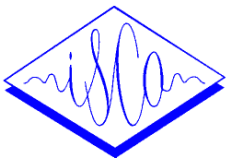

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

INTERFERENCE OF SPEECH RECOGNITION FEEDBACK DURING DIAGNOSTIC TASKS

E.J.A. Verheijen[1], F.L. van Nes[1], L.M. de Bruyn[2], A.Hasman[2], J.W. Arends[3]
e-mail: ellenv@prl.philips.nl

1. Institute for Perception Research/Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
2. Department of Medical Informatics, University of Limburg, P.O. Box 616, 6200 MD Maastricht, The Netherlands.
3. Department of Pathology, Academic Hospital of Maastricht, P.O. Box 5800, 6202 AZ Maastricht, The Netherlands.

10.21437/Eurospeech.1995-334

ABSTRACT

The difficulty of detecting errors in one's own reports may be the bottleneck in enabling a widespread application of automatic speech recognition (ASR). Although quite a lot of research concerning correction procedures has been conducted, detection strategies have received little attention. As ASR systems still produce errors a lot depends on the detectability of errors especially in the medical environment, as reports have to be error-free. This paper presents the results of an experiment which investigated error detection performance during a diagnostic task. No differences in detection performance could be found between auditory or visual feedback except for the detection of substituted words. For this category of errors, visual feedback proved to be better.

1. INTRODUCTION

In many diagnostic departments specialists such as pathologists and radiologists dictate their findings during their examination of a microscopic section or x-ray. Legal regulations demand authorized reports and therefore these recordings are typed out by secretaries and signed by the specialist responsible. Large laboratories can be confronted with more than a thousand examination requests a day. Departments have to get the reports to the requesters on time to enable them to start treatment and so need quite a few (expensive) secretaries to manage this number of recordings.

Application of Automatic Speech Recognition (ASR) appears to be a good and 'cheap' alternative. Unfortunately ASR systems still produce quite a lot of errors. Especially in the medical environment it is important for reports to be error-free. Medical specialists therefore need to correct each report thoroughly. An advantage of ASR is that it will enable the specialists to correct their reports sooner. The specialists would remember their dictation better. In a previous experiment [7], this hypothesis was tested. Pathologists were asked to correct reports which they had dictated the same day and reports which they had dictated the day before. Simulated speech recognition errors were deliberately put in the pathologists' reports. The hypothesis was rejected but what the experiment did show is that it is very difficult for pathologists to detect speech recognition errors. Two-thirds of the errors remained uncorrected.

An explanation for this finding is that the pathologists fail to remember their dictation or the details of the medical image. It could also be the case that the problems are caused by the fact that the reports are dictated (spoken) whereas the pathologists have to correct written text. The written words may be an insufficient cue to allow recollection of what had been dictated or what had been examined. A third reason has to do with the similarity of the material that the pathologists are confronted with. In the time between dictation and verification of the report, pathologists examine and report on other, often similar cases. The cases probably interfere.

At this moment the exact reason for the pathologists' inability to correct these ASR errors is still unclear. Presenting the speech recognition results directly to the pathologist might be a good solution. The modality of the feedback may also be of influence. There seems to be a controversy between Schurick et al. [6] who conclude that visual, word-to-word feedback with history seems most optimal and Frankish & Noyes [3] who suggest better performance with auditory feedback. Baber et al. [1] wisely conclude that a determining factor in the use of feedback is its function in the task and that different tasks have different optimal feedback media. The present experiment will make it possible to re-examine both the Schurick et al. and Frankish & Noyes findings in the light of the diagnostic tasks of pathologists.

2. HYPOTHESES

It is thought that a number of factors influence the detectability of the simulated speech recognition errors. The modality in which the feedback is presented is of importance. As previously mentioned, visual feedback of the dictated text may not be an effective cue for the pathologists to recall what they dictated. It is believed that auditory feedback is more similar to the dictation and that detection performance should therefore be better in the case of auditory presentation.

Another factor that influences the detectability of errors is the complexity of the microscopic sections that the pathologists need to diagnose. The more complex the sections the more the pathologists will have to devote their attention to them and the less they can concentrate on detecting errors.

A third factor that seems interesting to investigate is the

level of experience of the pathologists. For trainee pathologists the information they see and the words they dictate are much less familiar than for the pathologists, who are experts. It is because of this that trainee-pathologists are expected to detect more errors than the more experienced pathologists. To test these hypotheses the following experimental setup was designed.

3. METHOD

3.1 Subjects

Five pathologists and three trainee-pathologists of an academic hospital in the Netherlands participated in the experiment. All subjects had some experience with computers and some had seen demonstrations of speech recognition equipment but none had worked with it. The subjects were asked to examine microscopic sections and to dictate a report containing their findings. After each utterance the recognized utterance was reported back to them. The experiment took approximately one hour.

3.2 Material

Eight microscopic sections were used: four cases of liver excidum and four naevus (skin) cases. The cases were selected from the hospital archive based on their age (all were approximately one year old) and on the similarity of the diagnostic reports that had been made about them.

Three types of errors were introduced into the reports: deletions, substitutions and inversions as a special case of substitutions[1]. Due to the fact that the pathologists were free as to what they dictated, the number and type of modifications could not be determined beforehand. It was decided to change specific words with a certain probability in such a way that the resulting 'recognized' report contained a number of errors similar to those produced by ASR systems. The type of words was extracted from [4] and [5].

3.3 Set-up

The experiment was setup in one of the offices of the pathology department of the hospital. A computer containing a videocard and a speech synthesis card was connected to a PC in a neighbouring room via an RS-232 cable. In this room a typist was seated who had been trained to type the relevant medical terms. He typed out what the subjects dictated.

3.4 Design of conditions

Last but not least the details of the two feedback conditions must be mentioned. The visual feedback was presented on a computer monitor. The computer contained a Video Blaster card and corresponding software to enable the projection of the microscopic image onto the monitor screen via a camera in the microscope. The subjects could use the handles on the microscope to scan through the sections. The visual feedback was displayed in a bar at the bottom of the screen, resembling the way subtitling is displayed on television. A maximum of two lines could be displayed, the letters were 24 points large and remained on the screen for 6 seconds corresponding to a subtitling situation (see e.g. [8]). The monitor was located at a distance of approximately 1.5 metres from the subject. This made it possible for the subjects to look at both the display of the section and the text simultaneously, as these were located in the same visual field.

In the auditory condition subjects looked at the same computer monitor and could scan through the section in the same way as in the visual condition. Instead of presenting the recognition result in a bar on the screen the 'recognized' utterance was reported back to the subject by the Polyglot diphone speech synthesis system. The lexicon was extended with 140 medical terms from liver and naevus reports. The length of the utterance reported back to the subject was similar to what was reported in the visual condition but the time between dictation and feedback was somewhat longer (a mean difference of 1.5 seconds) due to the fact that the synthesis program had to start up again for each utterance.

In both conditions the subjects responded to errors by pressing a key on the keyboard and by saying out loud what the error was exactly. Logfiles were made of the text before the error generation and after the errors were put in, and of the subjects keypresses. Auditory tape recordings were made as well.

3.5 Preparatory analysis of speech synthesis

A small experiment was designed to test whether the synthetic speech was sufficiently good to detect errors at all. Four subjects were each given six reports. A number of changes had been made in the reports that were spoken by the Polyglot system. The subjects' task was to check the synthesis and to mark all differences in the text. The results are presented in table 1. The bottom row are the

Table 1: Results preparatory analysis

Subject	DEL	SUB	INV	Total
1	11	11	3	25
2	9	7	3	19
3	9	15	3	27
4	8	10	3	21
Entered	**11**	**20**	**3**	**34**

1. Although the most common categorization of ASR errors is in deletions, substitutions and *insertions,* insertions often co-occur with a substitution. A relatively large word is substituted by two small words. For this reason these types of errors are merely counted as substitutions in this research. An example from the experiment is the change of "worden" into "wordt een". This is counted as one substitution instead of one substitution and one insertion.

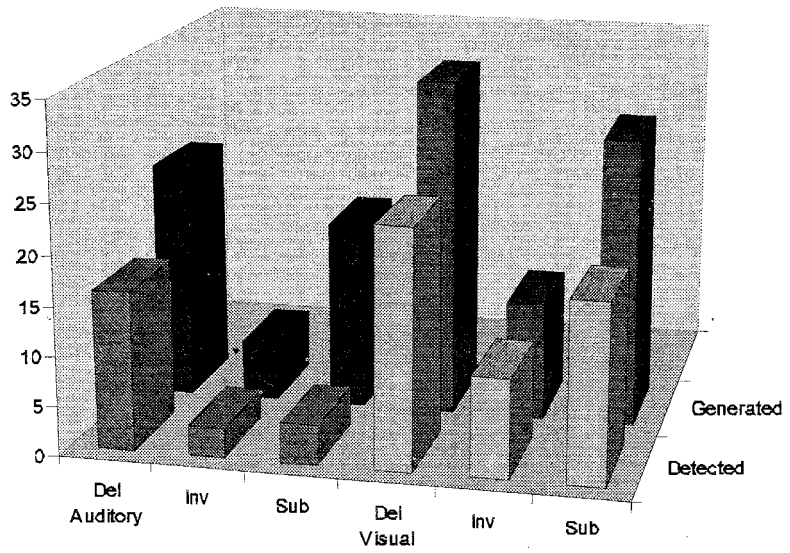

Figure 1. Generated and detected errors in each condition.

number of errors entered. For each subject the number of detected errors is given per error-type. The total number of detected errors per subject is presented in the column on the right.

4. PROCEDURE

A repeated measures experimental design was employed, with subjects working through the two feedback conditions in the same experimental session. The order of conditions and reports was counterbalanced across subjects.

Subjects were first given brief instructions mentioning the goal and procedure of the experiment. The way they should react to errors and what was to be counted as an error was made clear to them. Each condition was practised just before the subject experienced the condition for the first time. During this practise session the experimentator pointed out the errors and reminded subjects how to react.

Each subject received two complex cases (liver excidum) and two easy cases (naevus) randomly selected from the eight available microscopic sections. The sequence was either Auditory (A)-Visual (V)-Visual-Auditory or V-A-A-V.

5. RESULTS

As previously mentioned, errors were randomly introduced into the reports. The resulting numbers of generated errors are displayed in figure 1. For both conditions the numbers of generated errors of each type are displayed; this is represented by the bars at the back (the second) row. The same figure also shows the number of detected errors across the subjects. The bars in the front show the number of detected errors for both conditions (two groups of three) and for each error-type. A t-test for independent samples indicated that there was no significant difference between the number of errors entered in each condition, $t(25)=.57$, $p>.05$ for the deletions, $t(11)=.80$, $p>.05$ for the inversions and $t(22)=.53$, $p>.05$ for the substitutions. These results indicated that the conditions were comparable in terms of error generation.

In total, subjects corrected 23 of 49 errors in the auditory condition and 52 of 75 errors in the visual condition, equalling 47% and 69% respectively. No statistically significant difference was found between the detection performance in the auditory and visual condition. When each error-type was tested separately, however, it was found that detection performance for substitutions did differ between the conditions. A t-test, with LOG transformations in order to be able to handle proportions, revealed $t(21)=4.02$, $p=.001$. With respect to substitutions it can be concluded that the modality of the feedback influences that detection performance. Subjects are better able to correct substitutions when these are presented to them visually.

Regarding the complexity of the reports and the detectability of errors, no significant difference could be found. For the inversions it was very close to significance, $t(7)=2.16$, $p=.06$ but the number of cases was very small. It seems that there is a trend towards less detection when the reports become more difficult but more research is necessary to confirm this.

Learning effects or effects due to fatigue were not found. There was no significant difference between the errors detected in the first two reports and the errors detected in the last two reports.

Finally the detection performance in the auditory condition was compared to the results in the preparatory analysis. For deletions and inversions there was no significant difference but for the detection performance of substitutions there was, $t(9)=4.18$, $p=.05$. Detection performance for substitutions differed for the two groups.

6. DISCUSSION

It was found that the modality of feedback influenced how well subjects were able to detect substitutions in their own text. The overall tendency was that subjects detected more errors in the visual condition but the difference was not significant.

For substitutions a difference between the two conditions is found and one can imagine that the quality of the speech synthesis is the reason for this finding. Deletions become apparent in a similar way for both conditions: there is a word missing. The situation is similar for inversions. The subjects know that when they dictate these words that they have to pay attention. The subjects may not be able to understand the speech synthesis completely and could have given the computer the benefit of the doubt: they may have presumed that the feedback was correct.

With respect to this quality issue the prepatory analysis was performed. As mentioned in the results, statistical analysis revealed a significant difference between detection performance of substitutions in the prepatory analysis and in the auditory condition of the experiment. Finding could be attributed to the fact that the subjects were merely more strict but it could also be the case that they were able to concentrate better on the task. In the experiment subjects received feedback which they had to check during their examination of microscopic sections. Because of this they had to divide their attention which could have caused their lesser performance.

Another explanation regarding this finding has to do with the subjects' familiarity with the text. In the prepatory set-up subjects received a text which was not their own. In the experiment the subjects were asked to compose their own reports. Daneman and Stainton [2] have shown that extreme familiarity with text (like one's own text) is likely to result in less corrected errors. This could explain why subjects miss the errors in the experiment. On the other hand one would expect that also more deletions and inversions would then be missed in the experiment and this is not the case.

As pointed out in the introduction the subjects' failure to detect errors could be caused by memory problems. Recall of the exact structure of a sentence is very short and the time between dictation of an utterance by the subject and presentation to the subject may have been to long and could explain the overall low performance. Inversions are content related changes and should according to this explanation have been detected better than the other two categories. This was not the case.

It was expected that the complexity of the microscopic sections would also influence detection performance and especially when the attention explanation is true this should be the case. No significant difference was found but it is possible that the subjects did not take their task of examining the sections seriously as they knew it was an experiment.

7. CONCLUSIONS

Regarding the stated hypotheses the following can be concluded. Not auditory presentation but visual presentation seems to be better to detect errors, especially for substitution errors. The complexity of the microscopic sections did not significantly influence the detection performance and also the level of experience of the subject made no difference. For the time being it is believed that the detection difficulty lies in the fact that the subjects have to divide their attention between checking the auditory feedback and visually inspecting the microscopic section. This dual-task situation could be the reason for the poorer performance in detecting substitution errors.

From this finding it can be concluded that visual feedback should be given when the main, most attention demanding task is a visual one. Future research should focus on the question if the level of attention is of importance and it would be interesting to see if the opposite is also true: is auditory feedback is better when the main task is auditory in nature?

8. REFERENCES

[1] C. Baber, D.M. Usher, R.B. Stammers and R.G. Taylor. Feedback requirements for automatic speech recognition in the process control room. *International Journal of Man-Machine Studies*, 37: 703-719, 1992.

[2] M. Daneman and M. Stainton. The generation effect in reading and proofreading: Is it easier or harder to detect errors in one's own writing?, *Reading-and-Writing*, 5(3): 297-313, 1993.

[3] C. Frankish and J. Noyes. Sources of human error in data entry tasks using speech input. *Human Factors*, 32(6): 697-716, 1990.

[4] K.F. Lee. *Hidden Markov Models:Past, Present and Future. In: EuroSpeech 89, 1, 148-155,* 1989.

[5] S.Pauws and M. Ceelen. *MARS: Medical Applications for Recognition of Speech.* Eindhoven, Technical University Eindhoven, 1993.

[6] J.M. Schurick, B.H. Willigies and J.F. Maynard. User feedback requirements with automatic speech recognition. *Ergonomics*, 28 (11): 1543-1555, 1985.

[7] E.J.A. Verheijen, L.M. de Bruijn, F.L. van Nes, A. Hasman and J.W. Arends. Automatic speech recognition in diagnostic environments: will it improve report reliability? Submitted to *Behaviour and Information Technology*, 1994.

[8] G. d'Ydewalle, L. Warlop and J. Van Rensbergen. Television and Attention. *Medienpsychologie*,1: 42-57, 1989.