Mobile Dictation With Automatic Speech Recognition for Healthcare Purposes

Tuuli Keskinen[1], Aleksi Melto[1], Jaakko Hakulinen[1], Markku Turunen[1], Santeri Saarinen[1], Tamás Pallos[1], Riitta Danielsson-Ojala[2], and Sanna Salanterä[2]

[1] School of Information Sciences, University of Tampere
Kanslerinrinne 1
FI-33014 University of Tampere, Finland
{firstname.lastname}@sis.uta.fi

[2] Department of Nursing Science, University of Turku
Lemminkäisenkatu 1
FI-20014 University of Turku. Finland
{firstname.lastname}@utu.fi

ABSTRACT

This paper introduces a mobile dictation application with automatic speech recognition for healthcare purposes, and its evaluation in a real hospital environment. Our work was motivated by the need for improvements in getting dictated patient information to the next treatment step and the complexity of patient information systems. We designed, implemented and evaluated the application as a close collaboration between human-computer interaction and nursing science researchers. The application was evaluated as a Wizard-of-Oz scenario where two nurses used the application as part of their work routines and a researcher acted as the wizard, i.e., checked the recognition results before sending them back to the nurse. The nurse was then still able to edit the text and then copy it to the patient information system. Our main focus was to gather subjective feedback, and we gathered both user expectations and experiences from the participants. The results show true potential for our mobile dictation application.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces And Presentation**]: User Interfaces – *Input devices and strategies, Interaction styles, Haptic I/O, Voice I/O.*

General Terms

Measurement, Performance, Design, Experimentation, Human Factors, Languages.

Keywords

Speech recognition, healthcare dictation, evaluation, user expectations, user experience.

1. INTRODUCTION

Spoken language has traditionally been heavily used in healthcare field, where doctors commonly dictate information on patients. Manual typing of these dictations is still common but utilizing speech recognition is increasing. Through our discussions with professionals working in the healthcare area, we see problems in getting patient information effectively to the next treatment step: e.g., in the ward we piloted in, the dictated statements may take up to several days before they are available in writing. These are usually statements that are not so urgent, but there are queues and unnecessary delays also with critical dictations and their transcription.

According to Parente et al. [1] first speech recognition systems for healthcare reporting were developed almost twenty years ago, but still they are not widely used, especially within a language like Finnish, which is spoken only by 5.5 million people. One reason behind this is the fact that data for building speech recognition is not as readily available. This is particularly so for healthcare field, where language is very specific for each subfield and separate language models are often necessary, e.g., for doctors working in different fields. For Finnish language, the language modeling is challenging since it is a morphologically rich language. Thus, the recognition method cannot be based on fixed vocabularies because they would grow too big and be practically impossible to create. One example of utilizing speech recognition in Finnish healthcare is presented by Koivikko et al. [2], who followed radiologists changing from conventional cassette-based reporting to speech recognition based dictating.

Motivated by the paucity of using dictation applications with speech recognition in Finnish healthcare, we have developed a mobile dictation application for healthcare purposes to be used by doctors, nurses and other professionals in the field. While many studies on speech recognition in the area of healthcare have been presented, e.g., [1], [3], [4] and [5], these studies focus more or less on objective qualities, e.g., dictation durations and speech recognition error rates. Our main goal was to study the subjective user expectations and experiences of the mobile dictation application and automatic speech recognition from HCI perspective. In addition, the application features a mobile device in the form of a tablet computer, which is designed to support dictation during the regular work and enables not only dictation but also review and editing of both the recording and recognition result on the go. Our primary target user group for the application has been nurses. Most dictation applications in healthcare area are aimed for doctors, whose needs and types of dictations differ from those of nurses. The language nurses dictate is often closer to regular spoken language but still contains a lot of special vocabulary. Nurses also have more often a need for the mobile style of dictation, since they usually work and interact with numerous patients, often in short durations at the time. The work is done as a multidisciplinary collaboration between researchers

from the field of human-computer interaction (HCI) and researchers from nursing science. In this paper, we report results from a pilot study in real-life environment.

The rest of the paper is organized as follows. First, we describe the mobile dictation application. Then, we present the evaluation in detail, including descriptions of methodology and data collection. Finally, we conclude by presenting and discussing the results, and their implications of future potential.

2. SYSTEM

The mobile dictation service is based on "MobiDic" system presented by Turunen et al. [6]. It consists of a mobile client and a server that communicate with speech-to-text recognition engines and M-Files document management system. The system is compatible with Nuance's Dragon Mobile Dictate speech recognition service and Lingsoft's speech recognition service. The system uses XML based Lightweight Dictation Model (LD-Model) from MobiDic to manage and model text counterparts for dictations.

The client application is used for recording dictations and for browsing and editing recognized text. Recordings and text counterparts are stored locally in the client and uploaded to the server. Server communication is done using Java SSL sockets and running them in threads in background. Therefore server communicant is transparent to users, as long as there are no network problems. After each recording, the audio is sent to server that redirects it for speech-to-text recognition service. After the recognition finishes the results are sent to the client and shown to the user. If recognition service provides n-bests for words in the results, they are represented by highlighting the words in red, as can be seen in Figure 1. The user can tap any word and type a replacement or choose an alternative word from the n-best list. While recording audio there is also the possibility to add punctuation marks into the text counterpart for the current time point. During audio recording an energy meter shows the current recording level and voice activity detection visualization is used to provide a simple view of the recorded audio. It is also possible to listen to parts of the audio by clicking on the bars on the view.

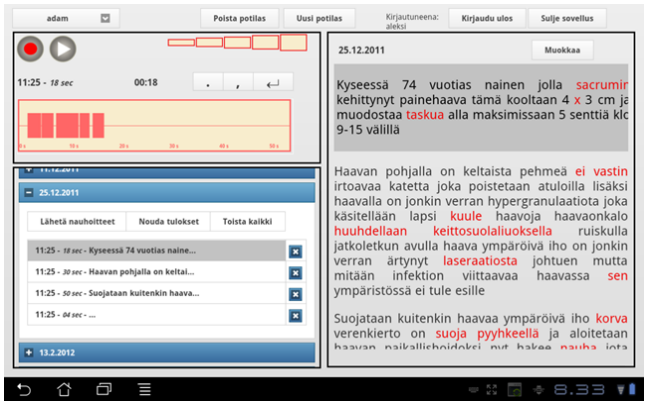

Figure 1. The graphical user interface of the mobile dictation application.

The client is an Android tablet application with a WebView-based user interface that uses JQuery Mobile framework. WebView contains HTML5 and JQuery Mobile elements and events, CSS3 style sheets and simple JavaScript runtime operations.

The server solution consists of five Java Standard Edition services and M-Files document management system running on Windows 2008 server. The Java services allow the client to upload audio and document metadata, which are stored and passed to speech recognition service. When the recognition finishes the results are exported into LD-Model. During this process, the server can pass the result to proof reading component, testing different n-best combinations and add new alternate suggestions to words based on proofing service suggestions. N-best results can also be sorted based on history information of users' previous corrections with tablet UI. After that the client will automatically download text counterpart for the audio. The server publishes the recognition result as a text document also into the M-Files document management system. Files in the M-Files system can be accessed with secured browser interface, but the general case is that user's files in PC are synchronized automatically over the Internet with files accessible by her profile in the M-Files, or M-Files is integrated into the patient data management system. The text counterpart, which can be modified in the tablet client, is kept synchronized with server backups and with M-Files system. Further, the M-Files system keeps the text counterpart synchronized with users connected to M-Files. Therefore it is possible for a user to edit text results in a tablet while another user, with given access to first user's files (e.g., a supervisor), sees the changes in the corresponding document with her own device that could be any other device such as PC laptop.

Two modifications for the system were done for the evaluation. Lingsoft's recognizer was exclusively used, because at the time for tests only that had a Finnish medical language model available for us.

While a medical language model was available, it was a generic one, based on doctors' dictations. Since our target users for the first evaluation were nurses specialized in wound care, the language of their dictations differs quite much from doctors' language. The most challenging difference is that there are many special products commonly used only in this field, and thus they were mostly missing from the language model. On our preliminary tests for recognizers with the medical language model and texts from the target user group, the word error rate average was varying between 28% and 50% depending on the user. Even though the nature of the errors was commonly a phrasing error or a letter missing from the end of the word making the context usually understandable, we considered there were still too many vital words for the scenario missing from the language model. Decrease in error rate and fixing the issue of missing words could be achieved with modern speech recognition techniques and engines with appropriate training material, but it was not possible to update the language model by the time of our test. In order to achieve the recognition level of a present day we ended up using a variation of Wizard-of-Oz technique.

The recognized text counterpart is partly corrected by a researcher before it is sent to participant's tablet application. The researcher makes the corrections with the tablet UI on her own tablet with separate privileges and then sends the text back to the server. Then it is sent to participant's tablet where it may be further edited as necessary. The wizard does not aim for fixing all the errors but filling the missing words and correcting significant substitution errors. The participants are not aware of corrections made by the researcher. They are only told that the speech is recognized into text on the Internet and the process takes some time. As a result to the WoZ technique, the time for recognition

progress will increase but the word error rate apparent to user will drop to acceptable level, thus allowing us to focus on the user experience aspects, while the language model is being improved.

The equipment for the evaluation was an Android tablet computer and a headset enabling recording. The integrated microphones in the tablets we tested did not achieve an acceptable level of audio quality for the recognition. We also implemented logging for the system in order to gather objective data and find possible user patterns and support the findings of subjective data. The logging is accurate enough to re-construct the whole use.

3. USER EVALUATION

We conducted a user evaluation in real context with real users in one of the university hospitals in Finland. Here, we present the user evaluation in detail.

3.1 Methodology

The methodology was selected and modified taking into account the three main factors of user experience: system, context and user [7]. The data collection was planned so that it would benefit both research fields, i.e., HCI and nursing science. The core of gathering user expectation and experience data is based on SUXES methodology [8], but experiences after the use were collected also with the System Usability Scale (SUS) [9]. In addition to the more subjective data, we gathered background information and log data to support the analysis and findings.

3.1.1 Background interview

Before the actual test phase, the participants were verbally interviewed with a structure consisting of almost 40 questions. They were asked basic questions, such as age and working experience, but the main focus was on their practices on dictating or making entries into the patient information system. They were asked how frequently they do either of these, what information about the patient they record, and what systems they use. The participants were also interviewed about their habits considering making the dictations or writing the entries, e.g., when do they make them (during the treatment situation or at the end of their work shift) and do they make notes for the entries. We were also interested of frequencies, needed time, and the easy and the hard things in making the entries or dictations. As background information, the participants' previous experience with tablet PCs and speech recognition was inquired as well. Further, they were asked about the potential of utilizing speech recognition in their work.

3.1.2 User Expectations and Experiences

We gathered subjective data from the participants utilizing SUXES [8] which is a method for gathering pre-usage expectations and post-usage experiences from users of an interactive system. In SUXES subjective opinions from the users are asked with a set of statements on properties or qualities of the system or, e.g., individual modality, and a seven-step scale ranging from low to high. Expectations before the usage are reported by giving two values for each statement: an *acceptable level*, i.e., the lowest acceptable level required for even using the system, and a *desired level*, meaning the highest level that can even be expected of the system or property. After the usage the users report their experiences giving only one value, *perceived level*, on exactly the same statements. The two expectation values, acceptable and desired levels, form a gap, where the experience value, perceived level, is expected to rank. The nine statements in the original form of the SUXES relate to speed, pleasantness, clarity, error-free use, error-free function, easiness to learn to use, naturalness, usefulness and future use. A statement can be structured, e.g., "*Using the application is fast*" and the users report their expectations/experiences by marking the levels the higher the faster they expect/experienced the application to be.

In order to suit the data collection for this case, we made some modifications to the original SUXES. For example, considering the great amount of time it takes to make the patient information system entries, in this context we wanted to gather user expectations and experiences not only on the dictation application, but also to compare the dictation application to the usually used entry practice of the participants. Thus, we asked the users' opinions on the following comparative statements in addition to the "original" SUXES statements: "Dictating with the application is 1) faster, 2) more pleasant, 3) more clear, 4) easier than with the entry practice I normally use"; and "5) I would rather make the entries with the dictation application than with the entry practice I used before." These statements were naturally included both in the expectation and experience questionnaires. The questionnaires were in electronic form and could be filled in using a typical web browser on a PC. The experience questionnaire included open questions in addition to the statements: the participants were asked how the dictation application changed their working practices, how speech recognition or the application could be developed, and they were provided with a chance to give free-form feedback.

Due to the multidisciplinary nature of the project, we gathered subjective experiences from the participants also with the System Usability Scale, SUS [9]. The SUS is originally designed to measure usability, but it has a strong subjective approach as the users themselves report the answers. Thus, the results gained by SUS can be considered as subjective user experiences of usability-related properties. In this article we will focus on the SUXES results, though.

3.2 Participants

In the first phase evaluation of the mobile dictation application we had two female nurses as participants. Both of them worked in a outpatient wound clinic: one (P2) of them worked there two days a week, and the other (P1) one day every two weeks. Participants' background information, work practices and earlier experience on tablet PCs and speech recognition can be seen in Table 1. This data was collected before the start of the pilot.

Table 1. Participants' background information and usual work practices.

	P1	**P2**
Age	30 years	36 years
Work experience in nursing/current unit	8/3 years	13/8 years
Do you dictate or write nursing entries?	Write.	Dictate.
How often do you dictate nursing entries?	Not at all.	Weekly.
How often do you write	Several times in	Weekly.

nursing entries?	a work shift.	
Do you make notes for the nursing entries?	Yes.	Yes.
How many patients do you treat in a work shift?	4–7	5–8
How much time dictating or writing nursing entries takes in a work shift?	About 80–100 minutes.	About 60 minutes.
In what kind of situations speech recognition might be useful in your work?	In making the nursing entries.	In making it faster and easier to dictate and see the text.
Could you dictate during the care situation while treating the patient?	Yes.	Yes.
How much do you have experience on speech recognition?	I've heard/read about it.	No experience at all.
How often do you use speech recognition (e.g., in a device or service)?	Not at all.	Not at all.
How much do you have earlier experience on using a tablet computer?	I've tried one a few times at most.	I've seen one.

3.3 Procedure

Before the pilot started, the participants were asked about their background information and work practices. The application was also introduced to the participants. The basic functionality was taught and they were able to ask questions concerning the application. After the introduction, the participants were asked to fill in their expectations as described earlier. Then, using the application the participants first dictated everything they would normally record directly to the patient information system. As mentioned earlier, Wizard-of-Oz approach was used and the human "wizard" checked and fixed the recognition results at this point. After the wizard had corrected the text, it was "published" and the original dictator, i.e., our participant, was able to see the recognized text in her tablet application. She was also able to edit the text if needed. Finally, she accessed the M-Files system with a web browser on a PC and copied the saved nursing text to be pasted into the patient information system. This was vital as our system was not communicating with the patient information system, and not missing any patient information was obviously our top priority.

After the pilot the participants filled in their experiences on both the SUXES and SUS questionnaires. The pilot lasted in total three months. During this time we gathered 30 dictations from participant 1 and 67 dictations from participant 2.

4. RESULTS AND DISCUSSION

User expectations and experiences on the application, i.e. the results on the "original" SUXES statements, are presented in Figure 2 (A). The participants had high expectations about the dictation application: the desired level is 6 or 7 on all statements. Despite the high hopes, almost all of these expectations were met. Not only did the participants feel the application was fast, pleasant, clear and natural to use, but they also felt it was easy to learn. When considering we are talking about introducing new technology in a working environment, usefulness and willingness to use the new technology again are probable the most important properties measured here. Our participants experienced the mobile dictation application to be highly useful and they would clearly like to use it again. It should be noted that experienced usefulness alone is not always enough: if the users have the option to choose whether to use a new or an old way of doing things, they most probably will choose the familiar and safe option if they do not have a subjective desire to choose the new way.

Practically the only negative experiences can be seen considering error-free functioning, which was in addition experienced differently by our participants. These negative, or modest, responses are rather well explained by the fact that there were technical problems with the wireless Internet connection during the evaluation. Due to strict regulations, our pilot usage was dependent on the hospital network connection, and unfortunately we were unable to address the network connection problems during the evaluation.

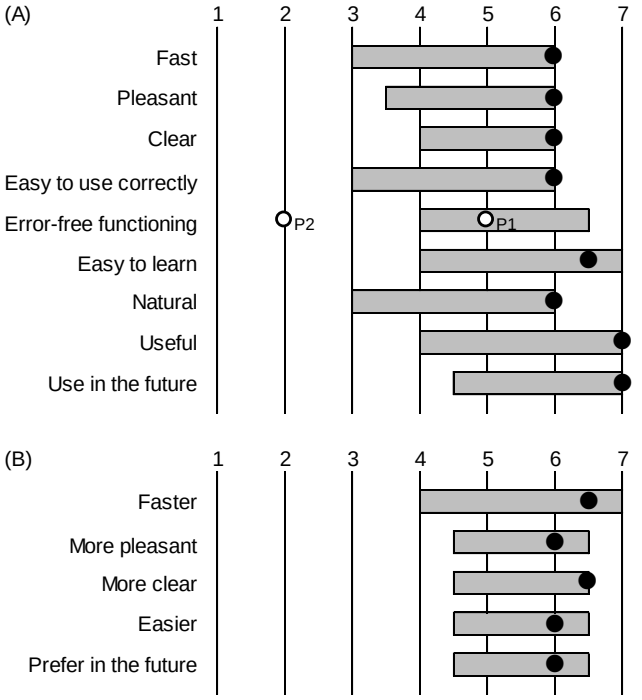

Figure 2. User expectations and experiences on the mobile dictation application (A), and compared to the normally used entry practice (B). Grey boxes represent the median expectations (acceptable–desired levels), and black circles represent the median experiences (perceived levels).

Results concerning the dictation application compared to the normally used entry practice can be seen in Figure 2 (B). It is obvious that the participants had high expectations towards the application from this point of view. In fact, their expectations were even higher than when judging the application alone. This suggests that in order for them to be willing to change their work routines, they would require the new approach to be clearly better. The experienced levels on the comparative statements are positively high, and even more so considering that our other participant (P1) was not even used to dictate as her normal daily work routine.

Further, open questions revealed that the participants did not find the headset interfering with the dictating. In fact, they were ready to use it daily if it was a prerequisite for using the application. By introducing speech recognition and dictation application they could now check the text at that moment, while before it took about a week before the text was available for the participant who normally dictated her nursing entries. Neither of the participants reported missing speech commands or buttons. When asking for development areas, the participants wished for a better recognition for compound words. The other participant (P2) also mentioned that the unreliability of the Internet connection took some unnecessary extra time when sending the files.

Obvious willingness to use our application in the future combined with other positive responses, shows a great potential for introducing such a system for Finnish healthcare – not only for dictation purposes, but also as a true option for writing the nursing entries. Be it these are experiences of only two users, they were professionals working in the field, and thus, the application shows a good starting point for further development.

5. CONCLUSIONS

We have presented a mobile dictation application with automatic speech recognition for healthcare. While a more accurate language model for nurses' purposes is being developed, we evaluated the application using a Wizard-of-Oz scenario: medical language model based on doctors' dictations was used for the speech recognition, the results were then finished by a researcher, and finally, sent to the participant's tablet application. The user experiences received from the nurse participants indicate that introducing such an application for Finnish healthcare is warmly welcome: the nurses get a transcript of their dictations almost immediately as opposed to at worst a week, they now have to wait for the text counterpart. Our results show true potential for the approach, thus making our further development and evaluation plans towards a pleasant, useful, and fully automated dictation-to-text process very relevant for Finnish healthcare.

6. ACKNOWLEDGEMENTS

This work was supported by the Finnish Funding Agency for Technology and Innovation (TEKES) in the project "Mobile and Ubiquitous Dictation and Communication Application for Medical Purposes" (grant 40056/11). We thank Lingsoft and M-Files, and other project partners, for collaboration.

7. REFERENCES

[1] Parente, R., Kock, N., and Sonsini, J., "An analysis of the implementation and impact of speech-recognition technology in the healthcare sector". Perspectives in Health Information Management, 1(5), 2004.

[2] Koivikko, M., Kauppinen, T., and Ahovuo, J., "Improvement of report workflow and prodictivity using speech recognition – a follow-up study". Journal of Digital Imaging, 21(4), 378–382, 2008.

[3] Devine, E., Gaehde, S., and Curtis, A., "Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports". Journal of the American Medical Informatics Association, 7(5), 462–468, 2000.

[4] Borowitz, S., "Computer-based speech recognition as an alternative to medical transcription". Journal of the American Medical Informatics Association, 8(1), 101–102, 2001.

[5] Mohr, D., Turner, D., Pond, G., Kamath, J., De Vos, C., and Carpenter, P., "Speech recognition as a transcription aid: a randomized comparison with standard transcription". Journal of the American Medical Informatics Association, 10(1), 85–93, 2003.

[6] Turunen M., Melto A., Kainulainen A., and Hakulinen J., "Mobidic – A Mobile dictation and notetaking application". In Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech), 500–503, 2008.

[7] Hassenzahl, M., and Tractinsky, N., "User experience – a research agenda". Behaviour & Information Technology, 25(2), 91–97, 2006.

[8] Turunen M., Hakulinen J., Melto A., Heimonen T., Laivo T., and Hella J., "SUXES – User Experience Evaluation Method for Spoken and Multimodal Interaction". In Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech), 2567–2570, 2009.

[9] Brooke, J., "SUS – A quick and dirty usability scale". In P. W. Jordan, B. Thomas, B. A. Weerdmeester, and A. L. McClelland (Eds.), Usability Evaluation in Industry. London: Taylor and Francis, 1996.