Improving user-friendliness by visually supporting speech recognition

Juliette Waals, Frank L. Kooi, & Sjaak Kriekaard
TNO Human Factors
PO Box 23, 3769 ZG Soesterberg, the Netherlands
(+31).3463.56242 / kooi@tm.tno.nl

ABSTRACT
While speech recognition in principle may be one of the most natural interfaces, in practice it is not due to the lack of user-friendliness. Words are regularly interpreted wrong, and subjects tend to articulate in an exaggerated manner. We explored the potential of visually supported error correction (speech recognition integrated in a graphical interface) to improve the user-friendliness of speaker independent speech recognition. We tested five schemes (Words only, Words via menu, Number-correction, Background color-correction, and Number + color-correction) in four noise environments (Office, Radio, Cafeteria, and Car noise) on eight subjects. Background color correction fits within the design trend to develop "low attention interfaces" because it does not require the user to fixate the display accurately. The results show that a visual support significantly improves the speech recogniser's recognition rate, on average by 10% in all four noise environments, and each subject benefits. In the car noise condition, task success is highest with Color+number support; in the other three noise environments Color-support works best. Most subjects show the tendency to articulate less clearly (= more naturally) in the conditions that include a correction step. A correction step that involves the pronunciation of the background color therefore makes the system more effective and more natural to use.

KEYWORDS: Speech recognition, error correction, color coding, color naming.

INTRODUCTION
Following many others [1, 2, 3, 4, 5, 6], we have set out to improve the user-friendliness of speech recognition, focussing on situations where the user is viewing a computer display. The type of use we envisage includes mobile phones, in-car displays, PDA's, and dual task locations like the home environment. A common feature is the potential presence of noise, making the need for correction support much more acute than is the case in office environments.

METHODS
In the experiment the user is instructed to select the desired command by pronouncing it. We then visually highlight all commands with a recognition score above a pre-set level as illustrated in the middle and right columns of Figure 1. In essence this makes the commands the user might have said available for selection. The user says either the superimposed number (middle column) or the background color (right column) of the desired command. The user therefore not only receives immediate feedback on the status of the recogniser, we also expect the recognition score to improve because the correction vocabulary is small and phonetically distinct. Number-coding is an established technique available in most dictation systems. To our knowledge, color-coding is new, with three potential advantages over number coding: 1) Background colors are much easier to detect from a distance or while looking away from the display, 2) the choice of colors in a palette is free and may be chosen to ease pronunciation and optimise phonetic discriminability, and 3) Color-coding does not take up extra space in the graphical user interface.

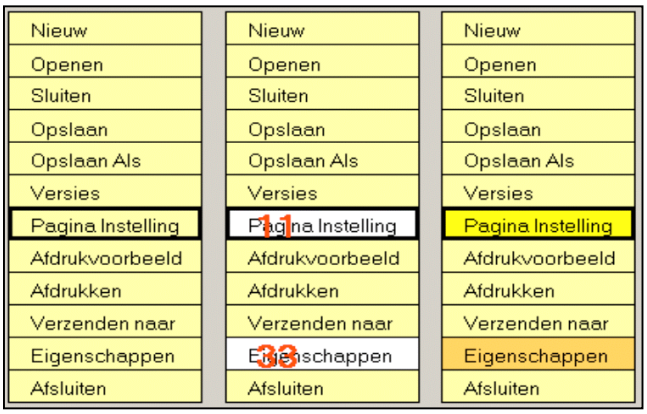

Figure 1. Illustration of the visual support experiment. Leftmost column: the user is instructed to pronounce the boxed command (*Pagina instelling*). In case the speech recogniser is unsure whether *Pagina instelling* was said or *Eigenschappen*, each is marked, either by a number (middle column) or by a background color (right column). The user is instructed to have a second go by saying the number or colour that corresponds to *Pagina instellingen* (in this example '11' or 'yellow', which is 'geel' in Dutch).

The control condition ("Words only") measures the recognition score without correction. Eight subjects ran each correction condition at four noise environments (Office: 35 dBA, Radio: 50dBA, Cafeteria: 57dBA, car: 72 dBA).

RESULTS

Figure 2 provides a compact summary of much of the data. The horizontal axis contains the errors made prior to correction for the "Words", "Numbers", and "Colors" conditions. On the vertical axis, the error rate after correction is shown. For the control condition ("Words only"), the initial and final error rate by definition are equal. Color correction leads to a 10% decrease in errors for all noise conditions, Number correction is slightly less effective. It appears that subjects prefer to pronounce the word more naturally if it it is followed by a number or color correction. If the word is not followed by a correction modality, subjects choose to pronounce the word as clearly as possible in order to avoid errors. This result indicates that, even in the Office condition, the speech recogniser forces people to articulate carefully. Thus, while speech recognition in principle may be one of the most natural interfaces, in practice it is not. Schapira and Sharma [7] also reached this conclusion: "even though Point and Speak is one of the most natural strategies, repeating spoken words constantly is not; users were tired after the first few trials and finished being relatively annoyed by talking to the screen for several minutes".

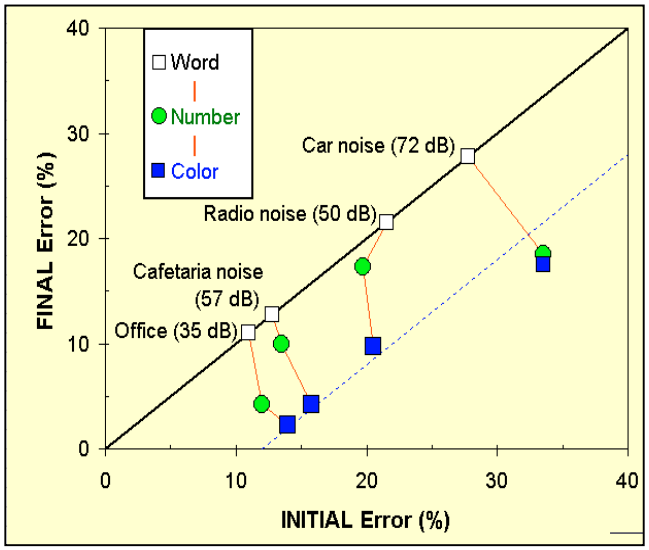

Figure 2. Initial error rates (before correction) and Final error rates (after correction) for the four Noise Conditions, two correction schemes (Number & Color), and the control scheme (Word). Scores below the diagonal benefit from the correction step. Color correction reduces the error rate by about 10% (dotted blue line).

DISCUSSION & CONCLUSIONS

The present results show that the addition of a correction modality makes the speech interface more user-friendly and improves the recognition rate by approximately 10% in all four noise conditions. Our color coded design requires users to only loosely view the display while interacting with the computer by way of speech recognition, congruent with a "low attention interface". This is particularly useful in situations that people will want to interact with the computer while at the same time doing something else. The prime example are in-car systems, combining the driving task with what we now consider as office tasks.

REFERENCES

1. Karat, J., D. Horn, C. Halverson & C. Karat (2000). Overcoming Unusability: Developing efficient strategies in speech recognition systems. Proceedings of the International Conference on Computer-Human Interaction, The Netherlands: Amsterdam.

2. Mankoff, J., S. Hudson & G. Abowd (2000). Providing Integrated Toolkit-Level Support for Ambiguity in Recognition-Based Interfaces. Proceedings of CHI 2000, pp. 368-375.

3. Oviatt, S. & R. vanGent (1996). Error resolution during multi-modal human-computer interaction. Proceedings of the International Conference on Spoken Language Processing 1996, pp. 204-207.

4. Oviatt, S., P. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson & D. Ferro (2000). Designing the User Interface for Multi-modal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions. Human-Computer Interaction, Vol. 15, pp. 263-322.

5. Suhm, B., B. Myers & A. Waibel (2001). Multi-modal Error Correction for Speech User Interfaces. ACM Transactions on Computer-Human Interaction, Vol. 8, No.1, pp 60-98.

6. Terken, J. & S. te Riele (2001). Supporting the Construction of a User Model in Speech-only Interfaces by Adding Multi-modality. Proceedings of Eurospeech 2001, pp. 2177-2180.

7. Schapira, E. & R. Sharma (2001). Experimental Evaluation of Vision and Speech based Multimodal Interfaces. *Proceedings of the Workshop on Perceptive User Interfaces*. Orlando, FL. USA.