Study Of Effect Of Speaker Variability And Driving Conditions On the Performance of an ASR Engine Inside a Vehicle

S. Kadambe

HRL Laboratories, LLC, 3011 Malibu Canyon Road, Malibu CA 90265[1]

E-mail: skadambe@hrl.com

Abstract

Spoken dialogue based information retrieval systems are being used inside vehicles. The user satisfaction of using such a system depends on how an ASR engine performs. However, the performance of an ASR is affected by speaker variability, driving conditions, etc.. Here, we report the study that we performed to analyze these effects of speaker variability, different driving conditions and the effect of driving task on the ASR performance. This study consists of experimental design, data collection and systematically testing an ASR engine using this data. From the obtained results, it can be observed that (I) the ASR performance exhibits (a) significant speaker variability since the stress of driving task varies from speaker to speaker, (b) significant performance degradation across driving conditions since the noise type and level varies and (c) significant effect of driving task on recognition performance, and (II) the effect of live noise on recognition performance is not same as adding car noise to the pre-recorded speech data. The former observations are important since by just training an ASR engine on lots of speech data will not help and it is essential to include stress factors and cognition load in ASR engines to improve its performance.

1. Introduction

Spoken dialogue information retrieval applications are becoming popular in automobiles. Due to the typical presence of background noise and multi-tasking (i.e., speaking while driving) the speech recognition accuracy is affected significantly. This in turn affects the overall performance of the information retrieval system and the user's satisfaction. Hence, there is a need for a systematic study of the effect of speaker variability, different driving conditions and the driving task on the performance of an ASR engine which would help in improving the user's satisfaction. In [1], the effect of speech recognition performance on driving task of a driver is studied. However, in this paper, the effect of driving task on recognition performance is studied. To the knowledge of this author such a study has not been reported before. The study in [1] indicates that the recognition accuracy significantly affected the driving task across drivers of different ages. The study reported in this paper helps in understanding how the speech of a driver is affected by the stress of the driving task which in turn affects the ASR performance. By understanding this effect and noting from [1] that by improving the ASR accuracy driving task can be improved, it is hoped that the ASR performance can be improved by including the stress and driver's work load while training an ASR system. This results in improved driving task and thus the driver's safety. The rest of the paper is organized as follows. In the next section, the experimental design in terms of estimating the required sample size that is needed to perform the proposed study is described. In section 3 the details of data collection is provided. Section 4 provides the results of testing an ASR engine using the collected data. Finally, in section 5, we conclude and discuss the future work.

2. Experimental design

The experiment of studying the effect of speaker variability driving conditions and driving task on the ASR performance is designed first by estimating the minimum sample size that is needed to collect enough data that provides statistically significant results. For this sample size estimation, we use word error rate (WER) of an ASR engine as a measure to compare the different datasets.

For the sample size computation the following equation is used:

$$\Phi^2 = \frac{s' \left(\sum_i (\mu_i - \mu)^2 \right)}{a \sigma^2}$$ where Φ is a parameter that relates to statistical power i.e., the probability of correctly rejecting the null hypothesis, a is the number of experiments, μ_i the expected average mean WER for each experiment (or dataset), μ is the overall average WER, σ is the pooled within groups error standard deviation and s' is the estimated sample size.

We assume that the mean WER for each experiment varies from the overall mean by 5 for all datasets by keeping in mind a difference of 3-4 % is statistically significant and the standard deviation is assumed to be 10. We use $\alpha = 0.05$. By substituting these values in the above equation we get:

$$\Phi^2 = \frac{s'}{4} \;\&\; \Phi = \frac{\sqrt{s'}}{2} .$$

Using the curves plotted in [2] for power versus Φ, for $\alpha = 0.05$, degrees of freedom for numerator = $(a - 1) = 2$, ($df_{denominator}$ = infinity) for a power of 0.8 which corresponds to $\Phi = 3$ we get an estimated sample size $s' = 36$. Note that a is selected as three as we conduct three different experiments. So the minimum sample size of 40 for each experiment (five sentences/per speaker and total of 8 speakers) seems to have good enough statistical power.

3 Data collection

Based on the above described sample size estimation, we decided to collect data from twelve speakers with each speaker uttering twenty sentences. Note that this sample size is more than the estimated sample size and hence, our ASR performance results should be statistically significant. We chose all twelve speakers as male to avoid the effect of inter i.e., male versus female speaker variability. Since our overall goal is to improve the user's satisfaction of using the information retrieval systems inside a vehicle, we selected twenty utterances related to queries on weather information at different cities. These twenty sentences are listed in Table 1. The driving conditions under which data was collected are mentioned in Table 2. The following datasets were collected using 12 male subjects.

1. Data set1: the selected sentences spoken by each subject in a vehicle (SUV) under each driving condition mentioned in Table 2.
2. Data set2: the recordings of selected sentences spoken by the subjects inside the vehicle when it was stationary. This is equivalent to data collected in a laboratory.
3. Data set3: the play back data of the recorded sentences (data set2) inside a vehicle under different driving conditions mentioned in Table 2.

These three datasets were collected using two microphone arrays and one clip microphone which was clipped to a shirt in front below the chin of a subject. The two microphone arrays are: 1. CSLR [3] and 2. Andrea. The microphone array built by CSLR has five microphones with five output channels where as Andrea microphone array has only one output channel. Seven channels – 5 channels of CSLR array, 1 channel of Andrea and 1 channel of clip microphone of data was recorded using an 8 channel Fostex's DAT recorder.

Note that the conditions 1 & 2 in table 2 are referred as "quiet" and "noisy" in recognition accuracy tables and in figures in the next section whereas "stat" correspond to the data that was collected when the vehicle was stationary (data set2).

4 Simulation details and ASR results

Recognition experiments were conducted using the three datasets mentioned above using only clip microphone channel data since the speech data quality from microphone arrays were poor and recognition accuracies were bad. The ASR engine that is used in this study is a continuous speech recognizer. The vocabulary size of this engine is 2000 words. In Table 3, the ASR's performance in terms of percentage of word recognition accuracy is provided for each subject under "stat", "quiet" and "noisy" conditions (data set1 and data set2). In addition, the mean and the standard deviation of word recognition accuracy across conditions for each subject are provided. The recognition accuracy across three conditions that was computed by pooling the speech data of all subjects under each condition is tabulated in Table 4. In Figure 1, the histogram plots for the recognition accuracy for these three conditions are provided. In addition, the bar plot of recognition accuracy across conditions is provided. From this bar plot and Table 4, it can be seen that the recognition accuracy degrades by 5 % for "quiet" as compared to "stat" and by 11 % for "noisy" as compared to "stat".

In Figure 2 row 1, the histograms of mean and standard deviation across conditions for each subject are plotted. From Table 3, Figure 1 and Figure 2 row 1, it can be seen that the speaker variability across conditions is very significant.

In Table 5, the ASR's performance in terms of percentage of word recognition accuracy is provided for each subject under "quiet" and "noisy" conditions for play back data (data set3). In addition, the mean and the standard deviation of word recognition accuracy across these two conditions for each subject are provided. The recognition accuracy across two conditions that was computed by pooling the play back speech data of all subjects under each condition is tabulated in Table 6.

In Figure 3, the histogram plots of recognition accuracy for the "quiet" and "noisy" conditions are provided. Further, the bar plot of the recognition accuracy across these two conditions is also plotted in Figure 3. From these, as in the live speech data case, it can be seen that the speaker variability across conditions is very significant. *Further, comparing Tables 3 and 5, it can be seen that the recognition accuracy is better for play back data for both "quiet" and "noisy" conditions indicating the effect of driving task on recognition performance. This difference in performance also indicates that live noise effects are not same as adding car noise to the pre-recorded speech data.* In Figure 2 row 2, the histograms of mean and standard deviation across conditions for each subject for play back data are plotted.

In summary, the data analysis indicates:

- Significant speaker variability
- Significant performance degradation across conditions
- Effect of driving task on recognition performance
- Effect of live noise on recognition performance is not same as adding car noise to the pre-recorded speech data.

5 Conclusions

In this paper, as part of user satisfaction in using information retrieval systems in vehicles a systematic study on various factors that affect the performance of an ASR engine which is an integral part of an information retrieval system is reported. The sample size that provides required statistical power is estimated. While collecting data this estimated sample size is considered. Three different datasets were collected and analyzed using a continuous speech recognition engine. From the results it can be seen that, that (I) the ASR performance exhibits (a) significant speaker variability since the stress of driving task varies from speaker to speaker, (b) significant performance degradation across different conditions since the noise type and level varies and (c) significant effect of driving task on recognition performance, and (II) the effect of live noise on recognition performance is not same as adding car noise to the pre-recorded speech data. The latter observation is important since generally

noise effect is studied by adding noise to clean speech. Former observations indicate that to improve the ASR performance and thus driver's safety, it is important to consider the effect of driving task on speech in terms of stress and Lombard effects which vary from speaker to speaker. It is also equally important to enhance speech quality by applying source separation techniques. Our future work will consider adding speech features that relate to stress and Lombard effect in our ASR engine and will apply HRL's blind source separation technique [3] to enhance speech data to further improve the recognition accuracy that is reported in [4].

6 References:

1. A. W. Gellatly, "The use of speech recognition technology in automotive applications," PhD dissertation, Virginia Polytechnique Institute and State university, 1997.
2. G. Keppel, Design and Analysis: A researcher's handbook, Prentice Hall Inc., Englewood Cliffs, NJ, 1973.
3. M. Peterson and S. Kadambe, "A probabilistic approach for Blind Source Separation of underdertmined convolutive mixtures," in Proc. Of ICASSP, pp. VI-581-583, Hong Hong, April 6-10, 2003.
4. "Robust ASR inside a vehicle using using blind probabilistic based under-determined convolutive mixture separation technique", In DSP in Mobile and Vehicular Systems, edited by H. Abut, K. Takeda and J. H.L. Hansen, Published by Kluwer Academic Publishers, April 2004. (**Invited by Prof. H. Abut**).

1. i want to know if it will rain in seattle on sunday
2. what will the weather be like in phoenix
3. can you repeat that
4. how is the weather in hartford connecticut
5. i+m looking for the extended forecast for houston
6. what is the forecast this weekend for buffalo
7. what about saturday
8. please tell me the weather today in portland oregon
9. thank you
10. what cities do you know in california
11. how+s the weather in miami
12. please tell me what the weather will be like tomorrow in new york city
13. can you give me the forecast for boston
14. is it raining in pittsburgh
15. what was the high today in pasadena
16. will it rain in denver today
17. what is the current temperature in fargo north dakota

Table 1: Selected weather related Speech utterances

stat	Quiet	noisy	μ across conditions	σ across conditions
92.7	87.5	81.3	87.17	5.7073

Table 4: The recognition performance in terms of word recognition accuracy in percentage for live speech data collected in a vehicle across conditions by pooling all subjects' data for each condition

Test Condition	Road Surface 4 (Freeway, 50 –60 MPH)	Windows Up	Windows Down	Fan Off	Fan On	Tape (radio) Off	Tape (radio) On	Turn Signal Off	Turn Signal On	Windshield Wipers Off	Windshield Wipers On	Comments
1.	•	•		•		•		•		•		Baseline on freeway("quiet")
2.	•	•			◎		◎		◎		◎	Combined effect("noisy")

Table 2: Conditions under which sample speech data were collected

subject	Stat	quiet	Noisy	μ(across conditions /subject)	σ(across conditions/subject)
1	92	86.9	84.8	87.9	3.703
2	95.7	95.6	95.5	95.6	0.1
3	83.8	82.7	78.4	81.63	2.854
4	94.9	93.4	89.8	92.7	2.621
5	91.1	89.7	84.9	88.57	3.252
6	95.6	78.8	73.9	82.77	11.38
7	98.4	94.5	90.6	94.5	3.9
8	98.4	81.2	70.9	83.5	13.89
9	98.4	90.6	87.4	92.13	5.658
10	89.8	87.6	87.4	88.27	1.332
11	94.1	85.6	75.8	85.17	9.158
12	95.8	95	90	93.6	3.143
μ(across subjects)	94	88.47	83.82		
σ (across subjects)	4.2819	5.6513	7.9392		

Table 3: The recognition performance in terms of word recognition accuracy in percentage across 12 subjects and across three driving conditions for live speech data collected in a vehicle

Quiet	Noisy	μ across conditions	σ-across conditions
93.0	89.0	91.0	2.8284

Table 6: The recognition performance in terms of word recognition accuracy in percentage for play back speech

data collected in a vehicle across conditions by pooling all subjects' data for each condition

Subject	Quiet	noisy	μ(across conditions/ subject)	σ(across conditions/subject)
1	94.1	92.3	93.2	1.2728
2	96.6	93.3	94.95	2.3335
3	83.9	81.5	82.7	1.6971
4	94.1	93.4	93.75	0.4950
5	91.1	89.7	90.4	0.9899
6	93.2	92.9	93.05	0.2121
7	97.6	96.9	97.25	0.4950
8	98.4	81.9	90.15	11.6673
9	96.9	93.7	95.3	2.2627
10	86.6	83.5	85.05	2.1920
11	95.8	93.2	94.5	1.8385
12	98.3	96.6	97.45	1.2021
μ(across subjects)	93.88	90.74		
σ(across subjects)	4.6211	5.4343		

Table 5: The recognition performance in terms of word recognition accuracy in percentage across 12 subjects and across two driving conditions for play back speech data collected in a vehicle

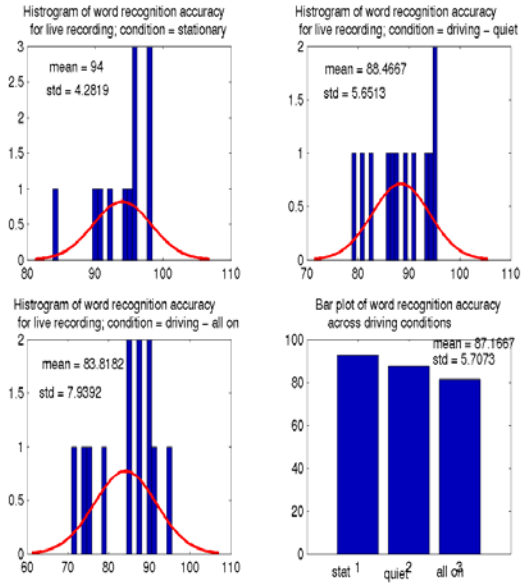

Figure 1: The histogram and bar plots of word recognition accuracy for live speech data collected in a vehicle for different driving conditions

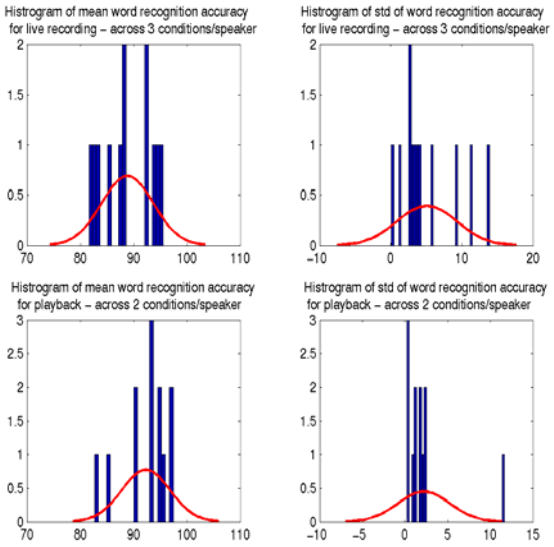

Figure 2: The histogram plots of word recognition accuracy across conditions by pooling all subjects' data for live speech (row 1) and for play back speech (row 2)

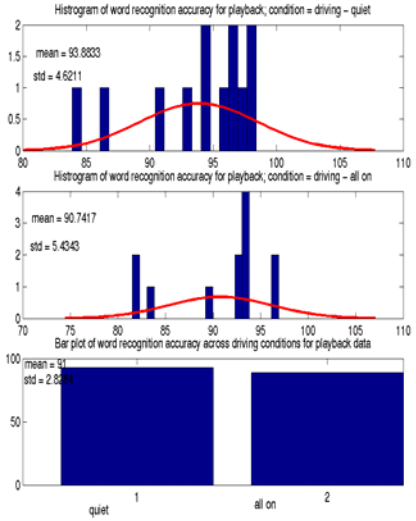

Figure 3: The histogram and bar plots of word recognition accuracy for play back speech data collected in a vehicle for different driving conditions