in Musashino, Tokyo. He completed his Ph.D. in speech processing in 1972, writing his dissertation on "Speech Analysis and Synthesis System based on a Statistical Method." He worked on isolated word recognition in the Acoustics Research Department of Bell Labs under James Flanagan from 1973 to 1975. Between 1975 and 1981, he researched problems in speech analysis and synthesis based on the Line Spectrum Pair [LSP] method. In 1981, he was appointed as Chief of the Speech and Acoustics Research Section at NTT. He left this position in 1984 to take a professorship in communications theory and signal processing at Nagoya University. After 20 years of teaching and research at Nagoya University, he retired from Nagoya University and joined Meijo University in Nagoya. His major contributions include theoretical advances involving the application of stationary stochastic process, linear prediction, and maximum likelihood classification to speech recognition. He patented the PARCOR vocoder in 1969 the LSP in 1977. His awards include the IEEE ASSP Senior Award, 1975, an award from Japan's Ministry of Science and Technology, 1977, the 1986 Morris N. Liebmann Award (with B. S. Atal), the 1997 IEEE Signal Processing Society Award, and the IEEE third millennium medal. He is a fellow of the IEEE, a fellow of the Institute of Electronics and Communication Engineers of Japan, and a member of the Acoustical Society Japan.