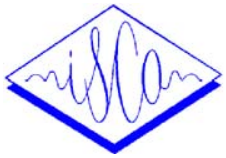
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Analysis of In-car speech recognition experiments using a large-scale multi-mode dialogue corpus

Hiroshi FUJIMURA, Katunobu ITOU, Kazuya TAKEDA, Fumitada ITAKURA

Graduate School of Information Science
Nagoya University, JAPAN
fujimura@itakura.nuee.nagoya-u.ac.jp

10.21437/Interspeech.2004-684

Abstract

The dependency of conversational utterances on the mode of dialogue is analyzed. A speech corpus of 800 speakers collected under three different modes, i.e., talking to a human operator, an WOZ system and an ASR system, is used for analysis. Some characteristics such as sentence complexity loudness of the voice and speaking-rate are found to be significantly different among the dialogue modes. Linear regression analysis results also clarify the relative importance of those characteristics on speech recognition accuracy.

1. Introduction

Although speech dialogue is expected to be the most convenient interface while driving, current technologies are not natural and reliable enough so that the driver does not suffer from any distractions under a complex task. One of the most important issues in in-car speech recognition is that the driving conditions and thus the driver's attitude to the system changes frequently. Current speech recognition technology, however, is sensitive to changes of utterances in both acoustic and linguistic aspects, such as vocabulary, complexity of the sentence, speaking-rate and voice loudness (SNR). Therefore, for the development of in-car dialogue systems, modeling and prediction of utterance variation is indispensable.

The goal of this study is to control the mode of dialogues depending on the driving conditions. This control would allow the driver to talk to the system as naturally as to a human navigator while stopped, but with a more restricted voice interface when making a turn, passing, etc. As the first step research on such flexible interface, in this paper, we will analyze a spontaneous utterances made under different dialogue modes, which is a part of CIAIR in-car behavioral signal corpus.[1]

The Center for Integrated Acoustic Information Research (CIAIR) at Nagoya University has been collecting a large scale corpus of the in-car speech and other behavioral signals. In this paper, using the CIAIR in-car speech corpus, we will analyze and clarify how the driver utterances change across dialogue modes, i.e., talking to a human navigator, to a Wizard-of-OZ system and to an ASR system.

The rest of this paper consists of the following sections. First,we give an overview of the CIAIR in-car speech corpus in Section 2. In Section 3, we detail corpus analysis from both viewpoints of language complexity and acoustical characteristics will be detailed. In Section 4, we show the results of speech recognition experiments using language and acoustic models trained for three dialogue modes . we also show the results of the linear regression analysis between word accuracy and characteristics of the utterances for three dialogue modes. Finally, we offer some conclusions in Section 5.

2. CIAIR IN-CAR SPEECH CORPUS

The CIAIR in-car corpus consists of more than 1.4 TB of multi-media signals recorded over 800 drivers[1]. A specially designed recording system built in a Data Collection Vehicle (DCV) has been used for synchronous recording of multi-channel audio data, multi-channel video data and the driving data. The audio signals are captured through 16 microphones that are distributed in the various positions in car including close-talking microphone and a four-linear-microphone-array. The microphone referred as visor-microphone in the later sections is located at the visor position and is nearest (less than 40cm) to the speaker's mouth except for the close-talking microphone.

During about one hour driving sessions around Nagoya university campus, the driver performed dialogues with three different systems, i.e., a human navigator, a Wizard of OZ system and an ASR system. We refer those dialogues as the HN session, the WOZ session and the ASR session, respectively. For the WOZ sessions, a speech response system equipped with a touch panel and speech synthesizer was installed in the DCV so that a human navigator can quickly make appropriate system responses. The ASR based dialogue system was also installed on the car, which is based on a bi-gram language model with 1500 word vocabulary, a dialogue controlling scenario written by FSN and the Julius decoder [2].

Table 1: Corpus statistics

(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

SPR[msec/mora]	144.3	143.5	149.1

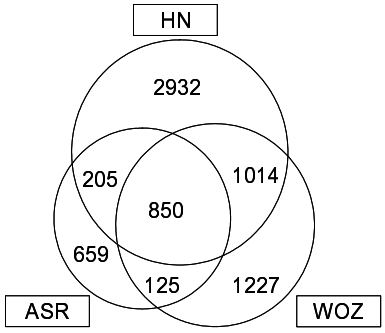


Figure 1: The overlap between the vocabularies in difference sessions.

3. Corpus Analysis

3.1. Corpus Size

The characteristics of the corpus that used in this study is summarized in Table 1 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 silence of longer than 200ms, therefore, in this corpus, most of "utterances" consist of a "speech unit". The boundaries between silences and utterances are given manually. The average length of the drivers' speech unit of HN session is longer than the other two sessions in both duration, 2.26sec., and number of morphemes, 8.72.

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

3.2. Complexity of the utterance

As seen from the table, the vocabulary size of the ASR session is half of human navigator session and that of WOZ session is in-between the two sessions. The overlap between the vocabularies in different sessions is shown in Figure 1. The vocabulary of ASR session can be regarded as a subset of 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 bi-gram entropy of the utterances are almost proportional to the vocabulary size, i.e., 18.1, 14.1 and 9.1 for HN, WOZ and ASR, respectively.

Table 2: The text data for language model training

	HN	WOZ	ASR
number of subjects	535	586	575
#Speech Unit	22240	19044	21289
#Morph	149213	117250	66612
#vocabulary size	5532	3694	2083
#bi-grams	35095	22277	9850
#trigrams	67972	44322	18403

3.3. 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 [3]. The SNR of the utterances in ASR session is better than HN and WOZ sessions by about 2dB. Since the driving conditions, therefore noise conditions, are designed to be same across sessions, the driver speaks in a louder voice to the ASR session.

3.4. Speaking-Rate

We define the speaking-rate(SPR) as the duration per mora. It is calculated using the result of forced alignment of the reference monophone label. The speaking-rate of the utterances in ASR session is the shorter than HN and WOZ sessions by about 5msec/mora. This indicate that the driver talks slowly to an ASR system.

4. Speech Recognition Experiments

4.1. Experimental Setup

In order to compare the difference between three sessions, language models were constructed for each session. The statistics of the text data for language model training are shown in Table 2. The text data was transcribed without any correction for speech error and disfluency. It was tagged by the morphological analyzer ChaSen, and the morphological labels were manually corrected. The morphological label includes part-of-

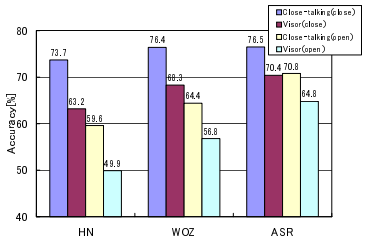


Figure 2: Recognition performance for the utterances in three sessions.

speech information. We constructed forward word bigram models and backward word trigram models without any cutoffs. Language models were trained for each speaker from all of the test set without dialogues by oneself, and these models are refered as "open." For comparison, a closed language model was also trained from all of the test set, and this model is refered as "closed." The training text data were manually tagged with morphological labels, which include part-of-speech information.

Acoustic models were trained from close-talking microphone captured data and distant microphone captured data. Training sets were made from all of the test set except short speech units which consists less than ten syllables, because most of them are the same yes/no answer. The total amount of the training data is 22.4 hours which includes 11,746 units, 11.4 hours from HN sessions, 8,550 units, 7.84 hours from WOZ session, and 4,878 units, 3.10 hours from ASR sessions. All speakers of the test set were included in the training data, however the amount of training data depended on the individual speakers. Acoustic models were prepaired for each microphone, the visor and the close-talking. The test set was divided into five parts for each microphone. Acoustic models were trained from four sets for the evaluation of the other set. In total, we trained five acoustic models for each microphone, and these models are refered as "open." For comparison, a closed acoustic models was also trained from all of the test set, and this model is refered as "closed." The bandwidth was limited from 250Hz to 8000Hz. Acoustic models are three-state continuous density triphone HMMs and 2000 shared states with 32 Gaussian mixture components, which was trained using HTK. These are Only the short pause model has skip transition from start state to final state. As a decoder, Julius was used.

4.2. Performance Comparison over sessions

The recognition performance for the utterances collected in three sessions are shown in Figure 2. In the open case, the performance is worst for HN session utterances and best for ASR session. Especially in the visor-microphone case, the performance differece between HN and ASR becomes more than 15%. From this result, it can be seen that the acoustic condition mostly governs the recognition performance in the visor-microphone case.

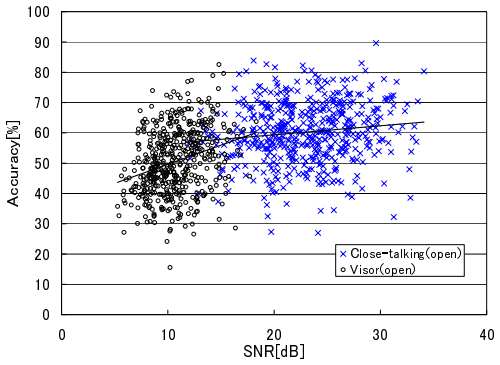


Figure 3: Word accuracy score and SNR value averaged over one speaker. In total 435 speakers results are plotted.

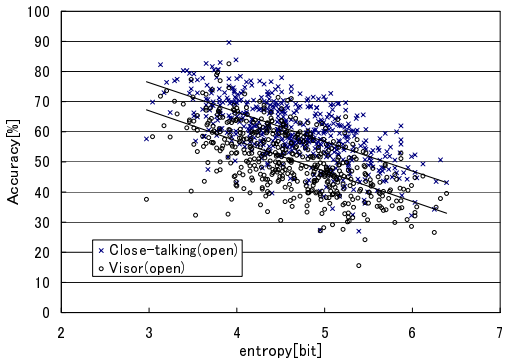


Figure 4: Word accuracy score and entropy value averaged over one speaker. In total 435 speakers results are plotted.

4.3. Performance Comparison over SNR, entropy and speaking-rate

In Figure 3, 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., approximately 0.31%/dB, in all sessions and cross speaker variability seems to be much larger. For the visor-microphone speech, however, the dependency gets higher, to be about 1.25%/dB.

In Figure 4 the average word accuracy and entropy are plotted for all drivers. There is no difference between close-talking microphone and visor microphone on the dependency of the recognition accuracy on the entropy. The higher the entropy is ,the lower the recognition accuracy becomes.

In Figure 5 the average word accuracy and speaking-rate are plotted for all drivers. There is no dependency which has been described by other reference of the recognition accuracy on the speaking-rate[4]. Then the standard deviation for speaking-rate per a speaker is also investigated. In Figure 6 the average word accuracy and standard deviation for the speaking-rate are plotted for all drivers. The dependency of the average word accuracy on standard deviation for the speaking-rate is little stronger

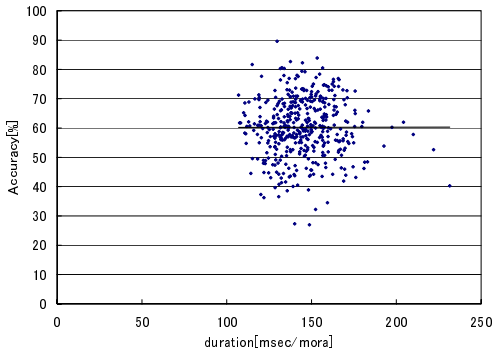


Figure 5: Word accuracy score and speaking-rate value averaged over one speaker. In total 435 speakers results are plotted.

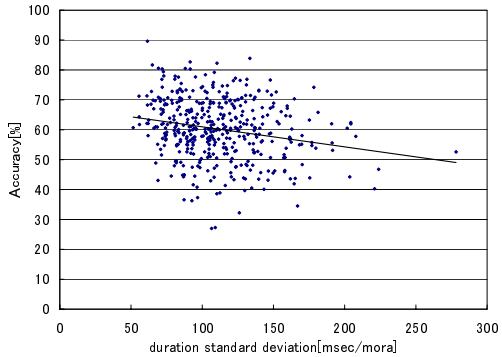


Figure 6: Word accuracy score and standard deviation for speaking-rate over one speaker. In total 435 speakers results are plotted.

than the dependency of the average word accuracy on the speaking-rate itself.

4.4. Regression Analysis

In Figure 7, the correlation coefficients between word accuracy and entropy, SNR and speaking-rate, are plotted. The highest linear correlation is found between accuracy and entropy, especially for the ASR session, correlation coefficient of about 0.77 was obtained. The correlation between word accuracy and SNR is high in visor-microphone speech cases where the correlation coefficient was about 0.28. Unlike the lecture speech case[4], the correlation between speech rate and word accuracy is significantly low.

5. Summary

In this paper, how conversational utterances vary depending on the mode of the dialogues is reported. A large amount of driver's speech collected through different di-

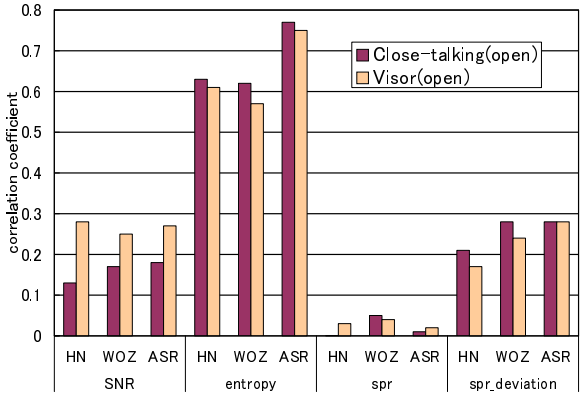


Figure 7: The correlation analysis results between recognition accuracy and various corpus characteristics.

alogue systems are characterized by vocabulary size, average entropy of the sentences, SNR and speech rate. The most clear and important result of this analysis is that the driver talk to an ASR system in a louder voice and with less complex sentences so that the system can recognize the utterances. The average loudness of the voice is 2dB higher than in human navigator and WOZ systems and the complexity of the sentences are lower than the utterances to a human operator. To an ASR system the driver talks more slowly. The regression analysis of the speech recognition accuracy also gives important suggestions on how to balance the SNR improvement and task restriction for improving the speech recognition performance.

In this paper, the discussions are mainly based on the averaged values. However, it was also clarified from the scatter analysis that the speaker variability is quite large. Therefore, controlling the mode of the dialogues depending on speakers must be the most important next research step.

Acknowledgment

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

6. References

[1] N.Kawaguchi,et.al.,"Construction of speech corpus in moving car environment", Proceedings of International Conference on Spoken Language Processing,pp.1281-1284,2000.

[2] A.Lee,et.al.,"Continuous Speech Recognition Consortium — an Open Repository for CSR Tools and Models —", Proc International Conference on Language Resources and Evaluation (LREC2002), pp.1438-1441, 2002.

[3] M.Kondo,K.Takeda,F.Itakura,"Predicting the Degradation of Speech Recognition Performance from Sub-band Dynamic Ranges", IPSJ Journal,pp.2242-2248,2002.

[4] T.Shinozaki,S.Furui,"Analysis on Individual Differences In Automatic Transcription Of Spontaneous Presentations", Proc. ICASSP2002, Orlando, U.S.A., vol.1, pp.729-732,2002.