ANALYSIS OF A LARGE IN-CAR SPEECH CORPUS AND ITS APPLICATION TO THE MULTIMODEL ASR

Hiroshi Fujimua†*, Chiyomi Miyajima*†*, Katsunobu Itou*†*, Kazuya Takeda*†*and Fumitada Itakura*‡

†Graduate School of Information Science, Nagoya University, Nagoya 464-8603 Japan
‡Faculty of Science and Technology, Meijo University, Nagoya 468-8502, Japan

ABSTRACT

In-car ASR performance improvement utilizing a large in-car speech corpus, consisting of the utterances of more than five hundreds drivers under real driving conditions is discussed. A subset design method for efficient cross validations in large-scale speech recognition experiments is proposed. The factor analysis of the results of the recognition experiments show the relationship between word accuracy and utterance characteristics, i.e., SNR, entropy and speaking rates. Based on the factor analysis results, a multimodel approach which uses the utterance duration and subband SNRs as the model selection measures for acoustic and language models, respectively, is proposed. By the proposed multimodel approach, a relative error reduction of 16% is obtained.

1. INTRODUCTION

Providing a human-machine interface in a car is one of the most important applications of speech signal processing, where conventional input/output methods are unsafe and inconvenient. 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 and disfluency while driving, must be overcome[1, 2, 3]. 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. 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 [4, 5, 6].

The authors constructed a large corpus of in-car speech communication by recording a dialogues of more than 500 drivers under real driving conditions [7]. Using the corpus, in this work, large-scale speech recognition experiments are performed, where more than 18,000 utterances are recognized using open models. In order to do this, a subset design method for effective cross validation for the large-scale speech recognition experiment is proposed. This paper also presents a factor analysis on the variability of recognition accuracy based on the linear regression model. The results of the analysis show that entropy and SNR of the utterance are the major factors, in this order.

Finally, we propose multimodel approaches, namely, an "utterance-length-dependent language model" and a "subband-SNR-dependent acoustic model," for improving the modeling by fully utilizing the variabilities stored in the corpus.

The structure of this paper is as follows. In the next section, we briefly describe the data collection procedure. In Section 3, we show the results of speech recognition experiments using the corpus. The subset design method for effective cross validation testing and the results of factor analysis of the recognition performance are also described in this section. In Section 4, a multimodel approach for improving the recognition accuracy is proposed.

2. DATA COLLECTION PROCEDURE

A specially designed Data Collection Vehicle (DCV) which has multichannel (16 ch of spatially distributed microphones) as well as multimedia (audio, video and car-related signals) data recording capability is used for the collection [7].

For the data collection, during approximately, a one hour drive around the campus of Nagoya University, each driver makes various types of utterances including a dialogue with the human operator, the Wizard of Oz system and the ASR system. The task domain of the ASR system is a slot-filling dialogue for restaurant retrieval, which is controlled by a simple state-transition dialogue model. The vocabulary size of the task is 1,500 and bigram model is used for the language model of the ASR system. A continuous Speech Recognition Consortium (CSRC) standard triphone HMM and a Julius decoder are used for the acoustic model and the LVCSR engine [8], respectively. Although the close-talking microphone is used for the input to the ASR system, the utterances recorded at the microphone placed at the visor position (approximately 40cm distant from the driver's mouth) are used for this research.

Among the three modes of the dialogue sessions, the dialogue with an ASR system is used for the recognition experiment in this paper. Therefore, language models are trained using the utterances recorded in the ASR session only, whereas the acoustic models are trained using utterances from all of three sessions as well as other utterances.

The total size of the utterances used for model training is summarized in Table 1. Note that short (less than 10 mora) utterances, most of which contain a 'yes/no' answer only, are not used for acoustic modeling.

3. DATA ANALYSIS

3.1. Subset Design

In order to analyze the collected data in terms of speech recognition performance, both acoustic and linguistic models are trained using the collected data. For efficient cross validation testing, we

This was partially supported by a Grant-in-Aid for COE Research (No. 11CE2005) and the MECSST Leading project for the e-Society

Table 1. Training sentences for language models used for recognition experiments.

	model	
	language	acoustic
number of subjects	505	
male	321	
female	184	
# of utterances	18243	47463
# of morphemes	58240	-
vocabulary size	1947	-
# of bigrams	9009	-
# of trigrams	16600	-

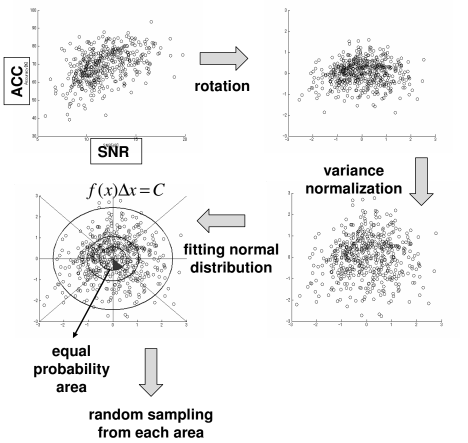

Fig. 1. Subset design procedure for efficient cross validation.

have designed a subset of the corpus. The grouping is designed so that the average recognition accuracies and SNR values in each group have equal distribution. For this purpose, as shown in Figure 1, a two dimensional Gaussian is fit to the distribution of speakers in the SNR-accuracy space so that to define equal probability areas. By uniformly sampling the speakers from all equal probability areas, we can make arbitrary numbers of speaker subsets. By comparing with the least variance case among 50 trials of the random sampling, we have confirmed that this method can reduce the variance of accuracies among speaker groups from 2.69% to 1.73%, when applied to design 20 groups.

3.2. Recognition Experiment

After dividing the speakers into five groups, five sets of HMMs trained using four of the five groups can be used for an open evaluation of the acoustic modeling. In contrast, an open language model is trained for each speaker, for which the utterances made by all speakers except the evaluated speaker were used.

Throughout the recognition experiments, the feature parameters for the HMM acoustic model were fixed to 12MFCC, 12ΔMFCC and Δ log power. Although the original signal is sampled at 16 kHz, the bandwidth was limited to the range from 250Hz to 8000Hz. In order to focus only on the modeling, we did not perform any speech enhancement other than low cut filtering. The basic structure of the HMM is also fixed as 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 the start state to the final state. Julius [8] was used as the decoder.

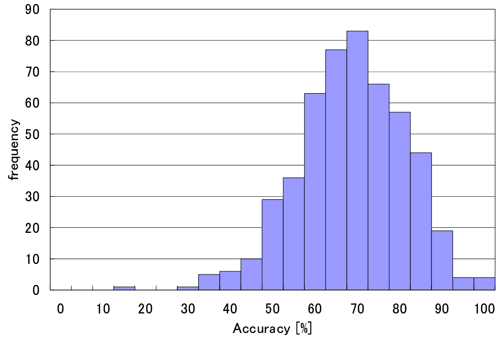

Fig. 2. Histogram of word accuracy scores.

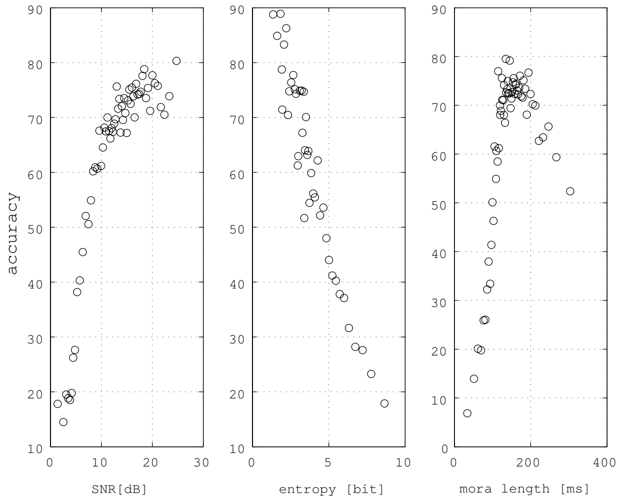

Fig. 3. Correlation between averaged word accuracy and SNR, entropy and mora length of the utterances. The averaged word accuracy is the average of the accuracy scores of the utterances that have the same values for each factor.

In this evaluation framework, 18,243 utterances were recognized using open acoustic and language models. The histogram of the word accuracy averaged for each speaker is plotted in Figure 2. The wide variabilities of the utterances collected under real in-car conditions can be seen from the figure, where the mean and the standard deviation of the accuracy are 66.1% and 12.5%, respectively.

Table 2. Correlation between average word accuracy and SNR, entropy and speaking rate of the utterance.

	SNR	entropy	mora len.
correlation	0.86	-0.93	0.57

3.3. Factor analysis

Among various factors reported to affect the ASR performance [9, 10, 11], we have tested the SNR, the complexity of uttered sentences and the speaking rate. Figure 3 shows the averaged word accuracy, i.e., the average of the accuracy scores of the utterances that have the same value for the factor, as a function of SNR, complexity and speaking rate.

The SNR is estimated without VAD by fitting a two-mixture Gaussian model to the distribution of frame log powers [12]. The complexity of each sentence is measured by the cross-entropy between the utterance and the trigram open language model. The Speaking rate is measured in terms of the average mora length after forced alignment.

Unlike read text or monologues, the entropies of the sentences used in the dialogues have a wide distribution, i.e., 2 to 1024, in terms of perplexity. Therefore, the highest linear correlation is found between entropy and accuracy. Due to the high variability of the acoustic conditions of in-car speech, the distribution of SNR is also wide and thus its degree of correlation with accuracy is also high. The speaking rate is also correlated with the accuracy, however, compared with SNR and entropy, it has less effect. Correlation between the averaged word accuracy and SNR, entropy and mora length are summarized in Table 2.

4. MODEL REFINEMENT BY MULTIPLE MODELING

In order to match the training and recognition conditions well, multimodel approaches that train multiple models and use them selectively have been proposed [13, 14]. In this section, we cluster the collected utterances and generate multiple models, each of which takes the particular context of the training utterances into account. Because the variability of utterances in the real world can only be covered by a large scale corpus, clustering utterances in this corpus can generate a more effective model set than that trained for artificially generated variabilities, e.g., the addition of noise to clean speech with a predetermined SNR [15].

For the clustering and model selection measure, we propose the duration of the utterance for the language model and the subband SNRs for the acoustic model. Because both measures can be calculated directly from noisy speech, consistent decisions can be made at the training and recognition stages, and therefore, performance degradation due to uncertainty of the optimal class selection can be avoided.

4.1. Utterance-length-dependent language model

In a simple spoken dialogue system for an information retrieval task, most utterances consist of the queries to the system and the responses to the system confirmation. In general, an utterance for the query sentence is longer than the response to the confirmation. Therefore, the duration of an utterance is expected to be related to the pattern of the sentence. As shown in Table 3, by dividing the training utterances into three groups of durations, i.e., -0.5 sec., 0.5-1 sec. and 1- sec, we can reduce the test set entropy by 0.12 bits on average. Since there is no uncertainty in selecting the language model after determining the utterance duration, the same amount of entropy reduction can be expected for the unknown input utterances.

A drawback of this method is the latency of measuring the speech duration, i.e., we have to wait for the end of the utterance, in order to select the language model. However, the latency may be eliminated by a delayed decision strategy where multiple decoding processes run in parallel; once decoding reaches the end of an utterance, the final result will be selected according to the sentence duration.

Table 3. Entropy of the length-dependent language models. Entropy (single) shows the average entropy of the single language model calculated for each utterance group. Entropy (multi) shows the average entropy of the language model trained for the utterance group.

	utterance-length (sec.)			
	-0.5	0.5 - 1.5	1.5 -	average
# of utterances	10,834	17,378	9766	18,243
entropy (single)	2.77	3.10	3.85	3.14
entropy (multi)	2.60	2.98	3.77	3.02

Table 4. Result of the 8-class clustering based on subband SNRs of five channels. Centroids for eight clusters are listed. The average word accuracy of the utterances in each group is also listed for both single model and multimodel cases.

		SNR at subband in kHz[dB]					acc.[%]	
id	utt.	0.2-0.6	0.6-1	1-2	2-3	3-4	single	multi
1	6008	2.8	3.1	2.9	2.9	4.5	20.0	16.9
2	6690	4.8	6.7	6.5	7.2	10.1	33.3	36.9
3	9476	9.7	12.2	10.0	7.5	9.2	55.6	59.7
4	8186	9.4	12.2	11.0	11.5	15.0	72.7	73.8
5	10686	14.4	16.3	14.0	10.6	12.6	67.5	71.4
6	6089	13.4	16.3	15.1	15.7	19.7	76.9	79.5
7	6581	18.4	20.4	18.4	14.7	16.7	73.3	76.0
8	2963	20.3	23.9	22.4	21.3	24.1	84.0	86.3

4.2. Subband-SNR-dependent acoustic model

In order to characterize the acoustic conditions under which an utterance was made, we use subband SNRs. A subband SNR is the log power ratio between speech and noise at the given frequency region, therefore, both the signal level and the spectral shape of the background noise can be represented with subband SNRs in multiple frequency regions. Subband SNRs can also be robustly estimated through the blind SNR estimation method [12].

We have calculated the subband SNRs for the five subbands of 200-600Hz, 600-1kHz, 1-2kHz, 2-3kHz and 3-4 kHz. All utterances are clustered in the five-dimensional space into eight groups using the K-means algorithm, where the number eight is decided as the best condition in the preliminary experiment. In Table 4, the

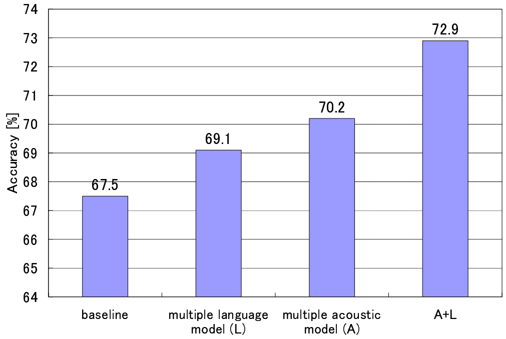

Fig. 4. Word accuracy improvement using multiple models, i.e., sentence-length-dependent language model (L), subband-SNR-dependent acoustic model (A), and their combination (A+L), from left to right.

centroids obtained for the eight clusters are listed.

Multimodels are then generated by transforming a original standard HMM through MLLR adaptation using all of the utterances in each cluster. In our experiment, no significant improvement was obtained simply by training an independent model using the utterances in the group, because the training data for each model became fewer.

4.3. Experimental Evaluation

The effectiveness of the multimodeling based on the utterance length and subband SNRs has been evaluated through recognition experiments using the evaluation framework described in 3.2.

The recognition results are shown in Figure 4. By selectively using language models trained for different lengths of utterances, we obtained a 1.6% improvement in word accuracy. The introduction of multiple acoustic models, on the other hand, achieved a 2.7% improvement in word accuracy. In addition, by combining multiple acoustic and linguistic models, further improvement is obtained. Finally the 72.9% word accuracy , i.e. a 16% error reduction from the baseline performance, is achieved. On the basis of this result, the effectiveness of multiple model approach based on utterance duration and subband SNRs is confirmed.

5. SUMMARY

In this paper, we described a large in-car speech corpus obtained under real driving conditions, and its application to improve in-car ASR performance. After proposing a subset design method for efficient cross validation, five hundred drivers' utterances are analyzed through large-scale speech recognition experiments. From the results of the experiments, the SNR and entropy of the utterances are found to be the two major factors governing the recognition accuracy for in-car dialogue utterances.

Based on the above investigation, a multimodel approach for both acoustic and language models is proposed. In the proposed method, the duration of the utterance and subband SNRs are used as the multiplication and selection criteria for language and acoustic models, respectively. Since both criteria can be directly calculated from a noisy speech waveform, the loss of information due to the uncertainty in model selection is small, and therefore, performance improvement from precise modeling is fully utilized. Finally, the effectiveness of using multiple models is confirmed through recognition experiments, where a 16% improvement in speech recognition accuracy is confirmed.

6. REFERENCES

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

[2] P. Gelin and J.C. Junqua, "Techniques for robust speech recognition in the car environment," Proc. EUROSPEECH '99, pp.2483–2486, 1999.

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

[4] P.Geutner, L.Arevalo and J.Breuninger, "VODIS - Voice-operated driver information systems: a usability study on advanced speech technologies for car environments," Proc. ICSLP2000, pp.IV378–IV381, 2000.

[5] J.H.L.Hansen et al., "CU-Move: Robust Speech Processing for In-Vehicle Speech Systems," Proc. ICSLP2000, pp.I527–530, 2000.

[6] A.Moreno et al., "SpeechDat-Car: A Large Speech Database for Automotive Environments," Proc. of LREC 2000, pp.373–378, 2000.

[7] N.Kawaguchi et al., and Y.Inagaki, "Construction of Speech Corpus in Moving Car Environment," Proc. ICSLP2000, pp.362–365, 2000.

[8] A.Lee, et al., "Continuous Speech Recognition Consortium — an Open Repository for CSR Tools and Models —," Proc. of LREC2002, pp.1438-1441, 2002.

[9] B.H.Juang, "Speech Recognition in Adverse Environments," Computer Speech and Language 5, pp. 275–294, 1991

[10] S. Nakagawa and I. Murase, "Relationship among phoneme/Word recognition rate, perplexity and sentence recognition and comparison of language models," Proc. ICASSP'92, pp. 589 - 592, 1992.

[11] J. Zheng et al., "Word-level rate of speech modeling using rate-specific phones and pronunciations," Proc. ICASSP'00, pp. 1775 - 1778, 2000.

[12] T.H Dat, K. Takeda and F. Itakura, "Robust SNR estimation of noisy speech based on Gaussian mixtures modeling on log-power domain," Proc. Robust2004, CDROM Proceedings, 2004.

[13] J. Ming et al., "Improving speech recognition performance by using multi-model approaches," Proc. ICASSP'99, pp.161–164, 1999.

[14] M. Westphal and A. Waibel, "Model-combination-based acoustic mapping," Proc. ICASSP'01, pp. 221–224, 2001.

[15] C. Huang, H. Wang, and C. Lee, "An SNR-incremental stochastic matching algorithm for noisy speech recognition," IEEE Trans. Speech Audio Processing, vol. 9, pp. 866 - 873, November 2001.