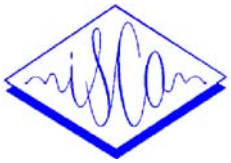

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

THE EFFECTS OF SPEAKER TRAINING ON ASR ACCURACY

Stephen Anderson, Natalie Liberman, Larry Gillick, Stephen Foster, Sahoko Hama
Dragon Systems
320 Nevada St.
Newton, MA 02460

{stevea, natalie, larry, stephenf, shama}@dragonsys.com

ABSTRACT

Do experienced speech recognition users achieve high accuracy rates because their systems have taught them successful speaking styles? We report an experiment to quantify this "speaker training" effect.

In our experiment, 30 computer-literate elderly speakers (15 male, 15 female) with no previous ASR experience were given 2 hours of intensive training in using a speech recognition system. Before and after this training session, they were asked to read separate 520-word texts. Measuring the word error rates (WERs) on these "before training" and "after training" recordings, we find a small but statistically significant improvement. Before training, speakers had an average WER of 20.9%, and after training, 19.8%. We examine changes in speaking rate, phrase length, and SNR and their impact on WER.

This improvement is surprisingly small; anecdotal evidence suggests that experienced ASR users have substantially higher accuracy than novices. The effect may be larger for more extensive training.

10.21437/Eurospeech.1999-104

1. INTRODUCTION

It is widely believed that users of interactive automatic speech recognition (ASR) systems learn, over time, to modify their dictation style to maximize recognition accuracy. A novice user might speak too rapidly, mumble, or run words together, while advanced users alter their speaking style (enunciation, volume, rate, inflection, etc.) and achieve better recognition. If this "training effect" were large and consistent across users, it would be of great commercial interest.

We have conducted an experiment to quantify this effect. Our findings confirm the existence of a "training effect"; recognition accuracy does improve as users become increasingly familiar with the system. However, we find that the magnitude of this effect is small, about a 5% relative improvement (20.9% without training *vs.* 19.8% with).

30 computer-literate volunteers (15 male, 15 female) with no previous speech recognition experience were contacted through local senior citizen computer clubs. The first task for each subject was to dictate one of two 520-word (3.5 minute) scripts. They were then given 2 hours to learn to dictate effectively, with Dragon personnel present to suggest practice exercises and answer questions. This practice session was effective; afterwards, users were able to dictate paragraphs and correct recognition errors by voice without assistance. Finally, users were asked to read a second 520-word script. (Half the speakers read script *A* before training and *B* after, while the other half reversed this order).

We ran automatic recognition of both the "before" and "after" recordings, and could therefore determine the WER improvement due to experience with the system. Our results show a small but statistically significant improvement due to training; 20.9% WER before training, 19.8% after, using speaker-adapted models. We conclude that the training effect is real but small, at least over the first few hours of user experience.

To characterize the changes in speaking style due to the training, we calculate speaking rate, phrase length, and signal-to-noise (SNR) ratios for each speaker, for both their "before training" and "after training" recordings. As expected, those speakers who learned to speak more slowly and more loudly increased their recognition accuracy.

2. TEST DESIGN

In this section we describe the experimental design, the participants, the structure of each session, and the acoustic/language models used.

2.1. Participants

Over a two-month period (July-August 1998), we enlisted 30 senior citizens (15 male, 15 female) to participate in this study. Contact was made via local senior centers with computer clubs or classes. The participants ranged in age from 60 to 84, with median age being 70-74 for both men and women, as seen in Table 1.

M/F	60-64	65-69	70-74	75-79	80+	All
M	3	4	4	3	1	15
F	4	4	6	0	1	15
Tot.	7	8	10	3	2	30

Table 1: Participant age and gender.

As a group, the participants reported an average of 5-10 years computer experience. Of the 27 subjects providing

this data, the distribution of speakers by reported years of computer experience was:

OS	Average Years Experience
DOS	18 speakers: 7.5 years
Windows	22 speakers: 3.3 years
Apple	6 speakers: 4.3 years

Table 2: Participant computer experience

The post-training recording for one speaker (a 70-74 year old male) was not usable. This reduced our participant set to 29 speakers (14 male, 15 female), but did not change the overall demographics.

2.2. Study Design

Each session began with participants completing a demographic questionnaire, including questions about their previous computer experience. None had previously used speech recognition software.

With no explanation that there would be a "before" and "after" aspect to the study, participants were then asked to read aloud a 2-page (520 word) document to the computer, with instructions to "dictate so that the computer can write down what you say". This reading was recorded for later analysis. This text, and the post-training text, were excerpted from a story about a deposited "junk mail" check. It was easy and engaging reading for all participants.

Participants were then given two hours of training and practice in how to dictate to a computer. To simulate an ASR user's experience, the participant used the Dragon NaturallySpeaking® (version 2.5) user interface as a front end to a research recognizer. Training consisted of:

1) Recording 10 minutes of acoustic adaptation data.
2) Viewing the "Quick Tour", a series of 10 introductory dialogs with dictation/correction tips for the new user.

As they worked through the "Quick Tour", we gave tips, explanations, and led them through a series of practice exercises designed to quickly teach the basics of computer dictation. This training took on average 1.5 hours. No explicit advice about speaking style was given.

At the completion of training, subjects practiced dictating for a further 10 to 15 minutes, using the skills and commands they had acquired.

Having learned the basics of dictation, participants then read aloud a second two-page text (also 520 words). Since this reading immediately followed the training session, we anticipated that their speaking style would reflect their newly acquired experience with ASR.

Each participant was assigned a unique ID number. Participants with even-numbered IDs read text *"A"* before training and text "*B*" after training, while odd-numbered speakers read "*B*" first and "*A*" second. This alteration balanced the slight differences in text difficulty.

The entire test took from 2.5 to 3 hours to complete.

2.3. Acoustic and Language Models

All recognition was done with a research acoustic model built from 80 hours of elderly speech and 80 hours of WSJ speech, as described in [1]. It includes 6300 output distributions, each of which has up to 6 multivariate gaussian components.

The language model was built from general English, as described in [1]. Perplexities on this task were 341 for text *A* and 298 for text *B*.

3. RESULTS

Using speaker-dependent acoustic models (adapted to 10 minutes of the participant's speech) in recognizing both "before" and "after" recordings, we found the following word error rates:

	Before	After	Change
Group 1	21.6 %	22.5 %	0.86 %
Group 2	20.3 %	17.3 %	-3.01 %
Total	20.9 %	19.8 %	-1.14 %

Table 3: Speaker-adapted acoustic models: WERs: *before* and *after* training.

Note that, on average, speakers had a 1.14% improvement in accuracy (5% relative) after being trained to use an ASR system.

When the unadapted acoustic models were used, the average word error rates before and after training were:

	Before	After	Change
Group 1	27.5 %	29.1 %	1.61 %
Group 2	26.7 %	23.1 %	-3.57 %
Total	27.1 %	26.1 %	-0.98 %

Table 4: Speaker-independent acoustic models: WERs *before* and *after* training.

3.1. Quantifying Changes in Speaking Style

How did our participants change their speaking styles to obtain better recognition performance? We can measure how they altered their speaking rate (words per minute), their speaking volume (SNR dB), and their phrase length (words per phrase). We can then see if these altered

dictation parameters correspond to improved or degraded recognition.

We find two results:

1) Speaking more *slowly* or *loudly*, (or both) after training led to improved recognition.

2) Just over half of our participants discovered these phenomena.

Our results show strong correlations between making appropriate speaking style changes and improved accuracy. But not enough of our participants correctly perceived these trends for us to observe a large average improvement.

Speaking Rate

For each participant, we calculate the speaking rate both before and after training. In Figure 1 we plot the change in WER (WER after - WER before) against the change in speaking rate (rate after - rate before). There is a clear trend: speakers who learned to speak more slowly achieved a lower word error rate.

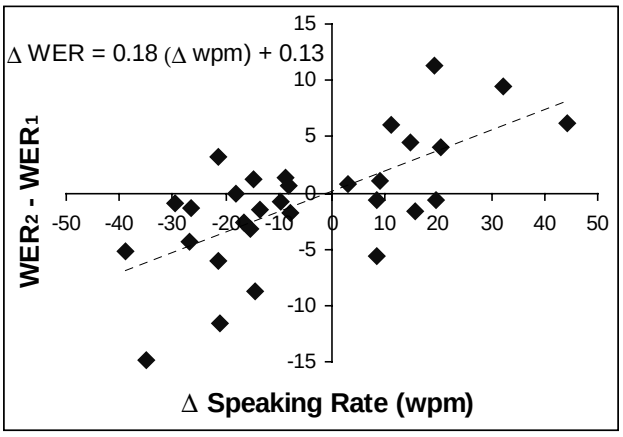

Figure 1: WER Change (after training - before training), *vs*. Speaking Rate Change.

SNR

We also calculate and plot the change in WER against the change in SNR (Figure 2). We see that the speakers who learn to speak more loudly tend to decrease their WER.

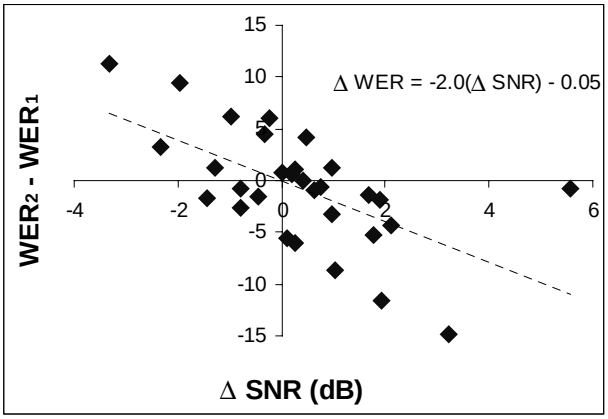

Figure 2: WER Change *vs.* SNR Change

Effect of Changing Phrase Length

Similarly, we calculate the phrase length (words per phrase) both before and after dictation training. Figure 3 does not display as clear a trend, but the participants who shortened their phrase length tended to have improved accuracy as well. A separate analysis shows that long phrases are strongly correlated with faster speaking rates, which may explain the lower accuracies.

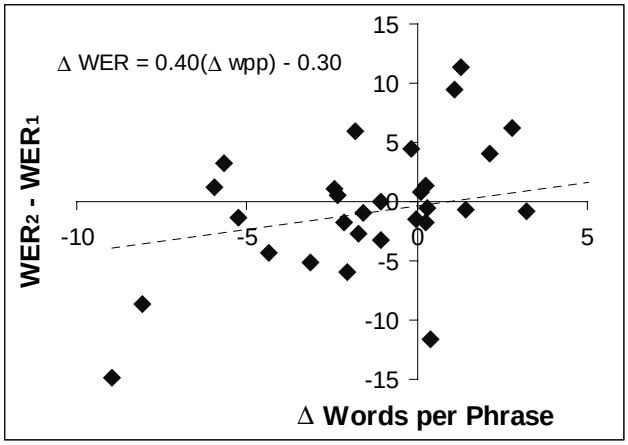

Figure 3: WER Change *vs*. Phrase Length Change. Longer phrases lead to lower accuracy.

4. STATISTICAL ANALYSIS

We divide the participants into two groups:

Group 1: Speakers read text *A* before training and text *B* after.

Group 2: Speakers read text *B* before training and text *A* after.

Let there be n_1 speakers in group 1, and n_2 speakers in group 2. We have n_1=14 and n_2=15.

We define X_{ij} to be the WER of speaker i of group 1 reading script j. Similarly, we define W_{ij} to be the WER of speaker i of group 2 reading text j. In each case, j = *A* or *B*.

We can then define the WER improvement after training for speakers in group 1 or 2 to be y_i or z_i, respectively:

$$y_i = X_{iA} - X_{iB}$$

$$z_i = W_{iB} - W_{iA}.$$

If we define w_A and w_B to be the mean word error rates for readings *A* and *B* (before training), and α to be the improvement due to training, we can rewrite the changes in WER y_i and z_i as:

$$y_i = \alpha + w_A - w_B + \delta_i$$

$$z_i = \alpha - w_A + w_B + \varepsilon_i$$

where we assume that δ_i and ε_i are independent normal random variables drawn from $N(0, \sigma^2)$. (Note that the normality assumption allows us to use the *t*-test below. Alternatively, we could have proceeded with the weaker assumption of constant variance.)

With these definitions, and the definition of Δ as the difference in before-training word error rates,

$$\Delta = w_A - w_B,$$

we can write the average WER change for both groups as

$$\bar{y} = \alpha + \Delta + \bar{\delta}_i$$

$$\bar{z} = \alpha - \Delta + \bar{\varepsilon}_i.$$

The least square estimates of α and Δ are then:

$$\hat{\alpha} = (\bar{y} + \bar{z})/2$$

$$\hat{\Delta} = (\bar{y} - \bar{z})/2$$

To evaluate the significance of our results, we also need to know the standard error of $\hat{\alpha}$. To do so, we use the estimates

$$s_1^2 = \left(\frac{1}{n_1 - 1}\right)\sum_i (y_i - \bar{y})^2$$

$$s_2^2 = \left(\frac{1}{n_2 - 1}\right)\sum_i (z_i - \bar{z})^2$$

of the variances of *y* and *z*, and define

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The variances of $\hat{\alpha}$ and $\hat{\Delta}$ are then

$$Var(\hat{\alpha}) = Var(\hat{\Delta}) = \frac{1}{4}\left(\frac{s^2}{n_1} + \frac{s^2}{n_2}\right)$$

To decide whether the observed improvements in WER are significant, we perform a two-sided t-test for

$$z = \frac{\hat{\alpha}}{\sqrt{Var(\hat{\alpha})}}$$

and n_1+n_2-2 degrees of freedom: $t(z, n_1+n_2-2)$.

5. CONCLUSIONS

We find a small (1.14%) point improvement in WER after training. This improvement is statistically significant, at the P=0.014 significance level.

This level of improvement due to speaker training is surprisingly small. There are several possible explanations. One is speaker fatigue: the speakers were tired after a 3-hour session, and we were effectively measuring the combined effects of fatigue and training. Secondly, participants only used the system for 2 hours. One might expect their WERs to further improve with more practice.

Several changes in speaking style were correlated with this (small) increase in recognition accuracy. Speakers who decreased their speaking rate, who spoke more loudly, or who used shorter phrases increased their recognition accuracy.

6. ACKNOWLEDGEMENT

We wish to thank the NIST-ATP program for supporting this work through grant 70NANB5H1181.

7. REFERENCES

[1] S. Anderson, N. Liberman, E. Bernstein, S. Foster, E. Cate, and B. Levin, *"Recognition of Elderly Speech and Speech-Driven Document Retrieval"*, IEEE International Conference on Acoustics, Speech, and Signal Processing, Phoenix, AZ, March 1999, p. 145.

[2] R. Roth *et. al. "Dragon System's 1994 Large Vocabulary Continuous Speech Recognizer"*, Proceedings ARPA Spoken Language Systems Tech. Workshop, Austin, 1995, p. 116.