Visual Error Resolution Strategy for highly-structured text entry using Speech Recognition in FP6-ALLADIN project

Xavier Ricco, Stéphane Deketelaere, Jo De Lafonteyne, Alexandre Girardi

Multitel ASBL, Speech and Signal Department,
Parc Initialis - Avenue Nicolas Copernic 1 7000 Mons, Belgium
{ricco, deketelaere, delafonteyne, girardi}@multitel.be

Abstract. Man-Machine Interaction using only speech input is not well received by users, even for high performance recognizers (WER of about 2%). In most free text dictation application, attaining users intention is more important than specific speech tools performance, and low transaction success rate results in user's rejection to speech interfaces [6]. For highly-structured text entry, users will better accept speech technology when it is combined with a good multimodal error resolution strategy maximizing the usability of the system. This paper describes an innovative "multimodal" interface component developed in the scope of the ALLADIN project and called Speech Transcription Manager (STM). The purpose of this component is to offer this efficient multimodal interface combining speech recognition and visual error correction strategy in an application for physiotherapists using spoken sentences to produce diagnosis and enter repetitive information. STM integrates different modes of data entry: speech (using recordings coming from a Personal Digital Assistant), keyboard strokes, mouse gesture and visual feedback.

1 Introduction

An innovative user interface was developed in the scope of the ALLADIN project ALLADIN focuses on the development of a user-friendly natural language based decision support software for neuro-rehabilitation, in particular in stroke. ALLADIN will provide an adequate and fast solution for a client centred practice, for discharge planning and for utilization of rehabilitation resources. The need of a faster and easier way to fix transcriptions in the scope of the ALLADIN project was the main motivation of this work.

Our task in the project is the development of an efficient multimodal interface combining Speech Recognition and visual error correction to allow physiotherapists using spoken sentences to produce diagnosis and enter repetitive information. The particularity of our approach versus a classical free language dictation system is the highly structured language model. It is highly structured because it is associated with medical information in the context of the physiotherapy language describing human bodies and its potential symptom. We took initially the decision to exploit a visual feedback based on a multimodal error resolution method, because several

studies [2],[3],[6],[7],[8] show that it is more efficient than unimodal correction. For example, using speech for input and combining visual feedback and mouse click (or touch screen) for correction is more efficient than using speech for both input and correction. This advantage is still more important if a highly structured language model can be exploited. If an error occurs, alternative candidates can be easily proposed.

Finally, the multimodal solution superiority is mainly inferred from an easier user feedback. Unimodal systems requires very quick, repetitive and irritating tasks to manage error correction strategies. Indeed, unimodal systems do not take the best advantage of user's short term memory[1] and lack an easy barge-in[2] method. Our solution to the above problem, developed in the framework of the ALLADIN project, is the Speech Transcription Manager and is described in next paragraph.

2 Speech Transcription Manager User Interface

The Speech Transcription Manager window (Fig. 1 shows a session to be verified by the user) presents by default two panes: the left pane is a tree, which contains recordings stored hierarchically per session/patient folders; the right pane shows the speech transcriptions generated by the speech recognition process. On the user interface, you can also find some miscellaneous information in the upper-right corner and a menu containing buttons in the bottom-right corner. It allows speech to be synchronized from a PDA stored as digital audio in a SD Card, then sent off for transcription. The progress of the transcription can be followed with the progress bars.

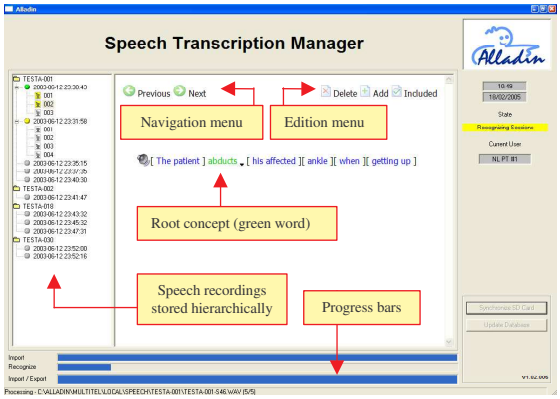

Fig. 1. The speech transcription manager presents sessions to be verified by the user.

The main pane shows the speech transcriptions resulting from the speech recognition process. A sound icon lets the user play the recorded sentences corresponding to the translated transcriptions. In order to take advantage of the speaker adaptation module, the user is invited to mark all sentences where the speech recognition output doesn't match the text on the screen.

[1] It is demonstrated by cognitive research that it is very difficult for a normal user to remember more than 7 alternatives together

[2] You must wait for the correct answer to be proposed, thus waiting for the last alternative to be prompted, which is very frustrating …

A *choice mediator* based on a pie menu [5] was implemented to facilitate the correction of errors. According to Hopkins, D. [4] "Pie menus are faster and more reliable than linear menus, because pointing at a slice requires very little cursor motion, and the large area and wedge shape make them easy targets. Pie menus are easy to learn, fast to use, and provide a gestural style of interaction that suits both novices and experts".

Our contextual menu allows selecting a value from a list, which is linked to a predefined template. Contrary to the pie-menu technique used by Kurtenbach *et al.*[5] and referenced by Mankoff *et al.*[6] as non-contextual, we decided to take contextual information into account when popping up the pie-menu. The menu is composed of a remove icon, blue words and black words. The blue words in the contextual menu represent the speech recognition hypothesis. They are sorted from the higher scored result to the lower one coming from the speech recognition engine ranked by a *n*-best list. The user can type any letters to filter the pie menu and the linear menu.

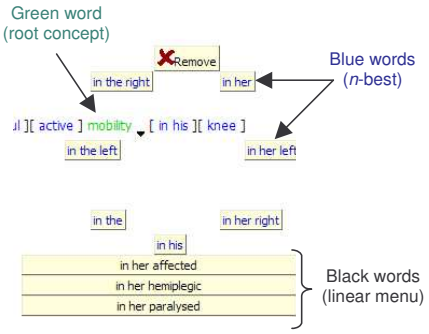

Fig. 2. One pie menu containing *n*-best results coming from the speech recognition engine.

The hypothesis with the highest score is located at the right of the remove button (e.g., in Fig. 2, *in her* is the top-scored value). The black words in the contextual menu represent the list of allowed words in the current sentence reflected by the grammar format. This format consists in including a specific root concept in each sentence, which is represented by the green colour in Fig. 2.

When the user selects another root concept, the speech recognition is re-launched using a smaller constrained grammar restricted to the root concept in question, taking into account the strong word hypothesis provided by the user in a relevance feedback mechanism. This approach will provide to the speech recognition a smaller and therefore easier "second chance" (grammar perplexity is reduced because several best candidate hypothesis become irrelevant due to user's choice), with the opportunity to correct the full sentence at once avoiding the cumbersome task of retyping the entire sentence. We also avoided sorting the linear menu based on the ranking score, because cognitive studies show that user can quickly retrieve the right information from a list if an inherent a priori human learned order, like alphabetical, exists.

3 Conclusion and future work

This paper presented an innovative "multimodal" speech centric interface developed in the scope of the ALLADIN project. It integrates in a single interface speech

recognition, pie-menus, and the concept of root words. Root words facilitate the correction of important deviations from the main user intended speech input, while keeping menu size within a reasonable size. The developed STM is producing logs that will provide us important information to understand the value of pie-menus and root words concepts for error resolution strategies evaluation in the second phase of the ALLADIN project. The clinical partners of the project which are using the device and the software have already made first assessment. The feedback is very positive about usability of this kind of interface [9].

In the future, we wish to explore error resolution strategies using approaches such as: Naïve Bayesian Networks [1] for rapid learning of user's corrections (In this case, the most probable correction may be enhanced to still simplify and accelerate the correction phase) and background speaker adaptation using all the stored verified audio recordings to adapt the acoustic model to the end-user voice.

Acknowledgements

This work is a mid-term result from a project supported by IST. We thank the members of the ALLADIN Research Group who used the system we developed for the design and execution of this project.

Many thanks to Sophie Roekhaut and Richard Beaufort for the minimization of the language set used in the Speech Transcription Manager.

References

1. Androutsopoulos, I., et. al.: "Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach", in Proc. 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2000), (Sep. 2000), pp. 1-13.
2. Berglund, A., Qvarfordt, P.: "Error Resolution Strategies for Interactive Television Speech Interfaces". Human-Computer Interaction, INTERACT'03 M. Rauterberg et al. (Eds.) Published by IOS Press, (c) IFIP (2003), pp. 105-112
3. Deketelaere Stéphane, "How to take advantage of voice in wearable computer context", Proceedings of 1st IFAWC forum, TZI-bericht Nr 30, (2004), pp 85-96
4. Hopkins, D.: "The Design and Implementation of Pie Menus. There're Fast, Easy, and Self-Revealing". In Dr. Dobb's Journal, lead article, user interface issue (Dec. 1991), pp. 16-26.
5. Kurtenbach, G., Buxton, W.: "User learning and performance with marking menus". In Proc. of CHI (1994), pp. 258-264.
6. Mankoff, J., Hudson, E.H., Abowd, G.D.: "Interaction techniques for ambiguity resolution in recognition-based interfaces", in Proceedings of the 13th annual ACM symposium on User interface software and technology (2000), pp. 11-20.
7. Oviatt, S. "Mutual disambiguation of recognition errors in a multimodal architecture". In Proceedings of the Conference on Human Factors in Computing Systems, CHI, (1999), pp. 576–583.
8. Suhm, B., Myers, B., and Waibel, A. "Multimodal error correction for speech user interfaces". ACM Transactions on Computer-Human Interaction 8, 1, (2001), pp. 60–98.
9. Lamson, J., Ruijter, Sd., and Vaerenbergh, JV., "ALLADIN: A helping hand for making the right choice in neuro-rehabilitation", http://www.alladin-ehealth.org/publications, (2005)