The Effect of Time Stress on Automatic Speech Recognition

Accuracy When Using Second Language

Fang Chen, Assistant Professor,
Swedish Center for Human Factors in Aviation,

Jonas Sääv, Application Engineer
Virtual Technology in Linköping AB,

Linköping University, SE- 581 83 Linköping, Sweden

Abstract

The purpose of the present study is to compare the ASR performance when Swedish people speaking Swedish and English under time-stress and due-task performance. Fifteen university students (20 to 40 years of age, native Swedish language speaking) participated in the experiment. Three factors were studied: time-stress, which was manipulated by PWSP program. Two models of presenting the commands, one is by displaying the text on the screen and another is by headphone voice. Swedish and English languages were tested on Philips FreeSpeech 2000 speech recognition system. There is no individual voice file training and pre-designed grammar file for the speech recognition system. The results show that there are no interactions between any of the factors. The individual differences are large. There is a significant decrease of recognition accuracy ($p<0.05$) for both languages during stress. The recognition accuracy on Swedish language is significant higher ($p<0.01$) than English Language due to the Swedish accents.

Introduction

The way that the stress affect the use of automatic speech recognition (ASR) is not a well defined concept in a number of ways, such as by changing the speech frequency, loudness, or by affecting cognitive functions for spoken commands [1]. Some studies have found consistent patterns that can be identified in the voice under stress [2-4]. Still, it is generally considered that both physical and task-induced stress represent a major obstacle to the success of speech recognition applications [5-7].

Another well-known factor that degrades the performance of speech recognition systems is speaker's accent [8]. Training the system on that accent [9-11], selecting an appropriate language model [12], or adapting the accent/speaker [13] did not really solve the problem. Each of these approaches has trade-offs in terms of training complexity [7].

Degradation in recognition performance due to accent is a concern not only to commercial applications such as applied on the telephone network and on personal computers but is also a concern in military applications with the multinational forces and in air traffic control [7]. This area is getting increased attention because of the significant benefits that will be derived in commercial applications. The unique military aspects will be the effects on speech recognition performance with combinations such as accented speech in a stressful, high-noise environment [7].

With the forming of the European Union, and the JAA (Joint Aviation Association), there is a definite need to identify issues relating to speech and the utility of English as the communication language. The purpose of present study is to compare the ASR performance when Swedish people speaking Swedish and English under time-stress and due-task performance.

Methods

Subjects

There were 15 males university students (native Swedish language speaking) between 20 to 40 years of age as volunteers to the experiment. They were all native Swedish speaker and have good English level.

Equipment

Time stress during the experiment was manipulated by using Pilot Workload Simulation Program (PWSP) (the interface is shown as Picture 1). PWSP is a modified NASA-TLX program ported from DOS to Windows 9x, or higher, with Borland C++ Builder 5.0. The reason not to use NASA TLX in the first place was the wish to use Philips FreeSpeech 2000 as a speech recognition engine and there were also needs to have more control over the workload (wanted higher) and the output data format.

Philips FreeSpeech 2000 is a phoneme based natural speech recognition software. PWSP uses the SDK (a C-language application programmer interface) provided to implement the speech command functionality. In the present experiment, we used 55 vocabularies to build up 60 commands (both in Swedish and English of each language) without any pre-designed grammar. In another word, any of the word has equal probability to combine with another word. The subjects were asked to press the fire button on the joystick before they started to speak, and they should release the button as soon as they finished speaking.

Picture 1. Interface of PWSP

The independent variables were:

a. Two level stress: One is not stress at all; the subjects do not need to do any other things than speaking to the computer. The other is high stressed. PWSP was used to set up the mental and manual workload. The subjects were asked to press different keys on the keyboard by their left hand to keep different displays on PWSP interface in the right level. At the same time, the right hand was asked to control the joystick to keep the small circus inside the middle of the tracking display. The level of the workload was design in such way that during high stress condition, the subjects' left hand will be fully occupied on the key board and the right hand have to hold on the joystick and moving it to adjust the tracking position.

b. Two ways of presenting the spoken sentences: One was to present it by text on the same computer screen, another way was to present it by male voice through the headphone. The interval between presenting of two sentences was randomly arranged within 10 seconds. Thirty commands, which was randomly selected from the total of 60 commands in each language was presented to the subject.

c. Two languages: All the sentences were presented in both Swedish and English.

Experimental process:

When the subjects came to the laboratory, the experimenter explained the process of the experiment to them with a detailed instruction of the experiment on paper. They had the right to withdraw from the experiment whenever they wanted. They were informed that performing of PWSP was equally important as speech. Each subject tried the demo-PWSP (5 minutes of each) three times as training process. They were paid by two lunch tickets for being the subjects.

All the conditions have been randomly arranged during the experiment. The subjects only need to come to the laboratory for once. There was a short pause (about 5 minutes) between stress and non-stress condition. The order of the stress and non-stress was also randomized between subjects.

Measurement:

Four different measurements were taken during the experiment:

a. General questionnaires: It includes the general background of the subjects, specially regarding to their English level and experience of using different ASR systems.

b. Recognition accuracy: Recognized and misrecognized sentences are automatically logged by PWSP.

c. Reaction time of PWSP performance. It records the time between the signal went wrong and the subject pressed the right key on the keyboard to correct the signal.

d. NASA-TLX stress self-rating scales [14]. This scales asked mental, physical and temporal demands of the subjects, at the same time, self-satisfaction of the total performance (including PWSP and speech recognition). Effort and frustration level are also asked. The scale was provided to the subject after they finished one of the workload tasks.

Results

General information about the subjects:

The subjects were asked to rate their fluency in English speaking from 1 (fairly good) to 9 (very good). The average value was 7±1.6. All of them did not have experience on living in English Speaking countries for more than 2 months. They learn English through general school education in Sweden. Naturally they spoke Swedish better than English on daily activities. Among the total 15 subjects, there were three of them that felt it was equally easy to speak English as to speak Swedish. None of them has experience of using ASR system excepted one.

ASR accuracy from different experimental conditions

The results of recognition accuracy by the Philips Freespeech 2000 from different experimental conditions is showed in Figure 1. The bar show the average data with its standard deviation

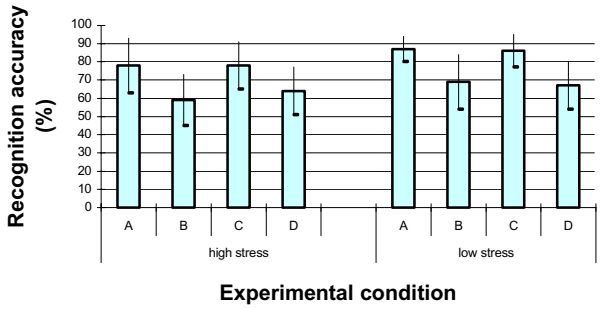

Figure 1, A: Commands was displayed in Swedish text on the screen; B: Commands was displayed in English text on the screen; C: Commands was displayed in Swedish by male voice and D: Commands was displayed in English by male voice. The first four bars on the left side represents the results from high stress condition and the last four bars on the right side represent the results from low stress condition.

Multiple ANOVA analysis

There are no any interactions between any of the factors. The individual differences are very significant ($p<0.05$). Stress has significant effect on the recognition accuracy ($p<0.05$). There is no significant difference between text display on the screen or voice input from headphone. The recognition accuracy on Swedish language is significant higher ($p<0.01$) than English Language in both high stress and low stress situation. Comparing the two input models, there is no significant difference regarding the reaction time of performing PWSP.

NASA-TLX stress self-rating scales

The average and standard deviation of NASA-TLX stress self-rating scales is showed in Table 1. There were significant differences ($p<0.05$) between low stress and high stress conditions according to the scales except the performance satisfaction.

Table 1. The average and standard deviation of the results of NASA-TLX stress self-rating scales.

	Mental Demand	Physical demand	Temporal Demand
Low stress	2.3±0.6	1.7±1.1	2.1±0.8
High Stress	7.0±1.5	6.1±1.9	7.2±1.1
	Performance Satisfaction	Effort	Frustration Level
Low stress	6.0±1.9	3.0±1.9	3.0±2.0
High Stress	4.2±1.7	7.1±1.3	5.6±1.7

Discussion and Conclusion

All of the subjects were at least university students. It means that they were able to read, write, listen and speak English fluently, but with Swedish accents. All the subjects were very confident for their English level. The time stress introduced by PWSP program has significant effect to the subjects due to the results of NASA-TLX stress self-rating scale (see Table 1).

The total recognition accuracy is not very high (from 59% to 89%, see figure 1). There were two reasons for the low accuracy level: the first reason, Philips FreeSpeech 2000 is basically a speaker dependant system. In this experiment, we did not make any individual training due to the small vocabularies. The second reason, we did not use pre-designed language grammar for the commands. These was the purpose to make the system more sensitive to the variance of pronunciation and accents of the subjects.

By the capacity of the products, Philips FreeSpeech 2000 is able to recognition equally well for native Swedish as well as for native American English. The results shows that the recognition accuracy is significant lower on English than Swedish. This is due to the Swedish accents when the subjects speaking English.

Misrecognition comes also from other aspects, such as human error. For instance, saying wrong (insertion and submission), pressong the button in wrong time (too late to press the fire button or too early to release the button). Another type of human error is time out. This is due to the effects from the time stress.

From Figure 1 one can see that during stress situation, the recognition accuracy for English comments is slightly better when it is voice inputting comparing with text inputting, but the difference is not significant. It may due to reason that some of the subject's English pronunciation was affected strongly by the inputting voice. Among the 15 subjects, there were 10 of them that increased the recognition accuracy during voice inputting condition. There is not such pattern found in Swedish language speaking. Since the individual difference is high, so from statistically, there is not significant increase from present study.

Present study imply some interesting aspects which calls for further study:

1. Training individual's voice file for FreeSpeech. This may increase the recognition accuracy but it would be interesting to see to what degree it can be improved.
2. Comparing different strategies of pre-designed grammar files for the commands, for the trade-off between the flexibility and the accuracy of recognition.
3. Studying the effects of different accents of spoken headphone voice to the subject's accent of output voice when using ASR during stress.

Reference

1. Baber, C., *Automatic speech recognition in adverse environments.* Human factors, 1996. **38**: p. 142-155.
2. Howells, H., *Verbal Protocols and Indices of Task Loading*. 1982, Royal Aircraft Establishment, Farnborough, Hants : HM50.: London.
3. Cairns, D.A., Hansen, J.H.L., *Nonlinear analysis and classification of speech under stressed condition.* J. Acoust. Soc. Am, 1994. **96**(6): p. 3392 - 3400.

4. Brenner, M., Doherty, T., Shipp, T., *Speech Measures indicating workload demand.* Aviation, Space, and Environmental Medicine Journal, 1994. **65**: p. 21-26.
5. Baber, C., Mellor, B., Graham, R., Noyes, J.M., Tunley, C., *Worload and the use of automatic speech recognition: The effects of time and resource demands.* Speech Communication, 1996. **20**: p. 37-53.
6. Baber, C. *The effects of workload on the use of speech recognition systems*. in *NATO/ESCA Work-shop on Speech Under Stress: European Speech Communication Association*. 1995. Lisbon.
7. Anderson, T.R. *Applications of speech-based control*. in *RTO Lecture Series on "Alternative Control Technolgies: Human Factors Issues"*. 1998. Brétigny, France.
8. Pallett, D.a.F., J. *1996 Preliminary Broadcast News Tests*. in *Proc. of DARPA Speech Recognition Workshop*. 1997.
9. Kudo, l., Nakama, T., Watanabe, T., and Kameyama, R. *Data collection of Japanese dialects and its influence into speech recognition*. in *Proc. of Intl. Conf. on Spoken Language Systems*. 1996.
10. Hansen, J.H.L., Arslan, L.M. *Foreign Accent Classification using source generator based prosodic features*. in *Proc. of Intl. Conf. on Acoust., Speech, and Signal Processing*. 1995.
11. Teixeira, C., Trancoso, I., and Serralheiro, A. *Accent Classification*. in *Proc. of Intl. Conf. on Spoken Language Systems*. 1996.
12. Humphries, J.J., Woodland, P.C., and Pearce, D. *Using accent-specific pronunciation modelling for robust speech recognition*. in *Proc. of Intl. Conf. On Spoken Language Systems*. 1996.
13. Diaoloukas, V., Digalakis, V., Neumeyer, L., and Kaya, J. *Development of dialect?specific speech recognizers using adaptation methods*. in *Proc. of Intl. Conf. on Acoust., Speech and Signal Processing*. 1997.
14. Hart, S.G., Staveland, L.E., *Development of NASA-TLX (task Load Index): Results of Empirical and theoretical research*, in *Human Mental Workload*, P.A.H.a.N. Meshkati, Editor. 1988, Elsevier Science Publishers B.V., North-Holland.