GOATS TO SHEEP: CAN RECOGNITION RATE BE IMPROVED FOR POOR TANGORA SPEAKERS?

Catalina M. Danis
User Interface Institute
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598

ABSTRACT

This paper reports on a study of recognition performance for a group of new users during their first month of experience with the TANGORA system. TANGORA is a 20,000 word, speaker dependent, isolated-word system which transcribes speech input into text in real-time. Twelve users, six males and six females, participated in 21 sessions each, during which they read aloud unrelated sentences selected from a corpus of office correspondence. Their goal was to develop a speaking style which minimized TANGORA's recognition error. Hypotheses were generated about users' speech habits which may have lead to increased recognition error and suggestions were made to them on how to modify their speaking style accordingly. On average, recognition errors decreased by 33% from the first to the fourth week. Some characteristics of successful speakers have been identified.

INTRODUCTION

There is a great deal of variability in the accuracy with which users of large vocabulary automatic speech recognition (hereafter ASR) systems are recognized. In a typical finding, Brown, Vosburgh & Canetti (in preparation) reported that recognition error for a group of first time users of a 20,000 word ASR system varied from 2.0% to 14%. Two conclusions may be drawn from such results. First, the technology is good enough to produce a high degree of recognition accuracy for some speakers. Second, there are some speakers who encounter severe problems and, for them, the technology is probably not usable. This research was motivated by the latter group. It is concerned with whether recognition accuracy can be improved through behavioral means for speakers who are initially poorly recognized by an ASR system. Can a user modify his or her speaking style in ways which will be acceptable to the user and will result in a significant improvement in recognition, thereby making ASR systems more widely usable?

IBM's experimental TANGORA system, implemented on the Personal Computer AT, was used in this investigation. This system functions in real-time and has the capacity to recognize 20,000 words. TANGORA is an isolated word system; this requires that users pause briefly between words. Further, it is a speaker-dependent system and must be "trained" to the user's voice. Such a system is most accurate when it has a description or model of the acoustic characteristics of a user's voice. This speaker model is generated by TANGORA from a sample (1200-2400 words) of the user's speech, collected during a "training session." A description of the TANGORA system can be found in Averbuch et al., (1986).

This investigation had four general goals. The first was to investigate recognition performance for a group of new users during their first month of experience with the TANGORA system. The focus was on determining the rate and amount of improvement, if any, in recognition accuracy. It is an important, but unanswered question, whether poor ASR speakers can improve substantially with experience. The second goal was to determine whether re-training TANGORA after users have had experience speaking in isolated-word mode is a useful strategy for improving recognition performance. One might expect that experience with an ASR system leads users to modify their speaking style. Consequently, use of an up-to-date speaker model which reflects these changes might result in improved recognition accuracy. The third goal of this study was to identify those aspects of a user's speaking style which resulted in errors by

the TANGORA system. A description of these problems would serve as the basis for suggestions to the user on how to modify his or her speaking style in order to produce more accurate recognition performance. The final goal was to characterize speakers who are recognized accurately by TANGORA.

METHOD

Twelve users (six males, six females) participated in 21 sessions each. Their task was to produce speech which would be recognized by TANGORA with a high degree of accuracy. To this end, users were encouraged to experiment with their speaking style and to use the feedback provided by recognition errors to shape their speaking style. In the first session, users were given a basic explanation of how their speech would be recognized by the TANGORA system. The importance of clear and consistent speech was stressed. In addition, they were given 30 minutes of experience talking in isolated-word mode with another user's model.

The remaining 20 sessions consisted of four iterations of a five session sequence (see Figure 1). The first session in each sequence was devoted to training the system. The user read aloud, in isolated-word mode, a 2400 word (171 sentence) text. A model of the speaker's voice was computed from the speech sample collected during these training sessions. Each session lasted approximately one hour.

CONTENT	SESSION #
Introduction	1
Training of System - model 1	2
Practice A	3
Practice B	4
Test 1 - model 1	5
Test 2 - model 1	6
Feedback; Training of System - model 2	7
Practice A	8
Practice B	9
Test 1 - model 2	10
Test 2 - model 2	11
Feedback; Training of System - model 3	12
Practice A	13
Practice B	14
Test 1 - model 3	15
Test 2 - model 1	16
Feedback; Training of System - model 4	17
Practice A	18
Practice B	19
Test 1 - model 4	20
Test 2 - model 2	21

Figure 1. Sequence of experimental sessions.
Order of sessions 15 & 16 and 20 & 21 was counter-balanced across users.

In the first two weeks of the study, the speaker model which resulted from a training session was used by TANGORA to decode the speech produced during the following four sessions in the five session sequence.

In each of the final two weeks, the newly created speaker model was used in next three sessions only. The fourth session was decoded against a model generated during an earlier week, as described below.

Two practice sessions followed a training session. Users were given lists of 20 unrelated sentences, selected from a corpus of office correspondence, to read aloud as input to the system. They were instructed to experiment with their speaking style and to try to develop a style which would be successfully recognized by TANGORA. In order to facilitate this process, users immediately re-read a sentence if it was not perfectly recognized, up to a total of four times. They attempted to use the feedback from misrecognized words to selectively modify their speaking styles. The final two sessions in each sequence were devoted to tests: Users were given 40 or 50 sentence lists (also office correspondence) to read aloud to the system and were instructed to use what they had determined to be a "good" speaking style in an effort to produce perfect recognition. They read each sentence only once. It should be noted that all words in both the practice and the test sentences were included in TANGORA's vocabulary and that both practice and test sentence sets were carefully controlled for sentence length and perplexity.

Prior to the second, third and fourth training sessions, each user's performance was analyzed by the experimenter who generated hypotheses about the user's speech habits which may have caused him or her to be poorly recognized by TANGORA. These hypotheses were described to the user and suggestions were made on how the user might modify his or her speaking style.

In order to determine whether re-training the system would improve recognition accuracy, decoding of each user's speech was done against both the current and an older speaker model during weeks 3 and 4. Thus, during the third week, users completed one test session with the newly generated speaker model and one with the model generated at the beginning of the first week. Similarly, the model from the fourth week was compared against the one generated during the second week. If training (which includes the effect of practice) rather than practice alone is the means whereby accuracy is improved, then the following results should be obtained: (1) accuracy during the third and fourth weeks should be better with each week's current model than with the model which had been generated two weeks earlier, and (2) accuracy during the third week with the third week's model should be better than accuracy from the first week with the first week's model and similarly, better during the fourth week with the fourth week's model than in the second week.

RESULTS

Initial performance for this group of users was comparable to previous results with the TANGORA system. Error rate for the first day of practice with the system ranged from 4.5% to 14.0%, with an average of 8.6%. This replicates the findings for first day performance reported by Brown et al. (in preparation).

To address the issue of changes in recognition performance over the four week period, error rate for each **week** was computed by averaging data from all sessions obtained with a given speaker model, when that model was current. That is, for weeks one and two, averages were taken over Practice Session 1, Practice Session 2, Test Session 1 and Test Session 2. Whereas in Weeks 3 and 4, averages were taken over the two Practice sessions and only one Test session. These data were then averaged over all 12 users. The resultant average error rate for the first week was 8.4%. It dropped to 7.5% in the second week, to 6.4% in the third week and to 5.6% in the final week. Thus, there was a 33% reduction in error rate from the first to the fourth week (see Figure 2). There was no reduction in error within a week. That is, performance was constant over the four (Weeks 1 & 2) or three (Weeks 3 & 4) sessions in which each speaker model was used as a **current** model. Data were collapsed over all four weeks and all users to produce an average error rate for the first day, for the second day, for the third day and for the fourth day. These error rates were, respectively, 6.1%, 5.9%, 6.3% and 6.2%.

Recognition accuracy for 11 of the 12 users improved from the first to the fourth week. Figure 3 shows first week error rate plotted against final week error rate for each user. The diagonal line represents no

improvement. As can be seen from the figure, only one user fell at or above this line. Further, eight of the 12 users obtained an error rate under six percent by the end of the study. Whereas, during the first week,

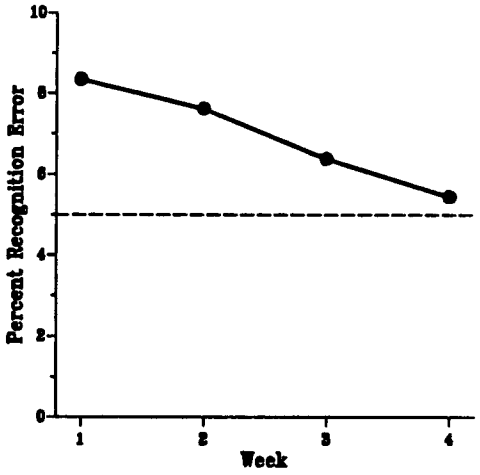

Figure 2: Recognition error at the four weeks, averaged over all users.

only three users had error rates between 5.0% and 6.0%; the error for the remaining nine users fell between 6.0% and 16.0%. Four speakers completed the study with an error rate between 6.8% and 8.3%.

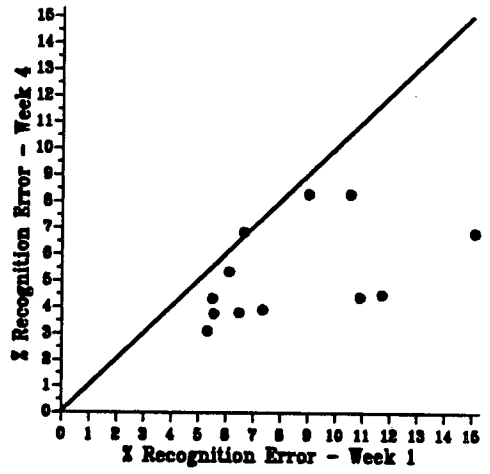

Figure 3: First week error rate vs fourth week error rate for each user.

Re-training proved to be a successful technique for decreasing error rate (see Figure 4). During the third week, error rate for the one test session in which the current (i.e., third week) speaker model was used was 6.9%. However, use of the older speaker model (i.e., the one generated during the first week) during the third week produced an error rate of 10.3%. A similar pattern was observed during the fourth week. Error rate with the current, fourth week, model was 5.7%. It increased to 8.0% when the model from the second week was used.

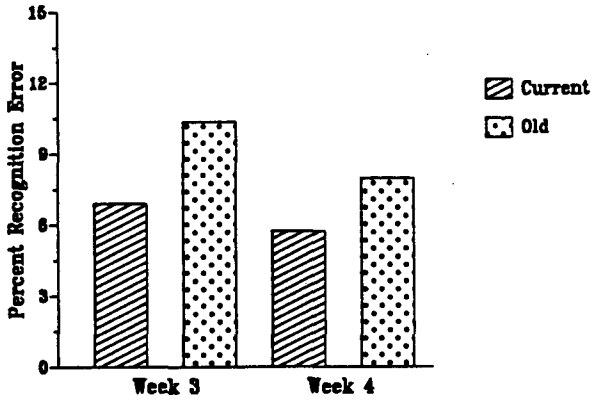

Figure 4. Effect of re-training manipulation.

A number of speech habits brought by users to the ASR situation were identified as contributing to poor recognition by TANGORA. These included: (a) a too fast speech rate, (b) failure to pause between words, (c) hyper-correct articulation of the final phoneme in words, and (d) incomplete articulation of the first phoneme in words. Suggestions on how to modify speaking style were made to users and, in some cases, dramatic decreases in recognition error followed. Speaking at a too fast rate, which resulted in the production of words which were not clearly articulated, was frequently observed with these users. The error was easily corrected by slowing down. Only one user had pervasive difficulties pausing between words. In trying to overcome this problem, in response to feedback, he developed the habit of emphasizing the final phoneme in words. This, referred to here as hyper-articulation, resulted in substitution errors composed of the word spoken by the user plus an unintended word ending (e.g., hearT - > hearts; detaiL - > detailed). Unclear or too short articulation of the first phoneme in words, particularly for words with unstressed initial syllables, frequently resulted in errors as well. Most users were able to modify their speaking style in response to feedback about types of recognition errors the TANGORA system was making. These behavioral interventions were less successful for the four speakers who completed the study with the highest error rate (between 6.8% and 8.3%). Further analysis of their speech is needed to determine why this was the case.

DISCUSSION

The data presented here speak to four points. First, the large range in initial error rate observed for the new users in this and previous studies suggests that large vocabulary ASR systems such as the TANGORA are not "walk-up and use" systems. Most users in this study had to modify the way they spoke in order to be recognized well. It is encouraging, however, that considerable improvement was realized by making changes at the behavioral level. Users discovered some of the changes themselves and were also able to implement suggestions made by the experimenter. This resulted in a decrease in recognition error of 33% from the first to the fourth week.

Re-training on a weekly basis was instrumental in decreasing error rate. The relative contribution of practice and of a better speaker model in the observed improvement can not be determined from these data since the later models were produced by more practiced speakers. It is clear, however, that practice alone, without re-training will not result in performance as good as that which results from practice plus re-training. As the Week 3 and Week 4 test pairs have shown, decoding the speech from a user in a more practiced state against a model generated earlier, when recognition performance was worse, produces a decrement in performance relative to decoding against a current model. The failure to find any improvement in performance within a week (i.e., with the same speaker model) provides further support for the importance of re-training in decreasing recognition error.

The tentative characterization of a "good" TANGORA speaker which has emerged from this study is one who speaks at a fairly slow (approximately 70 words per minute), evenly paced rate and articulates clearly, particularly at the beginning of words. The fact that the speech of some users was still not being recognized well at the end of the fourth week suggests that some modifications at the system level may be needed as well in order to make TANGORA generally usable.

From the standpoint of developing an application using large-vocabulary ASR technology, this study argues for the importance of including tests with "live" users in investigations of recognition accuracy for such systems. It is necessary to provide systems designers with an accurate, realistic characterization of their system and there are aspects of speaking style which emerge when naive users interact with an ASR system, which probably cannot be captured using "canned" speech. First, users clearly modify their speaking style in response to the feedback provided by recognition accuracy in the testing situation. The extent and nature of such adaptation by the user will provide valuable feedback to the designers of the system. Second, it is important to sample users under conditions of everyday use in order to capture the variability in the quality which exists in normal speech. Factors such as alertness, anxiety, and illness affect voice quality and ASR technology must be robust with respect to them if it is to be usable.

Acknowledgments

I would like to thank Norman Brown for many discussions at all stages of this project and Linda Jaynes for her assistance in data collection. My thanks to the Speech Recognition Group at the T. J. Watson Research Center, particularly Ken Davies, for providing technical support as well as informative discussions about TANGORA.

References

Averbuch, A., Bahl, L., Bakis, R., Brown, P., Cole, A., Daggett, G., Das, S., Davies, K., DeGennaro, S., de Souza, P., Epstein, E., Farleigh, D., Jelinek, F., Katz, S., Lewis, B., Mercer, R., Nadas, A. Nahamoo, D., Shichman, G., & Spinelli, P. (1986). *An IBM PC-based large-vocabulary isolated-utterance speech recognizer.* (Research Report RC No. 58679). Yorktown Heights, NY: IBM, Thomas J. Watson Research Center.

Brown, N. R., Vosburgh, A. M., & Canetti, S. (in preparation). *Factors affecting the recognition accuracy of a large-vocabulary voice recognition system.*