On speech recognition during anaesthesia

Alexandre Alapetite

NOVEMBER 2007 ROSKILDE UNIVERSITY COMPUTER SCIENCE RESEARCH REPORT #119

Alexandre Alapetite

Computer Science
Department of Communication,
Business and Information Technologies
Roskilde University
P. O. Box 260
DK–4000 Roskilde
Denmark

Telephone: +45 4674 3839
Telefax: +45 4674 3072
Internet: http://www.ruc.dk/dat_en/
E-mail: datalogi@ruc.dk

ISSN 0109–9779

Research reports are available electronically from:

http://www.ruc.dk/dat_en/research/reports/

On speech recognition during anaesthesia

Alexandre Alapetite

Revision 2007-10-30

On speech recognition during anaesthesia

Alexandre Alapetite

PhD thesis, July 2007

Computer Science, Roskilde University, Denmark

Systems Analysis, Risø National Laboratory, Technical University of Denmark

ADVISES, European Research Training Network

Abstract in English

This PhD thesis in human-computer interfaces (HCI, informatics) studies the case of the anaesthesia record used during medical operations and the possibility to supplement it with speech recognition facilities.

Problems and limitations have been identified with the traditional paper-based anaesthesia record, but also with newer electronic versions, in particular ergonomic issues and the fact that anaesthesiologists tend to postpone the registration of the medications and other events during busy periods of anaesthesia, which in turn may lead to gaps and inaccuracies in the anaesthesia record.

The thesis first studies the role and the importance of the anaesthesia record as a work tool during operations. Related work procedures are also described in detail. Some small-scale surveys are conducted, which corroborate the observations mentioned above.

Supplementing the electronic anaesthesia record interface with speech input facilities is proposed as one possible solution to a part of the problem. A discussion paper made with a socio-ergonomist describes some of the short and long-term consequences if such an idea is to be deployed.

The thesis then investigates the possibilities and technical limitations of the most widely used speech recognition system in Danish for medical applications. Of particular interest is the deleterious effect of various background noises found in medical operation theatres. While loud noises in the operating room can have a predominant negative effect, recognition rates for common noises are found to be only slightly below performances obtained in an office environment. Other factors have a major impact as well, such as the words to be recognised, participants, the type of speech recognition system (natural or constrained language) and the type of microphone. Finally, a proposed redundant architecture succeeds in improving the reliability of the recognitions.

After that, a prototype of electronic anaesthesia record interface with speech input facilities is developed on the basis of the knowledge gained at the previous steps as well as interviews with some anaesthesiologists.

The next phase is based on full-scale anaesthesia simulations involving the prototype, to compare it with the traditional touch-screen and keyboard interface. Inspired from the mathematical queuing theory, a special metric for characterizing differences in mental workload is developed to compare the two interfaces. Results show that the speech interface leads to much shorter registration delays and to a greater accuracy of the information than the traditional electronic interface.

The simulation-based experiments also permitted the testing of some speech input strategies chosen for the prototype (hands-free vocal interface activated by a keyword; combination of command and free text modes), which were successful, even with the ambient noise. Speaking to the system while working appeared feasible, although improvements in speech recognition technologies are still necessary.

The above experiments form the main results of the thesis. They are followed up by secondary investigations.

An opportunity is taken to study *via* questionnaires and other indicators the deployment, acceptance and success of a speech recognition system – sharing technological similarities with the above-mentioned prototype – used to produce patient records in a Danish hospital. Physician satisfaction with the use of the system is modest, yielding *a posteriori* an approximately even balance between those in favour of, and those against the introduction of speech recognition to transcribe medical records. One of the main reasons for users' dissatisfaction is the new work procedure introduced simultaneously with the speech recognition technology, which requires physicians to spend more time on producing the records.

In order to get more objective data on the effect of introducing this speech recognition system on the quality of the medical records, a blinded comparison is done between former and new work procedures. The results show that records produced with the new work procedures involving possible use of speech recognition contain more errors than the ones produced with the former method where a secretary is in charge of the transcription. However, the difference between speech recognition and secretary based transcription is relatively small in terms of number of transcription errors, and does not apply to records that follow a fixed and recurrent pattern. These results may therefore not be construed as showing that speech recognition does not bring advantages when considering the gains, *e.g.* in total turnaround time.

The conclusion is that speech recognition is a very interesting modality that should be used when appropriate and only for tasks for which it is efficient when compared to other alternatives.

Résumé en français (abstract in French)

Cette thèse en interaction homme-machine (IHM, informatique) étudie le cas du journal d'anesthésie utilisé durant des opérations médicales, et le bien fondé d'enrichir son interface avec des fonctions de reconnaissance vocale.

Des problèmes et des limitations ont été identifiés dans le cas de la version papier traditionnelle du journal d'anesthésie, mais aussi avec les nouvelles versions électroniques, en particulier à propos de considérations ergonomiques et le fait que les anesthésiologistes ont tendance, durant les périodes chargées, à délayer la saisie des médicaments et autres événements, ce qui entraîne des manques et des inexactitudes dans les journaux d'anesthésie.

La thèse étudie en premier lieu le rôle et l'importance du journal d'anesthésie en tant qu'outil de travail pendant les opérations. Les procédures de travail concernées sont décrites en détail. Des enquêtes à petite échelle sont menées et corroborent les observations mentionnées ci-dessus.

L'enrichissement du journal d'anesthésie avec des fonctions de reconnaissance vocale est proposé comme une solution à une partie du problème. Un article de réflexion écrit avec un socio-ergonome décrit les conséquences envisagées à court et long terme si une telle idée venait à être déployée.

La thèse examine ensuite les possibilités et les limitations techniques du système le plus utilisé de reconnaissance vocale en danois pour les applications médicales. Un intérêt particulier est accordé aux effets délétères des différents bruits de fond d'une salle d'opération. Tandis que les bruits forts dans la salle d'opération peuvent avoir un effet prédominant, les taux de reconnaissance en présence de fonds sonores communs sont mesurés comme seulement légèrement inférieurs aux performances obtenues dans un environnement de bureau. D'autres facteurs ont eux aussi un impact majeur, comme les mots à reconnaitre, les participants, le type de reconnaissance vocale (texte libre ou contrôlé) ainsi que le type de microphone. Enfin, une architecture redondante est proposée et réussit à améliorer la fiabilité des reconnaissances.

Après cela, un prototype d'interface pour un journal d'anesthésie électronique avec reconnaissance vocale est développé en se basant sur les connaissances acquises aux étapes précédentes, et sur des entrevues avec des médecins anesthésistes.

L'étape suivante se compose d'expériences dans un simulateur d'anesthésie complet, destinées à comparer le prototype avec l'interface électronique traditionnelle constituée d'un écran tactile et d'un clavier. En s'inspirant de la théorie mathématique des files d'attente, une métrique spéciale est développée pour caractériser les différences de charge mentale de travail, et pour permettre *in fine* de comparer les

deux interfaces. Les résultats montrent que l'interface vocale permet des délais d'enregistrement bien plus courts et une meilleure précision de l'information en comparaison avec l'interface électronique traditionnelle.

Les expériences de simulation permettent aussi de tester des stratégies d'entrée vocale choisies pour le prototype (interface vocale main-libres activée par un mot clef ; combinaison de texte libre et contrôlé), qui ont été satisfaisantes, même en présence de bruit de fond. Le fait de parler au système est apparu envisageable, malgré des progrès nécessaires de la part de la technologie de reconnaissance vocale.

Les expériences présentées ci-dessus forment les résultats principaux de la thèse. Ils sont suivis d'investigations secondaires.

Via des questionnaires et autres indicateurs, une opportunité est prise d'étudier le déploiement, l'acceptation et le succès d'un système de reconnaissance vocale – partageant des caractéristiques communes avec le prototype susmentionné – utilisé pour produire des journaux médicaux de patients dans un hôpital danois. La satisfaction des médecins envers le système est modeste, avec une répartition *a posteriori* approximativement égale entre ceux en faveur et ceux opposés au déploiement de la reconnaissance vocale pour transcrire les journaux médicaux. Une des raisons principales expliquant l'insatisfaction des médecins est l'introduction de nouvelles procédures de travail simultanément au déploiement de la technologie de reconnaissance vocale, et qui requièrent un plus grand investissement en temps de la part des médecins dans la production des journaux.

Afin de collecter des données plus objectives sur l'effet de l'introduction du système de reconnaissance vocale sur la qualité des journaux médicaux, une comparaison en aveugle est effectuée entre les anciennes et les nouvelles procédures de travail. Les résultats montrent que les journaux produits avec la nouvelle méthode, impliquant l'utilisation possible de la reconnaissance vocale, contiennent plus d'erreurs que ceux produits avec l'ancienne méthode où une secrétaire est en charge de la transcription. Cependant, la différence négative entre la reconnaissance vocale et les secrétaires est relativement faible en terme de nombre d'erreurs de transcription et ne concerne pas les journaux qui suivent un modèle fixe et utilisent des phrases routinières. Ces résultats ne doivent en conséquence pas être interprétés comme montrant que la reconnaissance vocale ne fournit pas d'avantages, surtout en considérant les autres bénéfices, comme par exemple dans la réduction des délais totaux de production des journaux.

La conclusion est que la reconnaissance vocale est une modalité très intéressante qui devrait être utilisée dès lors que cela est approprié mais seulement dans le cas de tâches pour lesquelles cette technologie est efficace face à d'autres alternatives.

Resumé på dansk (abstract in Danish[1])

Denne ph.d.-afhandling i menneske-maskine interaktion (MMI, informatik) undersøger brugen af anæstesijournaler under kirurgiske operationer samt muligheden for at supplere disse med en talegenkendelsesfunktion til journalregistrering.

Problemstilling og -afgrænsning er blevet identificeret og fastlagt på baggrund af eksisterende papirbaserede anæstesijournaler såvel som gennem nyere elektroniske systemer, herunder særligt fastlæggelsen af de ergonomiske forhold og det faktum, at anæstesilæger under travle perioder af anæstesiforløbet har en tendens til at udsætte registreringen af medicinering og andre hændelser, hvilket kan føre til mangler og unøjagtigheder i anæstesijournalen.

Afhandlingen indledes med en undersøgelse af anæstesijournalens rolle og betydning som arbejdsredskab under operationer. Arbejdsgange, der relaterer sig til operationer, er også behandlet. Enkelte mindre analyser er blevet udført og bekræfter rigtigheden af de ovenfor beskrevne observationer.

Det foreslås at supplere grænsefladen til den elektroniske anæstesijournal med en talegenkendelsesfunktion som en mulig løsning af dele af problemerne. Overvejelser udarbejdet i samarbejde med en ergonom-sociolog beskriver de kortsigtede og langsigtede konsekvenser, hvis en talegenkendelsesfunktion indføres.

Afhandlingen undersøger herefter mulighederne og de tekniske begrænsninger i det mest udbredte talegenkendelsessystem på dansk til medicinske applikationer. Af særlig relevans er konsekvensen af forskellig baggrundsstøj på operationsstuer. Mens meget støj på operationsstuen kan være af afgørende negativ betydning for graden af genkendelse, er genkendelsesgraden ved almindelig støj kun en smule lavere end den, der opnås i et kontormiljø. Andre faktorer, såsom de ord, der skal genkendes, deltagerne, typen af talegenkendelsessystem (fri eller begrænset tale) samt hvilken mikrofontype, der benyttes, har ligeledes stor betydning for genkendelsesgraden. Endelig medfører en foreslået redundant opbygning af systemets arkitektur en større pålidelighed i talegenkendelsen.

Herefter er en prototype af en grænseflade til en elektronisk anæstesijournal med en talegenkendelsesfunktion blevet udviklet på baggrund af den viden, der er opnået i de tidligere faser og gennem interview med anæstesilæger.

I den efterfølgende fase er foretaget fuld-skala simulationer i anæstesiologi, hvori prototypen er anvendt til sammenligning med de eksisterende touch-screen og

[1] Grateful acknowledgements to my fiancée Rikke for her help in doing the Danish translation of this summary

tastatur-grænseflader. Inspireret af den matematiske køteori er en særlig metrik til at karakterisere den psykiske arbejdsbelastning ved brugen af de to forskellige grænseflader blevet udviklet for at kunne sammenligne disse. Resultaterne viser, at en stemmegrænseflade medfører en væsentlig kortere tid mellem en hændelse indtræffer og registreres, samtidig med at informationerne bliver mere nøjagtige sammenlignet med den eksisterende elektroniske grænseflade.

Disse simulationsbaserede forsøg giver endvidere mulighed for at teste de strategier der er valgt for talegenkendelsesfunktionen til prototypen, herunder håndfri stemmestyret grænseflade aktiveret af et tastatur og kombinationen af kommando- og fri tale-indstillinger, hvilket fremstod som vellykket, selv med omgivende støj. Det viste sig endvidere muligt at tale til systemet under arbejdet, selvom forbedringer i talegenkendelsesteknologien stadig er nødvendige.

De ovennævnte simulationsforsøg danner grundlaget for hovedresultaterne i denne afhandling. Sekundære undersøgelser er derefter foretaget som en opfølgning.

Det har været muligt gennem spørgeskemaer og andre indikatorer at analysere brugeropfattelsen af indførelsen og brugen af et talegenkendelsessystem – der er teknisk sammenligneligt med den ovenfor nævnte prototype – anvendt til indføring af notater i patientjournaler på et dansk sygehus. Lægernes tilfredshed med brugen af systemet er beskeden og viser en næsten ligelig fordeling mellem de, der efterfølgende var for og de, der var imod indførelsen af stemmegenkendelse til transskribering. En af hovedårsagerne til brugernes utilfredshed er begrundet i nye arbejdsgange – introduceret på samme tid som talegenkendelsesteknologien – der kræver, at lægerne bruger længere tid på at udarbejde journalerne.

For at opnå mere objektive resultater af effekten af indførelsen af dette talegenkendelsessystem og kvaliteten af patientjournalerne er der foretaget en blindet vurdering til sammenligning af de tidligere og nye arbejdsgange ved registrering af notater. Resultaterne viser, at journaler udarbejdet med de nye arbejdsgange med muligheden for brug af en talegenkendelsesfunktion indeholder flere fejl end de, der er udarbejdet ved den tidligere anvendte metode, hvor en sekretær er ansvarlig for transskriberingen. Imidlertid er forskellene mellem talegenkendelses- og sekretærbaseret transskribering relativt små for så vidt angår antallet af transskriptionsfejl, og der er ingen forskelle for journaler, som følger et fast og stadigt tilbagevendende mønster. Resultaterne må således ikke tages som et udtryk for, at talegenkendelse ikke medfører fordele, særligt henset til gevinsterne som talegenkendelse indebærer, herunder eksempelvis i forhold til det samlede tidsforbrug ved udarbejdelsen af en journal.

Konklusionen er, at talegenkendelse er en meget brugbar og relevant grænseflade, som bør bruges når situationen er egnet hertil men kun til opgaver, hvor den er effektiv sammenlignet med andre alternativer.

Acknowledgements

This work was supported by the Fifth Framework Programme of the European community[2], within the ADVISES[3] Research Training Network about "Analysis Design and Validation of Interactive Safety-critical and Error-tolerant Systems". This network (2003 – 2006) was started by Professor Chris Johnson[4] (University of Glasgow, Scotland, United Kingdom) and Professor Philippe Palanque[5] (Université Paul Sabatier, Toulouse III, France). The main research objective of this network was to provide a multi-disciplinary research training that could combat the impact of human error during the design, operation and management of safety-critical, interactive systems. Another aim of this research training network was to strengthen international contacts between research entities; the so-called "young researchers" were therefore offered scholarship in foreign EU countries. The ADVISES network formed a friendly and yet scientifically inspiring group of people.

I was recruited as a French informatics engineer[6] in November 2003 by the Danish node of this network, Risø National Laboratory, organised by senior scientist Hans H. K. Andersen[7], PhD. I spent most of my PhD in the Systems Analysis Department, Research Programme Safety, Reliability and Human Factors[8], where my main supervisor was senior scientist Henning Boje Andersen[9]. Prior to starting my PhD, I participated from April 2003 in the EU project Safesound in the same department, developing requirements to a speech recognition system for airliner cockpits, with senior scientist Steen Weber[10], PhD. This was a precious experience for my PhD.

---

[2] [http://ec.europa.eu/research/fp5.html]

[3] [http://www.cs.york.ac.uk/hci/ADVISES/]

[4] [http://www.dcs.gla.ac.uk/~johnson/]

[5] [http://liihs.irit.fr/palanque/]

[6] [http://alexandre.alapetite.net/cv/]
Master of engineering in mathematics and informatics from Université des sciences et techniques du Languedoc, Montpellier II, France, 2002; Master in artificial intelligence, pattern recognition, robotics from Université Paul Sabatier, Toulouse III, France, 2003.

[7] [http://www.risoe.dk/sys/Staff/SPM/hakr.htm]

[8] [http://www.risoe.dk/sys/spm/]

[9] [http://www.risoe.dk/sys/Staff/SPM/hebq.htm]

[10] [http://www.risoe.dk/sys/Staff/SPM/stwe.htm]

During the whole PhD period, I was a student at Roskilde University (Denmark), in the Department of Computer Science[11], where my supervisor was Associate Professor Morten Hertzum[12].

I would like to thank the European commission, RUC and Risø for the ideal research environment they have offered me, and my supervisors for their close and valuable monitoring.

This project took advantage of many fruitful collaborations, such as with the members of the ADVISES network, the Danish Institute for Medical Simulation[13] at Herlev Hospital (in particular doctors Doris Østergaard, Ann Moller, Nini Vallebo), Køge Hospital[14] (in particular doctor Viggo Stryger), Vejle Hospital[15] (in particular Aase Andreasen), and the industrial partner for speech recognition, the Danish company Max Manus[16] (in particular Peter Damm).

Further credits are given after each of the papers collected in the thesis.

More personally, I would like to thank primarily my fiancée Rikke – and hopefully future wife, as we will get married in August 2007 – for her constant support. This is without forgetting friends and family who contributed to my happiness and sometimes provided inspiration for my work (often through the magical Internet), to name only a few, Guillaume Fraysse (Université de Montpellier II, France), Michel Jaumes, Stephanie Ropenus (Risø), Jens Skåning (formerly at Risø), my brothers Didier & Frédéric, and my father Xavier. To my mother Alice: you remain in my heart…

[11] [http://ruc.dk/dat/]

[12] [http://akira.ruc.dk/~mhz/]

[13] [http://herlevsimulator.dk]

[14] [http://rs-roskilde.dk/ras/koge/koge.asp]

[15] [http://vejlesygehus.dk]

[16] [http://maxmanus.dk]

Contents

General introduction

1 Presentation of the topic

This PhD thesis is rooted in the HCI (human-computer interaction) field of informatics, with some considerations about ergonomics and other human factors. Test cases are conducted in the medical domain.

Common human-computer interfaces are screens, speakers (for output); keyboards, mouses, microphones, webcams (for input); or touch-screens (input/output).

This thesis deals with multimodal interfaces, which are human-computer interfaces with a combination of input and output possibilities, in safety critical work environments. More specifically, the focus is set on supplementing existing electronic anaesthesia records with some speech input facilities during the operations. Notably, a prototype was developed to test various hypotheses.

Anaesthesia is the process of inducing, controlling and reverting the loss of some perception – and of consciousness during full anaesthesia – to allow patients to receive surgery and other medical operations that otherwise would be painful or traumatising, but also to ensure sufficient muscular relaxation for those medical acts to be possible. An anaesthesia record is a document for the reporting, either manually or automatically, of most of the vital signs during anaesthesia (*e.g.* pulse, oxidation), medications and gases, together with the most important diagnosis, observations and various events on a time line.

1.1 Brief history of the topic

In collaboration with a hospital, the first project proposals were supposed to address limitations and problems with the paper-based version of the anaesthesia records. The main idea was to design an electronic interface with touch-screen and basic speech input in Danish. It was only after a few months of literature survey and the beginning of some interviews in other hospitals that I discovered that electronic versions were already in use in about half of the hospitals in Denmark. One semester after the beginning, the project mostly dropped the paper-based issues to concentrate on building upon the electronic version of existing anaesthesia record systems, since the latter were already tackling many of the concerns regarding paper-based systems.

Similarly, although some academic speech recognition engines in Danish (from Aalborg University) were known at the beginning of the project, it is only a few months later that I discovered an industrial speech recognition engine in Danish that freshly hit the market. The project consequently dropped planned efforts on prototyping simple speech input in Danish to favour a partnership with the company providing the commercial system.

2 Thesis outline

This thesis aggregates some of the articles published during the two years of my PhD studies dedicated to this topic (3 years in total from October 2003 to December 2006, minus one semester of studies and another semester of duties). The articles are included in their chronological order, arranged in such a way that they offer a natural and logical progression with minimal overlap, reinforced by additional transition chapters.

The first part introduces the main notions, the topic, the rationale, the problem, the research question and the envisaged solutions (partially based on [Alapetite 2005.a][1]).

The following paper [Alapetite & Gauthereau 2005][2] aims at preparing the work, based on an extensive literature review, interviews and direct observations. Written in collaboration with a socio-ergonomist from the ADVISES network, we carefully studied the work practices in anaesthesia, such as the use of the anaesthesia record. A few surveys were conducted to verify the reality of the problems. We then envisaged the possible consequences of introducing a speech recognition interface in this theatre. Finally, we sought comments from experts by presenting this work at a conference.

Thereafter, it was deemed necessary to study experimentally in a laboratory the possibilities and limitations of the available speech recognition technology in Danish [Alapetite 2006][3] that was to be used in the next prototyping phase. Of particular concern was the background noise found in the operation room and its direct effects on speech recognition by altering the audio channel, and indirect effects by affecting users. Strategies to cope with background noise were also tested at this stage.

Based on the analysis of the laboratory experiments and lessons learned from previous considerations and literature survey, a prototype of speech recognition interface for an electronic anaesthesia record was developed. The graphic user interface is a mimic of

---

[1] [Alapetite 2005.a] Alexandre Alapetite. Voice recognition in multimodal systems: the case of anaesthesia patient journal. *In: Proceedings of the first ADVISES Young Researchers Workshop. Hans H.K. Andersen, Asmatullah Nayebkheil (eds.), Risø National Laboratory (DK), Systems Analysis Department. Risø-R-1516(EN), 2005, pp. 5-9.*

[2] [Alapetite & Gauthereau 2005] Alexandre Alapetite & Vincent Gauthereau. Introducing vocal modality into electronic anaesthesia record systems: possible effects on work practices in the operating room. *Proceedings of EACE'2005 (Annual Conference of the European Association of Cognitive Ergonomics) 29 September - 1 October 2005, Chania, Crete, Greece; section II on Research and applications in the medical domain, 189–196. ACM International Conference Proceeding Series, vol. 132. University of Athens, 197-204,* ISBN:9-60254-656-5.

[3] [Alapetite 2006] Alexandre Alapetite. Impact of noise and other factors on speech recognition in anaesthesia. *International Journal of Medical Informatics, available online December 2006*. doi:10.1016/j.ijmedinf.2006.11.007

an existing anaesthesia workstation in use in one of our partner hospital. Subsequent interviews with anaesthesia physicians were carried out to reach a working prototype.

Once the prototype had been developed, it was possible to undertake experiments [Alapetite 2007][4]. In order to ensure control and reproducibility, these experiments were conducted in a full-scale anaesthesia simulator with real anaesthesia teams. The methods involved and the analysis of those experiments form the main results of the thesis.

A few questions were naturally raised by this prototype experiment, for instance regarding the possible deployment of such as system. Although it was not possible to envisage larger scale experiments given the time and budget affected to the project, there was a chance to study – *via* questionnaires and other indicators – the deployment, acceptance and success of a speech recognition system (sharing technological similarities with the above mentioned prototype) used to produce patient records in another hospital [Alapetite, Andersen, Hertzum 2007][5].

Given the mitigated acceptance of the speech recognition system in the later hospital, there were some doubts that the physicians responding to our survey were not entirely objective in estimating the impact of the new work procedure involving speech recognition on the quality of the produced medical records. Therefore, a dedicated quality survey was conducted [Andersen, Alapetite *et al.* 2007][6], comparing samples produced with the traditional system using Dictaphones and transcriptions by secretaries, and samples produced with the new work procedure where the physicians are producing the documents themselves directly on a computer, with the possible help of speech recognition.

Finally, a general conclusion summarises the results and highlights the main findings and recommendations of the thesis.

---

[4] [Alapetite 2007] Alexandre Alapetite. Speech recognition for the anaesthesia record during crisis scenarios. *International Journal of Medical Informatics, accepted August 2007.* doi:10.1016/j.ijmedinf.2007.08.007

[5] [Alapetite, Andersen, Hertzum 2007] Alexandre Alapetite, Henning Boje Andersen, Morten Hertzum. Acceptance of Speech Recognition by Physicians: A Survey of Expectations, Experiences, and Social Influence. *Submitted to the International Journal of Human-Computer Studies, 2007.*

[6] [Andersen, Alapetite *et al.* 2007] Henning Boje Andersen, Alexandre Alapetite, Peter Ivan Andersen, Aase Andreasen, Per Hølmer, Stig Jørring, Claus Varnum. Blinded comparison of quality of medical records produced with speech recognition or traditional dictation and transcription. *To be submitted, 2007.*

3 Publications not included

During this PhD, about a semester of work in total is expected from the student for his institution(s), namely ADVISES European training network, Risø National Laboratory and Roskilde University. While most of it is not relevant for this PhD thesis and therefore not included in the body of the thesis, some selected publications are cited in this section to provide the reader with the scientific background knowledge and experience that have been gained during the PhD period. Undeniably, this has contributed to the achievement of this thesis.

As part of my work for the TRENDS European project (The Resource Network facilitating HSEQ "health, safety, environmental and quality" Development for a Sustainable Energy Industry), I developed an Open Source PHP migratory library[7] for PHP4/domxml to work under the newer PHP5/dom, which is used in at least a hundred different products at the time of the writing and was the topic of an invited publication [Alapetite 2004][8].

As a member of the COGAIN European network of excellence on Communication by Gaze Interaction, I have made a publication [Alapetite 2005.b][9] on the accessibility of Web documents.

The result of my participation in the SEE European project on Sight Effectiveness Enhancement, regarding infrared vision experiments in aircraft cockpits, is published in [Andersen & Alapetite 2006][10].

---

[7] [http://alexandre.alapetite.net/doc-alex/domxml-php4-php5/]

[8] [Alapetite 2004] Alexandre Alapetite. XML en PHP5 avec la bibliothèque interne DOM (XML in PHP5 with the DOM internal library). *Direction|PHP, special issue 1 Tout sur PHP5 (All about PHP5), September 2004, Nexen Services SA (France)*. ISSN:1765-2634. http://www.directionphp.biz/a_la_une.php?mois=2004-h1

[9] [Alapetite 2005.b] Alexandre Alapetite. Content accessibility of Web documents: Overview of concepts and needed standards. *In: COGAIN (European Network of Excellence,* http://www.cogain.org*) deliverable D6.1 "State of the art report of evaluation methodology", pages 28-34, September 2005. Long version in Risø-R-1576(EN),* ISBN:87-550-3546-9*, Risø National Laboratory, October 2006*.

[10] [Andersen & Alapetite 2006] Henning Boje Andersen & Alexandre Alapetite. SEE – Sight Effectiveness Enhancement, Results of the Aeronautical Evaluation. *SEE (European project) deliverable D6.3, December 2005. Risø-R-1573(EN), Risø National Laboratory, November 2006*. ISBN:87-550-3541-8

Finally, as part of the compulsory collaboration between the various nodes of ADVISES, a joint publication [Dokas & Alapetite 2006][11] was written with the University of Paderborn (Germany), mutualising our respective experience on experts systems, and human computer interfaces with a focus on Web standards and accessibility.

4 Note about digital references

When available, the scientific references in this thesis provide a DOI[12] (Digital Object Identifier System) aimed to easily locate a given article in a persistent manner. For instance, an article with a DOI such as `doi:10.1093/jhered/esh074` can be retrieved at the following URL [`http://dx.doi.org/10.1093/jhered/esh074`].

Hyperlinks to Web pages are most of the time provided to the original location of the resource (without the superfluous `www` subdomain prefix whenever possible), even if the resource had already been removed at the time this thesis was submitted.

To see an archived version of a given Web page, the "Internet Archive Wayback Machine"[13] may be used. For instance, if my Web page at Risø is removed [`http://www.risoe.dk/sys/Staff/SPM/alal.htm`], an archived copy can be retrieved by appending `http://web.archive.org/` in front of the URL, giving [`http://web.archive.org/http://www.risoe.dk/sys/Staff/SPM/alal.htm`].
An explicit link to the archived version is sometimes provided when there is a need to point to a precise version rather than to the last available.

[11] [Dokas & Alapetite 2006] Ioannis Dokas & Alexandre Alapetite. A view on the Web engineering nature of Web based expert systems. *Poster paper in the proceedings of ICSOFT'2006, the 1st International Conference on Software and Data Technologies, 11-14 September 2006, Setubal, Portugal, pages 280-283*. A development process meta-model for Web based expert systems: the Web engineering point of view. *Long version in Risø-R-1570(EN),* `ISBN:87-550-3536-1`*, Risø National Laboratory, October 2006*.

[12] [`http://dx.doi.org`]

[13] [`http://webarchive.org`]

Speech recognition in multimodal systems: the case of the anaesthesia patient record[1]

1 General background

1.1 About speech recognition

Automatic recognition of human speech started in the 1950's. One of the first serious attempts is from [Davis *et al.* 1952] with a circuitry capable of recognising isolated digits in English after being tuned to one given individual's voice. Over the years, the size of the vocabulary has increased, together with recognition accuracy. Then the need to separate each word has disappeared, thus allowing continuous speech. Although constrained language (based on a grammar describing the possible sentences) is still extensively used nowadays in command & control, teleservice applications, or when there is a need to get very high recognition rates, the more challenging natural language recognition (free text) tends to supplant the former mode in many areas. Free text mode, commercially available since the beginning of the 1990's, offers possibilities not achievable in command mode, mainly automatic transcription of dictated text. Natural language recognition typically requires much larger dictionaries (more vocabulary) and more complex high-level processing.

Speech recognition engines are based on various layers of processing [Zafar *et al.* 1999]. They typically have a signal-processing layer to discretize and prepare the signal (*cf.* Figure 1), followed by an acoustical model to split the audio signal into a probabilized sequence of phonemes.

A lexical model is then required to make a probabilized matching between phonemes and real words in a given natural language. At this layer, many issues remain unsolved, such as homonyms, in particular the ones that are homophone but not homograph, together with oronyms (two distinct sequences of words with a similar pronunciation):

"*mint spy*" or "*mince pie*"?

"*If two witches would watch two watches, which witch would watch which watch?*"

[1] This section uses some material from the following workshop article: Alexandre Alapetite. Voice recognition in multimodal systems: the case of anaesthesia patient journal. *In: Proceedings of the first ADVISES Young Researchers Workshop. Hans H.K. Andersen, Asmatullah Nayebkheil (eds.), Risø National Laboratory (DK), Systems Analysis Department. Risø-R-1516(EN), 2005, pp. 5-9.*

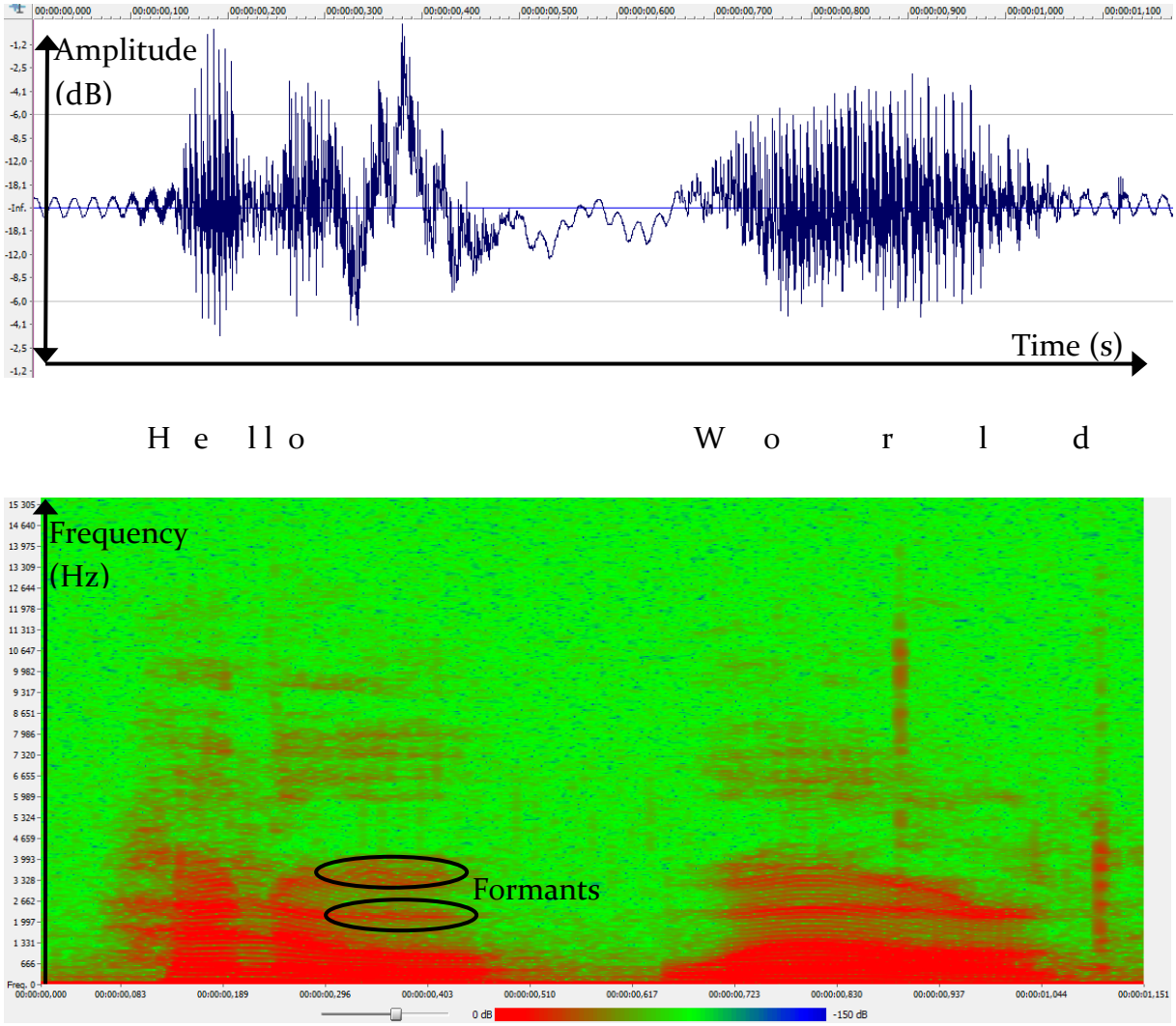


Figure 1: Wave signal and Spectrum analysis (44.1 KHz, Hanning window 1024) for a speech sample of "Hello World", with ellipses around examples of formants for the 'O' vowel.

To address those issues, a statistical language model provides probabilities of transitions between words. Some speech recognition engines may include higher level processing including grammar rules of the targeted natural language, and other types of syntaxic or semantic knowledge. The final decision is a complex combination of the probabilities of each level of processing.

Most of these layers can be trained with a corpus of data, and can learn from the user himself/herself, but the tendency is to provide speech recognition systems that work out of the box with less and less training required from the user. Speaker-independent speech recognition (*i.e.* without the need to train the system to recognise each user's voice) has been available for a while for constrained language, and is spreading in natural language recognition as well.

Although technologically obsolete, there has been a regain of attention towards simpler and lighter speech recognition engines over the last few years, as many electronic devices such as mobile phones and GPS navigation systems in cars offer basic speech recognition facilities.

Speech recognition is now available for many natural languages. In the case of Danish, which is a relatively small language spoken by about 5.5M people mainly in Denmark, the first commercial system appeared in 2000 with a language pack for Philips SpeechMagic[2], followed in 2001 by a product from Nuance[3] [Dybkjær & Dybkjær 2002].

Speech recognition is still under a very active development, both in academic research and in industrial solutions.

In this thesis, when referring to "speech recognition", it should be understood as the later "large vocabulary continuous speech recognition". The main system used in the experiments is in Danish and still speaker dependent.

1.2 About multimodality

Multimodal systems are characterised by human-machine interfaces that go beyond the traditional screen, keyboard and mouse. The use of the various input channels (keyboard, mouse, speech recognition, eye tracking, gesture recognition, etc.) and output channels (screen, sounds, speech synthesis, force feedback, etc.) are the so-called modalities. It should be noted that some channels are inherently bidirectional such as touch-screens, haptic interfaces (force feedback) and some brain-computer interfaces. Multimodality is the combination of multiple input and/or output modalities in the same user interface, together with additional software components such as fusion, fission and synchronisation engines.

Multimodality has been studied since the 1980's. The famous "Put-That-There" [Bolt 1980] concept was probably the advent of the field, by combining gesture recognition and speech input in his "Media room". The field has reached a new dimension with technologies such as augmented reality, which makes an extensive use of multimodality as well as providing interfaces and experiences that would not be at all possible without it.

[2] [http://speechrecognition.philips.com]

[3] [http://nuance.com]

There are different types of multimodality. Mainly, in the case of input channels:

1. Overlapping or redundant input modalities are cases when those modalities can be used to achieve the same type of input; in this situation, it is often up to the user to decide which modality to use. In the case of long inputs, some systems may offer the possibility to switch between those redundant modalities. Switching between modalities at any time during input requires a synchronisation engine so that the user does not have to restart the input from the beginning when picking another modality.

2. Complementary input modalities are situations when two modalities or more are used simultaneously or sequentially to generate an input message, which would not have been possible to express by using only one of the modalities. This type of behaviour requires a fusion engine that will combine the information coming from the various channels given some synchronisation rules.

Similarly, there are overlapping or complementary output modalities. Interfaces using multimodal outputs may need a fission engine to split the information and/or select the most appropriate channel.

Multimodal interfaces enrich human-computer interaction possibilities and provide advantages such as an increase in robustness and flexibility.

However, there are some drawbacks such as an increased complexity, together with more hardware and software needed. Even more than for traditional non-multimodal interfaces, there is a need to better grasp and model user activity, therefore calling for some notions from *e.g.* cognitive psychology and ergonomics, especially in safety critical systems [Andersen & Andersen 2003]. Of particular interest are frameworks such as Cognitive Systems Engineering (CSE), especially in relation to modelling human information processing; for instance, CSE identifies three classes of users, namely "novices", "intermediate" and "experts", whose behaviour is respectively mainly based on the three levels of control that are "knowledge", "rules" and "skills" [Rasmussen 1986:93-106]. Those are important concepts to be able to choose the appropriate modality in a given situation.

Some computer frameworks (*e.g.* W3C MMI[4]) and languages have been developed especially to facilitate the development of multimodal interfaces. An example of such language is XHTML+Voice[5] (X+V), which is a combination of XHTML[6], VoiceXML[7],

[4] W3C Multimodal Interaction Framework [http://www.w3.org/TR/mmi-framework/]

[5] XHTML+Voice [http://www.w3.org/TR/xhtml+voice/],
[http://www.voicexml.org/specs/multimodal/x+v/12/spec.html]

JavaScript[8], CSS aural style sheets[9] or speech module[10], with some additional facilities such as automatic synchronisation between the modalities.

X+V has been used during the early phases of this PhD, to get familiar with multimodality, test some concepts, and even during lectures given about multimodality on the Web[11]. As part of this preliminary work, I have demonstrated the implementability of complementary multimodality with a simple "Put-That-There" test case using combined speech recognition and mouse allowing user inputs such as "I want a new blue square … there".

When systems with a higher complexity are envisaged, component-based software engineering approaches have emerged to tackle the problem [Bouchet & Nigay 2004]. Finally, multimodality can be combined with multi-platform systems (*e.g.* the user can switch between a desktop computer and a mobile phone) to offer even richer possibilities [Paternò 2004].

In this thesis, when referring to "multimodality", it is often redundant input multimodality, that is to say interfaces offering overlapping ways of entering information where it is up to the user to select the most appropriate solution between, say, touch screen, mouse, keyboard or speech input.

1.3 About anaesthesia

Anaesthesia, which is the process of controlling pain and relaxation, can either be local when the patient is conscious, or general when the patient is asleep. Under general anaesthesia, the physician is also in charge of maintaining the vital functions such as respiration, hemodynamics (heart rate, blood pressure) and muscle tonus.

In occident, opium and alcohol were among the first anaesthetics, for instance applied by the French surgeon Ambroise Paré in the 16th century, sometimes called the father of modern surgery. After primitive techniques using the anaesthetic effect of cold, chemical anaesthetics for controlled local anaesthesia appeared in the late 19th century, with *e.g.* cocaine soon replaced by safer derivatives such as procaine and later lidocaine.

---

[6] XHTML Modularization: Modularization of the Extensible Hypertext Markup Language [http://www.w3.org/TR/xhtml-modularization/]

[7] VoiceXML: Voice Extensible Markup Language [http://www.w3.org/TR/voicexml20/]

[8] ECMA-262: ECMAScript [http://www.ecma-international.org/publications/standards/Ecma-262.htm]

[9] CSS 2 (Cascading Style Sheets), aural style sheets [http://www.w3.org/TR/CSS21/aural.html]

[10] CSS 3 (Cascading Style Sheets), speech module [http://www.w3.org/TR/css3-speech/]

[11] [http://alexandre.alapetite.net/phd-risoe/mxml/]

The first scientifically controlled general anaesthesia techniques were developed at the beginning of the 19th century, mainly with the use of gazes such as carbon dioxide and later diethyl ether, itself replaced by chloroform. With the spreading of the techniques came the first fatalities often due to untrained practitioners.

The advent of anaesthesiology has revolutionised surgery, considerably improving the working conditions, patient comfort, and allowing types of operations otherwise impossible. Nevertheless, safety considerations had very soon to be taken into account. Thanks to a better understanding of human physiology, new drugs and new anaesthesia apparatus including improved monitoring devices, the risk of death imputable to anaesthesia has significantly decreased during the second part of the 20th century, being about 70 times lower in ~1990 than in ~1950; however, many errors are still happening during anaesthesia, with human error responsible for about 70-75% of the cases [Chopra *et al.* 1992].

Therefore, in occident at least, most countries require anaesthesia to be supervised by a specialised doctor. In some countries, such as Denmark, the anaesthesia doctor in charge can follow several anaesthesias at a time, helped by specialised anaesthesia nurses that must stay close to the patient. Finally, more research is required to further decrease the number of (human) errors during anaesthesia.

The anaesthesia procedure and work practices are described more in details in the next article [Alapetite & Gauthereau 2005] with a focus on the topic of the thesis. In this thesis, when referring to anaesthesia, it is most of the time general anaesthesia.

1.3.1 About the anaesthesia record

During a medical operation involving anaesthesia, the anaesthetic record is important; not only as a legal document, but also because it is used during the operation – in particular if a new physician is joining the team – as a communication tool, and to make recall to the anaesthesiologists what occurred previously. The fact that the document is an indispensable source of information during the operation is the main reason for maintaining a real-time system: the information entered into the anaesthesia record cannot be just recorded (audio/video) and eventually transcribed after the end of the operation.

In operation rooms, registration of the patient record during anaesthesia has been done manually on paper for a long time. Today, some anaesthesia departments have switched to electronic systems. While electronic anaesthesia record systems are aimed at solving most of the issues encountered with paper-based recording, there is still a room for improvement, especially in emergency situations. In the case of electronic records, the comments, in particular, are not described as precisely as they could be, partly due to the use of a keyboard, which is not a convenient input device in such an environment.

1.3.1.1 Mail survey on types of recording systems

A survey was launched in Denmark in October 2004, targeting by paper mail almost all the anaesthesia departments of the country (N = 47). 35 answers were received, showing that 13 of them (⅓) did not use any form of electronic system and 14 used a complete electronic system (⅓). The 22 departments that used partial or complete electronic systems were using about 12 different systems. On the 29 paper models or printouts received, only 3 of them were identical, while all the others were specific to just one anaesthesia department. This situation appeared to have an historical explanation: the different departments progressively built their own system, with little communication among each other.

2 Rationale

The aim of this research project is to study how multimodal interfaces, especially with the addition of vocal interaction, could make recording during anaesthesia more accurate, flexible and robust.

2.1 Context

Speech recognition engines have significantly improved over the last decade thanks to the introduction of new techniques and increased computer power. When used carefully, as an alternative or a supplement to other more conventional modalities (buttons, touch-screen, keyboard, mouse, etc.), speech input can now be considered in various safety-critical environments. Speech recognition technology has actually already been used for some years in safety critical domains such as medicine [Devine *et al.* 2000], military aviation and air traffic control [Lechner *et al.* 2002].

2.2 What is the problem?

The ideal anaesthesia setup is yet to be found, as current typical workplaces face ergonomic problems such as the difficulty for the anaesthesiologist to see at the same time the patient and the anaesthesia electronic record.

Furthermore, in emergency situations during anaesthesia when physicians and nurses are busy and maybe stressed, the registration process is delayed. This is a problem, because postponing the registration often leads to uncertainty, inaccuracy and other errors. Furthermore, even in the case of adverse events, physicians will not always report and document the issues once the operation finished, for various reasons including time resources or bad feeling in the case of error reports [Andersen *et al.* 2002].

Besides, if speech recognition is to be considered, as with other types of equipments in safety critical domains, it has to match requirements such as reliability, accuracy, robustness, fault tolerance, etc. Even with state-of-the-art technology, there is still a great need for improvement of these systems in order to match strict safety critical criteria. Strategies such as redundancy should be investigated. Last but not the least, human factors when using of speech recognition under emergency and safety critical situations should be investigated carefully.

Those issues are further described in the next section [Alapetite & Gauthereau 2005].

2.3 Why is this research needed?

There are some beliefs that adding some modalities to existing electronic patient record systems, such as voice, could be beneficial for the quality of recording. Making multimodal interfaces enables practitioners to choose between different ways of registering, depending on the current situation (touch-screen, keyboard, voice, etc.). Indeed, the different modalities do not have the same requirements (using hands, standing close to the machine, noise and light conditions, etc.) and capacities (accuracy, robustness, etc.). Speech input could be very valuable for anaesthesia electronic record interfaces, as well as for commanding – in some cases – other anaesthesia equipments [Schmitz & Weiss 2004]. This could be a good solution for improving recording even in crises, when the registration is most of the time delayed with current interfaces and work practices.

There have been a few attempts to integrate speech recognition into electronic anaesthesia record systems or monitors [Smith *et al.* 1987; Smith *et al.* 1990; Sanjo *et al.* 1999; Jungk *et al.* 2000]. While their early results are interesting and promising, the authors call for more research, in particular on the human factors [Smith *et al.* 1987], on real time experiments [Smith *et al.* 1990], on identifying task areas where vocal interaction can be beneficial, on evaluating ergonomic designs [Jungk *et al.* 2000] as well as the effects on vigilance and contact with the patient [Sanjo *et al.* 1999]. This thesis seeks to respond to some of these calls.

A broader literature survey shows that some research has already been done with speech recognition in the medical domain, mainly as an alternative to medical transcription [Lai & Vergo 1997; Zafar *et al.* 1999; Devine *et al.* 2000; Borowitz 2001; Zick & Olsen 2001; Al-Aynati & Chorneyko 2003; Mohr *et al.* 2003]. Another example is [Detmer *et al.* 1995], who made some "Wizard of Oz" experiments with speech recognition as an interface to a decision support system. However, most existing applications are targeted at non-real-time environments where physicians provide dictation, perhaps in an office, where input and subsequent review and correction may be made in batch mode. Consequently, there is little literature about real-time speech input during operations or anaesthesias, when speech recognition is not the primary task.

3 Methodology

The chosen approach is based on action research [Baskerville 1999], and is partially technology-driven; this is actually needed to investigate the potential of speech recognition in applications for which it is not entirely mature [Danis & Karat 1995].

The work is conducted in collaboration with expert users in three hospitals in Denmark (Herlev hospital; Køge hospital; Vejle hospital). The goal is to improve existing systems, to do some experiments and to validate some hypothesis, but not to validate or to create a new theoretical framework. Part of the research is a loop of prototyping and analysis. Some measurements of the impact of the new system on work procedures and quality of recording are done and reported.

The project started by an extensive literature review, and several meetings with anaesthesia experts (physicians, nurses, engineers). They agreed on the fact that paper-based recording suffers from many problems. On the other side, I made interviews in anaesthesia departments where electronic anaesthesia records are used: even if there is still some improvement needed, many of the traditional paper problems appeared to have been solved. This seems consistent with a survey that yielded the following result: "46% of medication errors occur on admission or discharge from a clinical unit/hospital when patient orders are written, and they drop by 90% when they are electronic" [Pronovost 2003]. Nevertheless, remaining problems together with new ones introduced by electronic versions of the anaesthesia record have to be addressed. This first phase of dialog with experts in the field convinced me of the relevance of the final topic and methodology, after they had been refined.

At this stage, the final goal is thus to create a prototype and test it in an anaesthesia simulation environment, involving physicians and nurses performing simulated anaesthesias, and with measurements of the benefits of the prototype compared to existing solutions will be made. To reach this goal, preliminary investigation and intermediary experiments were known to be necessary, as detailed later in the thesis.

An analysis of current practices is also needed in order to establish the *phraseology*, which is the definition of what could be said to the system and how. This is required to build the grammar, which is the formal basis of what the system can accept, and understand. Free speech, within a limited context, is also used to allow practitioners to put some more detailed comments in the patient record. There are various technical possibilities for implementing the expected prototype, and there is therefore a need to think and to test various architectures and strategies.

3.1 Partners and contacts about speech recognition technologies

When doing applied research, it is important to ensure contacts with up-to-date technologies. Therefore, starting in March 2004 from Vejle hospital, pioneer in Denmark on speech recognition adoption, I analysed some reports and visited some places where using speech input in the medical domain had been successful or had failed.

A contact was then established with Max Manus[12], a Danish company providing speech recognition in Danish based on Philips SpeechMagic[13] technology. At this point, this system had indeed been put into daily use at the radiology department of Vejle Hospital, and tested with less success at the pathology department of Aalborg hospital and the pathology department of Sønderborg hospital [Hvidberg 2003].

In Denmark, apart from the Max Manus technology, there was also an apparently successful use of voice commands at Hvidovre hospital (demonstrated to the press in October 2004), with the HERMES[14] system [Luketich *et al.* 2002] from Stryker company, but limited to English (instead of Danish) and to short commands.

References

[Al-Aynati & Chorneyko 2003] Maamoun M. Al-Aynati & Katherine A. Chorneyko. Comparison of Voice-Automated Transcription and Human Transcription in Generating Pathology Reports. *Archives of Pathology and Laboratory Medicine, 2003, 127(6):721–725.*

[Andersen *et al.* 2002] Henning Boje Andersen, Marlene Dyrløv Madsen, Niels Hermann, Thomas Schiøler, Doris Østergaard. Reporting adverse events in hospitals: A survey of the views of doctors and nurses on reporting practices and models of reporting. *In: Investigation and reporting of incidents and accidents. Workshop (IRIA 2002), Glasgow (GB), 17-20 June 2002. Chris Johnson (ed.), (University of Glasgow, Department of Computing Science, Glasgow, 2002) (GISTTechnical Report, G2002-2) p. 127-136.*

[Andersen & Andersen 2003] Hans H. K. Andersen & Verner Andersen. Establishing user requirements in HCI - A case-study in medical informatics. *Proceedings of HCI International'2003, International conference on human-computer interaction, Vol. 1, Theory and practice, Part 1, Crete (GR), 22-27 Jun 2003, J. Jacko & C. Stephanidis (eds.), (Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 2003) p. 611-615.*

[Baskerville 1999] Richard Baskerville. Investigation information systems with action research. *Communications of the Association for Information Systems, Volume2, Article 19, October 1999.*

[12] [http://maxmanus.dk]

[13] [http://speechrecognition.philips.com]

[14] [http://www.europe.stryker.com/index/st_pag_medic-home/st_pag_detailed-product-info/st_pag_endoscopy-systems/st_pag_endo-hermes-prod-res.htm]

[Bolt 1980] Richard A. Bolt. "Put-that-there": Voice and gesture at the graphics interface. *Proceedings of the 7th annual conference on Computer graphics and interactive techniques, Seattle, Washington, USA, pp. 262-270*. doi:10.1145/800250.807503

[Borowitz 2001] Stephen M. Borowitz. Computer-based speech recognition as an alternative to medical transcription. *Journal of the American Medical Informatics Association, 2001, 8(1):101-102*.

[Bouchet & Nigay 2004] Jullien Bouchet & Laurence Nigay. ICARE: a component-based approach for the design and development of multimodal interfaces. *CHI'2004 Conference on Human Factors in Computing Systems, pp 1325-1328, Vienna, Austria*. doi:10.1145/985921.986055

[Chopra *et al.* 1992] V. Chopra, J.G. Bovill, J. Spierdijk, Floor Koornneef. Reported significant observations during anaesthesia: a prospective analysis over an 18-month period. *British Journal of Anaesthesia, 1992, 68:13-17*.

[Danis & Karat 1995] Catalina Danis & John Karat. Technology-driven design of speech recognition systems. Proceedings of DIS'1995, Symposium on Designing Interactive Systems, pp. 17-24. doi:10.1145/225434.225437

[Davis *et al.* 1952] K. H. Davis, R. Biddulph, S. Balashek. Automatic Recognition of Spoken Digits. Journal of the Acoustical Society of America, November 1952, 24(6):637-642. doi:10.1121/1.1906946

[Detmer *et al.* 1995] William M. Detmer, Smadar Shiffman, Jeremy C. Wyatt, Charles P. Friedman, Christopher D. Lane, Lawrence M. Fagan. A Continuous-speech Interface to a Decision Support System: II. An Evaluation Using a Wizard-of-Oz Experimental Paradigm. *Journal of the American Medical Informatics Association, Jan–Feb 1995, 2(1):46–57*.

[Devine *et al.* 2000] Eric G. Devine, Stephan A. Gaehde, Arthur C. Curtis. Comparative Evaluation of Three Continuous Speech Recognition Software Packages in the Generation of Medical Reports. *Journal of the American Medical Informatics Association. 2000,7(5):462–468*.

[Dybkjær & Dybkjær 2002] Hans Dybkjær & Laila Dybkjær. Experiences from a Danish Spoken Dialogue System. *Proceedings of the 2nd Danish HCI Research Symposium, 7 November 2002, DIKU technical report 02/19, pp. 15-18, Erik Frøkjær & Kasper Hornbæk (Eds.), University of Copenhagen, Denmark,* ISSN:0107-8283.

[Hvidberg 2003] Jens Hvidberg. Talegenkendelse - muligheder og barrierer for anvendelse til klinisk dokumentation (In Danish). *Master's thesis, Aalborg University, May 2003*.

[Jungk *et al.* 2000] Andreas Jungk, Bernhard Thull, Lutz Fehrle, Andreas Hoeft, Günter Rau. A case study in designing speech interaction with a patient monitor. *Journal of Clinical Monitoring and Computing, 2000, 16:295-307*. doi:10.1023/A:1011456205786

[Lai & Vergo 1997] Jennifer Ceil Lai & John George Vergo. MedSpeak: Report creation with continuous speech recognition. *In: Proceedings of the CHI'1997 Conference on Human Factors in Computing Systems. ACM Press, 1997, New York, pp. 431-438*. ISBN:0-89791-802-9

[Lechner *et al.* 2002] Alicia Lechner, Kevin Ecker, Patrick Mattson. Voice recognition – Software solutions in realtime ATC workstations. *Aerospace and Electronic Systems Magazine, IEEE, 2002, 17(11):11-16*. doi:10.1109/MAES.2002.1047373

[Luketich *et al.* 2002] J.D. Luketich, H.C. Fernando, P.O. Buenaventura, N.A. Christie, S.C. Grondin, P.R. Schauer. Results of a randomized trial of HERMES-assisted versus non-HERMES-assisted laparoscopic antireflux surgery. *Surgical Endoscopy, 2002, 16:1264-1266*. doi:10.1007/s00464-001-8222-7

[Mohr *et al.* 2003] David N. Mohr, David W. Turner, Gregory R. Pond, Joseph S. Kamath, Kathy B. De Vos, Paul C. Carpenter. Speech Recognition as a Transcription Aid: A Randomized Comparison With Standard Transcription. *Journal of the American Medical Informatics Association, 2003, 10(1):85–93*. doi:10.1197/jamia.M1130

[Paternò 2004] Fabio Paternò. Multimodality and Multiplatform Interactive Systems. *Proceedings of WCC'2004, the 18th IFIP (International Federation for Information Processing) World Computer Congress, Toulouse, August 2004, Kluwer Academic Publishers, René Jacquart (ed.), "Building the Information Society", pp. 421-426,* ISBN:1-4020-8156-1

[Pronovost *et al.* 2003] Peter Pronovost, Brad Weast, Mandalyn Schwarz, Rhonda M. Wyskiel, Donna Prow, Shelley N. Milanovich, Sean Berenholtz, Todd Dorman Pamela Lipsett. Medication reconciliation: a practical tool to reduce the risk of medication errors. *Journal of Critical Care, 2003, 18(4):201-205.* doi:10.1016/j.jcrc.2003.10.001

[Rasmussen 1986] Jens Rasmussen. Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering. *Book published by Elsevier Science Inc., 1986*. ISBN:0444009876

[Sanjo *et al.* 1999] Yoshimitsu Sanjo, Tetsuo Yokoyama, Shigehito Sato, Kazuyuki Ikeda, Reiko Nakajima. Ergonomic automated anesthesia recordkeeper using a mobile touch screen with voice navigation. *Journal of Clinical Monitoring and Computing, 1999, 15:347-356*. doi:10.1023/A:1009972223750

[Schmitz & Weiss 2004] Achim Schmitz & M. Weiss. Are we going to talk with our anaesthesia monitors in the future? *Acta Anaesthesiologica Scandinavica, 2004, 48(2):255-256*. doi:10.1111/j.0001-5172.2004.00295b.x

[Smith *et al.* 1987] N. Ty Smith, M. L. Quinn, A.J. Sarnat. Speech Recognition for the Automated Anesthesia Record. *In: The Automated Anesthesia Record and Alarm Systems, Chapter 11, Jonathan S. Gravenstein, Ronald S. Newbower, Allen K. Ream, N. Ty Smith (eds.), Butterworths, 1987, pp. 115-134.*

[Smith *et al.* 1990] N. Ty Smith, Robin A. Brien, Daniel C. Pettus, Brian R. Jones, Michael L. Quinn, Andrew Sarnat. Recognition accuracy with a voice-recognition system designed for anaesthesia record keeping. *Journal of Clinical Monitoring, 1990, 6(4):299-306.*

[Zafar *et al.* 1999] Atif Zafar, J. Marc Overhage, Clement J. McDonald. Continuous speech recognition for clinicians. *Journal of the American Medical Informatics Association, 1999, 6(3):195–204.*

[Zick & Olsen 2001] Robert G. Zick, Jon Olsen. Voice Recognition Software Versus a Traditional Transcription Service for Physician Charting in the ED. *American Journal of Emergency Medicine, July 2001, 19(4):295-298.* doi:10.1053/ajem.2001.24487

Transition 1

This first part (partially based on [Alapetite 2005.a]) was dedicated to the familiarisation with the topics, the related literature, partners and other contacts, together with the state of the art.

Consequently, this allowed the establishment of a more precise and stable research question, namely on investigating the potential of speech input – as a supplement to the traditional touch-screen and keyboard – to address some of the apparent problems and limitations of human-computer interfaces to electronic anaesthesia record systems.

It was then required to gain knowledge on this specific research question. Furthermore, in the ADVISES research network, there was a rich pool of competences in ergonomics and sociology. This was giving insight to use their background to deeper understand the current work practices around the anaesthesia record, as well as planning the research to come, by thinking of the possible direct and indirect consequences of the envisaged modifications, such as the introduction of speech recognition.

This primary reflection was considered important, not to embark on conceptually wrong directions. This was also necessary to identify points of interest to keep in mind when progressing towards a prototype and when doing some evaluations.

Written in collaboration with a socio-ergonomist from the ADVISES network, the following paper [Alapetite & Gauthereau 2005] is undertaking this work, based on an extensive literature review, interviews and direct observations.

Introducing vocal modality into electronic anaesthesia record systems: possible effects on work practices in the operating room

Alexandre Alapetite[1,2], Vincent Gauthereau[3]

1. Risø National Laboratory; Systems Analysis Department; Research Programme Safety, Reliability and Human Factors; P.O. Box 49; DK-4000 Roskilde; Denmark
2. Roskilde University; Computer Science; Universitetsvej 1; P.O. Box 260; DK-4000 Roskilde; Denmark
3. Université de Liège; LECIT Laboratory; Boulevard du Rectorat, 5 (B32); B-4000 Liège; Belgium

This section is a modified version of the following conference article:

Alexandre Alapetite & Vincent Gauthereau. Introducing vocal modality into electronic anaesthesia record systems: possible effects on work practices in the operating room. *In the proceedings of EACE'2005, annual conference of the European Association of Cognitive Ergonomics, 29 September - 1 October 2005, Chania, Crete, Greece; Section 2 on "Research and applications in the medical domain", pages 189-196. University of Athens. In the ACM International Conference Proceeding Series; Vol. 132, pages 197-204, ISBN:9-60254-656-5.*

Abstract

The work reported in this paper is part of a project aiming at introducing vocal modality into the electronic anaesthesia record in Denmark. The purpose of the paper is to offer a basis for comprehending the use of anaesthesia records in work practice, to list the current main issues and possible improvements, and finally to foresee the impact of the addition of a new voice interface. The present paper is the result of a collaboration between an engineer, involved in making prototypes of the system described above, and a socio-ergonomist. The analysis is based on a literature review, interviews and direct observations.

Keywords

Anaesthesia; electronic records; patient; voice; speech

1 Introduction

Problems in the way paper-based and electronic anaesthesia records are filled during anaesthesia have been observed, such as anachronisms, temporal defects and lacking entries. In response, it has been suggested that a way to surmount these deficiencies is to offer anaesthesiologists a voice-based input modality to an electronic record [Schmitz & Weiss 2004]. Some attempts have been made [Sanjo *et al.* 1999; Jungk *et al.* 2000], and are calling for more research, on identifying precise subtasks where voice input can be beneficial, on the specific human-computer interaction, such as feedbacks, and on the trade-off between vigilance and visual contact with the patient. Some studies have demonstrated the usefulness of activity modelling in anaesthesia interfaces [Beuscart-Zéphir *et al.* 2001], but they focussed on the pre-operative consultation. On top of technical difficulties, other issues and a number of research questions need to be considered, such as the different roles of the anaesthesia record, and the prediction of possible impacts of the new technology on the activity. A detailed analysis of the activity should allow us to extract some guidelines to be used in future experimentations of the new tool. The implications of the modification of tools on work practices are also taken into account.

2 Focus of this paper

We leave aside a first set of questions, which has to do with the underlying assumption behind the problem as defined above, and raises the issue of whether a "good" anaesthesia record is a one that is fully and correctly filled. To say a word about this, we often take for granted that increasing the amount of information in the record, and allowing this information to be filled synchronously to the operational reality is a good objective. However, interviews have shown the difficulty to differentiate between important and unimportant information in electronic anaesthesia records that were 'fully' filled, while in hand-written records, that were visibly incomplete, important information was easier to identify. Moreover, while extending the file, anaesthesiologists do not focus on the patient. This area of issues around the question of what a "good" record is thus raises a set of questions, like the roles of record, during an operation or outside the operating room, and the interests of anaesthesiologist in filling the records 'fully'.

A second set of issues concerns the changes that the new technology will have on work practices, both in the operating room and outside. For instance, in the operation room, we can expect that the new modality might affect the existing communication schemes. In order to start comprehending these issues, we need to have a reasonably complete picture of the activity of the anaesthesiologist, and the role of the record, at least in the activity around the patient.

The present conference article will focus on the second set of questions. The aim of the present work is to describe the activity of the anaesthesiologist and of the roles of the anaesthesia record in this activity. In fact, while we believe that the first set of questions (the one left aside) is central when seeking to validate the new technology in relation to patient safety [Gauthereau 2004], we also believe that we first need to comprehend the activity itself, if we ever wish to understand the mechanisms behind its evolution over time [Lave 1993].

3 Electronic anaesthesia records (EAR)

3.1 Introduction to anaesthesia

Anaesthesia is a medical act aimed at reducing the pain and consciousness of a patient, in order for him to receive a medical act such as surgery. There are various kinds of anaesthesia; some of them are only targeting an area of the body, with the patient still awake. In this paper, we will mainly focus on general anaesthesia, which is applied to the whole body and keeps the patient asleep using various kinds of drugs administration, such as induction agents to produce unconsciousness, analgesics to reduce pain, muscle relaxants, inhalation agents to keep the unconsciousness, etc. In many countries, general anaesthesia can only be conducted by a specialist doctor. In some other countries however, nurse anaesthesiologists may deliver anaesthetics, normally under the supervision of a specialist doctor. In this paper, "anaesthesiologist" refers to the practitioner, doctor or nurse, directly in charge of the patient. During anaesthesia, many choices have to be made by the practitioner, based on knowledge, monitor trends as well as direct observations. This activity is reported in the anaesthesia record (AR).

3.2 Importance of the anaesthesia record (AR)

During anaesthesia, the main task is to take care of the patient, so the AR is a secondary task. This means that anaesthesiologists do not necessarily have much time to do it. They may be stressed or not fully concentrated on the record keeping; they sometimes postpone it after the operation and have to rely on their memory. But the AR is important, not only because it is a legal document, but also because it is used during operations to communicate and make available what has occurred previously, especially to support a quick oral briefing if someone joins the team. Indeed, at the organisational level, some hospitals base their incidents recuperation strategy on experienced anaesthesiologists joining the medical team in a minute [de Keyser & Nyssen 1993].

An analysis has shown that 70% of reported anaesthesia incidents were related to human errors [Chopra *et al.* 1992], and a study of some accidents shows a lack of functional communication in the medical team [de Keyser & Nyssen 1993].

The fact that the document is an indispensable source of information during the operation is the main reason for maintaining a real-time system: the information entered into the anaesthesia record cannot be just recorded (audio/video) and eventually transcribed. It is also used as a verification mechanism; for instance, some anaesthesiologists believe that it is better to fill the anaesthesia record before transfusing some blood to the patient, in order to ensure that the codes are checked correctly before any critical administration [de Keyser & Nyssen 1993].

What is actually recorded in the AR and how, reflects local customs, but the AR must at least contain the main vital signs (*e.g.* heart rate), time, techniques, route and dose of the administrated drugs, as well as the main events (*e.g.* surgery started).

3.3 From paper templates to electronic systems

In operation rooms, registration of anaesthesia records during anaesthesia has been done manually on paper for a long time. However, it is well known that hand-written documents in the medical domain are a common source of communication mistakes. A survey yielded the following result: "46% of medication errors occur on admission or discharge from a clinical unit/hospital when patient orders are written, and they drop by 90% when they are electronic" [Pronovost 2003]. Moreover, handwriting is quite time consuming and forces the practitioner to leave the current task to use pen and paper. Therefore, especially during busy and perhaps emergency phases, staff will sometimes defer writing down the information in the anaesthesia record. In turn, this may lead to the risk that practitioners might forget or misremember data, which will produce misleading information with potential impact on subsequent phases of the anaesthesia.

Furthermore, in contrast to electronic systems, paper-based recording does not provide much barrier to ensure that the provided data is consistent; it is filled and used in various ways by the different practitioners, creating inconsistencies, and there is a lack of space to write the remarks or some other precisions.

As a result, it seems that a substantial percentage of anaesthesia paper-based records are incomplete or contain errors [Hamilton 1990]. This is in agreement with a rapid small-scale analysis we did in June 2004 at Herlev University Hospital (Copenhagen county, Denmark), which uses paper-based recording. 55 records were randomly chosen and computerised without correction by a highly skilled anaesthesia nurse. As examples, only 7 (13%) specified the ASA (American Society of Anaesthesiologists) physical status classification, which is recognised to be important, and 14 (8%) did not provide any information about the time when the operation or the anaesthesia ended. We then focused on obesity, as it is easy to establish inconsistencies automatically. Out of 55 files, 42 (76%) contained valid weight of the patient, which is a required information. 15 files (27%) provided additional information about height, which allowed us to calculate the body mass index (BMI). When BMI >= 30, it is likely a sign

of obesity; this was the case for 9 files, and out of them, 8 (89%) had not checked the obesity field, as they should have done.

Finally, the sometimes hard-to-read handwriting presents additional problems that are not always trivial. This makes those files difficult to use, especially when they have to be transmitted to another department or hospital. Some observers have therefore argued that there is a need for a complete electronic "patient data management system" (PDMS) [Schmitz & Weiss 2004]. Today, some anaesthesia departments have switched to electronic systems, including an electronic anaesthesia records.

We did a survey in October 2004 on almost all the anaesthesia departments in Denmark. Among the 35 responding departments, 13 (37%) did not use any form of electronic system, 14 (40%) used a complete electronic system, and the 8 (23%) left used a partially electronic system.

4 Description of the problem

4.1 Current problems with EAR

While electronic anaesthesia records (EAR) seek to solve most of the issues encountered with paper-based recording, there is still room for improvement. The comments, in particular, are not described as precisely as they could be, partly due to the use of a keyboard, not convenient in such an environment. When physicians and nurses are busy and maybe stressed, the registration process is often delayed, which can lead to omissions, uncertainty, inaccuracy, resulting in anachronisms. There are some events that do not require precision, and five minutes accuracy is fine for most cases, but this can be difficult to achieve with the current interface.

Moreover, observations in 3 other hospitals in Denmark (Køge, Frederiksberg, Bispebjerg) have shown that the touch-screen used in the current interface is often placed behind the anaesthesiologist, which is not especially convenient, as it makes difficult seeing the record and the patient at the same time (*cf.* Figure 1). In addition, no alternative pointing device has been observed, in case the touch-screen would fail, even though difficulties with the touch-screen have been noted, like when using menus. In addition, the small font size forces some users to change glasses to read or fill the record.

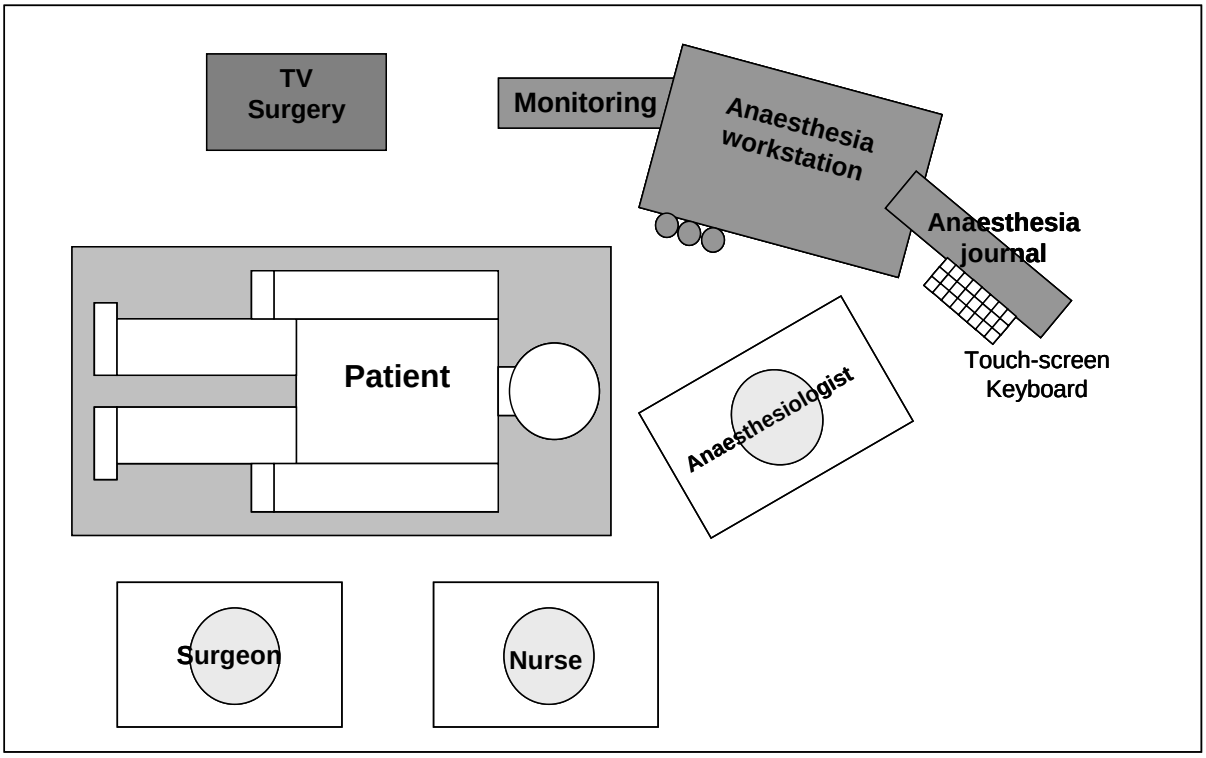


Figure 1: Anaesthesia theatre.

5 What to improve in the records?

Today, electronic patient data such as heart rate, blood pressure, etc. are automatically recorded. However, other information, such as patient current skin colour, is equally important to the anaesthesiologist. Furthermore, validating, labelling and commenting the data automatically recorded could be very useful for later interpretation. Even if this is already possible in the observed systems, it is rarely done. Improving the quality of the data being recorded during the operation should support different functions in which the records are being used during this operation.

5.1.1 Support for decisions

Anaesthesiologists like to see the parallelism between the vital signs and actions undertaken. While this is already true and efficient today, it could be improved with a more complete and accurate timeline of simultaneous actions and comments. Also, considering future possible developments, we can see that some efforts have been made to make anaesthesia monitors and alarms more "intelligent", in order to provide more concise information; but those systems are limited by the lack of relevant data: "not all information can be given by the monitors, and the anaesthesiologist is too busy" [de Graaf *et al.* 1997].

5.1.2 Support for memory

The AR is often used as a memory support, especially during long or difficult operations. Since gathering relevant information is laborious, it is important to improve the way it is recorded; otherwise, especially when workload is high, the anaesthesiologist will tend to rely only on memory, which can be a source of errors or time delays.

5.1.3 Support for communication

The AR is used as a support for verbal communication between the different actors involved in the anaesthesia. Observations have shown that they point to precise areas on the record while talking and explaining things. Insuring that the EAR is up to date is therefore crucial.

5.2 Why introducing vocal modality in EAR?

It has been thought that vocal modality could improve the way EAR is being filled out [Schmitz & Weiss 2004]. If it can provide a faster interface, it would be especially useful for short anaesthesias, like abortion or appendicitis, when the time spent to feed the record is sometimes longer than the operation itself. We believe voice input would be especially suitable to standard commands and remarks like “Intubation”. Other important benefits would be gained if voice can avoid postponing the registration, which creates a loss of precision, takes extra time and resources.

5.2.1 Speech recognition in the medical domain

Improvements in speech recognition have allowed successful project in the medical domain [Devine *et al.* 2000], such as voice commands to assist surgery at Hvidovre hospital (DK, 2004) (in English). With speech engines available in Danish, some systems have been put into daily use, such as diagnosis dictation at the radiology department of Vejle Hospital (DK, 2003) and are rapidly spreading to other departments: anaesthesia in October 2005 (Philips/Max Manus voice technology). However, most existing applications are targeted at non-real-time environments: physicians provide dictation, perhaps in an office, where input and subsequent reviewing may be made in batch mode.

Consequently, there is little literature about real-time speech input during operations or anaesthesias, when speech recognition is not the primary task and where there is a need of processing, interpreting and validating more complex speech in order to react, to do precise actions or verifications, to write in the correct fields and to move between them [Smith *et al.* 1990].

5.2.2 Differences with existing medical voice interfaces

Current speech recognition systems in daily use in the medical domain, such as for X-ray diagnosis in Vejle hospital, provide an efficient way to enter plain text into the system. However, this kind of application differs from the EAR in two respects.

First, as described before, the anaesthesia record is not the main task of the practitioner, while it can be considered as the main one for X-ray diagnosis. This situation creates additional difficulties for speech recognition with a noisy environment, possibly with other people speaking, and with variations in the speaker's voice because of stress, a mask covering the mouth, body movements and postures.

The second point is that in current systems, recognition is made in a free speech mode, which means that the user does not have many constraints in the way sentences are formulated. In return, the system delivers a block of plain text with no interpretation (the computer does not know what to do with the data), no verification (ranges, units) and almost anything could be said. When this method is perfectly adapted for writing a typical 15-line summary, it is not directly suitable for filling an anaesthesia record. Since the anaesthesia record is composed of several areas with fields that are meant to contain various kind of information, the speech recognition system has to be able to determine where to store the data, in order to use the correct format and to limit the range of what is acceptable (numbers, units, medications, etc.).

For filling in the anaesthesia record by voice, the anaesthesiologist will have to use a set of commands – based on keywords – to quickly navigate in the form, like moving between the fields. This phraseology (the way to speak to the system) can be extended with high-level sentences dedicated to the main events that occur during anaesthesia, such as "intubations" for example, for the anaesthesiologist not to have each time to explicitly specify the targeted field. Those commands and high-level sentences can be recognised by the speech recognition system and associated to a meaning.

Relying on a precise phraseology to address the system, the speech recognition engine is not only able to return some plain text, but also to react, to do precise actions or verifications, to write in the correct fields and to move between them.

6 Modifying work tools and its implications on work practices

As we are modifying work tools, it is crucial to be aware of their important role in human activity. In accordance with the activity-theory research tradition, we understand activity as basically mediated. That is to say, in order for a subject to perform an activity, there is always use of a mediator. This mediator can be either a physical artifact, or a symbolic one, or both simultaneously. The physical environment has structuring properties fundamental to cognition, as do artifacts whose structures are the products of a more or less long social-cultural process [Nardi 1996].

Given the complexity of tools usage in human activity, predicting all the implications of a technical change on work practice is almost impossible. Tools are used on different levels (instrumental or semiotic) and support different cognitive mechanisms. Moreover, their usage highly depends on the level of expertise of the user. While some implications of technological changes can be predicted, not all of them can. Validation of new technology while in need of in-depth studies of the actual activity, thus needs a stage during which the new technology will be introduced in a practice in order to study its actual effects on work practice.

Activity-analysis can support innovation but should not be used too much as a brake: not being able to predict all the implications of a technological change should not be used as a reason to stop the innovating process.

7 Use of electronic records in the activity

In order to highlight consequences of modifying anaesthesia work tools and to achieve a well functioning solution, we need to test prototypes in simulated environment. A good starting point is to test the principle of the new tool in “Wizard of Oz” experiments in which a perfect version of the tool will be simulated by humans. In order to prepare this set of experiments, we should have assumptions about the impacts that can be observed. The first step is thus to understand the current practice [Gravenstein 1989].

In this section, we are going to describe the activity of anaesthesia linked to surgery in an operating room. The main objective of anaesthesia in relation with surgical operation is to enable a situation that allows surgery: it is a facilitator’s role. However, while this is the assigned objective, anaesthesiologists have another main goal: to maintain the patient as close to consciousness as possible. It is thus a situation in which anaesthesiologists are dynamically controlling the level of consciousness of the patient in order to limit as much as possible the depth of anaesthesia while at the same time enabling the surgeon to perform his task on a patient that is non-reactive.

We can identify 6 mains phases in an operation with anaesthesia: a pre-operative phase, a pre-anaesthesia phase, an induction phase, a regulation phase, a post-surgical phase and a post-operative phase (*cf.* Table 1).

Table 1: Main phases of anaesthesia.

Anaesthesia process					
Pre-operative	Peri-operative				Post-operative
	Pre-anaesthesia	Induction	Maintenance	Recovery	

7.1 Pre-operative Phase

The main goal of this phase, which can be performed some days before the operation, is to establish the profile of the patient during an interview.

At that moment, the anaesthesiologist prepares the case, taking forth the patients data, in order to establish a risk level for each patient. This risk-level (ASA class) influences the procedures that will be in use during the operation. For planned operations, this evaluation is done through an interview of the patient, which is the occasion to establish the general health profile of the patient, current medications, etc.

Today, in Denmark, even when electronic records are in use, physicians generally use paper-based documents and will have to re-enter the information in the computer system. This appears to be mainly due to financial considerations and lack of interoperability among systems, and should be solved soon.

7.2 Pre-anaesthesia Phase

This phase usually takes place in the operating room itself, or in a preparation room in which everything will be settled and then moved all together in the operating room. The main goal is to prepare the anaesthesia: the patient, the equipment, the monitors, the drugs, etc. This phase begins a while before the arrival of the patient, to start preparing the drugs, and the equipment. For instance, the drugs are put forth and labelled. Once the patient arrives, the anaesthesiologist will first check the patient identity, and the kind of operation expected. During this phase, the anaesthesia record starts to be filled with information about the patient and the anaesthesia team. Sensors used to record the patient's vital signs are connected to the monitors and to the patient. The anaesthesiologist also explains to the patient what is going to happen. At this stage, the anaesthesiologist is typically assisted by another one.

7.3 Induction Phase

The induction phase starts with the administration of the first drugs. The anaesthesiologist is very active and needs to manage several tasks simultaneously, monitoring the consciousness level of the patient in order to intubate when it is appropriated.

During that phase, the anaesthesiologist needs to carefully follow the patient vital signs, both from the monitors and from the sight of the patient, while at the same time, more anaesthesia drugs must be administrated. Once the patient is intubated, the anaesthesiologist can rely a bit more on the artificial breathing system. Until then, the anaesthesiologist actually needs to support the natural breathing function of the patient using a manual breathing system.

In the observed situations in Denmark, two anaesthesiologists are present during this phase (typically an anaesthesia doctor and nurse), mainly communicating by looking at each other's actions, without much talking. The main events are reported in the EAR as soon as one of the anaesthesiologists has time to do it. Most likely, the one monitoring the patient and in charge of intubating the patient is not the one who will fill up the record. This step is quite short, as it lasts for about 5 minutes.

The next step is the intubation itself, and once the patient's state is stable, the surgery can start. This takes another few minutes. At the early stage that follows the intubation, the anaesthesiologists tend to verbalise quite much to the other nurses. With time, this will decrease to the benefit of verbalisation to surgeon. Focus is then more on the monitors and less on the patient.

7.4 Maintenance Phase

When surgery has started, there is usually only one anaesthesiologist left (typically the nurse), who constantly takes care of the patient, and administers drugs that will keep this one in an unconscious state. The anaesthesiologist carefully monitors the patient's health because of drugs side effects and surgery, like impact on blood pressure, heart rate and breathing.

During this phase, when surgery has started, the anaesthesiologist must be especially vigilant about vital signs of these basic functions. Since blood pressure and heart rate are not only impacted by the anaesthesia drugs, but also by the surgical act in itself, the anaesthesiologist needs, once changes in these vital signs are detected, to identify the specific cause behind these perturbations. The identification might, on the one hand, help the surgeon to detect an error (such as a cut of a wrong blood vessel), and on the other hand, alert the anaesthesiologist about a dangerous reaction of the patient to the drugs.

In order to monitor the patient's status, the anaesthesiologist looks at the monitors providing vital signs, but also at the patient (colour of the face, muscles' relaxation, hydratation level, pupil dilatation). Electroencephalogram (ECG) can sometimes be used, as they can help anaesthesiologists by providing data regarding the consciousness level of the patient.

Since this phase has normally a lower workload than induction and recovery, the anaesthesiologist usually takes time to complement the record for the previous phase. Today, when vital signs are automatically recorded, the anaesthesiologist must pay attention to the validity of this data, and also needs to enter data regarding the drugs used (changes in concentrations, injections, etc), and main data regarding the surgical act (at least start/end of surgery).

This phase is a lonely one for the anaesthesiologist. In case there was some help during the induction phase, the second anaesthesiologist has left the operation room shortly after the maintenance has started. Moreover, other nurses are usually more concerned by the surgical act than by the anaesthesia.

The vigilance and activity level of the anaesthesiologist may vary, during the different phases of the surgical act, but also from one patient to another one. For critical cases (*i.e.* high ASA classes), the anaesthesiologist will anticipate more on what could go wrong. In general, one could say that there is a constant need of anticipation; for instance: the inertia of the body's reaction to drugs requires proper temporal model. The anaesthesiologist also needs to follow the surgical act, as this information will be used in order to anticipate the beginning of the recovery phase [de Keyser & Nyssen 1993].

7.5 Recovery Phase

At this stage, surgery is about to be finished and a secondary anaesthesiologist has often joined the team. Anaesthetic gases have already been stopped, by anticipation. When recovery can actually start, the antidotes will be injected, especially in order to reverse the effects of muscle relaxants. By the end of the recovery phase – the patient still being unconscious –, the anaesthesiologist extubates the patient just before this one wakes-up. The main preoccupation of the anaesthesiologist at this moment is that the breathing function becomes natural again.

This transitory phase is complex for different reasons. Firstly, as we said, the patient needs to breathe on his own again. Secondly, the surgery being over, nurses will start cleaning up the patient, for instance by taking away compression points. In fact, since this is a more complex task with more simultaneous things to be done, the anaesthesiologist is often assisted, as in the first two stages. The division of tasks follows a traditional schema, so both the anaesthesiologist and the assistant know in advance who will do what. Normally, the anaesthesiologist and the assistant only discuss about sharing tasks when there is a need to modify the traditional division of labour, for instance when one of them wants to practice specific actions.

Here once again, the anaesthesiologist needs to enter data in the record, typically restricted to factual data such as time of extubation, injections, etc. Thanks to the anticipation of the anaesthesiologist, this recovery phase lasts approximately 10 minutes. It is the responsibility of the anaesthesiologist to decide when the patient is conscious enough – *e.g.* to reply orally – and can thus actually leave the operation room, to be handed over to the recovery room.

7.6 Post-Operative Phase

After a general anaesthesia, vital signs will continue to be monitored and reported in the AR. Together with the patient, the anaesthesia record is transferred to the recovery room. So far, the patient record contains both preoperative data as well as the AR with a description of what happened during the operation. Furthermore, the anaesthesiologist has put some specific comments into the AR, which will be used by the nurses in the recovery room to know how to handle the patient.

7.7 The case of crisis situations

In the previous description, we have not discussed the case of crises that may occur under a planned operation. Even if we have not yet observed such crisis situations, we know from interviews that under those circumstances, filling up the record has a low priority to the eyes of the anaesthesiologist. Even though this level of prioritisation decreases, a well-informed record is, in these cases, even more important. Indeed, these abnormal situations are the most interesting ones to analyse and comprehend. From that particular standpoint, records that are properly filled out are important. Not only it is interesting afterwards, but also during the operation itself; as crises are very demanding, the anaesthesiologist needs to take complex decisions that require good supports. In such a case, it is especially important to clearly see the relations between the vital signs, the drugs administrations and other undertaken actions. Being able to link these two sets of data should enable better decisions to be made.

8 A focus on timely constrained phases

Returning now to the study of the implication of the new tools, we can easily identify two major categories of impact. The first category is directly linked to the new modality: how does the use of this new modality influence the concurrent activities? The second category is linked to the product one wishes to obtain thanks to this modality: how can a better-filled record affect the upcoming activities. In order to analyse the potential impacts of the new interface, we assume that the vocal modality is used as intended, that is to say, data is recorded more or less in real-time.

8.1 Implications on concurrent activities

8.1.1 On the anaesthesiologist himself

During the induction phases, but also during recovery, the anaesthesiologist is quite active physically, and to record data by talking is yet another simultaneous task. On the one hand, one could argue that this could increase the workload of the anaesthesiologist, but the new task actually consists of verbalising current activities, that is to say, no new task is created that would be independent from the existing ones. On the other hand, the anaesthesiologist's self-consciousness might be improved.

During the maintenance, which does not require a lot of physical activity, the vocal modality is less needed. Regarding the impacts of the new technology on the anaesthesiologist himself/herself, differences with induction and recovery phases are minor, at least qualitatively.

8.1.2 On the interactions with other medical staff

During high-activity phases, the anaesthesiologist's audio channel might be less receptive to others. There is thus a potential impact on the communication from other staff to the anaesthesiologist. During these phases, the most probable person to interact with is the second anaesthesiologist. Nevertheless, oral communication between these two persons is kept low, at least under normal circumstances.

The other potential negative impact is on how others pay attention to the anaesthesiologist's talk. We have two alternative hypotheses: either people will not listen anymore, or in the contrary they will listen to everything said, even what is not of interest for them. It is also important to pay attention to the impact on the work of other medical staff. Choosing an appropriate microphone can reduce the negative impact by allowing quiet dictations.

8.2 Implications of verbalisation

According to Ericsson & Simon [Ericsson & Simon 1984], three levels of verbalisation can be identified.

The first level refers to situations where it is a matter of saying loud something without transformation, such as numbers or words displayed on monitors. This kind of verbalisation is very reliable and increases the cognitive load very slightly.

The second level requires creating dedicated sentences, such as a description of the patient's skin, or about the basic action that have been done. This is considered reliable, even if it increases a little the workload.

The third level, which is considered less accurate, implies additional cognitive processing, as it is about giving opinions, making inferences and filtering or using long-term memory. This is the case when reporting diagnosis, or reasons of specific past actions. This kind of verbalisation reduces the speed of the main task, and they are especially difficult when related to automatic actions with little consciousness, which are common for expert users.

The implications of verbalisation with speech recognition facilities will likely vary according to the level of consciousness of each reported fact. As people will naturally need to check if they are understood, an appropriate feedback is needed to limit the distraction.

8.3 Implications on upcoming activities

If voice facilities are deployed successfully, time should be saved for more careful monitoring, and the better quality of the EAR should support more effective diagnosis and actions.

8.3.1 Long term effects (of correct use)

During transition phases such as when a new actor is joining the medical team, there is the risk that a detailed EAR will lead to less communication, as the needed information will be wrongly taken for granted.

8.3.2 Long term potential drifts in the usage of the tool

There is a risk that new secondary tasks, not directly related to anaesthesia, will be assigned to the EAR, like recording more about the surgery act.

9 Developing the voice interface

In this section, we propose a methodology to build the first prototypes needed to answer the questions described above.

Based on action research, the development will be an iteration of prototypes and experimentation. The first step is to establish task requirements and user needs. A set of spoken commands has to be defined to control the system, such as to navigate between the different parts of the patient record. More generally, the phraseology – the way to speak to the system – has to be established before trying to implement it into a voice engine.

9.1 From natural speech

9.1.1 Part1

The first experiments are aimed to gather how anaesthesiologists would spontaneously express themselves to orally fill in an anaesthesia record. Experiments are conducted with minimum guidance, so they have a lot of freedom. Scenarios from anaesthesia simulation training can be used. A nurse or an anaesthesia secretary simply writes down what is being said by the anaesthesiologist for the anaesthesia record. After having done that with at least two different anaesthesiologists and the main types of anaesthesia, a nurse who has not participated in this scenario can try to fill out the anaesthesia record according to what has been written down. This will hopefully give a list of the main problems, such as ambiguities and contradictions.

9.1.2 Part2

Based on results from the first part, another set of simulations can be done. This time, anaesthesiologists receive a set of instructions and some guidance, to say their indications with less ambiguity, and to try not to forget important fields. In particular, anaesthesiologists start using keywords to make a difference between normal conversation and sentences targeted to the AR.

As a fallback alternative, it is possible to use a push-button to enable the speech recognition.

9.1.3 Part3

At this point, one can start establishing a phraseology, *i.e.* a set of rules about how to formulate the needed sentences, and how to process what has been said. Some discussions with various nurses, physicians, etc. are needed, as well as expertise from senior specialists.

9.2 Using list of existing fixed comments

There are lists of fixed comments that are currently used in EAR, and selected from a drop-down list on the touch-screen. They can be used as a starting point.

9.3 Wizard of Oz experiments

The next steps can be done by a succession of "Wizard of Oz" experiments: the speech recognition and the text entry are done by humans, perhaps a secretary. Such a testing is common with speech recognition applications in the early stages of design. This involves a human to play the part of the speech recognition computer, as a way of testing design prototypes before any actual programming is done. Most of the theoretical issues can be studied at this step. Then, the different tasks are progressively implemented in the computer.

9.4 Towards a full phraseology

New simulations will be conducted, keeping in mind that the computer cannot achieve the level of intelligence and expertise of a human, so most of the things have to be explicitly described, with phonetically distinct expressions. This time, the anaesthesiologist will try to conform to the phraseology when entering orally something in the anaesthesia record. Experiments and modifications of the phraseology will be made in loop, until finding a set of rules convenient for the anaesthesiologist and understandable by a machine.

9.5 Prototyping

The phraseology is then tested in a normal room, against an early prototype of speech recognition system, to be disambiguated, simplified and modified to improve the accuracy of recognition. Tests in anaesthesia simulators can then start. Volunteers try to address a fictive system during a normal simulation. Feedback tests should be made, in order to try various acknowledgment solutions for the recognitions, and interfaces as alternatives and complements to voice input. With the same kind of Wizard of Oz technique as before, a technician can remotely modify the screen to simulate an output from the computer.

Conclusion

This paper has presented a discussion and extracted a set of notions about introducing and testing voice-based electronic anaesthesia record. This can be used during the development and to evaluate integration tests of the new product, but also in a longer term. It illustrates the fruitfulness of collaborative efforts of engineers, sociologists and ergonomists early in the development process.

References

[Beuscart-Zéphir *et al.* 2001] M.C. Beuscart-Zéphir, F. Anceaux, V. Crinquette, J.M. Renard. Integrating user's activity modelling in the design and assessment of hospital electronic patient records: the example of anesthesia. *International Journal of Medical Informatics*, 2001, 64:157-171. doi:10.1016/S1386-5056(01)00210-6

[Chopra *et al.* 1992] V. Chopra, J.G. Bovill, J. Spierdijk, Floor Koornneef. Reported significant observations during anaesthesia: a prospective analysis over an 18-month period. *British Journal of Anaesthesia, 1992, 68:13-17.*

[Gauthereau 2004] Vincent Gauthereau. Emergent Structures in Drug Dispensing to inpatients: implications for patient safety. *Cognition, Technology and Work, 2004, 6(4):223-238.* doi:10.1007/s10111-004-0152-4

[Devine *et al.* 2000] Eric G. Devine, Stephan A. Gaehde, Arthur C. Curtis. Comparative Evaluation of Three Continuous Speech Recognition Software Packages in the Generation of Medical Reports. *Journal of the American Medical Informatics Association. 2000,7(5):462–468.*

[Ericsson & Simon 1984] K.A. Ericsson & H.A. Simon. Protocol analysis: Effects of verbalisation. *MIT Press (MA, USA) 1984.*

[de Graaf *et al.* 1997] P.M. de Graaf, G.C. van den Eijkel, H.J. Vullings, B.A. de Mol. A decision-driven design of a decision support system in anesthesia. *Artificial Intelligence in Medicine, October 1997, 11(2):141-53.*

[Gravenstein 1989] J.S. Gravenstein. The uses of the anesthesia record. *Journal of Clinical Monitoring, 1989, 5:256-265.* doi:10.1007/BF01618258

[Hamilton 1990] William K. Hamilton. Will we see automated record keeping systems in common use in anaesthesia during our lifetime? *Journal of Clinical Monitoring, 1990, 6(4):333-334.*

[Jungk *et al.* 2000] Andreas Jungk, Bernhard Thull, Lutz Fehrle, Andreas Hoeft, Günter Rau. A case study in designing speech interaction with a patient monitor. *Journal of Clinical Monitoring and Computing, 2000, 16:295-307.* doi:10.1023/A:1011456205786

[de Keyser & Nyssen 1993] V. de Keyser & A.S. Nyssen. Human errors in anesthesia. *Le Travail Humain, 1993, 56(2-3):243-266.*

[Lave 1993] Jean Lave. The practice of learning. *S. Chaiklin & J. Lave. Understanding practice: Perspectives on activity and context, 3-32. Cambridge (UK) University Press 1993.* ISBN:0521558514.

[Nardi 1996] B.A. Nardi. Context and Consciousness: Activity Theory and Human-Computer Interaction. *Cambridge MA, London: MIT Press 1996.*

[Pronovost *et al.* 2003] Peter Pronovost, Brad Weast, Mandalyn Schwarz, Rhonda M. Wyskiel, Donna Prow, Shelley N. Milanovich, Sean Berenholtz, Todd Dorman Pamela Lipsett. Medication reconciliation: a practical tool to reduce the risk of medication errors. *Journal of Critical Care, 2003, 18(4):201-205.* doi:10.1016/j.jcrc.2003.10.001

[Sanjo *et al.* 1999] Yoshimitsu Sanjo, Tetsuo Yokoyama, Shigehito Sato, Kazuyuki Ikeda, Reiko Nakajima. Ergonomic automated anesthesia recordkeeper using a mobile touch screen with voice navigation. *Journal of Clinical Monitoring and Computing, 1999, 15:347-356*. doi:10.1023/A:1009972223750

[Schmitz & Weiss 2004] Achim Schmitz & M. Weiss. Are we going to talk with our anaesthesia monitors in the future? *Acta Anaesthesiologica Scandinavica, 2004, 48(2):255-256*. doi:10.1111/j.0001-5172.2004.00295b.x

[Smith *et al.* 1990] N. Ty Smith, Robin A. Brien, Daniel C. Pettus, Brian R. Jones, Michael L. Quinn, Andrew Sarnat. Recognition accuracy with a voice-recognition system designed for anaesthesia record keeping. *Journal of Clinical Monitoring, 1990, 6(4):299-306*.

Transition 2

By presenting the previous article at EACE'2005, the annual conference of the European Association of Cognitive Ergonomics[1], we sought comments from experts in ergonomics in the medical domain.

The comments by the reviewers and then by the participants during and after the conference were mostly positive. The main reserve was some scepticism regarding the use of speech recognition in an operation theatre, given the diversity and level of background noise. Another concern was the capacity of anaesthesiologist to dictate entries in the anaesthesia record while working. Finally, there were doubts on the acceptance of the vocal modality by anaesthesiologists, but also by the rest of the medical team, such as the surgery team, that could be disturbed by some dictations.

This external input was enriching and helpful in refining upcoming research plans. Consequently, the rest of this thesis is devoted to the clarification of the points raised above.

The first point addressed in the next paper [Alapetite 2006] is the impact of background noise found in operation room, its direct effects on speech recognition by altering the audio channel and indirect effects by affecting users. Since the technologies are evolving, and speech recognition accuracy generally tends to become better, the impact of background noise should also be studied relatively to other factors known to affect speech recognition. Strategies to cope with background noise are finally tested to conclude on this first point.

The first preparations for the next part started in November 2005, with the experiments carried out from January to February 2006 and with an analysis phase until April 2006.

[1] [http://eace.info]

Impact of noise and other factors on speech recognition in anaesthesia

Alexandre Alapetite[1,2]

1. Risø National Laboratory; Systems Analysis Department; Research Programme Safety, Reliability and Human Factors; P.O. Box 49; DK-4000 Roskilde; Denmark
2. Roskilde University; Computer Science; Universitetsvej 1; P.O. Box 260; DK-4000 Roskilde; Denmark

This section is the long version of the following journal article:

Alexandre Alapetite. Impact of noise and other factors on speech recognition in anaesthesia. *International Journal of Medical Informatics, available online December 2006.* doi:10.1016/j.ijmedinf.2006.11.007

Abstract

Introduction: Speech recognition is currently being deployed in medical and anaesthesia applications. This article is part of a project to investigate and further develop a prototype of a speech-input interface in Danish for an electronic anaesthesia patient record, to be used in real time during operations.

Objective: The aim of the experiment is to evaluate the relative impact of several factors affecting speech recognition when used in operating rooms, such as the type or loudness of background noises, type of microphone, type of recognition mode (free speech versus command mode), and type of training.

Methods: Eight volunteers read aloud a total of about 3 600 typical short anaesthesia comments to be transcribed by a continuous speech recognition system. Background noises were collected in an operating room and reproduced. A regression analysis and descriptive statistics were done to evaluate the relative effect of various factors.

Results: Some factors have a major impact, such as the words to be recognised, the type of recognition, and participants. The type of microphone is especially significant when combined with the type of noise. While loud noises in the operating room can have a predominant effect, recognition rates for common noises (*e.g.* ventilation, alarms) are only slightly below rates obtained in a quiet environment. Finally, a redundant architecture succeeds in improving the reliability of the recognitions.

Conclusion: This study removes some uncertainties regarding the feasibility of introducing speech recognition for anaesthesia records during operations, and provides an overview of several parameters that are traditionally studied separately.

Original work

What was known before the study:

- Speech recognition is increasingly used for anaesthesia related applications (pre- and post-anaesthesia) and is now envisaged for real time use during operations.
- Background noise reduces speech recognition accuracy and there are various types of loud noises in an operating room.
- Several other factors have an influence on speech recognition rates, such as the type of microphone, participants, the type of training and recognition, etc.
- There are various known possible strategies to improve speech recognition rates.

What the study has added to the body of knowledge:

- The impact on speech recognition of various types of noises collected in an operating room has been measured.
- The relative effect of factors influencing speech recognition rates has been evaluated.
- A simple but original architecture has been tested in which two recognition engines and two microphones are used at the same time. This approach is especially interesting for safety critical applications such as real time medical applications.
- The author believes this is the first paper to be published about an experiment using a commercial speech recognition system in Danish.

Keywords

Anaesthesia; Anesthesia; Electronic medical records; Voice recognition; Speech recognition; Noise; Error rates

1 Introduction

This paper reports some preliminary experiment about the effects of various background noises in the hospital operating room (OR) environment on speech recognition. The envisaged audio interface would supplement existing electronic anaesthesia record systems with voice input facilities during the operation. This work is part of a project seeking to investigate [Alapetite & Gauthereau 2005] and further develop a prototype of such a system in Danish.

During the experiment, eight participants read aloud a corpus of typical anaesthesia comments to be transcribed by a continuous speech recognition system. The main goal of the study was to measure the respective impact on the recognition rate of various parameters, namely the type or loudness of background noises, the type of microphone (headset or handheld) and the type of recognition mode (free speech versus command mode). Additional parameters were also investigated, including the type of training (with or without background noise), the evolution of the performance over the sessions (learning effect, fatigue), and the gender of the participants. A logistic regression analysis was done to estimate the significance of each of the evaluated parameters.

As far as the author knows, this is the first study reporting the effect of background noises on speech recognition in Danish and the first to compare the relative impact of the above parameters, all known to separately affect speech recognition, but not yet studied in parallel. Finally, a redundant cross matching high-level architecture was tested and shown to improve recognition rates.

2 Methodology

In this part, I describe the methodology followed to conduct the experiments.

2.1 Preparatory work

To ensure the reproducibility of the background noises, it was decided to carry out the experiment in a laboratory rather than in the real-life context of a hospital OR.

2.1.1 Collecting background sounds

Although there are some noise databases available, I decided to collect some background noises, in order to be as realistic as possible. Some background noises were thus recorded in an OR (Herlev University Hospital of Copenhagen) during real anaesthesias with surgery and X-rays in November 2005.

This recording was done using a multi-directional microphone placed in the proximity of the anaesthesiologist, and recorded on a laptop computer (from now on called PC#1). The microphone was an omni-directional electret condenser microphone 1.7 kΩ model ECM-F8 from Sony Corporation (50 – 12 000 Hz). The laptop PC#1 was an IBM ThinkPad R32 type 2658 with an Intel Pentium 4m 1.6 GHz processor and 512 MB of memory under Microsoft Windows XP SP2. The open source Audacity[1] audio editor and recorder version 1.2.2 was used for the software part, recording in WAV PCM (Pulse Code Modulation) format at 44.1 KHz 16-bit mono channel.

[1] [http://audacity.sourceforge.net]

Simultaneously, an integrating sound level meter (from Brüel & Kjær, model 2225) was used to measure the peak level and fixed level in dB(A) of various sounds being recorded. The 60 s L_{eq}[2] in dB(A) was also calculated for the background noise made by the room ventilation. The measurements have been made from the place where the anaesthesiologist is usually standing, and by pointing the sound level meter toward the various sound sources.

This dB measure with an A-weighting (based on Fletcher-Munson loudness curves) is known not to be very consistent and not to reflect accurately the subjective loudness of all types of noise: dB(A) is mainly targeting pure tones, not too loud. It would probably have been better to use the standard ITU-R 468 noise weighting of the International Telecommunication Union, but it was convenient to use dB(A) since an integrating sound level meter in dB(A) was available at my laboratory.

2.1.2 Selecting samples

The collected sound files were edited and some samples of interest were selected, *i.e.* various isolated sounds for which I have some dB(A) information. Samples of the same type of noise were concatenated together to create longer sequences with the same type of noise. The nine sounds, or "background noises", were:

(1) "Silence": the laboratory background noise ~32 dB(A);

(2) "Ventilation1": the constant background noise in the OR, air conditioning and pulse beeps, 48 to 63 dB(A), slow measure 60 dB(A), peak 70 dB(A);

(3) "Alarms": a set of classic anaesthesia alarms using various tones, 57-68 dB(A), peak 80 dB(A);

(4) "Scratch": Velcro noise when opening anti X-ray suites 82 dB(A);

(5) "Aspiration": suction of saliva in the patient's mouth 65 dB(A);

(6) "Discussion": female voices, discussions between the surgeon 60 dB(A) and the nurse 70 dB(A);

(7) "Metal": various metallic clinks, 58 to 82 dB(A), peak 97 dB(A), this is the noise with the sharpest peaks;

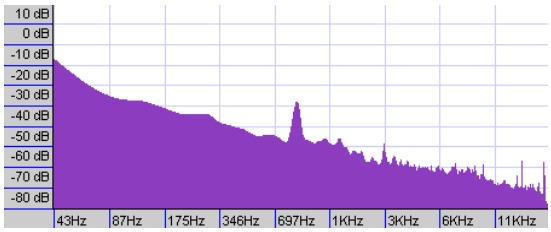
(8) "Ventilation2": Same as "Ventilation1" but 10 dB(A) louder, giving 61 to 73 dB(A);

(9) "Ventilation3": Same as "Ventilation1" but 20 dB(A) louder, giving 71 to 83 dB(A), slow measure 75 dB(A).

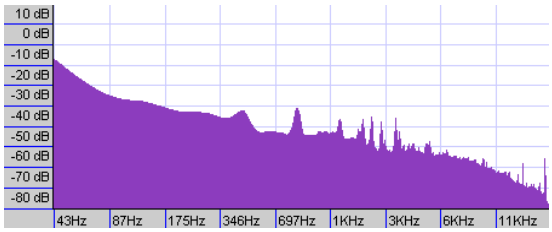
[2] L_{eq}: Equivalent continuous sound pressure

The level of the sounds 3, 7, 8 and 9 has only been measured when reproduced, using the same settings as other sounds reproduced.

Spectres of raw recorded background sounds (44.1 KHz, Hanning window 1024):

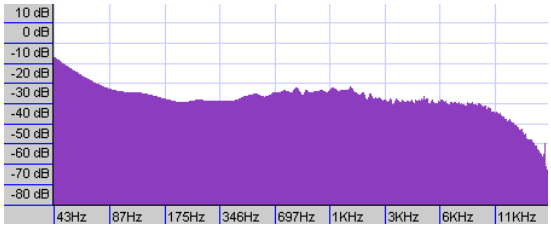


Spectre 2: ventilation 1. /

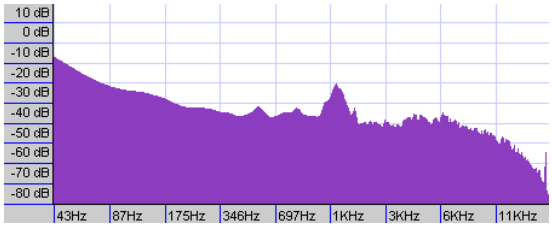


Spectre 3: alarms.

Frequencies below 50 Hz and above 12 KHz are not relevant due to microphones and speakers limitation. Spectre 2 reflects the constant background noises also present in all the following spectres. The pike at ~920 Hz is the cardiac pulse beep. On spectre3, the pikes from ~460 Hz to ~3 160 Hz are the various tones of the alarms.

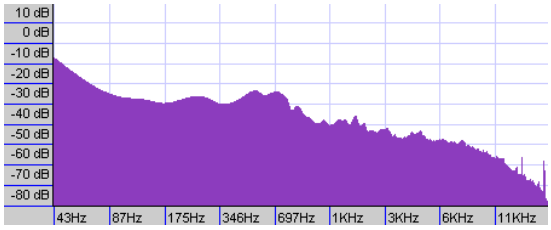


Spectre 4: scratch. /

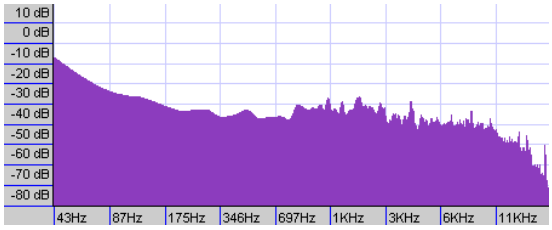


Spectre 5: aspiration.

Spectre 4 is the closest to white noise. On spectre 5, aspiration causes a pike at ~1 490 Hz.



Spectre 6: discussion. /



Spectre 7: metal.

Spectre 6 is influenced by female voices (typical fundamental frequency 165-255 Hz, plus harmonic series: 2×f, 3×f, …). Spectre 7 has many pikes caused by the various metal shocks.

2.1.3 Reproducing sounds

Samples were reproduced with a desktop computer (PC#3) plugged to an audio amplifier (Sony STR-GX290) with two loudspeakers (Jamo Compact 1000, 65 Hz-20 KHz, 90-120 W), positioned 1.5 meters apart and pointing toward participants about 2 meters away. This is similar to the distance from the anaesthesiologist to the noise sources in a real OR. The samples were played in a loop as long as needed.

2.1.4 Setting the volume

In order to replay the samples at the appropriate volume, the sound level meter was used again from the position where the participants would be sitting, pointing in the direction of the loudspeakers. The replay volume was adjusted to match as closely as possible the measured values in dB(A).

The distance between the hearing microphone and the various sounds has not been measured during the recording of the original background noises at Herlev hospital. This definitely useful information was however not crucial for the presented experiments, since the microphone was placed within proximity of the anaesthesiologist, and since the interesting value is the level of noise where the anaesthesiologist is standing, that is to say the point of measurement being used.

For sounds 2 to 7, I adjusted the replay gain both on PC#3 and on the amplifier in order to match as closely as possible the parameters in dB(A) that are the peak levels, fast levels and for some sounds the 60 s L_{eq}. I applied an amplification of +13 dB on the sound files, which was the maximal gain without saturation; the general and Wave volume settings in Windows staying at the maximum; I adjusted the volume on the amplifier until reaching the optimal level that was found at 5/10.

Sounds 8 and 9 were reproduced louder than reality, in order to see how the recognition rate evolves for one given background noise, when it becomes louder. For sound 8 “Ventilation2”, the volume on the amplifier was 6.3/10. For sound 9 “Ventilation3”, the software gain was +29.8 dB (just below saturation) and the volume on the amplifier was 6/10.

2.1.5 Subjective loudness

One problem I faced with these settings is that I subjectively perceived the reproduced sounds to be louder than what I recalled them to be when I recorded them. But it seemed to be an illusion, because the sounds were reproduced at the same decibel level as the original ones. This effect was particularly true with small desktop speakers, and was reduced a lot with the large Hi-fi speakers finally used.

I suspect this could be due to a combination of reverberation effects due to a different size and type of room [Gelfand & Silman 1979], of microphones and loudspeakers artefacts, and of physiological issues that make a sound appear much louder in a quiet environment than with a background noise such as the one found at the hospital. The subjective loudness scale is a known problem [Robinson 1957].

2.2 Experiments

2.2.1 Speech recognition software

The lab experiment was made with the speech recognition system Philips[3] SpeechMagic 5.1.529 SP3 (March 2003) and SpeechMagic InterActive (January 2005), with a package for the Danish language (400.101, 2001) and a "ConText" for medical dictation in Danish (MultiMed Danish 510.011, 2004) from Philips in collaboration with the Danish company Max Manus[4]. This medical module is not restricted to anaesthesia. The speech recognition workflow is the same as detailed in [Zafar *et al.* 2004].

For voice dictation in free speech mode, or "natural language", SpeechMagic is integrated with Microsoft Word 2003. At the time of writing this article, a similar speech recognition system was already in use and under further deployment at Vejle Hospital (Denmark), for pre- and post-operative tasks, but not during operations [Alapetite & Gauthereau 2005]. With this system, it is possible to record what is being said and to submit the WAV file for recognition afterwards; this was the process used for this experiment.

For voice commands, or "constrained language", SpeechMagic InterActive uses grammars [Giorgino *et al.* 2005] describing the set of possible commands. SpeechMagic InterActive comes with several examples and a software development kit to make *ad hoc* programs. Static grammars are written in the "Java Speech Grammar Format"[5] 1.0 (JSGF) that is in the Bachus Naur form, with some proprietary extensions. The grammar must contain the phonetic transcription of the terms used, in the "Speech Assessment Methods Phonetic Alphabet"[6] (SAMPA). It is possible to provide a list of alternatives, when a word or sentence can be pronounced in different ways. To assist in this operation, SpeechMagic InterActive contains a "Phonetic Transcriber component". Example for the Danish word for "point" with two alternatives:

```
punktum {PHONETIC="p O N t O m;p O N g t O m;";};
```

Since previous articles [Zaphar *et al.* 1999], Philips SpeechMagic has evolved, as it is now available in various languages (the present experiments have been made in Danish), is no longer batch only (*i.e.* documents can be navigated and corrected while dictated) and has an interactive mode combining free text and command mode.

---

[3] [http://speechrecognition.philips.com]

[4] [http://maxmanus.dk]

[5] [http://java.sun.com/products/java-media/speech/forDevelopers/JSGF/]

[6] [http://www.phon.ucl.ac.uk/home/sampa/]

2.2.2 Hardware

Two similar laptop computers were used, running identical software. The speech recognition engine was installed on PC#1. The whole system was duplicated to another laptop computer PC#2 using an image of the hard-drive with Symantec Norton Ghost[7]. PC#2 was a Dell Inspiron 1.6 GHz with 1 GB of memory. While it would have been better to use two identical computers for PC#1 and PC#2, there should be no difference for speech recognition since the processor speed was identical on both computers, both have enough memory, the free speech recognition is not made in real time, and microphones were using digital input through USB (universal serial bus) ports. USB connections were chosen for microphones, since the noise added when using the analog mini-jack input to the sound card of the laptop computers noticeably reduced speech recognition accuracy.

Two different microphones were employed, one per laptop, in order to evaluate the impact of these on the speech recognition quality. On PC#1, the microphone was a Philips SpeechMike Classic USB 6264[8] (Mic#1). This was the recommended model for the Philips SpeechMagic system. It is a Dictaphone-like device, held in one hand about 15 cm from the mouth. On PC#2, a headset microphone was used (Mic#2, ~2.5 cm from the mouth), model PC145-USB[9] from Sennheiser Communications (uni-directional, 80 – 15 000 Hz, -38 dB, ~2 kΩ). Sennheiser indicates that this model is suited for speech recognition. One of its earphones was removed, so that participants might hear the background noise properly and therefore be affected by the so-called "Lombard effect" [Lombard 1911]. This effect is the tendency to alter the voice in noisy environments, and is known to affect speech recognition performance [Hansen 1996].

Max Manus previously tested other models of Sennheiser headset microphones with good success. Another model, the PC120 using an omni-directional microphone, has been tried during the preparation of the experiments but due to its omni-directional nature, it gave not surprisingly poor results in presence of loud background noises.

[7] [http://symantec.com/sabu/ghost/ghost_personal/]

[8] [http://dictation.philips.com/index.php?id=1470]

[9] [http://www.oticon.com/eprise/main/SennheiserCommunications/com/Products/CNT05_VBLG?ProductId=PC145]

2.2.3 Experimental configuration

The experiment was made using the two microphones simultaneously; that is PC#1 and PC#2 ran in parallel, performing the same task but with two slightly different sound inputs due to the different positions and types of microphones. As visible on Figure 1 and Photo 1, the two laptop computers were on a desktop and the participant was sitting in front of them. The participant held the first microphone in one hand, and wore the second microphone as a headset. The loudspeakers (SP#1, SP#2) were two meters to the left of the participants. The two microphones were approximately at the same distance from the loudspeakers.

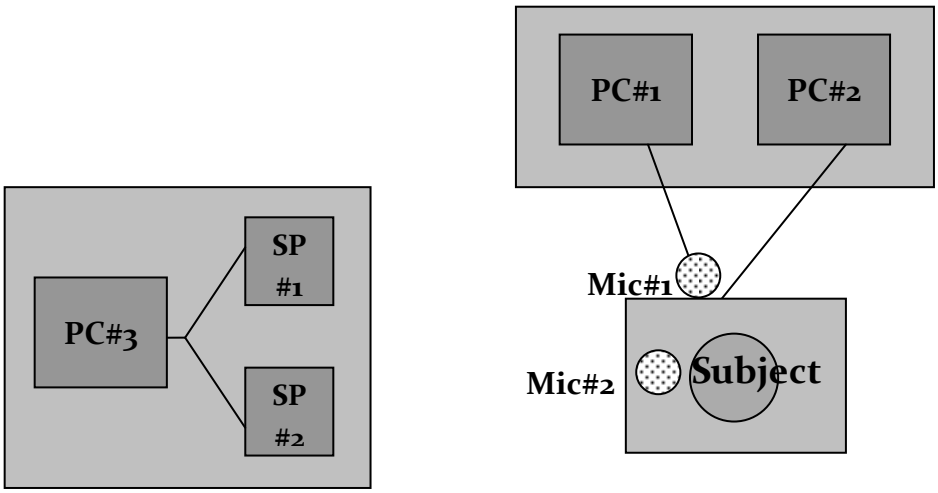


Figure 1: Experiments configuration.

2.2.4 Participants

Eight subjects participated in this experiment (4 males, 4 females, 27 to 62 years of age). The participants had no medical background. One of the participants had limited prior experience with speech recognition and the others had none. Prior to the experimental sessions the participants had the opportunity to familiarise themselves with the expressions and sentences to be dictated.

Photo 1: A participant during the experiments.

2.2.5 Test material

On 24 January 2005, I visited Køge Hospital (Denmark), where they use an electronic anaesthesia record system (Dräger Innovian[10]). I collected the list of fixed comments that are available thru their touch screen while the anaesthesia is running. I also took a transcript of the 600 most used comments out of the 12 009 different fixed or free comments typed by anaesthesiologists during the year 2004 (2004-01-01 to 2005-01-25). The distribution of frequencies is interesting: the most frequent comment was used 9 495 times, the 17th 1 135 times, the 43rd 105 times, the 146th 11 times, the 982nd 2 times and the rest only once.

The ~100 commands to be said during the experiments were taken from the top of this body of phrases, therefore covering most of the real life cases. While it is possible to easily add new words in the speech recognition software, the few unknown words were removed from the corpus, in order to avoid any bias due to variability when training new words for the various subjects.

The JSGF grammar for the command mode was especially developed to accept the selected comments from Køge corpus, followed by the Danish word for "point". The free text mode also accepts the Danish word for "point" as a special keyword that inserts a full stop. During dictations, each comment was followed by the Danish word for "full stop", so the same dictation could then be used both for command mode, and for a clean free text transcription.

[10] [http://www.draeger-medical.com/MT/internet/EN/us/Services/products/inform_tech/innovian/pd_innovian.jsp]

2.2.6 Training the speech recognition software

The Philips SpeechMagic system is speaker dependent and must thus be trained to recognise each speaker's voice. The enrolment phase was conducted with the configuration settings as described above. Each participant used the two microphones simultaneously and thereby trained the two computers PC#1 and PC#2 simultaneously. Training consisted of going through the training wizard, a module included in SpeechMagic. As the system learns every time it is used, especially when corrections are made, all the commands were then dictated once and corrected.

This training phase was done two times: once with a silent background ~32 dB(A), and once with the background noise "Ventilation1" collected at Herlev hospital (the automatic ventilation and the patient pulse ~60 dB(A)). Half of the participants trained first with the silent background and then the noisy one, the other half in opposite order. The first training was done before the recording sessions, while the second training was done after. The user profiles were deleted before the next participant – this in order to avoid most of the effects of the first training – and the system was set up not to improve its general model across users (called "context adaptation").

2.2.7 Dictation, recognition and transcription

During each session, each participant read a set of about 50 sentences. While speaking, the two computers worked in parallel, receiving the sound from their respective microphones. The computers did the recognition for the command mode in real time, and a text file containing the results was saved. The command mode was using the first profile only. Consequently, the command mode was done using a profile trained with background noise for half of the users, and using a profile trained in silence for the other half.

Simultaneously, each computer saved an audio file that was used afterwards for offline free text transcription, which produces a Microsoft Word document. The free text transcription was deferred to the end of the experiments, and that does not impact the quality of recognition, since this process is deterministic as soon as the sound file is digitalised and as long as the user profile is not modified. When the session was finished, the free text transcription was done twice, once with each of the two training profiles (with and without background noise).

2.2.8 Methodology memento

For each participant, there are nine sessions with various background noises. A session is composed of 50 sentences (± 1). In addition to the sessions, each participant trains the system twice: once in a quiet environment, once with background noise (two training profiles). The data thus comprise:

> 8 participants × 9 background noises × ~50 sentences ≃ 3 600 dictations

All dictations are in two audio files, recorded by the two microphones attached to PC#1 and PC#2. For each audio file, there are two recognition modes: the command mode based on a grammar and the free text mode using the medical context. The recognition in free text mode is done with both training profiles, while the recognition in command mode is done only with the first training profile (four participants with background noise, four without):

> ~3 600 dictations × 2 microphones × (1 command mode + 2 free text modes) ≃ 21 600 recognition samples

2.3 Statistics

The results and analysis presented below are based on descriptive statistics and regression analysis (Table 1, binary logistic regression where the dependent variable is binary: recognition is successful or not).

Work has been done with SQL queries under Microsoft Access 2003, graphs under Microsoft Excel, and regression analysis with SPSS[11] version 14.

2.3.1 Regression model

The core of the analysis done in this paper is a binary logistic regression model. Binary regression has been chosen in order to keep a high number of samples, instead of aggregating them to a percentage recognition rate. The regression model aims to show the relative impact of various parameters, or combinations of parameters, in a system where parameters are combined and difficult to isolate.

The model reported in Table 1 was obtained by testing many possible combinations of parameters and using the significance score to select the parameters. The presented model is thought to be representative of the experiments, both from a pure statistical point of view, and when taking into account the meaning of the data. None of the parameters or pairs of parameters left aside was significant when added to this model. In the following paragraphs, significance is noted ‘p’.

[11] Statistical Package for the Social Sciences [http://spss.com]

2.3.2 Recognition rates

There are various methods for calculating the recognition rate of any speech recognition engine, such as the percentage of words or sentences transcribed exactly as expected. One of the most commons is the word error rate (*cf.* section 2.3.2.1). Depending on the type of experiment, other metrics are sometimes used, such as the errors per word (EPW) [Sears *et al.* 2001].

2.3.2.1 Word recognition rate (WRR)

The word error rate (WER) or its complement, the word recognition rate (WRR) have limitations [McCowan *et al.* 2005] (*cf.* section 2.3.3), and it is only to facilitate comparisons with other articles that WRR will be reported for some of the results.

$$WRR = 1 - WER = \frac{N - (S + D + I)}{N} = \frac{N - L}{N}$$

Where *WER* is the word error rate, *N* is the number of words in the reference, *S* is the number of substitutions, *D* is the number of the deletions, *I* is the number of the insertions, *L* is the Levenshtein distance[12] [Levenshtein 1965] at the word level.

Here is the used PHP[13] implementation of the Levenshtein distance at word level:

```
function levenshteinWER($reference, $recognition)
{
 //Split the strings into words using a regular expression (PCRE[14])
 $words1 = preg_split('/\W+/',$reference);
 $words2 = preg_split('/\W+/',$recognition);

 //Initialise the matrices
 for ($i = 0; $i < count($words1) + 1; $i++) $d[$i][0] = $i;
 for ($j = 0; $j < count($words2) + 1; $j++) $d[0][$j] = $j;

 //Bottom-up dynamic programming
 for ($i = 1; $i < count($words1) + 1; $i++)
   for ($j = 1; $j < count($words2) + 1; $j++)
   {
     $cost = (strcasecmp($words1[$i - 1], $words2[$j - 1]) == 0) ? 0 : 1;
     $d[$i][$j] = min($d[$i - 1][$j] + 1, //deletion
                      $d[$i][$j - 1] + 1, //insertion
                      $d[$i - 1][$j - 1] + $cost); //correct or substitution
   }
 return $d[count($words1)][count($words2)];
}
```

[12] Wikipedia, The Free Encyclopaedia, "Levenshtein distance", revision 2 June 2006 10:26 UTC [http://en.wikipedia.org/wiki/Levenshtein_distance]

[13] PHP: Hypertext Preprocessor [http://php.net]

[14] PCRE: Perl Compatible Regular Expressions [http://pcre.org]

2.3.2.2 Concept-matching accuracy

In this paper, a semi-automatic measurement is favoured. This measurement is less impartial but more relevant to the targeted use: the percentage of sentences that can be understood "without ambiguity". The so-called "concept-matching accuracy" [Detmer *et al.* 1995; Jungk 2000] is considered more important than raw recognition accuracy. If a sentence is transcribed exactly as expected or with an alternate but correct spelling (*e.g.* "one" / "1") the sentence is accepted as a success (see "level 4" on Figure 3). If a sentence contains some mistake such as an incorrect plural mark (common in Danish speech recognition), the lack of a minor word (*e.g.* an article), or any alteration that does not prevent a skilled human reader from understanding its meaning without ambiguity, then this sentence is counted as a partial success.

Otherwise, sentences that are not recognised at all (deletions), sentences that were recognised while nothing was said (insertions) and sentences that cannot be understood are counted as failures. The final score is the percentage of accepted sentences (levels 3+4) compared with the number of sentences actually said plus the sentences wrongly inserted.

This method was decided before running the experiment, but had only a minor effect on the results since less than 2.4% of the samples are partial successes (only in free text mode, see "level 3" on Figure 3). It does not concern the command mode, which is either correct or wrong. Results at sentence level in this paper are therefore close to what they would have been without it.

2.3.3 Danish language

The natural language of this study was Danish, a language that, like German, joins compound nouns, which means that some words are glued together. For instance, "the general department" is written "stamafdelingen" so if "the child department" ("børneafdelingen") was recognised instead, that would give zero good recognition and one false recognition in Danish, but two good recognitions and one false recognition in English. This illustrates the fact that word error rate is less fair than command (sentence) error rate to compare recognition rates in Danish with those in English. Other metrics addressing variability in word length could be less sensitive to this problem, such as the errors per word (EPW) [Sears *et al.* 2001] where the number of words is calculated as the total number of letters divided by 5, but this is mainly suited for transcription tasks, at character level.

Furthermore, *"Danish has 21 monophthongs that are unevenly distributed in the vowel space, with a densely populated upper portion […]. British English, on the other hand, has only 11 monophthongs that are evenly distributed in the vowel space"* [Steinlen & Bohn 1999]. This makes Danish vowels, which in addition have long and short versions (total of 28) [Sobel 1981], potentially more difficult to distinguish than English ones, with a direct impact on current speech recognition engines that typically prioritise vowels because of the difficulty to recognise consonants and other voice parameters such as intonation. The context is also crucial in Danish, where many words differ very little phonetically, such as "department" ("afdeling") and "the department" ("afdelingen").

Additionally, since Danish is a relatively small language (~5.5M speakers), little research has been published about tuning speech recognition to its specificities (such as the glottal catch "stød").

Phonetic symbols in particular – which are used to transcribe sounds in speech recognition engines – have been shown to have small acoustic differences when articulated in English or in Danish [Steinlen & Bohn 1999]. The effect of this phonetic variation should however be minimal after proper training.

Finally, commercial recognition systems in Danish are still young (only the Philips/Max Manus system is on the market since 2002[15]), while research on speech recognition (in English) began in 1936 at AT&T's Bell Labs and first commercial products appeared in 1982 with Covost on personal computers[16].

Consequently, whatever the unit and for one same task, rates in Danish may be lower than in English anyway due to natural language differences. Experiments would be needed to confirm this assessment.

[15] North Denmark Invest , 2002-05-02,
[http://web.archive.org/20020903094647/http://www.northdenmark.com/frontpage/news.asp?idnr=63]

[16] Dragon Systems: A Timeline & History of Voice Recognition Software
[http://dragon-medical-transcription.com/history_speech_recognition_timeline.html]

2.3.4 Statistical methodology for the order effect

The raw recognition rates K_r for each rank cannot be used directly, as they are too much biased: since the order of sessions was chosen randomly, the sessions are not distributed equally among each rank and some ranks have more difficult sessions than others do. A correction factor F_r must therefore be used.

We know the overall recognition rate T (~78.54%) and the raw recognition rate per rank K_r ("rank" r from 1 to 10). Additionally, we know the averaged recognition rate R_s ("session" s from n=1 to 9, in percent) of each session across all the experiments, and the number of dictations N_{rs} in each session for each rank.

The first part F^1 of the correction factor F per rank r is the inverse of the averaged recognition rate R_s of the sessions involved at each rank, weighted by the number of dictations for each session N_{rs} done in this rank:

$$F_r^1 = 1 \div \left\{ \left[\sum_{s=1}^{s=n} \underbrace{\left(R_s \div 100\right) \times N_{rs}}_{\text{Recognition rate of this session type weighted by the number of dictations of this session type in this rank}} \right] \div \underbrace{\left[\sum_{s=1}^{s=n} N_{rs} \right]}_{\text{Total number of dictations for all sessions in this}} \right\}$$

Correction factor per rank

Sum for all session

By applying the correction factor F^1 to the recognition rate of each rank, one obtains a recognition centred on 100%, plus or minus the order impact of each rank.

To centre the recognition rate back to values closer to reality, we apply the averaged raw recognition rate of all the sessions, weighted by the number of dictation per raw that is also equal to the overall recognition rate T.

$$F_r = F_r^1 \times T$$

The corrected recognition rate K' per rank is obtained by applying the correction F_r.

$$K_r^{'} = K_r \times F_r$$

The corrected recognition rate K' can now be used to evaluate the order impact.

It would have been possible and even better to apply a correction factor taking into account the averaged difficulty of each sentence said, but the additional effort was not considered worse it.

2.3.4.1 Example of such a correction factor

Here is a simple fictive example illustrating the above described correction factor (see Figure 2). After applying the first correction F^1, the deviation from 100% is the impact of each rank. While this is enough to study the relative impact of each rank, the data is more palpable when recognition rates are provided. Therefore, the last part of the correction factor recentres the values on the average.

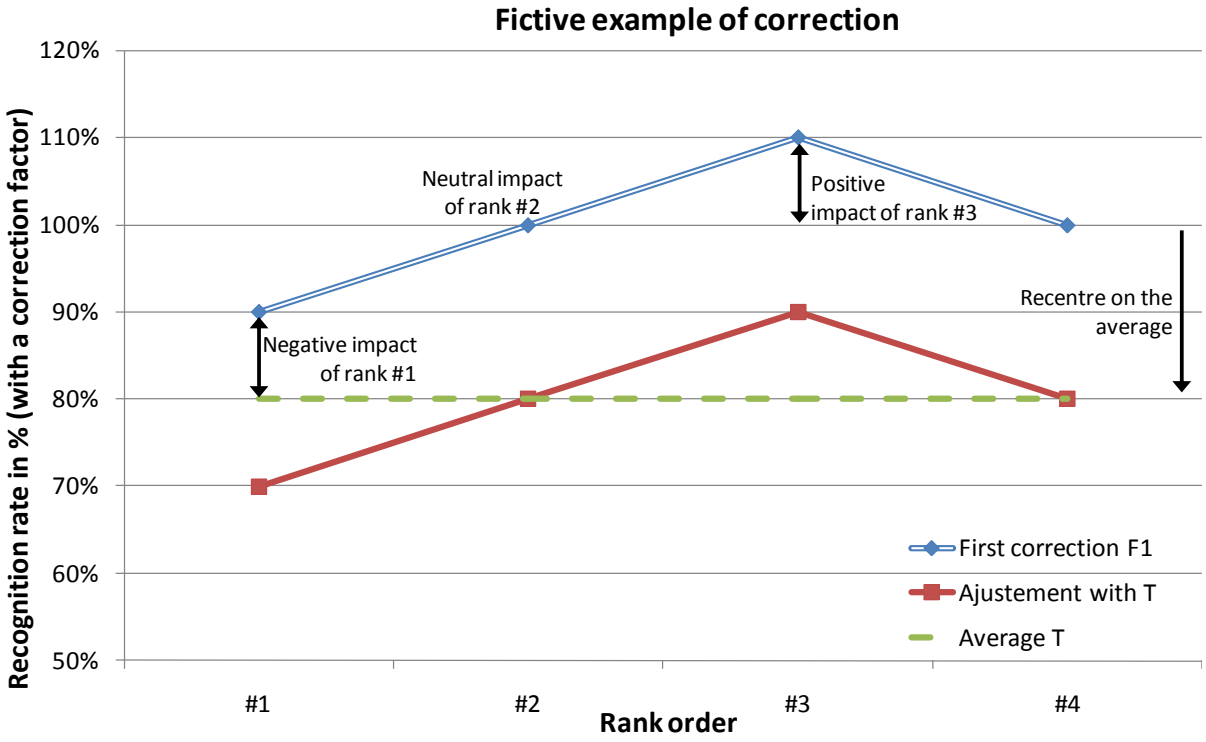


Figure 2: Correction factor.

3 Results

Figure 3 shows the percentages of recognition errors at sentence level, for free text and command mode, the two types of microphones, and overall. Results are discussed in details in the following sections.

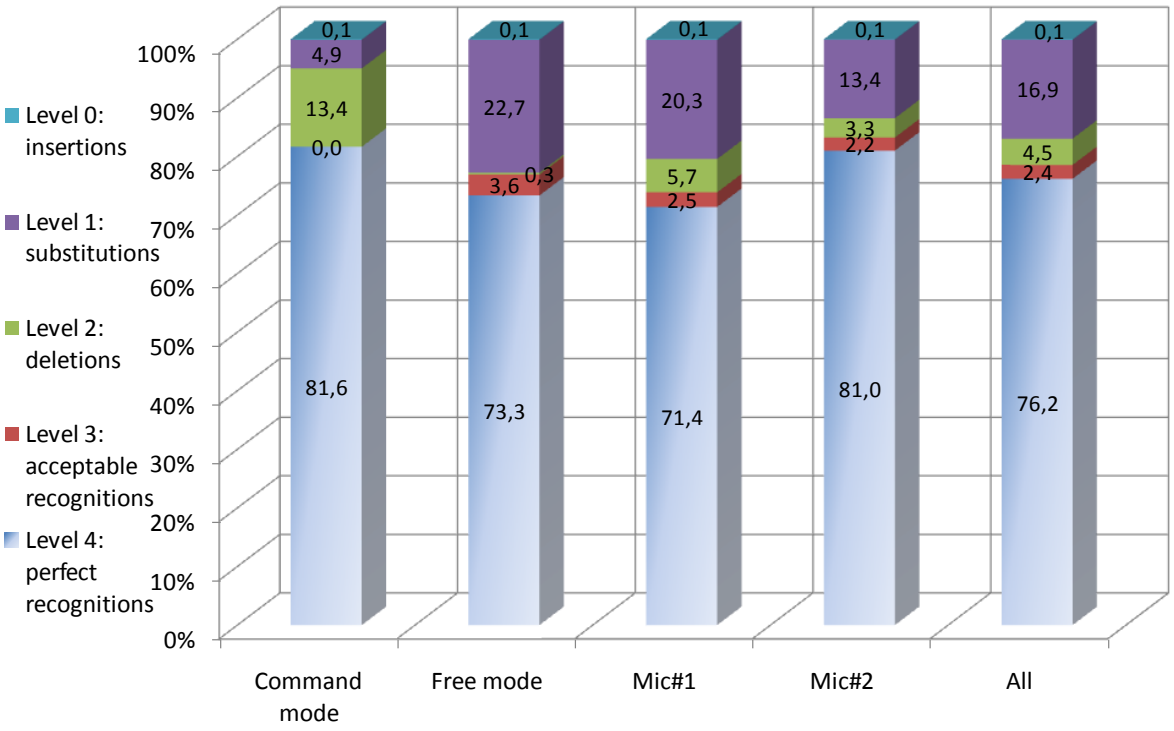


Figure 3: Classification of correct and failed recognitions (per sentence).

3.1 Regression model

In the resulting regression model, the types of microphone and recognition mode are used as combined parameters with other ones. Therefore, their significance must also be evaluated on the effects they have as combined variables. The same regression model excluding combined parameters shows both microphone and recognition mode as very significant ($p < 0.001$).

The regression model will be discussed in the subsequent chapters.

References in the following model: Microphone(0) is microphone 1, Mode(0) is the command mode, Training_with_noise(0) is training without background noise, Person_id(0) is a participant whose recognition rate was close to the average, and Session(0) is the session without background noise (silence).

Table 1: Binary logistic regression model: variables in the equation.

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	Mode(1) *Free text mode*	-.144	.114	1.588	1	.208*	.866
	Microphone(1) *Microphone 2*	.162	.119	1.845	1	.174*	1.176
	Training_with_noise(1) *With noise*	-1.039	.077	183.441	1	.000	.354
	Person_id *(Woman ~average)*			476.355	7	.000	
	Person_id(1) *Woman*	.178	.080	4.985	1	.026	1.195
	Person_id(2) *Man*	-.322	.075	18.335	1	.000	.725
	Person_id(3) *Man*	-.982	.071	193.092	1	.000	.375
	Person_id(4) *Man*	.264	.074	12.665	1	.000	1.302
	Person_id(5) *Woman*	.006	.070	.007	1	.932	1.006
	Person_id(6) *Man*	-.307	.065	22.554	1	.000	.735
	Person_id(7) *Woman*	-.625	.072	74.567	1	.000	.535
	session_id *(Silence)*			330.212	8	.000	
	session_id(1) Ventilation1	-.213	.117	3.305	1	.069	.808
	session_id(2) *Alarms*	-.503	.111	20.619	1	.000	.605
	session_id(3) *Scratch*	-1.520	.108	197.506	1	.000	.219
	session_id(4) *Aspiration*	-1.116	.107	109.486	1	.000	.327
	session_id(5) *Discussion*	-.751	.111	45.995	1	.000	.472
	session_id(6) *Metal*	-.475	.113	17.595	1	.000	.622
	session_id(7) *Ventilation2*	-1.051	.107	97.378	1	.000	.349
	session_id(8) *Ventilation3*	-1.414	.110	165.552	1	.000	.243
	session_order (*First session)*			146.413	9	.000	
	session_order(1)	.067	.125	.283	1	.594	1.069
	session_order(2)	.675	.139	23.492	1	.000	1.965
	session_order(3)	.762	.143	28.571	1	.000	2.143
	session_order(4)	1.310	.166	62.272	1	.000	3.705
	session_order(5)	.414	.142	8.469	1	.004	1.513
	session_order(6)	.727	.128	32.187	1	.000	2.069
	session_order(7)	.932	.130	51.699	1	.000	2.540
	session_order(8)	.585	.123	22.657	1	.000	1.794
	session_order(9) *Last sessions*	1.201	.154	61.207	1	.000	3.324
	Mode(1) by Training_with_noise(1)	1.214	.086	200.565	1	.000	3.368
	Mode * session_order			179.516	9	.000	
	Mode(1) by session_order(1)	-.070	.153	.209	1	.647	.932
	Mode(1) by session_order(2)	-.869	.161	28.976	1	.000	.419
	Mode(1) by session_order(3)	-.863	.162	28.523	1	.000	.422
	Mode(1) by session_order(4)	-1.737	.187	85.903	1	.000	.176
	Mode(1) by session_order(5)	-.524	.165	10.128	1	.001	.592
	Mode(1) by session_order(6)	-.745	.155	23.094	1	.000	.475
	Mode(1) by session_order(7)	-1.168	.152	58.844	1	.000	.311
	Mode(1) by session_order(8)	-.781	.147	28.040	1	.000	.458
	Mode(1) by session_order(9)	-1.541	.174	78.301	1	.000	.214
	Microphone * session_id			92.175	8	.000	
	Microphone(1) by session_id(1)	-.024	.164	.021	1	.884	.976
	Microphone(1) by session_id(2)	.038	.159	.057	1	.811	1.039
	Microphone(1) by session_id(3)	.778	.151	26.495	1	.000	2.176
	Microphone(1) by session_id(4)	.552	.154	12.890	1	.000	1.737
	Microphone(1) by session_id(5)	.533	.159	11.170	1	.001	1.704
	Microphone(1) by session_id(6)	.154	.161	.914	1	.339	1.166
	Microphone(1) by session_id(7)	.519	.154	11.365	1	.001	1.681
	Microphone(1) by session_id(8)	.908	.153	35.034	1	.000	2.480
	Constant	2.256	.135	278.933	1	.000	9.549

(a) Variables entered on step 1: Mode, Microphone, Training_with_noise, Person_id, session_id, session_order, Mode * Training_with_noise , Mode * session_order , Microphone * session_id .

*: Mode and Microphone are also used as combined variables; their effect is significant.

3.2 Microphones

As both microphones received the same material, it is possible to compare directly their average recognition rate. Microphone 2 (headset) has a higher recognition rate (83.2%) than microphone 1 (handheld, 73.9%), see Table 2 and Figure 3 (levels 3+4).

Table 2: Recognition rates per microphone.

Microphone * Is_recognised Crosstabulation

			Is_recognised		
			Bad recognition	Good recognition	Total
Microphone	pc1	Count	2940	8310	11250
		% within Microphone	26,1%	73,9%	100,0%
	pc2	Count	1886	9357	11243
		% within Microphone	16,8%	83,2%	100,0%
Total		Count	4826	17667	22493
		% within Microphone	21,5%	78,5%	100,0%

This advantage of microphone 2 is present for all sessions (*cf.* Figure 4). Part of this effect could be explained by the position of the microphones. Microphone 2 (headset, ~2.5 cm to the left of the mouth) is closer to the mouth than Microphone 1 (handheld, ~15 cm in front of the mouth).

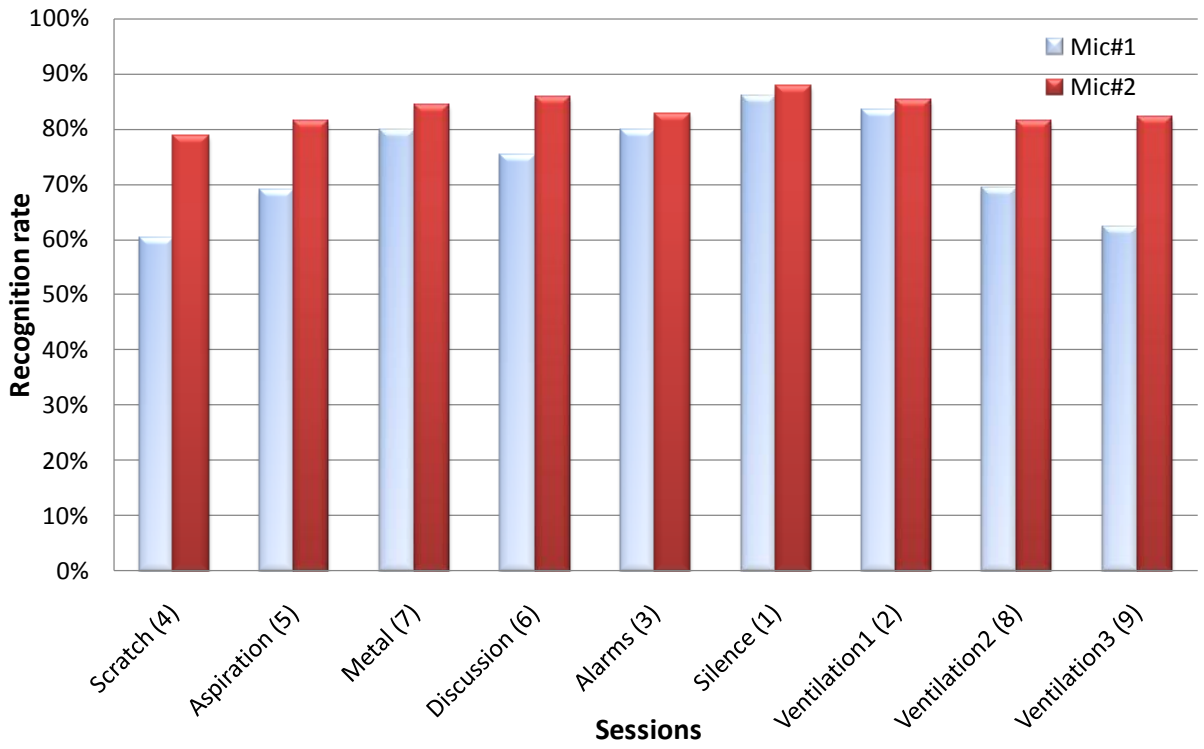
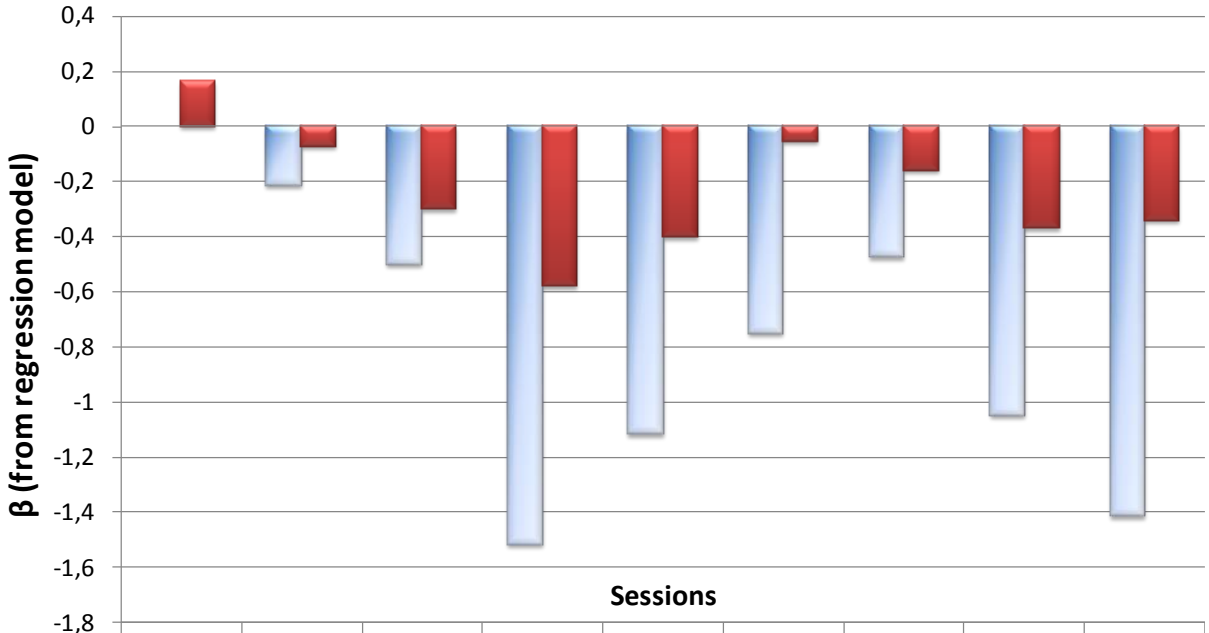


Figure 4: Recognition rates by microphones and by session types.

The regression model (Table 1) shows a significant difference for microphone type when combined with the type of background noise, as reported below (Figure 5); the combined effect of the type of microphone and the type of background noise is significant for most cases ($p < 0.001$).

While both microphones have similar recognition rates for silence and low background noise ("Ventilation1", "Alarms"), the advantage of microphone 2 becomes evident when the background noise gets louder ("Scratch", "Aspiration", "Ventilation2-3"). Microphone 2 is also less sensitive to a background with other people talking ("Discussion").



	#1	#2	#3	#4	#5	#6	#7	#8	#9
Mic#1	0	-0,213	-0,503	-1,52	-1,116	-0,751	-0,475	-1,051	-1,414
Mic#2	0,162	-0,075	-0,303	-0,58	-0,402	-0,056	-0,159	-0,37	-0,344

Figure 5: Effect of microphone type combined with session types (noises). The reference (0) is microphone 1 with a quiet background.

As visible on Figure 5, an analysis of the regression model using the combined variables *Microphone* and *Session* end up with the same trends as the above descriptive statistics. It tells that the combined effect of the type of microphone and the type of background noise is significant and more relevant than each of the two variables isolated.

The reference (0) is here the microphone 1 with a quiet background. We can see that Mic#2 performs always better than Mic#1, and that Mic#1 is much more impacted by strong background sounds than Mic#2. We can also see that the microphones are not equally impacted by the various types of background noises. For instance, Mic#2 is less sensitive to "Discussion" (6) than "Alarms" (3) while it is the opposite for Mic#1.

This contrasts with a recent study [Saastamoinen *et al.* 2005] that finds no significant difference between two types of microphone (unidirectional integrated to a Plantronics headset Audio 80, versus built-in omni-directional microphone of an Acer TravelMate 8000 laptop). A possible explanation may be that during the experiment reported in [Saastamoinen *et al.* 2005], some noises were mixed afterwards (*i.e.* not recorded simultaneously with the speech), and possibly not replayed at a sufficiently high volume.

3.3 Recognition mode (command versus free text)

With an average recognition rate of 81.6%, the command mode performed better than free speech mode (77.1%), as expected. Figure 6 shows that for some background noises, command mode performed considerably better than free text mode ("Scratch", "Aspiration", "Metal") and for some it is the opposite ("Ventilation2", "Ventilation1").

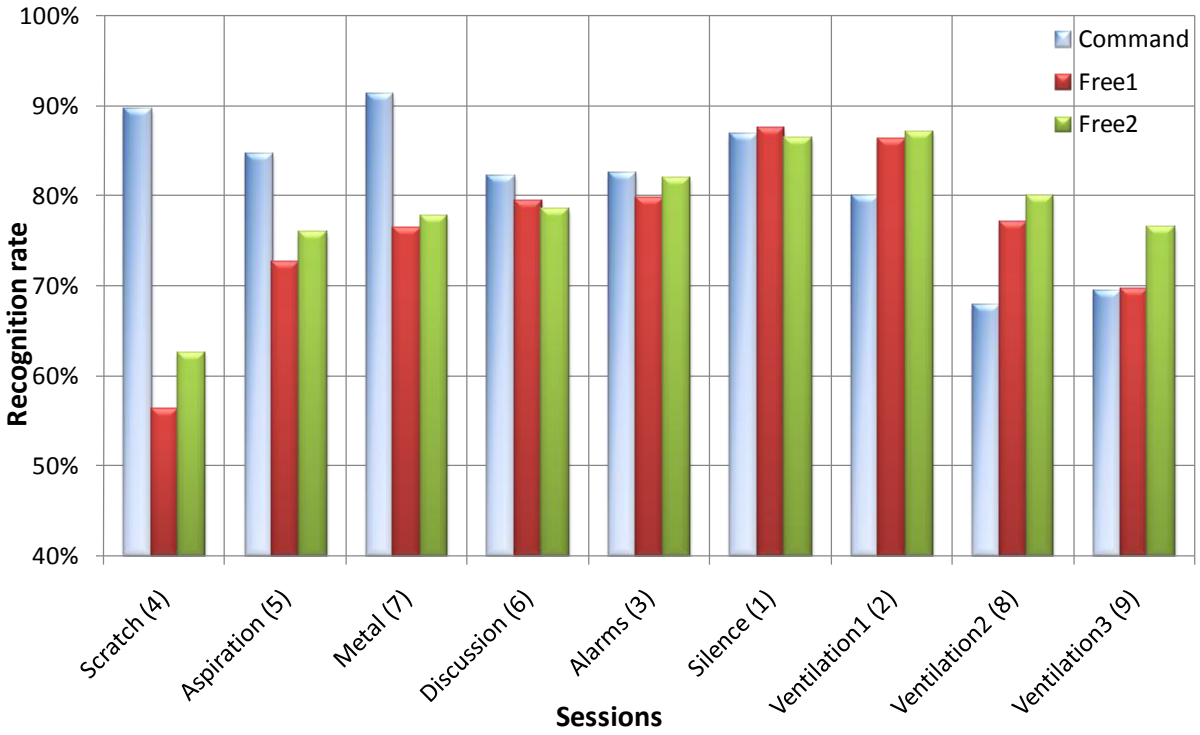


Figure 6: Recognition rates detailed per recognition modes and session types.

In command mode, false recognitions are rare (5%, insertions + substitutions + deletions). This is in agreement but adds some nuance to the results of [Gröschel *et al.* 2004] who had no false recognition at all during their experiments in German for out-of-hospital emergencies.

While the command mode had a better average performance than free text mode, there are some participants with an enormous difference in favour of the command mode (*e.g.* +23.2 points for Woman1, see Figure 7). In contrast, one participant shows the opposite effect (-12.55 points for Woman4).

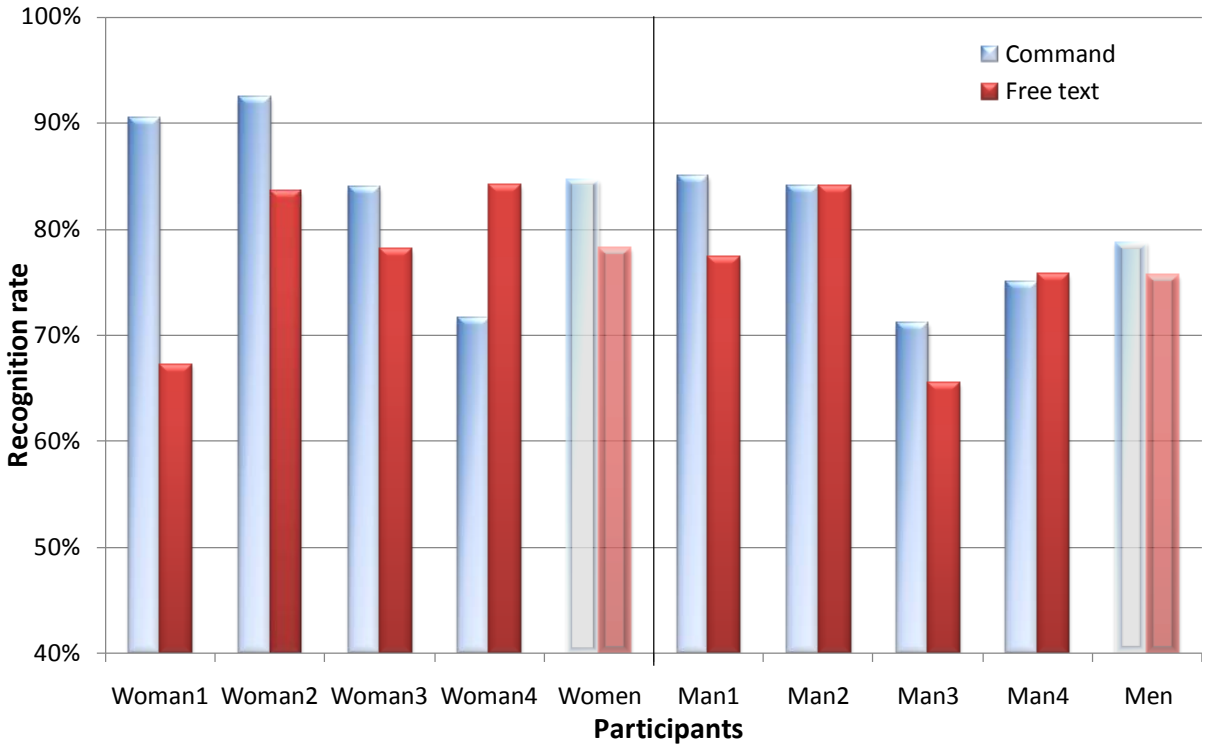


Figure 7: Recognition rates detailed per recognition mode and person (gender).

The regression model in Table 1 shows a significant effect of the recognition mode when combined with the type of training and the order of the sessions, $p < 0.001$ for most cases (the order of the sessions – see "time effect" on Table 5 – has only a very small impact).

3.3.1 Type of training: with or without background noise

Surprisingly, command mode trained without background noise performed better (85.5% recognition rate) than command mode trained with background noise (77.8%).

Table 3: Effect of training type combined with recognition mode.

β from regression model		Recognition mode	
		Command	Free
Training	Without noise	0	-0.144
	With noise	-1.039	0.031

This is confirmed by the regression analysis (Table 1 and 3); however, since in command mode there are only 4 participants for each type of training, this result should be treated with caution.

On average, free speech recognition performed a bit better when used with a profile trained with background noise (free2, 78.2% recognition rate) than when used with profiles without background noise (free1, 75.6%) (*cf.* also Table 3). The difference gets increasingly visible as background noises get louder ("Ventilation3", "Scratch").

As expected, in free text mode the best performances are achieved in silence with a system trained in silence (Table 4). When trained with background noise, the recognition rate is indeed lower for silent sessions. The second best performances are with a system trained with a given background noise in sessions with the same background noise. On other types of noises and at other levels of loudness, the system trained with background noise still performs better than the one trained in silence.

Table 4: Recognition rates in free text mode detailed per training type and sessions (noises).

Recognition rates	Free text mode	
Training / Session	Silence (Free1)	Ventilation1 (Free2)
Silence	87.53%	86.43%
Ventilation1	86.34%	87.09%
Other 7 sessions	74.09%	77.15%

These results are similar to previous studies [Hirsch & Pearce 2000]. A system using free text mode should therefore be trained with the type of background noise that will typically be present during use.

3.3.2 Confidence score in command mode

In command mode, a valuable indicator is the confidence score given by the speech recognition system for each recognised command. This score is between 0.0 and 1.0 and tells how confident the engine is that the command has been recognised correctly. The confidence score is especially valuable in settings where wrong recognitions may be dangerous and no recognition thus more desirable than a recognition that is likely to be wrong [Sears *et al.* 2003].

In the present experiment the confidence score was 1.0 in 5 813 (80.77%) of the 7 197 command-mode recognitions. For these 5 813 recognitions, the recognition rate was 98.16%.

As expected, the recognition rate falls rapidly when the confidence score gets lower (see chart 7). When the confidence is in [0.95, 1.0[the recognition rate is 59.78% (N = 179, see details on Figure 8), in [0.9, 0.95[the rate is 30.43% (N = 23), in [0.8, 0.9[the rate is 26.09% (N = 23), in [0.5, 0.8[the rate is 23.4% (N = 51), in [0.1, 0.5[the rate is 36.67% (N = 30), and with a confidence score below 0.1, the recognition rate is 22.81% (N = 114).

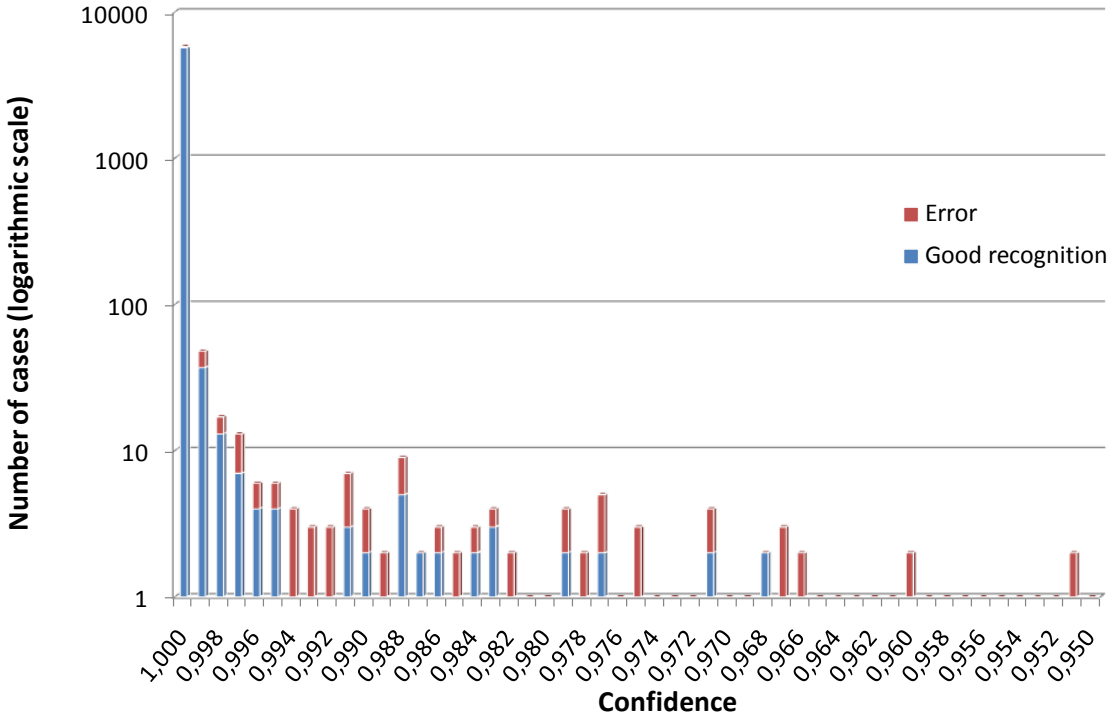


Figure 8: Recognition rate per confidence score.

The recognition rate decreases rapidly for lower confidence scores, showing that it can reliably be used as a threshold.

3.4 Background noises

As expected, decoding speech is less efficient in noisy conditions [Barker *et al.* 2005]. However, relative to the recognition rates obtained with a silent background (86.82%), the recognition rates obtained with "Ventilation1" are not significantly inferior (84.4%, β = -0.213 for mic#1, β = -0.075 for mic#2), see Figure 2. "Ventilation1" is the constant background noise observed in the OR environment and is also the one used when training with background noise (command2 and free2 modes). These results suggest that recognition rates in ORs may be close to the ones currently obtained in noise-free environments, provided no other type of noise intervene. This is in agreement with another study [Zafar *et al.* 1999], which reports that ambient noise (hospital ward, emergency room) had no effect on recognition accuracy.

The seven other types of background noises gave significantly lower recognition rates than the session with a silent background ($\beta \leq$ -0.475 down to -1.52 for mic#1, $\beta \leq$ -0.056 down to -0.58 for mic#2, with mic#1 in silence as reference). In Table 1, differences between most noises are significant ($p < 0.001$), also when combined with the type of microphone. The limited impact, for the best microphone, of people talking in the background is encouraging.

Ranking: The best recognition rate was with "Silence" ~32 db(A), followed by "Ventilation1" 48-63 dB(A), "Metal" 58-82 dB(A), "Alarms" 57-68 dB(A), "Discussion" 60-70 dB(A), "Aspiration" 65 dB(A) equally to "Ventilation2" 61-73 dB(A), "Ventilation3" 71-83 dB(A), "Scratch" 82 dB(A).

While the deleterious effect of background noises is to a large extent given by their loudness in dB(A), this can sometimes be misleading: "Metal" (slow measure 65-76 dB(A)) is louder than "Alarms" (slow measure 59-63 dB(A)) and nevertheless, "Metal" gives slightly better recognition rates (+1.5 points using microphone 2).

3.4.1 Background noise without speech

The speech recognition system comes with a customisable threshold intended to disable speech recognition when the microphone is not used. When the background noise gets louder, the threshold is eventually reached, enabling speech recognition even in cases when nothing is being said.

An additional experiment counting insertion errors has been made to illustrate this issue: a user profile was randomly chosen among the participants (it was a 27 years old female) and each of the nine types of background noises was produced for 1 minute, with the same experimental setup.

In the following Table 5 are reported the number of recognised commands per session (in 1 minute) together with their confidence score from 0.0 to 1.0, and the number of words recognised in free text mode. Those are insertion errors, *i.e.* something was recognised while nothing was said.

Table 5: Insertions errors when nothing is said.

Insertion errors	**Microphone #1**		**Microphone #2**	
Mode / **Session**	**Command**	**Free text**	**Command**	**Free text**
Ventilation1	0	0	0	0
Alarms	0	0	0	0
Scratch	8 (conf. ≤ 0.018)	0	8 (conf. ≤ 0.001)	0
Aspiration	5 (conf. ≤ 0.002)	29 words	0	10 words
Discussion	15 (conf. ≤ 0.994)	59 words	1 (conf. ≤ 0.311)	17 words
Metal	1 (conf. ≤ 0.001)	66 words	2 (conf. ≤ 0.000)	33 words
Ventilation2	0	38 words	0	0
Ventilation3	1 (conf. ≤ 0.148)	16 words	0	4 words

We can see on Table 5 that Microphone 2 always performed better than Microphone 1. In command mode, the recognised commands are, by nature, sentences allowed by the grammar, but their confidence was always low (conf. ≤ 0.994) and often very low, making most of them easy to discard. In free text mode, recognised words never formed a complete intelligible sentence.

During the experiment with participants, there were also some insertions errors (at sentence level). In command mode, the confidence score was never higher than 0.025 (N = 2) for microphone 2 but reached 0.975 for microphone 1 (N = 5). In free text mode, it is harder to tell due to alignment issues, but there were at least ~9 insertion errors for microphone 1 and ~5 for microphone 2. For a given background noise, it seems that there are fewer insertion errors when something is actually being said.

3.5 Participants

While women performed on average better than men (+3.5 points), the gender of the participants cannot be considered due to the high inter-subject variability ($p < 0.001$, 18.1 points, see Figure 7 and 9). A previous study [Mohr *et al.* 2003] reports inter-subject variability as large as 40 points (55 to 95% word accuracy) for 39 endocrinology authors.

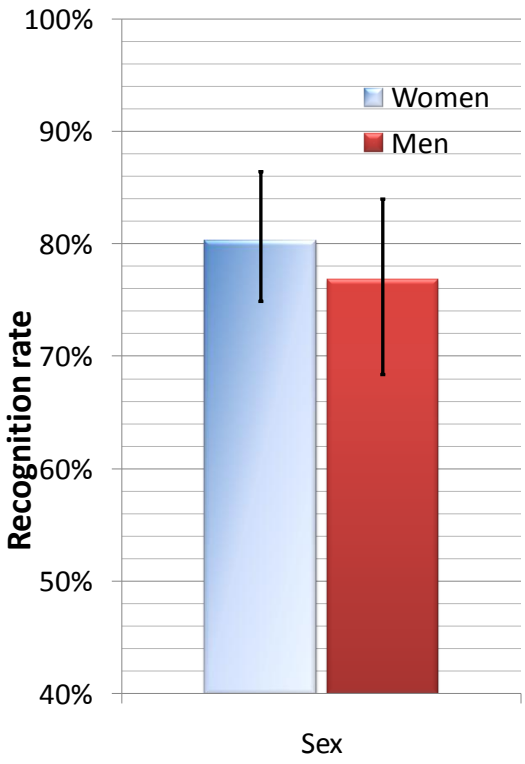


Figure 9: Recognition rate per sex.

The variability inter-subject was indeed very significant ($p < 0.001$) between some of them, even when taking into account other parameters such as the type of background noise or the order of training. Consequently, a couple of participants have a negative impact on the overall recognition rate. Without the lowest man and lowest woman, the recognition rate is increased by 2.2 points to 80.75% (N=17032). The age of the participants was not found to have any consistent impact.

There is not enough data to assess whether women perform better with some types of background noise, while men perform better with other types of background noise, although this is quite likely since men and women voices are not precisely in same frequency range.

3.6 Test material

The experiment has shown a very high inter-command variability of the recognition rate ($p < 0.001$ between many of them, even when taking a reference close to the mean).

As reported on Figure 10, the distribution of the recognition rate across the 108 different commands is interesting: while the best recognised command reaches a recognition rate of 97.7% (N = 218) (Danish word "tandskade"), the 31st command is below 90%, the 71st < 80%, 89th < 70%, 92nd < 60%, 95th < 50%, 101st < 40% and the 108th and last reaches a recognition rate of 13% (N = 198) ("lokal anæstetika"). Only 18% of the commands have a recognition rate below 70%.

It should be noted that the participants were not best suited to use this test material. Most of the commands were indeed specific to the medical domain, and many words were unknown to the participants. Even though the participants got time to familiarise with the corpus, there was some hesitation, mistakes and variability in their pronunciation that with no doubt impacted the system quite a lot. However, this does not affect too much the main goals of this work, which is to select the parameters to take into consideration during the next development and analysis phases.

This shows the importance of carefully designing grammars, by choosing words that are easily recognisable for the various users of the speech recognition engine and sufficiently distant from each other phonetically to avoid misrecognitions. "*Vocabularies should be natural to the task and sufficiently distinct to ensure recognition with few substitution errors*" [Pallett 1985].

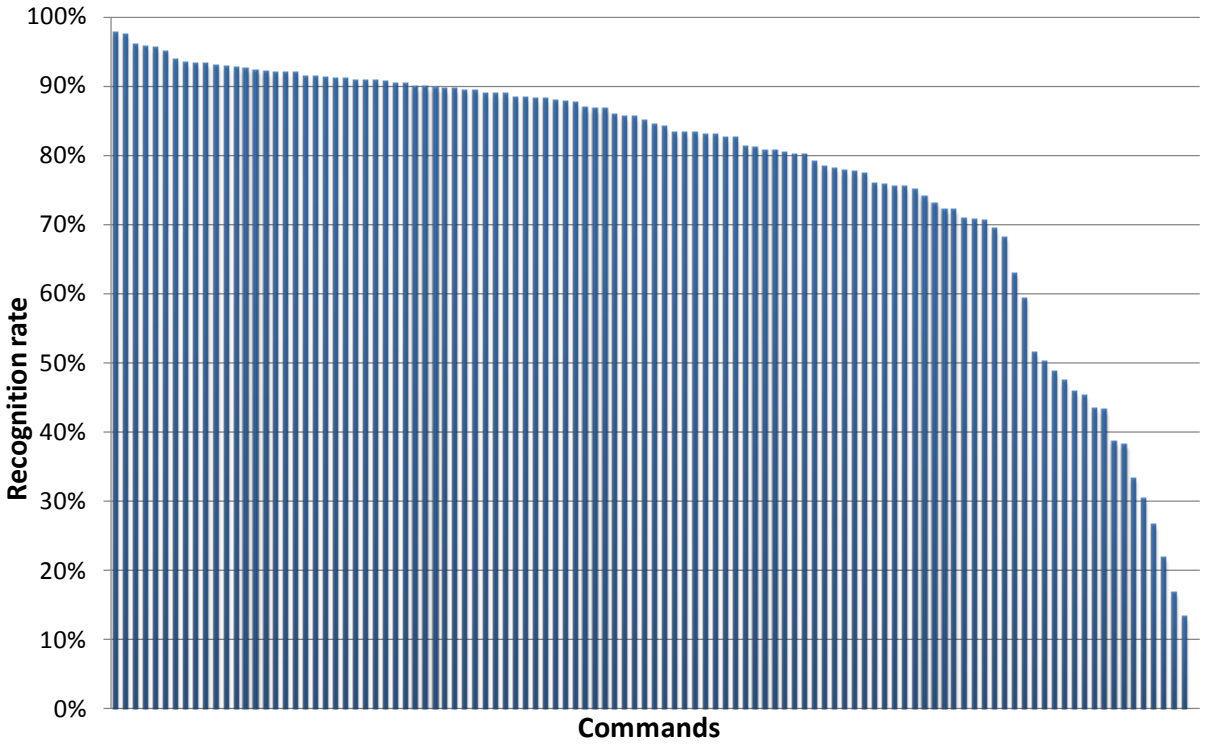


Figure 10: Recognition rates per commands.

In order to illustrate the impact of the commands that have a recurrent poor recognition rate, let us remove those below a defined threshold. Empirically for this example, I define the recognition rate as four times more important than the percentage of commands taken into account. Let us call α the recognition rate and β the percentage of commands taken into account, the criteria to maximise is the function $f(\alpha, \beta) = 4\alpha + \beta$. Having solved this basic multiple criteria optimisation problem, we get a new recognition rate $\alpha = 85.02\%$ with $\beta = 86.2\%$ of the commands taken into account (with this data, those are commands with a recognition rate over 52%). The recognition rate improvement is 6.48 points.

In the set of commands with the lowest recognition rates, we find one of the most difficult words for the participants to pronounce ("antitrendelenburg") and possibly the most difficult sentence to articulate ("svær intubation *via* larynxmaske"), no doubt also related to participants not being medically trained. More importantly, the set also includes all the long commands that are only distinguished by a number at their end ("journalen overført fra operationsstue {et, to…, otte}", which translates to "record transferred from operating room {one, two…, eight}"). Surprisingly, the names of the medications are not in this set, likelly because they are phonetically distant from anything else allowed by the grammar.

3.7 Order of sessions

The experiments were planned to reduce the impact of undesirable factors, among others the order in which the sessions were taken. This was done by randomising the order in which the sessions are completed with each participant. Consequently, since we have experiments in various orders, it is possible to study the impact of the ordering of sessions. Although there are 9 different types of sessions, there are 10 possible ranks due to an additional free text session not taken into account in this paper. Each rank contains between 1 794 and 2 511 dictations. After analysis, although this impact appears to be minor, results are interesting to report as they are in agreement with what was expected.

3.7.1 Results for the order impact

The order impact gives an indication of the evolution of the recognition rate over time. Those results are using the methodology described in the chapter "Statistical methodology for the order effect".

On Figure 11, the order impact expressed as a linear trend (least squares method) shows an improvement of 1.4 point (77.9% to 79.3% of recognition rate) along the experiments, which can be some learning effect of the participants (the system is not learning during the experiments). The participants had little feedback during the experiments (they could see the results of the command mode in real time), but they got trained in dictating and pronouncing some unusual words.

The order impact expressed as a quadratic trend (least squares method, polynomial order 2) shows a U curved, well known in human factors experiments [Schapira & Sharma 2001]. This is traditionally explained by a first phase dominated by the learning effect, an apogee, and a second phase where fatigue intervenes. Fatigue is known to affect speech recognition accuracy [Pallett 1985].

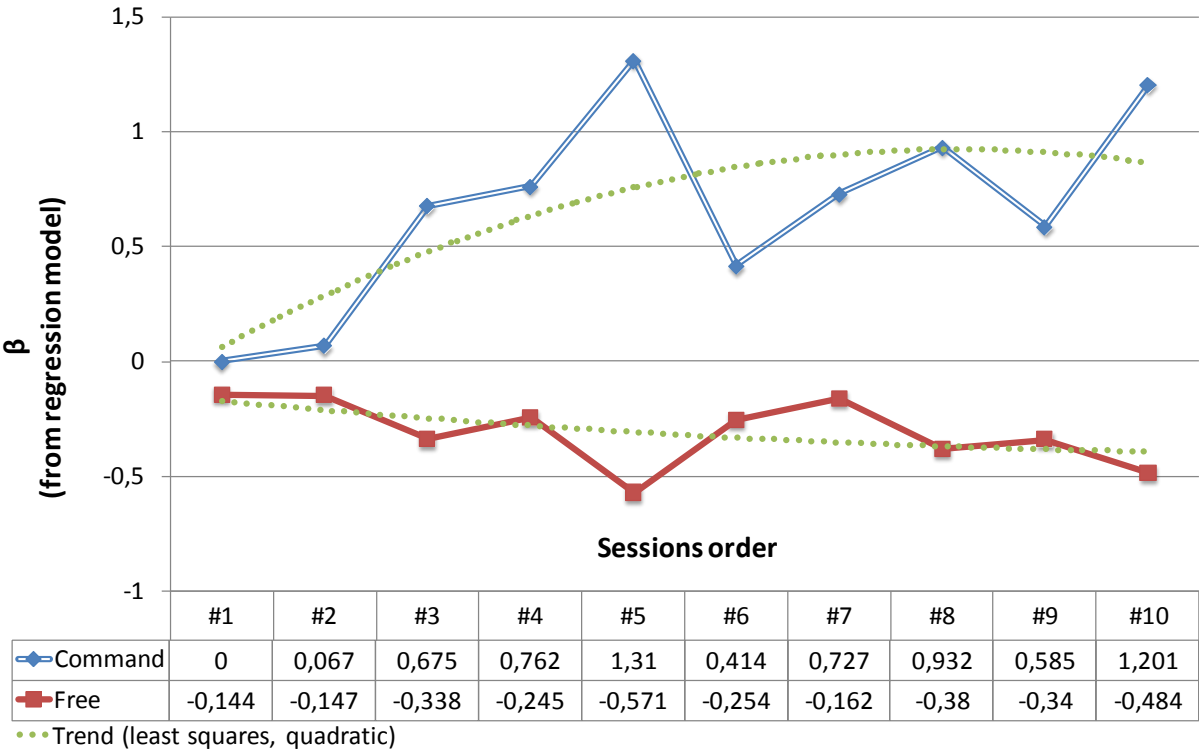
Figure 11: Recognition rate over time.

With only 10 points (10 ranks), the precision is not high enough to confirm the observed trends, nor to search for better models expressing the order impact.

Similar but more detailed results are achieved with another method, when studying the combined effect of the recognition mode combined with the session order from the regression model (Figure 12).

In addition, considering the quadratic trends and taking the first session in command mode as the reference, one can see on Figure 12 that the improvement over time is due to the command mode that improves over time and eventually stabilises then decreases, while the free text mode provides a smaller but negative effect over time.

This clear difference of reaction between the two recognition modes reveals the interest of studying the recognition mode and the session order as combined variables.



	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Command	0	0,067	0,675	0,762	1,31	0,414	0,727	0,932	0,585	1,201
Free	-0,144	-0,147	-0,338	-0,245	-0,571	-0,254	-0,162	-0,38	-0,34	-0,484

Trend (least squares, quadratic)

Figure 12: Combined variables recognition mode and session order.

3.8 Additional training

The experiment presented here has been done with minimal training. Max Manus reports that it requires 10 hours for the system to be fully trained. The results therefore only reflect the performance of the speech recognition engine "out of the box". There may be a potential performance improvement as the system learns the general task context and adjusts each user's profile. One study [Zaphar *et al.* 1999] reports that "*Accuracy improves with error correction by at least 5 percent over two weeks*". Another more detailed study [Al-Aynati & Chorneyko 2003] (using IBM ViaVoice Pro version 8 with pathology vocabulary support) reports that "*the lowest accuracy achieved* [...] *was on the first day of the study (87.4%* [word accuracy]*), and the highest was on the* [10th and] *last day (96%)*" with a plateau "*at approximately day 4–5 of the study (94%–95%)*". (See below the "Word accuracy" chapter to compare the recognition rates.)

To illustrate this learning effect, one participant did an additional training session (he read once more the 108 commands, which were then corrected and submitted to the system for adaptation). This participant was chosen randomly. He was male and achieved the sixth best recognition rate of the eight participants. His free speech recognition rate increased with 2.5 points (to 80.3%) on the same corpus by doing an additional ~5 minutes of training.

3.9 Other parameters

Some other parameters or combination of parameters have been used when defining regression models, but have not been found to have significant impact and therefore were left aside. This was the case for the order of trainings (half of the participants began with a training without noise, the others with noise), or for the combined effect of microphone and sex.

3.10 Redundant cross matching validation

Speech recognition in noisy environments is a long-standing problem, and many solutions have been tried both in upstream [Gong 1995] and downstream [Shiffman *et al.* 1995] of the recognition. In this paper, apart from the training with noise, no special improvement strategy has been used so far.

When redundant sources of information are available, such as through the two microphones in the present experiment, a post-processing system can be set up with the goal of obtaining better results than the best source alone. Such a concept has been described in, for instance, the ROVER system [Fiscus 1997] that is using an alignment and voting module. Previous experiments [Matsushita *et al.* 2003] combining various speech recognition systems demonstrated the usefulness of such an architecture. The positive gain of a combined system over the best system alone has been about 4 points out of a potential gain of 7 to 12 points if the voting was perfect. Other experiments have combined multiple microphones [Lai & Aarabi 2004] to improve the signal before sending it to a single speech recognition system.

The originality of the present experiment is an architecture made of multiple instances of speech recognition engines, each of them using a different microphone, and the combination of command mode with free text mode.

3.10.1 Cross matching with two microphones

For command mode: Out of 3 597 cases for microphone 2 in command mode, 3 056 were correct recognitions (84.96%). For microphone 1, the results were lower with 2 820 success out of 3 600 (78.33%). However, out of the 541 cases for which microphone 2 failed, 66 cases were correct for microphone 1 (12.20%). On the other hand, out of the 780 cases for which microphone 1 failed, 302 were correct for microphone 2 (38.72%). By selecting the best result between the two microphones, there are 3 122 correct recognitions out of 3 597 cases (86.79%) that is to say 1.83 points potentially better than microphone 2.

The problem is of course to choose between the results from the two microphones. The confidence score can be used, as it has been shown above to be quite reliable. We can see here again that this strategy introduced only a few errors: in 6 cases out of 3 597 (0.17%), the confidence score was higher for microphone 1 than microphone 2, even though microphone 2 was correct and microphone 1 was wrong; in 2 cases out of 3 600 (0.06%) it was the opposite.

It is therefore reasonable to choose to take the result with the higher confidence score, using microphone 2 in case of equally high confidence score. Using this very simple method, there are now 3 110 successful recognitions out of 3 599 cases (86.41%) which represents an effective improvement of 1.45 points from the best microphone.

For free text mode: Out of 7 646 cases for microphone 2 in free text mode, 6 301 were correct recognitions (82.41%). For microphone 1, the results were lower with 5 490 success out of 7 650 (71.76%). By selecting the best result between the two microphones, there are 6 492 correct recognitions out of 7 648 cases (84.88%) that is to say 2.47 points potentially better than microphone 2. The potential improvement is even larger than in command mode.

Unlike in command mode, the confidence score was not available in free text mode so here, no easy selection is proposed. However, the confidence score should be accessible in free text mode as well, when building *ad hoc* programs instead of using the standard user interface.

3.10.2 Cross matching with free text and command modes

Using only microphone 2, in free text mode trained with some background noise, there are 3 001 successful recognitions out of 3 597 (83.43%) while in command mode the ratio is 3 056 / 3 597 (84.96%). By selecting the best result between the two modes, there are 3 462 correct recognitions out of 3 595 cases (96.30%) that is to say a potential improvement of 11.34 points from the best mode. Of course, here also, the selection problem is not addressed.

3.10.3 Cross matching validation summary

Table 6: Recognition rates with cross matching validations.

	Microphone #1	Microphone #2	Best microphone (potential)	Best microphone (effective)
Command mode	78.33%	84.96%	86.79% (+1.83)	86.41%† (+1.45)
Free text mode	71.76%	82.41% 83.43%*	84.88% (+2.47)	N/A
Best mode* (potential)		96.30% (+11.34)	96.67%† (+11.71)	
Best mode* (effective)		N/A		

* Using free text mode trained with background noise.

† Using effective combination for command mode.

Table 6 summarises the results of cross matching validation. Horizontally it shows the improvement that can be achieved when combining the recognitions from the two microphones. Vertically it shows the combination of command mode with free text mode. The largest simple potential improvement is when combining command mode and free text mode, but combining the results from the two microphones is also beneficial. The combination of the two previous combinations is potentially even higher, reaching 96.67% of potential recognition rate if a perfect selection algorithm was used.

The "potential" improvement shows indeed an upper bound, as it is the ideal case where the best result is always selected, which is in practice not achievable. The "effective" improvement is real, as it uses the highest confidence score to select what is ultimately recognised, when two recognitions are not identical. The confidence score was only available for command mode, so the selection problem is not addressed for cases involving free text mode. The confidence score should be accessible in free text mode as well, when building *ad hoc* programs instead of using the standard user interface.

3.10.4 Discussion on cross matching validation

When combining cross matching principles, further improvement is potentially possible. Using on the one hand the command mode with the best confidence between microphone 1 and 2 (86.41% recognition rate) and on the other hand the free text mode trained with noise and with microphone 2, there are 3 476 successful recognitions out of 3 595 cases (96.67%).

Earlier in the paper, it has been shown that microphone 2 (headset) performed on average better than microphone 1 (handheld) for all types of background noises, for both command and free text mode, and for all participants. In the case of a system with multiple microphones, it would appear natural to use only headset microphones, or more generally, only the type that performs best. However, the best outcome from a multi-microphone system is likely achieved when microphones of different types are combined. Similarly, because the free text and command modes make different recognition errors there appears to be a considerable potential in combining these two types of recognition.

In this section, it has been shown that cross matching validation using redundant information could lead to a positive improvement. Using the confidence score available with the command mode is enough to already benefit from such an architecture. However, most of the potential improvement was not investigated, due to a lack of selection method in free text mode. Further experiments are needed with a confidence score for free text mode as well. Using more redundancy, with more than two microphones, would be an interesting continuation.

3.11 Word accuracy

In this study, recognition rates are reported at command level (*i.e.* per short sentence). To facilitate comparisons, the standard word recognition rate (WRR) was calculated for the silent session using free text mode trained in silence and taking into account the keyword for "full stop", which is the most typical scenario reported in the literature:

- Microphone 1 (86.78% accuracy on 401 sentences): 1 158 of 1 272 words recognised (91.04%), Levenshtein word distance of 155, WRR = 87.41%.
- Microphone 2 (88.30% accuracy on 401 sentences): 1 172 of 1 272 words recognised (92.14%), Levenshtein word distance of 145, WRR = 88.60%.

Keeping in mind that the experiment was made in Danish and that enrolments were very short (about 15 minutes), it is possible to compare the above reported recognition rate obtained with free text mode with a previous study [Devine *et al.* 2000] evaluating continuous speech recognition in the medical domain (in English, enrolment in less than 60 minutes). In this study, IBM ViaVoice 98 with General Medicine Vocabulary performed best (90.9% to 93% word accuracy) followed by the L&H Voice Xpress for Medicine, General Medicine Edition, version 1.2 (84.9% to 86.6%) and then Dragon Systems NaturallySpeaking Medical Suite, version 3.0 (84.8% to 14.1 to 85.9%). Another study [Mohr 2003] obtained an average of 84.5% word accuracy and another one [Happe *et al.* 2003] even reached 98% with one highly trained speaker in French and in a narrow medical field.

For information, here are:

- The WRR in the same condition for command mode for both microphones (86.82% accuracy on 402 commands): 1 110 of 1 262 words recognised (87.96%), Levenshtein word distance of 162, WRR = 87.16%.
- The overall WRR, including *e.g.* the two microphones and the sessions with noise: Total (78.61% accuracy on 22 472 commands): 60 508 of 71 508 words recognised (84.62%), Levenshtein word distance of 13 952, WRR = 80.49%.

4 Descriptive statistics summary

To provide an overview, Table 7 summarises the relative impact of 10 studied factors, giving recognition rates at command level. The "average recognition rates" are the overall average recognition rates of the two most extreme values of the studied parameter. The "largest observed impact" is the largest observed difference in recognition rates between two values of the studied parameter when combined with at most one other parameter. While the Table 1 provides the statistical analysis results, Table 7 gives a less precise but perhaps more illustrative overview. Of course, when reading it, one should keep in mind that there are *e.g.* some sampling (random) effects and interaction effects, so one must go back to the regression analysis table to get the formal information.

Table 7: Observed impact of studied parameters on recognition rates.

Parameter	Average recognition rates	Largest observed impact
Microphone type	73.9% / 83.2% Mic#1 / Mic#2	19.3 points for "Ventilation3" noise
Recognition mode	77.1% / 81.6% Free text / Command	30.19 points for "Scratch" noise
Training type (Free text mode)	75.58% - 78.19% Without / with noise	6.75 points for "Ventilation3" noise
Background noises	66.42% - 86.82% "Scratch" / "Silence"	25.72 points with Mic#1
Participants	68.39% / 86.48% Man#3 / Woman#2	21.29 / 38.81 points in command mode / for "Ventilation3" noise
Gender of the participants	76.81% / 80.32% Male / Female	12.11 points for "Ventilation3" noise
Commands	97.71% / 13.13% "tandskade" / "local anæstetika"	84.58 points
Time effect (learning/fatigue)	76.85% / 80.41% Session 2 / session 7	3.56 points
Training duration	77.5% / 80.3% with +5 mn training	2.5 points (potentially more)
Cross matching validation	84.96% / 86.41% command mode	1.45 points effective / 11.3 points potential

5 Discussion

Methodology: The methodology of the experiments is thought to be compliant with long-standing guidelines, such as defined in [Pallett 1985]. The parameters analysed in this paper are from a set of known factors influencing speech recognition. From this set, some parameters have not been investigated, such as dialect history, which can be quite strong in Denmark, even though this country is not that wide (43 094 km²).

Participants: The experiment would have been more realistic if participants had been medical staff. Undeniably, some medical words were not perfectly pronounced. Furthermore, errors that are due to mispronunciation and more generally any type of wrong dictation have not been removed from the statistics. However, the effect of those limitations is to decrease the recognition rate in a uniform way. Therefore, the main point of the experiment – to study the relative impact of various parameters – should not be affected.

Type of training: For the free text mode, the experiment shows an advantage of profiles trained with background noise, in agreement with the literature. However, there is a possible difference between constant and variable background noises. In the reported experiment, the background noise used for the enrolment was mainly constant (ventilation) but with an additional variable noise (a pulse beep). The author believes that constant background noise during enrolment will help when the system is afterwards used in a similar environment, while variable noises would only disturb the process. Additional experiments are needed to clarify this. Finally, a system such as Philips SpeechMagic, which learns every time it is used, should be evaluated for a longer period, and not only during the first session, to tell which type of training is ultimately the best for a given environment.

Laboratory: The reverberation observed in the small room where the experiment was conducted is known to affect speech intelligibility [Gelfand & Silman 1979] but that again should have only negligible effects on the relative impact of the studied parameters. While ORs are typically larger and therefore should suffer less from small room reverberation effects, some of them may have some even worse acoustics due to other factors.

Lombard effect: In a noisy environment, one modifies the tone and the loudness of one's voice. This is known as the "Lombard effect" [Lombard 1911], which is mainly due to the difficulty for a speaker to hear himself/herself. However, one tends to judge the actual loudness mainly by the physical effort rather than the perceived loudness, so called "autophonic response" [Lane *et al.* 1961]. The Lombard effect is known to reduce the accuracy of speech recognition systems. Therefore, it would be interesting to test the effect of providing the user with a supplementary audio feedback, for instance with an earplug in one ear, which could reduce the Lombard effect.

Metrics in the literature: When doing the literature study, several articles were found reporting recognition rates without stating clearly the definition used to calculate them. When such a definition was provided, it was commonly based on a notion of "recognition error", which in turn was not often accurately defined, even though it is not obvious and diverging interpretations can conduce in large differences. More attention should be set on rigorously providing the metrics definition.

Performance metric: Some differences have been shown between recognition rates at word level compared to rates at sentence level, keeping in mind that the sentences used in this experiment were short commands (two to seven words, mean 3.2). While the traditional word recognition rate (WRR) is a good measure of the raw performance of speech recognition engines, the author does not consider it relevant to measurements of the quality of speech recognition systems where the goal is a good semantic accuracy of short commands, avoiding "critical errors" [Zafar *et al.* 2004]. For the latter, the command recognition rate (CRR) should be favoured, possibly with a semantic layer that tolerates minor variations that do not alter the meaning. However, this CRR may not be suited for applications using long sentences

Conclusion

The above experiment has removed some uncertainties regarding the development of a voice-input interface for supplementing existing electronic anaesthesia record systems. Background noises have a strong impact on recognition rates, but common noises have been shown to cause only a slight degradation of performances, especially when combined with a suitable microphone, staying close to the performances that can be achieved in office environments.

When measuring the performances of a speech recognition system or comparing microphones in a noisy environment, a general advice would be to use various loudness levels. To get more precise results, several types of background noises should be tested and, in particular, not only "white noise".

When the loudness of background noises is above the threshold for automatic cut-off, for a given long timeframe (1 min), there are more insertion errors when nothing is said than when something is actually said. It is therefore especially important to have a way to pause speech recognition and an appropriately tuned filter for low confidence recognitions.

Apart from training, the major factor appears to be the words used in the commands. Therefore, the grammar for the command mode should be designed with care, avoiding words or commands that are hard to recognise or to distinguish from each other. Finally, it has been shown that a redundant architecture promises some interesting gains. There is indeed still a need for improvement before such speech recognition systems can be reliably deployed with only modest user effort.

Acknowledgements

This work was supported by the Fifth European Community Research and Development Framework Improving Human Potential Programme, within the ADVISES Research Training Network about "Analysis Design and Validation of Interactive Safety critical and Error-tolerant Systems". Special thanks to Viggo Stryger and Køge Hospital, Herlev University Hospital, my supervisors Morten Hertzum and Henning Boje Andersen, and to Max Manus for providing the Philips speech recognition software as well as the handheld microphone.

References

[Alapetite & Gauthereau 2005] Alexandre Alapetite & Vincent Gauthereau. Introducing vocal modality into electronic anaesthesia record systems: possible effects on work practices in the operating room. *Proceedings of EACE'2005 (Annual Conference of the European Association of Cognitive Ergonomics) 29 September - 1 October 2005, Chania, Crete, Greece; section II on Research and applications in the medical domain, 189–196. ACM International Conference Proceeding Series, vol. 132. University of Athens, 197-204,* ISBN:9-60254-656-5.

[Al-Aynati & Chorneyko 2003] Maamoun M. Al-Aynati & Katherine A. Chorneyko. Comparison of Voice-Automated Transcription and Human Transcription in Generating Pathology Reports. *Archives of Pathology and Laboratory Medicine, 2003, 127(6):721–725*.

[Barker *et al.* 2005] J.P. Barker, M.P. Cooke, D.P.W. Ellis. Decoding speech in the presence of other sources. *Speech Communication, 2005, 45(1):5-25* . doi:10.1016/j.specom.2004.05.002

[Detmer *et al.* 1995] William M. Detmer, Smadar Shiffman, Jeremy C. Wyatt, Charles P. Friedman, Christopher D. Lane, Lawrence M. Fagan. A Continuous-speech Interface to a Decision Support System: II. An Evaluation Using a Wizard-of-Oz Experimental Paradigm. *Journal of the American Medical Informatics Association, Jan–Feb 1995, 2(1):46–57*.

[Devine *et al.* 2000] Eric G. Devine, Stephan A. Gaehde, Arthur C. Curtis. Comparative Evaluation of Three Continuous Speech Recognition Software Packages in the Generation of Medical Reports. *Journal of the American Medical Informatics Association. 2000;7(5):462–468.*

[Fiscus 1997] Jonathan G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). *In: Proceedings of IEEE'1997 Workshop Automatic Speech Recognition and Understanding*. doi:10.1109/ASRU.1997.659110

[Gelfand & Silman 1979] Stanley A. Gelfand & Shlomo Silman. Effects of small room reverberation upon the recognition of some consonant features. *The Journal of the Acoustical Society of America, July 1979, 66(1):22-29*. doi:10.1121/1.383075

[Giorgino *et al.* 2005] Toni Giorgino, Ivano Azzini, Carla Rognoni, Silvana Quaglini, Mario Stefanelli, Roberto Gretter, Daniele Falavigna. Automated Spoken Dialog System for Hypertensive Patient Home Management. *International Journal of Medical Informatics, 2005, 74(2): 159-167*. doi:10.1016/j.ijmedinf.2004.04.026

[Gong 1995] Yifan Gong. Speech recognition in noisy environments: a survey. *Speech Communication, 1995, 16(3):261-291*. doi:10.1016/0167-6393(94)00059-J

[Gröschel *et al.* 2004] J. Gröschel, F. Philipp, St. Skonetzki, H. Genzwürker, Th. Wetter, K. Ellinger. Automated speech recognition for time recording in out-of-hospital emergency medicine – an experimental approach. *Resuscitation, 2004, 60:205–212*. doi:10.1016/j.resuscitation.2003.10.006

[Hansen 1996] John H.L. Hansen. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Communication, 1996, 20:151–173*. doi:10.1016/S0167-6393(96)00050-7

[Happe *et al.* 2003] André Happe, Bruno Pouliquen, Anita Burgun, Marc Cuggia, Pierre Le Beux. Automatic concept extraction from spoken medical reports. *International Journal of Medical Informatics, 2003, 70:255-263*. doi:10.1016/S1386-5056(03)00055-8

[Hirsch & Pearce 2000] Hans-Günter Hirsch & David Pearce. The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions. *In: Proceedings of ISCA ITRW ASR'2000 "Automatic Speech Recognition: Challenges for the Next Millennium", September 18-20 2000, Paris, France*.

[Jungk *et al.* 2000] Andreas Jungk, Bernhard Thull, Lutz Fehrle, Andreas Hoeft, Günter Rau. A case study in designing speech interaction with a patient monitor. *Journal of Clinical Monitoring and Computing, 2000, 16:295-307*. doi:10.1023/A:1011456205786

[Lai & Aarabi 2004] Calvin Yiu-Kit Lai, Parham Aarabi. Multiple-microphone time-varying filters for robust speech recognition. *In: Proceedings of ICASSP'2004, International Conference on Acoustics, Speech, and Signal Processing*.

[Lane *et al.* 1961] H. L. Lane, A. C. Catania, S. S. Stevens. Voice Level: Autophonic Scale, Perceived Loudness, and Effects of Sidetone. *Acoustical Society of America, 1961*. doi:10.1121/1.1908608

[Levenshtein 1965] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR, 1965, 163(4):845-848, (Russian). English translation in Soviet Physics Doklady, 1966, 10(8):707-710*.

[Lombard 1911] E. Lombard. Le signe de l'élévation de la voix. *Annales Maladies Oreille, Larynx, Nez, Pharynx, 1911, 31:101–119*.

[Matsushita 2003] Masahiko Matsushita, Hiromitsu Nishizaki, Takehito Utsuro, Yasuhiro Kodama, Seiichi Nakagawa. Evaluating Multiple LVCSR Model Combination in NTCIR-3 Speech-Driven Web Retrieval Task. *Proceeding of Eurospeech'2003, the 8th European Conference on Speech Communication and Technology, pp. 1205-1208*.

[McCowan *et al.* 2005] Iain McCowan, Darren Moore, John Dines, Daniel Gatica-Perez, Mike Flynn, Pierre Wellner, Hervé Bourlard. On the Use of Information Retrieval Measures for Speech Recognition Evaluation. *IDIAP (Institut Dalle Molle d'Intelligence Artificielle Perceptive), Martigny, Switzerland. IDIAP Research Report IDIAP-RR 04-73, March 2005*. http://www.idiap.ch/ftp/reports/2004/rr04-73.pdf

[Mohr *et al.* 2003] David N. Mohr, David W. Turner, Gregory R. Pond, Joseph S. Kamath, Kathy B. De Vos, Paul C. Carpenter. Speech Recognition as a Transcription Aid: A Randomized Comparison With Standard Transcription. *Journal of the American Medical Informatics Association, 2003, 10(1):85–93*. doi:10.1197/jamia.M1130

[Pallett 1985] David S. Pallett. Performance Assessment of Automatic Speech Recognizers. *Journal of Research of the National Bureau of Standards, September-October 1985, 90(5).* ISSN:0160-1741

[Robinson 1957] D. W. Robinson. The subjective loudness scale. *Acustica, 1957, Vol. 7.*

[Saastamoinen 2005] Juhani Saastamoinen, Zdenek Fiedler, Tomi Kinnunen, Pasi Fränti. On factors affecting MFCC-based speaker recognition accuracy. *International Conference on Speech and Computer (SPECOM'2005), Patras, Greece, pp. 503-506, October 2005.*

[Schapira & Sharma 2001] Emilio Schapira & Rajeev Sharma 2001. Experimental Evaluation of Vision and Speech based Multimodal Interfaces. *In: Proceedings of the 2001 Workshop on Perceptive User interfaces (Orlando, Florida, USA). PUI'2001, vol. 15. ACM Press, New York, 1-9.* doi:10.1145/971478.971481

[Sears *et al.* 2001] Andrew Sears, Clare-Marie Karat, Kwesi Oseitutu, Azfar Karimullah, Jinjuan Feng. Productivity, satisfaction, and interaction strategies of individuals with spinal cord injuries and traditional users interacting with speech recognition software. *Universal Access in the Information Society, 2001, 1(1):4-15.* doi:10.1007/s102090100001

[Sears *et al.* 2003] Andrew Sears, Jinjuan Feng, Kwesi Oseitutu, Clare-Marie Karat. Hands-Free, Speech-Based Navigation During Dictation: Difficulties, Consequences, and Solutions. *Human-computer interaction, 2003, 18:229–257.* doi:10.1207/S15327051HCI1803_2

[Shiffman *et al.* 1995] Smadar Shiffman, William M. Detmer, Christopher D. Lane, Lawrence M. Fagan. A continuous-speech interface to a decision support system: I. Techniques to accommodate for misrecognized input. *Journal of the American Medical Informatics Association, 1995, 2(1):36–45.*

[Sobel 1981] Carolyn Panzer Sobel. A generative phonology of Danish. *Ph.D. Thesis 1981, City University of New York.*

[Steinlen & Bohn 1999] Anja K. Steinlen & Ocke-Schwen Bohn. Acoustic studies comparing Danish vowels, British English vowels and Danish-accented British English vowels. *Collected Papers (CD-ROM) of the 137th Meeting of the Acoustical Society of America and the 2nd Convention of the European Acoustics Association, Forum Acousticum, Paper 2pSCb21, Technical University of Berlin, Germany. Abstract in the Journal of the Acoustical Society of America, 1999, 105(2):1097.* doi:10.1121/1.425143

[Zafar *et al.* 1999] Atif Zafar, J. Marc Overhage, Clement J. McDonald. Continuous speech recognition for clinicians. *Journal of the American Medical Informatics Association, 1999, 6(3):195–204.*

[Zafar *et al.* 2004] Atif Zafar, Burke Mamlin, Susan Perkins, Anne M. Belsito, J. Marc Overhage, Clement J. McDonald. A simple error classification system for understanding sources of error in automatic speech recognition and human transcription. *International Journal of Medical Informatics, 2004, 73:719-730.* doi:10.1016/j.ijmedinf.2004.05.008

Transition 3

The laboratory experiments reported in the previous paper [Alapetite 2006] were useful to define precisely the possibilities and limitations of the speech recognition technology in Danish used in this project.

Pursuing the goal to clarify the questions raised at the EACE'2005 conference (*cf.* Transition 2), and whose relevance had been since then confirmed, the next step was to study the capacity of anaesthesiologists to dictate entries in the anaesthesia record while working.

For this purpose, the first step was to define speech input strategies and a phraseology, *i.e.* the way to address the system when using command mode instead of free speech, given the capacities of the speech recognition system that were previously determined.

Therefore, time was allocated to the development of the prototype, with the major part being done in more than two months at full time between May and August 2006.

Since the overall problem of interest for this thesis is the tendency for anaesthesiologists to postpone the registration of events during time-constrained situations, it was vital to ensure that participants to the forthcoming experiments would face such situations. Therefore, two "busy" anaesthesia scenarios were chosen, *i.e.* scenarios that involved the patient developing complications, which would require the anaesthetic team to perform a number of tasks while keeping the patient under close observation.

Choosing time-critical scenarios was also crucial to magnify the differences between the traditional touch-screen and keyboard interface, and the envisaged one supplemented by speech input facilities. During normal full anaesthesia, there are typically long periods with minimal action and little time pressure (mainly in the maintenance phase) during which anaesthesiologists have time to register the past, present and prepare upcoming events. During those phases of lower activity, it would not make much sense to attempt a comparison of the efficiency of various human-computer interfaces based on the major criteria of interest for this chapter, namely their rapidity and capacity to be used in parallel with other tasks. Therefore, it was necessary to have difficult "busy" anaesthesia scenarios involving many medications and consequently many events to register.

It was not realistic to conduct experimentation during real and potentially severe operations involving real human patients. First of all, the uncertain nature of the experimentation and the risk that this might interfere with the ability of the anaesthetic and surgical teams to cope with events would, of course, ethically rule this out. Second, operations with complications are fortunately rare and hard to predict. Finally, experiments had to be reproducible at a reasonable degree of similarity to make measurements and statistical comparisons between parameters.

We had the good fortune of being offered access to the Danish Institute for Medical Simulation[1] at Herlev Hospital, Denmark, and we subsequently chose to run the experiment in their full-scale anaesthesia simulator in September 2006. The research group at Risø has prior experience in working with simulators to train or test abnormal safety-critical conditions, including aviation, maritime operations and anaesthesia (*e.g.* [Andersen *et al.* 2000; Weber *et al.* 2005]).

During the early thoughts about the experiment, the use of an eye tracking system was considered, to get more accurate information about the actions of the operators, and about what they are looking at, in a setup similar to the one proposed in [Andersen & Hansen 1995; Andersen *et al.* 2000]. However, considering the added complexity and the risk of affecting more important parts of the experiment, eye tracking was abandoned. Nevertheless, the planned subset of [Andersen & Hansen 1995] with the use of cameras from three different angles, central microphone, and screen recording of the main interface, can partially replace eye tracking, in particular regarding the interaction with the apparatus and the patient. During the video analysis, the information recorded by the cameras and microphone was enough to assess the actions of the physicians that were relevant to the study.

The following paper [Alapetite 2007] reports the experiments undertaken with the prototype to answer the research question and validate the solutions introduced above. The opportunity was also taken at this step to partially test the acceptance of the vocal modality by anaesthesiologists and the surgery team.

[1] [http://herlevsimulator.dk]

References

[Andersen & Hansen 1995] Henning Boje Andersen, John Paulin Hansen. Multi-modal recording and analysis of interaction among operators and work systems. *Proceedings of HMI-AI-AS'1995, the fifth international conference on human-machine interaction and artificial intelligence in aerospace, Toulouse, France, September 27-29 1995. Expended version in Risø-R-939(EN), Risø National Laboratory, Denmark, December 1996*.

[Andersen *et al.* 2000] Henning Bøje Andersen, C. R. Pedersen, Hans H. K. Andersen. Using eye tracking data to indicate team situation awareness. *In: Usability evaluation and interface design: Cognitive engineering, intelligent agents and virtual reality. Proceedings of HCI International'2001, International conference on human-computer interaction, 1(9): 1318-1322, New Orleans (LA, USA), 5-10 August 2001. D. Harris, M. J. Smith, G. Salvendy, R. J. Koubek (eds.), Lawrence Erlbaum Associates, Inc., Mahwah (NJ, USA), 2001.* ISBN:0-8058-3609-8

[Weber *et al.* 2005] Steen Weber, Jette Lundtang Paulsen, Henning Boje Andersen. Comparing of training effects in real and virtual environments. *In: Tagungsband: Virtual reality und augmented reality zum Planen, Testen und Betreiben technischer Systeme. 8. IFF-Wissenschaftstage, Magdeburg (DE), 22-24 June 2005. M. Schenk (ed.), (Fraunhofer-Institut für Fabrikbetrieb und -automatisierung IFF, Magdeburg) pp. 145-154*.

Speech recognition for the anaesthesia record during crisis scenarios

Alexandre Alapetite[1,2]

1. Systems Analysis Department; Risø National Laboratory; Technical University of Denmark; DK-4000 Roskilde; Denmark
2. Computer Science, Roskilde University, Universitetsvej 1; P.O. Box 260; DK-4000 Roskilde, Denmark

This section is the long version of the following journal article:

Alexandre Alapetite. Speech recognition for the anaesthesia record during crisis scenarios. *International Journal of Medical Informatics, accepted August 2007.* doi:10.1016/j.ijmedinf.2007.08.007

Abstract

Introduction: This article describes the evaluation of a prototype speech-input interface to an anaesthesia patient record, to be used in real time during operations. The evaluation of the prototype was conducted in a full-scale anaesthesia simulator involving six doctor-nurse anaesthetist teams.

Objective: The aims of the experiment were, first, to assess the potential advantages and disadvantages of a vocal interface compared to the traditional touch-screen and keyboard interface to an electronic anaesthesia record during crisis situations; second, to assess the usability in a realistic work environment of some speech input strategies (hands-free vocal interface activated by a keyword; combination of command and free text modes); finally, to quantify some of the gains that could be provided by the speech input modality.

Methods: Six anaesthesia teams composed of one doctor and one nurse were each confronted with two crisis scenarios in a full-scale anaesthesia simulator. Each team would fill in the anaesthesia record, in one session using only the traditional touch-screen and keyboard interface while in the other session they could also use the speech input interface. Audio-video recordings of the sessions were subsequently analysed and additional subjective data were gathered from a questionnaire. Analysis of data was made by a method inspired by queuing theory in order to compare the delays associated to the two interfaces and to quantify the workload inherent to the memorisation of items to be entered into the anaesthesia record.

Results: The experiment showed on the one hand that the traditional touch-screen and keyboard interface imposes a steadily increasing mental workload in terms of items to keep in memory until there is time to update the anaesthesia record, and on the other hand that the speech input interface will allow anaesthetists to enter medications and observations almost simultaneously when they are given or made. The tested speech input strategies were successful, even with the ambient noise. Speaking to the system while working appeared feasible, although improvements in speech recognition rates are needed.

Conclusion: A vocal interface leads to shorter time between the events to be registered and the actual registration in the electronic anaesthesia record; therefore, this type of interface would likely lead to greater accuracy of items recorded and a reduction of mental workload associated with memorisation of events to be registered, especially during time-constrained situations. At the same time, current speech recognition technology and speech interfaces require user training and user dedication if a speech interface is to be used successfully.

Summary points

What was known before the study:

- Studies have pointed out the limitations of the current anaesthesia record systems involving either a paper-based record or an electronic interface that typically cannot be seen by the anaesthesiologist when looking at the patient, and which are incomplete when things get busy, thus adding to the mental workload of the anaesthesiologist [Alapetite & Gauthereau 2005].
- Background noise and stress are among the factors having a negative effect on speech recognition rates [Alapetite 2006].
- Some experiments have been done to investigate the potential of speech recognition in anaesthesia, mainly during calm situations and not entirely realistic anaesthesia scenarios [Jungk *et al.* 2000]. Questionnaire surveys [Devos *et al.* 1991] and simulations [Detmer *et al.* 1995] have indicated that anaesthesiologists are largely in favour of introducing speech input to the anaesthesia record. Other experiments have elicited expressions of interest by anaesthesiologists in speech input during anaesthesia, but without comparing this option with traditional electronic interfaces [Sanjo *et al.* 1999].

- Medical records must be capable of containing both structured data and narrative text [Lovis *et al.* 2000]. Furthermore, due to current technical limitations, there is a need to find a balance between a large and therefore expressive grammar (and vocabulary) and a small and therefore less expressive one. Finally, a smaller grammar (and vocabulary) will tend to have a higher recognition rate [Shiffman *et al.* 1995]. Those are known factors between which a proper balance must be found.

What the study has added to the body of knowledge:

- This study is the first reported experiment, as far as the author has been able to ascertain, with a hands-free vocal interface used in real time during realistic and critical anaesthesias.
- The experiment has quantified the limitations of the typical touch-screen and keyboard interface during crisis situations in anaesthesia.
- A potential gain has been identified in reduction of mental workload, with a vocal interface supplementing a traditional one during crisis situations.
- The feasibility has been demonstrated of a hands-free vocal interface activated by a keyword during a real-time situation involving stress, background noise, extraneous oral discussions at normal level of loudness.
- The prototype used has shown the possibility of combining constrained (command based) and natural language (free text), giving a possibility to use both structured data and narrative text [Lovis *et al.* 2000].

Keywords

Anaesthesia record; quality; vocal interface; speech recognition; workload; secondary task

1 Introduction

While the primary task of anaesthesiologists during operations is to take care of the patient being anaesthetised, it is also important to devote resources to the secondary task of maintaining and thus continuously updating the anaesthesia record. This record has several uses: first, it serves as a legal document and must therefore contain a log of all important events and actions, second, it may also provide information for the patient medical record, and third and most importantly, it is used during the operation to help the anaesthesia team in remembering the medications given, what has been done, thus supporting decision making and briefing of new staff joining the operation [Alapetite & Gauthereau 2005]. While electronic anaesthesia records can automatically register a number of vital trends (*e.g.* pulse, oximetry measures, CO_2) – as opposed to paper-based anaesthesia records – anaesthesiologists still have to manually register a number of actions and observations, *e.g.* intubation, medications, or possible complications. During planned and smooth operations, there is usually enough time for anaesthesiologists to keep the anaesthesia record up to date. But during critical anaesthesias when acute attention must be focused continuously on the patient and vital signs, manual registrations will have to be postponed. Delaying recording, however, is a potential source of problems: due to well-known human memory limitations [Cowan 2000], anaesthetists will tend to forget some of the items, typically amounts, and times of repetitive medications actions. Moreover, the fact that anaesthesiologists during critical phases must remember all the medications and amounts may, it can be argued, impose an additional mental workload.

For the human computer interface of the anaesthesia record to be more capable of handling time critical situations, a few strategies have been reported in the literature, such as using bar codes on syringes and various multimodal interfaces. In this paper, the focus is on supplementing an existing touch-screen based electronic anaesthesia record system with speech input facilities, using a professional speech recognition software (in Danish). Some research has already been reported on this topic, calling for further work on identifying areas of interest in terms of work efficiency and on ergonomic design of speech interaction [Jungk *et al.* 2000]. Responding in part to that call, the aim of the experiment reported in this article was to estimate whether speech input for the anaesthesia record could be fitted into normal mode of working of anaesthesiologists even during crisis scenarios, and to test some HCI (Human Computer Interaction) choices about how to interact with the anaesthesia record by voice alone. In particular, a completely hands-free approach was evaluated that uses a keyword to activate speech recognition and another keyword to switch from constrained (command based) to natural language (free text). As this experiment did not aim at evaluating the quality of a given speech recognition engine, a partial "Wizard of Oz" setting was used to reduce potential disturbance in the flow of actions created by misrecognitions.

The experimental evaluation followed a partial cross-over design (within-group), in which two critical anaesthesia scenarios were conducted by six anaesthesia teams, each team composed of an anaesthesia doctor and an anaesthesia nurse. The scenarios were run in a full-scale anaesthesia simulator in two modes, one involving the traditional electronic anaesthesia record with touch-screen and keyboard interface with which the participants were familiar from their daily work, the other supplemented by a prototype speech recognition interface.

Several statistics are reported, but the major indicator is a metric inspired by queuing theory [Kozine 2007]: the average queue of events waiting to be registered. This metric is proposed as a useful way of measuring secondary task workload and therefore, in this case, the capacity to keep the record up to date and the associated mental workload imposed on anaesthesiologists when, in addition to their primary task of managing general anaesthesia to a patient, they must also devote attention and resources to the secondary task of maintaining the anaesthesia record.

2 Prototyping

In order to evaluate how a speech input interface would affect the ability of anaesthesiologists to keep the electronic anaesthesia record updated during crisis scenarios, it was decided to organise some repeated full-scale anaesthesia simulations.

At the same time, it was important to minimise the differences between the two work conditions to be compared: namely the speech input and the conventional (touch-screen and keyboard) input to the anaesthesia record. To achieve this, participants should be familiar with an electronic anaesthesia record. Participants were therefore recruited from an anaesthesia department, Køge Hospital (Denmark), which for several years has been running an electronic system to record real time monitoring data, medications and observations.

Since the full-scale anaesthesia simulator at Herlev University Hospital – in which the experiment was carried out – is not equipped with an electronic anaesthesia record system, it was decided to supply a mock-up of such a system. This mock-up was built by the author as a mimic of the electronic system used daily by participants in their usual workplace, which included an implementation of most functions of interest during execution of the two test scenarios.

2.1 The electronic anaesthesia record

The anaesthesia information management system in use at participants' home hospital, Recall[1] AIMS from Dräger Medical, includes an anaesthesia record component with a touch-screen and a keyboard (Picture 1). The Recall system is capable of automatically registering vital signs (*e.g.* pulse, oxidation), and the anaesthesiologist uses the touch-screen and keyboard to enter other information such as major events (*e.g.* intubation, surgery started), medications, and possible remarks (Picture 2). This system was used as a reference for the design of the mock-up.

Pictures 1 & 2: Dräger Recall electronic anaesthesia record.

2.2 Speech recognition software

For voice dictation in free speech mode, or "natural language", the speech recognition system Philips[2] SpeechMagic 5.1.529 SP3 (March 2003) was used. Voice command, or "constrained language", was done by Philips SpeechMagic InterActive (January 2005). The constrained language was extended with a package for the Danish language (400.101, 2001) and a "ConText" for medical dictation in Danish (MultiMed Danish 510.011, 2004) from Philips developed in collaboration with the Danish company Max Manus[3]. For each of the six participants who were assigned the task of managing speech input during the experiment, an individual voice profile had to be established, an exercise of around 30 minutes during which the speech recognition system is trained on the user's voice.

[1] [http://www.draeger.com/MT/internet/EN/us/prodserv/products/inform_tech/recall_aims/pd_recall.jsp]

[2] [http://speechrecognition.philips.com]

[3] [http://maxmanus.dk]

2.3 Speech interaction

To establish how the anaesthesiologist would interact with the anaesthesia record by voice, experience gained from a previous experiment with speech recognition in noisy operation rooms was used [Alapetite 2006]. In particular, the previous study suggested that since the "confidence" score given by the speech recognition engine after a potential recognition is fairly robust, a completely hands-free approach may be possible, using a keyword to activate speech recognition and another keyword to switch from constrained (command based) to natural language (free text). This means that the speech recognition engine is listening all the time, filtering out any speech not preceded by the activation keyword. In our case, each time the user says "Computer…", the system is alerted and then tries to recognise what follows, matching a predefined grammar (see below). If what is said cannot match the grammar with a high enough confidence, no action is taken, but an entry is added to a log in case of recognition with a low confidence below threshold.

To allow the user to enter unconstrained free text, a second keyword was introduced: when the user says in Danish, "Computer, bemærk…" (English: "Computer, remark…") the dictation that follows is processed by the speech recognition system until the user stops speaking for more than 2 seconds. If, perhaps through hesitation, the user has not completed the intended sentence before the two-second time-out, the user may simply repeat the keywords again and start on the sentence again. An audio feedback indicates the beginning and the end of the free text recognition, with two easily recognisable short sounds.

This keyword activation is a different approach than what has been reported so far in the literature for anaesthesia systems: [Detmer *et al.* 1995] used a button to activate the speech recognition system, [Sanjo *et al.* 1999] used a touch-screen to initiate the dialog, and [Jungk *et al.* 2000] did the dictations separately after the operations. It is to some extend similar to the activation of the prototype made by [Gröschel *et al.* 2004] for out-of-hospital emergencies, which was however limited to constrained language.

The possibility to choose between command and free text mode is also novel, it appears. Each of these two modes has its own advantages. Technically, command mode reaches higher recognition rates and is more robust [Alapetite 2006]. In terms of organisation, structured data (more suited to command mode speech recognition) can be automatically processed more easily, but more information can be kept using narrative text (only possible in free text mode speech recognition), so "both systems are needed in a tightly connected architecture" [Lovis *et al.* 2000].

To keep the voice interaction simple, users are allowed to make corrections of previously dictated entries by subsequent touch-screen and keyboard interface. This option is based on the repeated finding that hands-free speech-based navigation is less efficient using speech than traditional modes [Sears *et al.* 2003].

2.3.1 Speech grammar

The main principles of the syntax to follow when dictating commands to the system, which are formally written in "Java Speech Grammar Format"[4], were discussed with an anaesthesiologist from Køge Hospital[5]. The grammar was intended to be robust against background noise, finding a balance between a large and therefore expressive grammar (and vocabulary) and a smaller one but with higher recognition rates [Shiffman *et al.* 1995]. Furthermore, the grammar should be simple enough to be fast and quick to learn before proficient use. For the experiment, each of the six participants had indeed less than 20 minutes to learn how to address the system. In spite of its simplicity, the grammar was aimed to cover the main user needs.

Table 1: Syntax for speech commands (translated in English).

Type of speech command	Example	Range of possibilities
COMPUTER <fixed event>	COMPUTER Surgeon begins	181 fixed events
COMPUTER <medication> BOLUS <dosage>	COMPUTER Adrenalin BOLUS 0.5	88 medications 50 dosages
COMPUTER <medication> INFUSION (<dosage> | STOP)	COMPUTER Propofol INFUSION 60	
COMPUTER <liquid or gas> (START | STOP)	COMPUTER Oxygen START	3 liquids, 5 gases
COMPUTER REMARK {wait 1 s} <free text> {wait 2 s}	COMPUTER REMARK… Patient has fever between 38 and 39°C…	Unlimited

[4] [http://java.sun.com/products/java-media/speech/forDevelopers/JSGF/]

[5] Dr. Viggo Stryger

As reported in Table 1, there are 5 types of speech commands:

1) The fixed events are the ones traditionally selected by anaesthesiologist from Køge Hospital using the touch-screen interface.

2) The possible medications have been taken from the list of medications used at least two times in anaesthesia over the past two years at Køge Hospital. The dosages for the medications are simply a number or a decimal number, made by pronouncing, *e.g.* "zero point five"; for this experiment, only the 50 most used dosages between 0.1 and 1 000 were implemented.

3) For medications administered by "infusion" (*i.e.* over a long period of time, as opposed to "bolus"), it is possible to say "stop" instead of a dosage. To register a new infusion, the user states the dosage, and to modify the dosage of a running infusion, the new dosage is simply stated.

4) For liquids (such as NaCl) and gases (such as oxygen), no dosage was implemented, but only the "start" and "stop" keywords.

5) Finally, for everything else, it is possible to register some free text comments.

Having the speech recognition running continuously to be activated by a keyword is a challenging approach that calls for a few technical constraints on the grammar in order that it might succeed in noisy uncontrolled environment. The most noticeable constraint was on delays: a limit was set so that it was not accepted to pause during a speech command for more than around 200 ms. A speech command must therefore be said in one go, distinctively and without any dysfluency, or it will be rejected. During free text, pauses are accepted up to 2 seconds.

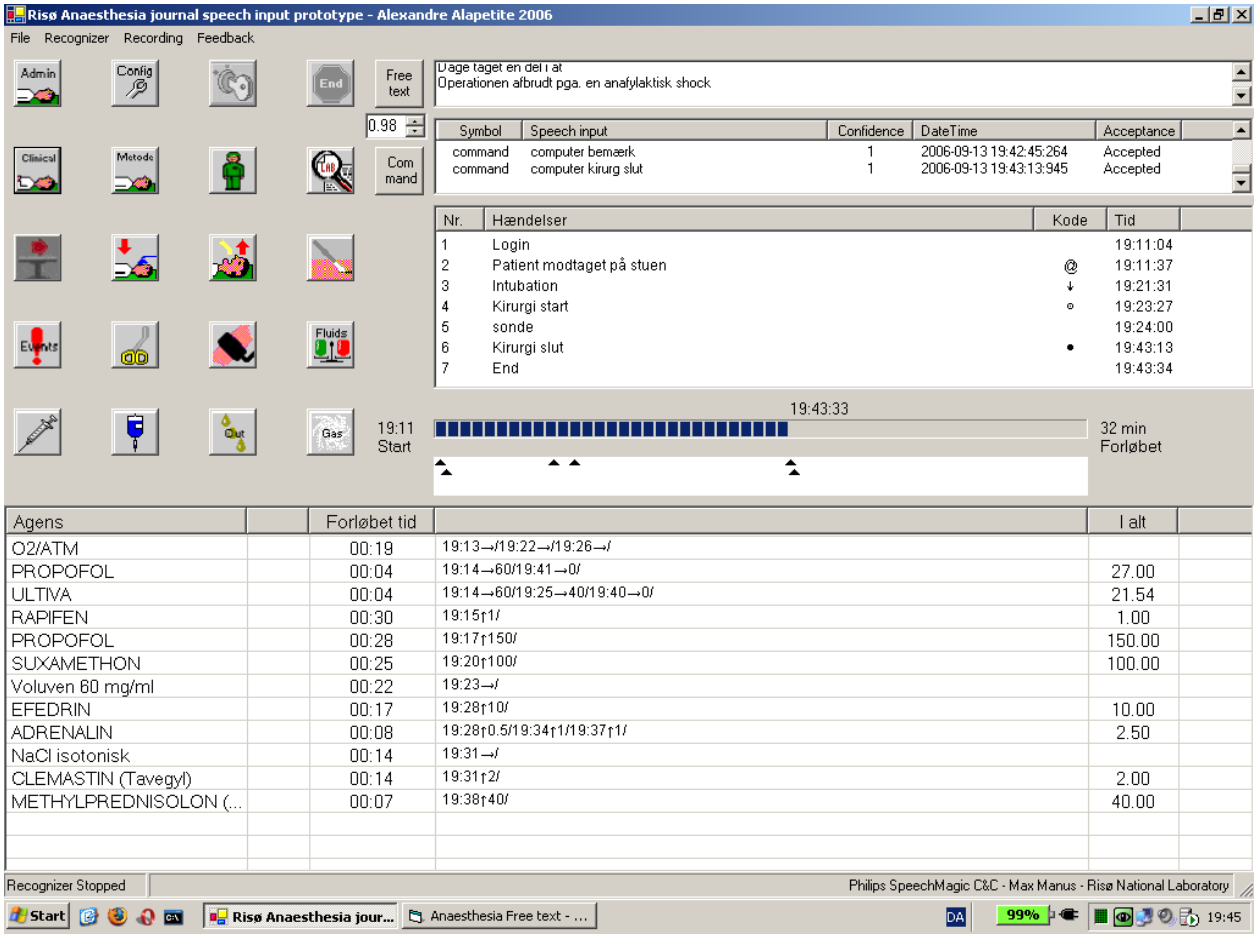
2.4 Audio feedback

While the main feedback is graphical and displayed on the touch-screen, there is also a need of another type of feedback for confirmations when participants are dictating without looking at the screen. In this prototype, there are two types of audio feedback. For the main fixed events (*e.g.* "intubation"), a pre-recorded voice is used to play back what was said. If this is found disturbing, there is a possibility to disable voice output and replace it by a short sound. For the other commands (*e.g.* medicaments), a short sound is used when something was recognised with sufficiently high confidence, and another sound when something was recognised but rejected due to too low confidence.

2.5 Prototype

The hardware of this multimodal prototype is composed of a laptop computer (IBM ThinkPad R32, Intel P4m 1.6 GHZ, 768 MB of memory, Windows XP SP2) linked to a touch-screen (3M MicroTouch M170 FPD 17″) and to a headset microphone (~2.5 cm from the mouth) model PC145-USB[6] from Sennheiser Communications (uni-directional, 80 – 15 000 Hz, -38 dB).

The main software part of the prototype, which is the graphic interface of the mock-up of the anaesthesia record (Picture 3), was developed with the programming framework Microsoft C# .NET 2.0[7], under SharpDevelop 2.0[8], an Open Source Development Environment. This part also controls the speech recognition in command mode, in particular the special keywords to activate recognition and to shift to free text mode.



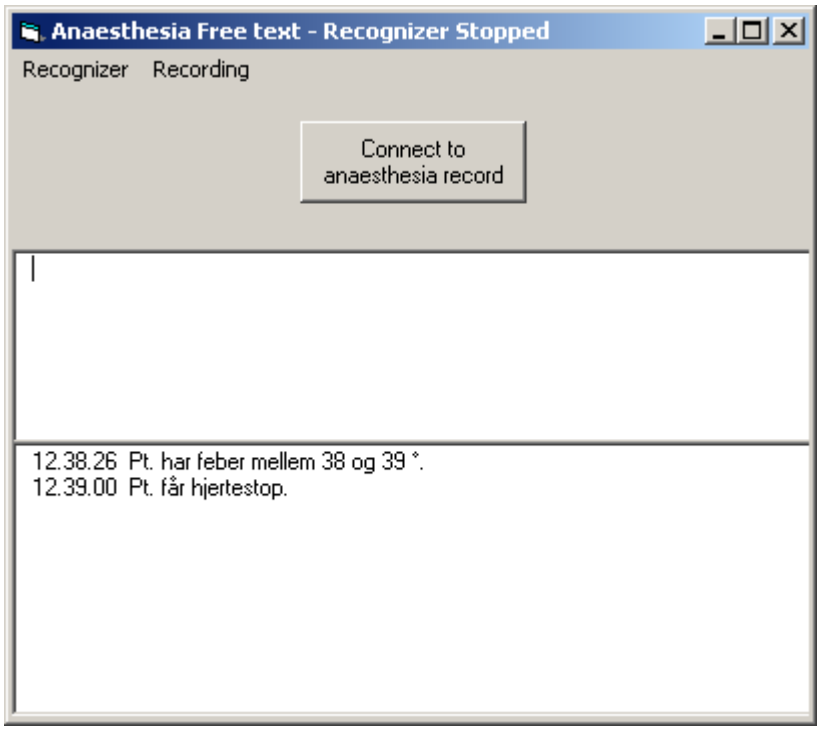
Picture 3: Mock-up of the anaesthesia record with speech commands.

[6] [http://www.oticon.com/eprise/main/SennheiserCommunications/com/Products/CNT05_VBLG?ProductId=PC145]

[7] [http://msdn.microsoft.com/net/]

[8] [http://icsharpcode.net/OpenSource/SD/]

The speech recognition in free text mode was developed as a separate program with Microsoft Visual Basic 6.0, running in the background and communicating with the main program through network sockets (Picture 4). The separation of the free text mode was chosen because it took too much processing power to switch between command and free text mode in one program. Having one program running for command mode and another one for free text mode allowed fast transitions between the two modes (about one second on the modestly powered laptop described above). In addition, this architecture was considered more resistant to software failure.



Picture 4: Independent module for free text speech recognition running in the background.

3 Methodology

3.1 Anaesthesia task

The general task of the anaesthesiologists has been described in detail in the literature, reflecting slightly different approaches in different countries. In Denmark, where this experiment was done, an anaesthesia doctor can be in charge of a few operations at a time, each operation being constantly monitored and managed by an anaesthesia nurse who remains with the patient during the whole operation. Therefore, for planned, non-complicated anaesthesias, an anaesthesia doctor is typically present only during the induction phase, sometimes during the recovery and will always be called in case of difficulty. The anaesthesia doctor will make the decisions regarding the strategy to follow, but the doctor and the nurse will often be rehearsing possibilities together. The nurse and the doctor may be replaced or supplemented by colleagues, especially during long operations; and during highly critical episodes where the patient's life may be at stake, the team will call for assistance from additional doctors and nurses.

While the main task of the anaesthesia team is clearly to take care of the patient, the anaesthesia record should be filled when possible, as a secondary task with lower priority. The general use of the anaesthesia record during the successive phases of anaesthesia is described in [Alapetite & Gauthereau 2005]. Filling in the record is typically done by the anaesthesia nurse, but sometimes the doctor will also enter remarks and medications into the record.

3.2 Experiment

3.2.1 The anaesthesia simulator

The experiment took place in September 2006 at the Danish Institute for Medical Simulation[9], Herlev University Hospital (Copenhagen region, Denmark) in one of the institute's full-scale simulators used for training anaesthesiologists [Østergaard 2004], following principles similar to but newer than those reported in [Gaba *et al.* 1988]. The simulation environment is organised around a mannequin on which the main anaesthesia techniques can be applied, such as intubation, ventilation, perfusions as well as auscultations. The operating room is equipped with classic anaesthesia apparatus including a choice between different brands of monitors. Adjoining the operating room there is a control room where an expert observer remotely modifies the state of the artificial patient (Picture 5), with the help of a dedicated software that is capable of automatically handling some of the simulation. During sessions, an instructor (an anaesthesiologist specialist) is present in the operating room.

[9] [http://herlevsimulator.dk]


Pictures 5 & 6: Herlev anaesthesia simulator DIMS.

For this experiment, the normal anaesthesia simulator setting was supplemented with the prototype electronic anaesthesia record system with speech input, with the touch-screen and the keyboard of the laptop computer being placed on the right side of the anaesthesia monitors, similar to the layout at participants' home department in Køge Hospital.

3.2.2 Audio-video recording

The anaesthesia simulator is equipped with two video cameras that record the simulations. Videos are normally used for the debriefing after sessions. For the purpose of this experiment, an additional camera was used to ensure detailed analysis of the sessions afterwards. A fourth video signal was used to record the screen of the anaesthesia record, using the Open Source Virtual Network Computing software UltraVNC[10] to forward the video screenshot to another computer that saved the video digitally (AVI[11] compressed with Xvid[12]), converting it to S-Video signal. The four video signals were mixed online by a "quad mixer" producing a single picture divided into 4 areas, thus avoiding all problems of synchronisation (Picture 6). A stereo microphone was placed in the middle of the operating room. The final audio-video signal was recorded on DVD (Picture 7).

[10] [http://ultravnc.sourceforge.net]

[11] AVI: Audio Video Interleave

[12] [http://xvid.org]

Picture 7: Recording sound and four videos at a time.

3.2.3 Participants

The 12 participants were volunteers from Køge Hospital. Their department was chosen because they had been using an electronic anaesthesia record for some years. There were 6 teams, each composed of a doctor and a nurse. Coming from the same department, all participants knew each other and had worked together during operations. After each session, each team received a debriefing on their handling of the difficult anaesthesia scenarios by the instructor of the anaesthesia simulation institute. As a compensation for spending their free time on the study, participants were offered a small gift.

For each team, the nurse was designated as the team member responsible for carrying out registration (following the common practice of their home department). Therefore, nurse members of each of the anaesthesia teams were equipped with a microphone with direct access to the speech recognition registration system.

As described above in the section about the speech recognition system, participants had to train the system. Due to their busy work schedule, each of the six nurses trained their voice profile a few days before the sessions for only about half an hour. This limitation was accepted, although the system is known to significantly improve its accuracy during the first days of use. Each nurse was briefly introduced to the concept of the experiment and speech commands, but they had only a few trials to test the voice commands by themselves before the real sessions.

3.2.4 Partial Wizard of Oz for speech recognition

Becoming confident with a phraseology and becoming used to speaking commands distinctively and without hesitation take more time than what was available. For this reason, and because the evaluation was not designed to test recognition rate of speech recognition, a partial "Wizard of Oz" approach was used. Participants were instructed to follow the syntax to address the system whenever possible, but to use their own words if they could not remember the syntax. Thus, the prototype would behave like a perfect recogniser, as described below. The choice of this technique was made because the goal of the experiment was to identify advantages and disadvantages of a speech interface in a realistic task environment, not to measure speech recognition rates.

In a Wizard of Oz experiment, users interact with a computer system that behaves as if it was autonomous but which is actually being wholly or partially operated by a human being. The idea of using this experimental paradigm on speech input to the anaesthesia record has already been reported in the literature [Detmer *et al.* 1995]. Indeed, the prototype was fully functional with respect to the tasks and goals of the experiment; but since participants could not be sufficiently trained to reach a satisfactory level of performance with the speech interface, the instances of non-recognition (or participants using an incorrect syntax) were neglected to ensure that the sessions would run smoothly. The Wizard of Oz technique used for the experiment had an experimenter (the developer of the prototype, the author) standing close to the keyboard and screen of the anaesthesia record and register manually any speech items that was not properly dictated or not correctly understood by the speech recognition system. During analysis, a distinction was made between "wizard" input and genuine user input, *i.e.* input recognised by the software. It was originally planned to use the VNC remote interface (*cf.* section on video recording) to do the Wizard of Oz, but a few tests had shown that this made it difficult for the anaesthesiologists to understand what was going on, especially when a few events were recorded in right after each other. Hence the choice of having an experimenter standing by the anaesthesia record.

3.2.5 Scenarios and sessions

Two anaesthesia scenarios had been prepared for the experiment: one in which the patient develops an anaphylactic shock (rapid allergic reaction) with ventricular fibrillation (cardiac arrhythmia), and another in which the patient exhibits increasing severe asthmatic symptoms (respiration problem) with asystole (cardiac arrest). The two scenarios are similar in several respects: they are difficult to manage, they are life threatening, they require the administration of several medications and proper actions are time critical. Such anaesthesia complications are rare at participants' department, which is mainly handling planned operations. However, anaesthesiologists should be capable of facing such events.

Each team did two sessions, each session lasting 30-45 minutes: the first session with only the traditional touch-screen based interface, and the second with the possibility to choose between the traditional touch-screen interface and speech input.

During the simulations, the anaesthesia team had the possibility to call for additional medications, the delivery of a defibrillator, etc. but they could not call for external assistance. There was a third person playing the role of the surgeon (and, on request, performing heart massage). The scenario started with the patient already on the operation table, and a few catheters already in place. The scenarios stopped after the crisis had been handled and thus did not continue until the full recovery phase and the patient was therefore not delivered to the wake-up room as normally.

The simulations were performed on three days, with two teams per day each doing the two scenarios. Due to simulator constraints, it was more convenient during a day to prepare the simulator for one scenario, to run the first scenario for two teams, then to modify the settings of the simulator, and finally to run the second scenario for the two same teams. Counterbalancing the scenarios has been made as much as possible: for two simulation days the first scenario was “anaphylaxis”, and for one day “asystole”.

This within-group experimental design where all teams perform the two sessions (as opposed to between-group design) was chosen first to reduce error variance associated to the natural variability between teams, and second to get the most data and the maximal statistical power given the time and the number of participants we could afford. The weaknesses of the within-group design, namely fatigue and learning effect, have been minimised by randomising the sessions and scenarios.

3.3 Statistics

The analysis of the sessions was primarily made with video analysis. Subjective data were supplied in the form of responses to a questionnaire filled out by respondents some days after the sessions.

While seeking to compare the two interfaces (with or without speech input facilities), it was not obvious how to identify an objective indicator of the completeness of the anaesthesia record and of the cognitive load related to this record. Statistics such as the average time between an event and its registration, or the time spent to fill the record are not good enough. Indeed many events are not registered during an anaesthesia crisis and are possibly handled afterwards. For those events, it was neither possible to assign a time when the registration was done, nor how much resources their registration required during the crisis.

A more robust and appropriate metric was inspired by queuing theory, *i.e.* the theory of waiting lines such as messages to be handled or tasks to be completed [Kozine 2007; Liu 1997]. For our application, the queue is the "average queue of events waiting to be registered". Each time an event that must be registered occurs, the queue (or stack) size is increased by one; when this event is registered, the queue size is decreased by one. The final measure is the averaged queue size over the simulation scenario.

$$W = \frac{\sum_{n=0}^{n=N-1} Q_n \times \left(t_{n+1} - t_n\right)}{t_N - t_0}$$

Where: W is the averaged queue of events to be registered (workload), t_n is the time in seconds of an event or a registration, Q_n is the queue size at time t_n (when t_n is an event, Q_{n+1} is increased; when t_n is a registration, Q_{n+1} is decreased), N is the total amount of events and registrations. Q is set to zero at the beginning of the simulation. A first event t_0 is added for the beginning of the simulation, and a last event t_n with $n=N$ for the end of the simulation.

In the cases for which a registration appends before its corresponding event, the queue is increased by one at the registration time, and decreased by one when the real event occurs.

3.3.1.1 Example of queue measurement

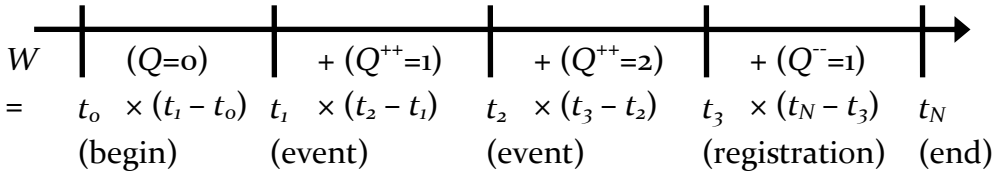


Figure 1: example of workload calculation using the proposed approach based on queuing theory.

In the example shown on Figure 1, lasting 40 seconds where each interval is 10 seconds, with two events and then one registration, the average queue size is:

$$W = [(0 \times 10\ \text{s}) + (1 \times 10\ \text{s}) + (2 \times 10\ \text{s}) + (1 \times 10\ \text{s})] / 40\ \text{s} = 1$$

3.3.2 Questionnaires

Table 2 reports the questions (translated in English) given to the anaesthesia nurses after the experiment. The anaesthesia doctors received similar questions, adapted to the fact that only nurses have dictated to the system. See Appendix for the full details.

Table 2: Questionnaire for the participants.

#	Questions	Scales of answer
q1.	What is the degree of similarity between the real Dräger system in use at Køge hospital for the electronic anaesthesia record and the prototype for the tasks needed during the simulation?	1: 0% similar 5: 100% similar
q2.	Based on your own experience, how useful is it to have an anaesthesia record up to date during the operation?	1: Not useful 5: Very useful
q3.	How often do you use the anaesthesia record to help you remembering what appended or as a support to take new decisions?	1: Never 5: Always
q4a.	In the first session with the traditional interface (without speech), how difficult was it to fill the anaesthesia record during the scenario?	1: Impossible 5: Very easy
q4b.	In the second session with speech interface, how difficult was it to fill the anaesthesia record during the scenario?	1: Impossible 5: Very easy
q5a.	To which extend filling the anaesthesia record with the traditional interface reduced the time you could use for the patient?	1: No reduction 5: Too much time
q5b.	To which extend filling the anaesthesia record with the voice interface reduced the time you could use for the patient?	1: No reduction 5: Too much time
q6a.	To which extend filling the anaesthesia record with the traditional interface disturbed your primary work or reduced your concentration?	1: No disturbance 5: Too much disturbance
q6b.	To which extend filling the anaesthesia record with the voice interface disturbed your primary work or reduced your concentration?	1: No disturbance 5: Too much disturbance
	If the traditional interface was supplemented by a almost perfect speech recognition system, how would that impact the quality of the anaesthesia record?:	
q7.	a. Up-to-date at any time during the operation	1: Clearly negative 3: Neutral impact 5: Clearly positive
q8.	b. Completeness at any time during the operation	1: Clearly negative 3: Neutral impact 5: Clearly positive
q9.	c. Completeness after the operation	1: Clearly negative 3: Neutral impact 5: Clearly positive
q10.	How would you rate the overall utility of having such a speech interface in addition to the current touch-screen and keyboard interface?	1: Not useful 5: Very useful

3.3.3 Video analysis

During the video analysis, the time stamps for most of the events of interest were recorded; for instance, the details of all the registrations in the record, all the medications given and major actions on the patient such as intubation or heart massage. In average, 74 events were transcribed per session. The exact transcription of what was dictated was registered together with what was actually recognised by the speech recognition engine, as exemplified in Table 3. This type of video analysis is common in HCI studies [Kushniruk & Patel 2004]. Afterwards, the events used for making the analysis and the statistics were selected. Particular attention has been made to use the precise same selection criteria between the two sessions (first without voice, second with voice) of a given anaesthesia team. In order to know if a given minor event should have been recorded in the anaesthesia record or not, some comparisons across teams have been made and if some other teams made the effort of registering a similar event, the registration was considered "required". Doing so, the expertise of the participants was used indirectly to make the classification of the events.

Table 3: Short excerpt from a transcript of session 12, translated into English. The code "ASR" stands for "Automatic speech recognition".

	Event 50	Event 51	Event 52	Event 53
Time begin	00:15:04	00:15:05	00:15:05	00:15:13
Time end			00:15:08	00:15:15
Time since event			00:00:04	00:00:03
Time accuracy of registration			00:00:04	00:00:03
Stack size	1	2	1	0
Nurse	Start "Voluven"	Stop "NaCl"	ASR "Computer Voluven infusion 500"	ASR "Computer sodium... [> 1 s pause] chloride stop"
Doctor				
Patient				
Speech recognition			OK "Computer Voluven infusion 500"	ERROR (Nothing recognised: too much delay)

3.3.4 Speech recognition rates

The main goal of the study was not to measure recognition rates, which were known in advance to be low, mainly due to the lack of preparation of the participants. However, during the data analysis, the author tried to distinguish the recognition errors due to the speaker from those due to the system. This process relies mainly on factual assessment and is therefore reasonably objective: the dictations with dysfluencies such as repetitions, "uh", noticeable hesitations, and incorrect syntax were categorised as speaker errors.

Once this categorisation done, the reported speech recognition rates indicate a "semantic accuracy" [Alapetite 2006], that is to say, the percentage of transcriptions that can be understood without ambiguity by a skilled human reader.

4 Results

4.1 General subjective data

We received questionnaire replies from 10 participants (6/6 nurses, 4/6 doctors) who rated the speech recognition interface and the realism of the experiment. Ratings were given on a 5-point Likert-type scale.

The average rating of the realism of the mock-up when compared with the original electronic anaesthesia record was 3.5 (question 1 = q1, potential range 1 to 5, where 1 is full disagreement, 3 is neutral and 5 is full agreement). They agreed positively on the utility of having an up-to-date record all along the operation (q2, 4.3/5), independently of the interface, should it be *e.g.* paper, touch-screen or voice. They reported to frequently use the anaesthesia record during operation as a support for memory and decisions (q3, 4.2/5). Those results (Figure 2) are close to what was expected. None of the questions were answered with a significant difference between nurses and doctors (Mann-Whitney U test; $p > 0.7$, $p > 0.2$, $p > 0.9$ for the three questions of Figure 2).

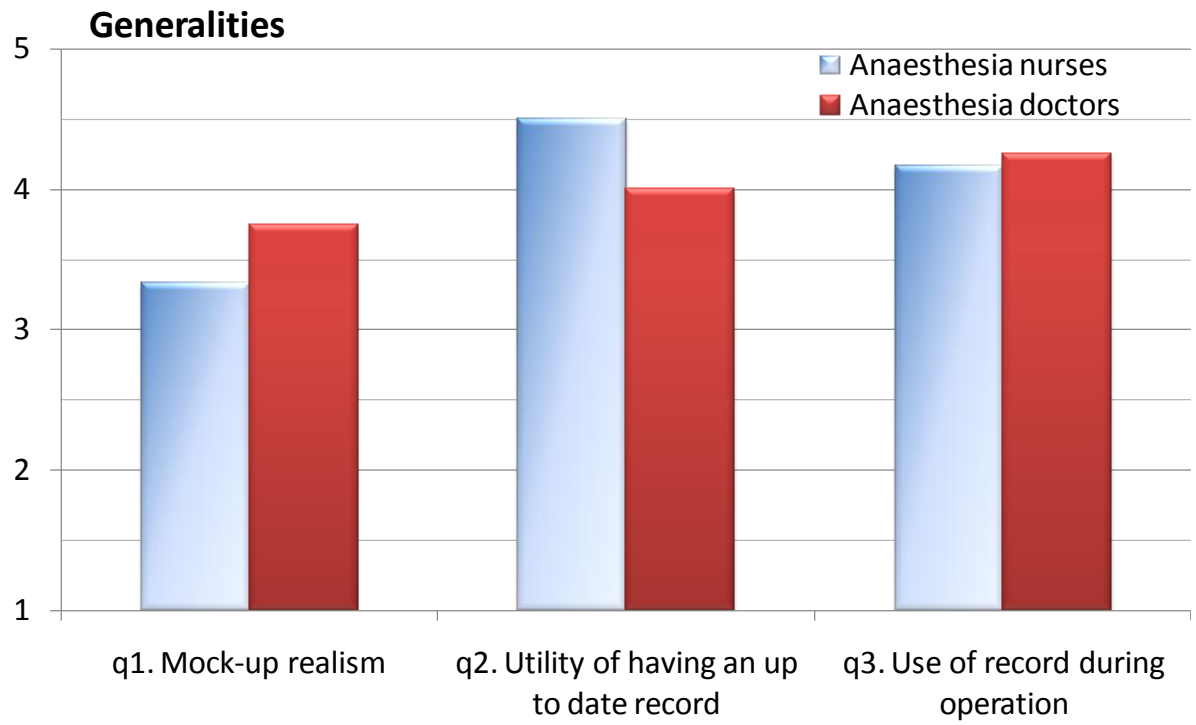


Figure 2: Questionnaire responses on general questions.

4.2 Record completeness and workload

In accordance with the objectives of the study, we have sought to identify indicators that can be used to reveal mental workload, comparing the two types of interfaces (with and without voice).

4.2.1 Subjective results

As shown by Figure 3, the participants found it slightly more difficult to update the anaesthesia record by voice (q4, 3.1/5 versus 2.8/5), and this modality required a little more concentration than the traditional interface (q6, 2.8/5 versus 2.5/5). Those small differences have been shown as not significant with a Mann-Whitney U test ($p > 0.4$, $p > 0.1$, $p > 0.6$ for the three questions of Figure 3), partly due to small samples. The small differences could at least be partially explained by the fact that the participants were accustomed to the traditional interface, but tried the speech interface for the first time.

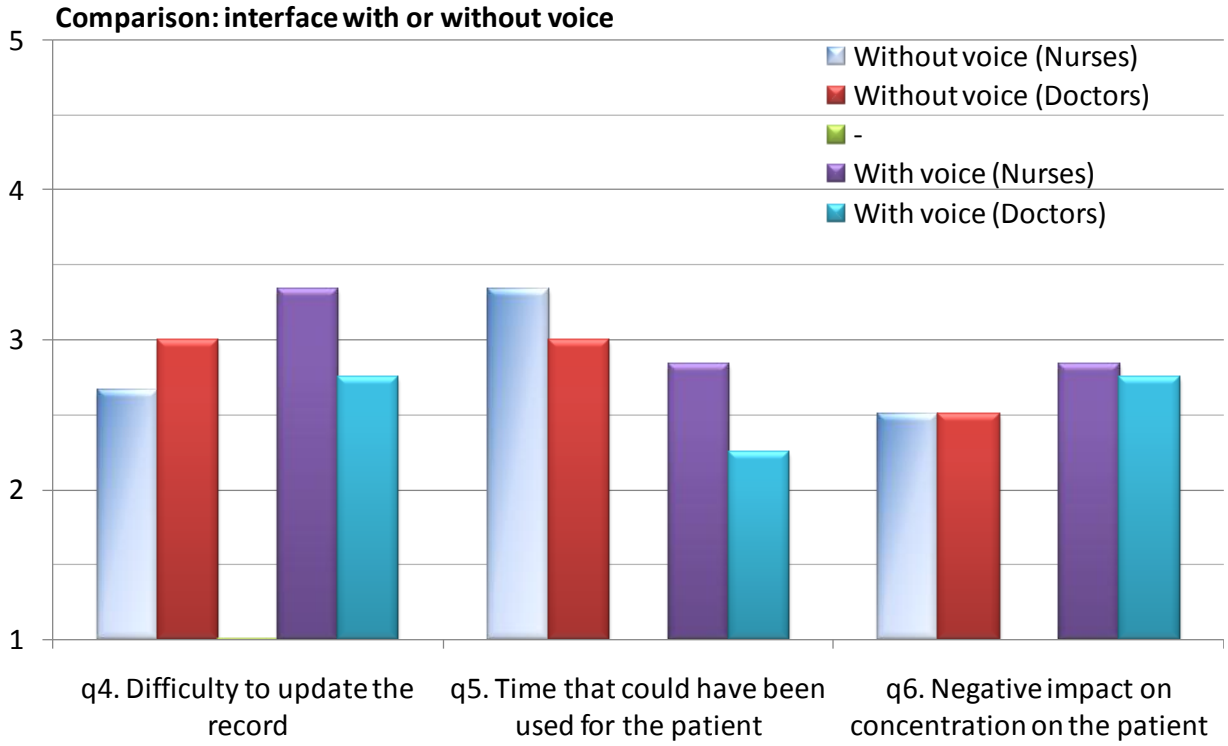


Figure 3: Questionnaire responses on time and difficulty to keep the anaesthesia record updated during the scenario, with or without voice. (See Table 2 for the full questions).

There is however the impression that the speech interface can save some time that can instead be used for the patient (q5, 3.2/5 versus 2.6/5) where 1 is when no time and 5 is too much time that instead could be used for the patient. See Appendix 1 for more details.

4.2.2 Quantitative measurements

While subjective results tend to be in favour of the traditional interface, objective results give a clear advantage to the speech interface – although it must be kept in mind that the speech interface was an ideal one, where failure of recognition was cancelled out by the Wizard of Oz setting, thus removing the negative effect of incorrect recognitions.

The sessions lasted on average 31 minutes without voice and 26 minutes with voice, but the differences are not significant ($p=0.14$, independent samples t-test).

As shown in Figure 4, the average "time spent to fill the record" is only slightly below with voice (2 min 42 s) than with the traditional interface (3 min 50 s, $p < 0.14$). However, this should be viewed in parallel with the fact that almost two times more registrations have been made in average with voice (26.5) than without (13.5, $p < 0.001$), as reported later in the study of the anaesthesia record quality (Table 4). This means it took on average 17 seconds per event registration with the traditional interface, and almost three times less with speech recognition, down to 6 seconds per registration ($p < 0.002$).

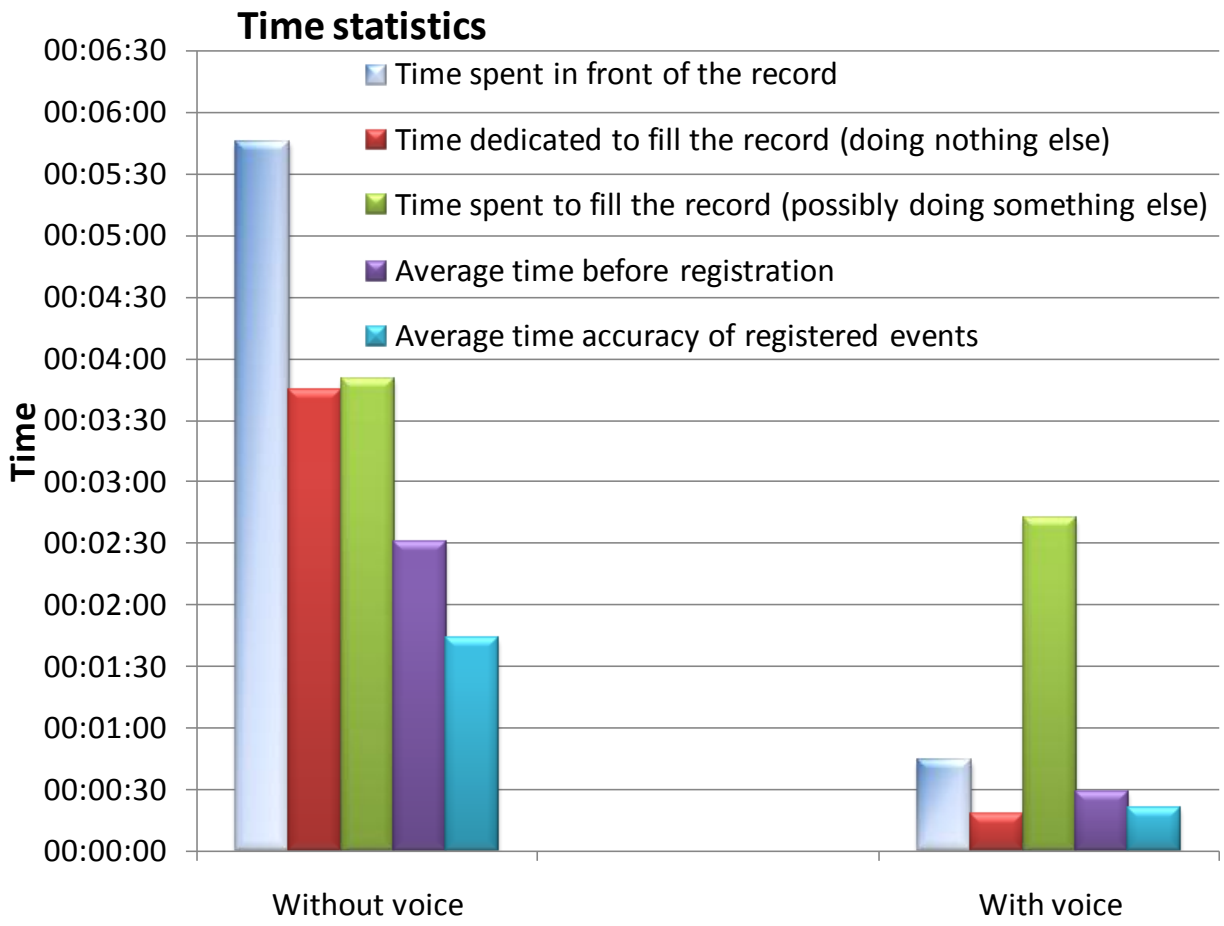


Figure 4: Measurements of delays and time used to fill the anaesthesia record, with or without voice.

On Figure 4, the "time dedicated to fill the record" (which means that the participant did nothing else in this period), is much reduced with the use of voice, from 3 min 45 s down to 18 s on average ($p < 0.003$). This is due to the fact that anaesthesia nurses could dictate some commands while performing what they were describing, such as manual ventilation, intubation, injection, etc. It should be noted that a few cases were observed where anaesthesia nurses could fill the record with the traditional interface using one hand while doing other things with the other hand. The difference between the time "spent" and the time "dedicated" to fill the record is an indicator of the time that was used for filling the record while possibly doing something else.

The "average time before registration" is the observed delay between one event and its registration in the record (Figure 4). As mentioned above, this indicator is afflicted by missing data, since events that had not been registered when the session was ended are not included. It shows, however, some clear differences between the two interfaces: when using voice, it took in average 2 min 31 s before registering an event, and they were registered more than 5 times quicker with voice ($p < 0.001$), on average 29 s later.

None of the measured parameters showed a statistically significant difference between the two scenarios ("anaphylaxis" and "asystole", $p > 0.3$, t-test), which supports the assumption that they were sufficiently similar for the purpose of this experiment.

With the traditional interface, the long delay before registration leads to queues of events that accumulate, as reported on Figure 5, and the queue increases all along the anaesthesia scenario. In contrast, the queue is kept small with the speech interface. As shown in Table 4, the average queue of events is 5.79 with the traditional (maximum at 11.67 on average) and is almost five times smaller with the vocal interface ($p < 0.001$), at 1.2 (maximum at 3.17 on average). Those results show also that it is possible for anaesthesiologists to verbalise their main actions even during difficult scenarios with emergency situations.

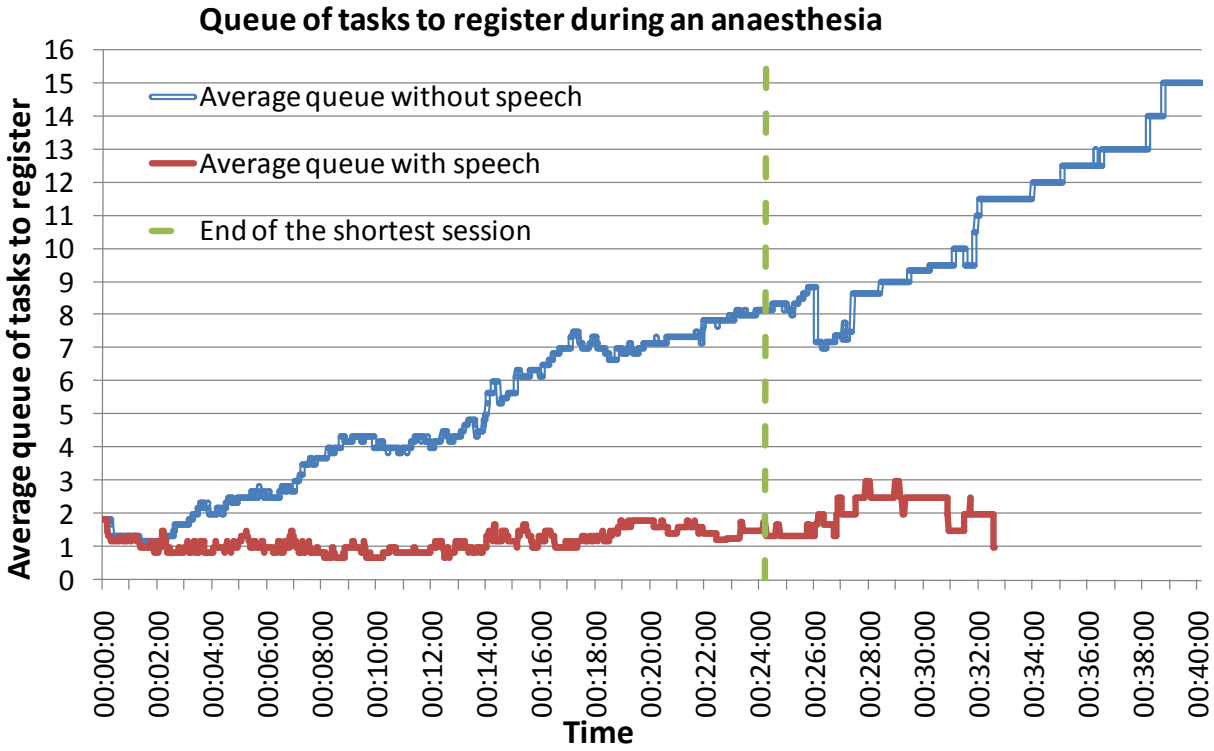


Figure 5: Evolution of the averaged queue of events to register during the anaesthesia scenarios, with or without voice.

In Figure 4, the "time spent in front of the record" is the time spent looking at the record, or walking toward it. With the traditional interface, it seems that anaesthesiologists had to think more and use more time in front of the record ($p < 0.001$) trying to reconstruct from memory what had happened and when. One of the salient differences that were revealed between the interactions with the two types of interface was that with the traditional interface the nurse had to spend time finding the correct category of medication. Medications are indeed organised in categories and the anaesthesiologist must know to which category a given medication to be registered belongs. For instance, four out of six anaesthesiologist nurses had difficulties (selecting at least one wrong category, or asking the doctor to help) or failed to find "Adrenalin", which is a medication well known to the nurses, but not often used in planned operations.

4.3 Anaesthesia record quality

Being capable of filling the record with minimal delay is only one of the considered parameters, but it is naturally of crucial importance to ensure the quality of the record.

Of particular importance is the percentage of medications recorded. As reported in Table 4, less than 56% of the administrated medications were registered *via* the traditional interface before the end of a scenario, while almost 99% of the medications were recorded in time with the vocal modality.

Table 4: Statistics measures from video analysis,
with or without voice, each condition averaged across 6 sessions.

	Without voice	**With voice**	**Independent-Samples t-test**
Number of fixed events registered	3.50 (84%)	5.67 (89.47%)	$p < 0.005$
Total number of fixed events	4.17	6.33	$p < 0.03$
Number of free text events registered	0.67 (40%)	4.33 (89.66%)	$p < 0.03$
Total number of free text events	1.67	4.83	$p < 0.03$
Number of medications registered	7.83 (55.95%)	13.00 (98.73%)	$p < 0.03$
Number of medications with error	0.83 (10.64%)	0.33 (2.56%)	$p = 0.3$; NS
Total number of medications	14.00	13.17	$p = 0.7$; NS
Number of air or liquids events registered	1.50 (56.25%)	3.50 (95.45%)	$p < 0.03$
Total number of air or liquids events	2.67	3.67	$p < 0.07$
Total number of registered events	13.50	26.50	$p < 0.001$
Average queue of events to register	5.79	1.20	$p < 0.005$
Max queue length	11.67	3.17	$p < 0.005$

In Table 4, the so-called "fixed events" are the common ones (*e.g.* surgeon begins, intubation) that can be selected from a list or dictated in command mode, while the "free text events" are the uncommon ones that must be typed using the keyboard or dictated in free text mode. Aggregating those two categories of events, it shows that 71.4% of events were recorded with the traditional interface, versus 89.5% with the vocal modality. With the traditional interface, the recorded events were mainly the very common ones (*e.g.* intubation, surgery started) while the uncommon ones were missed (*e.g.* defibrillation, heart stop). With speech recognition, there was a similar rate of recording between events that were available in the predefined list or not, both over 89%. The "air and liquids" (oxygen, glucose, NaCl, etc.) events were of less importance during the simulations, but show a similar advantage for the speech interface.

As shown in Figure 4, the time accuracy of the registered events was almost five times higher with the vocal interface (21 seconds accuracy) than with the traditional interface (1 min 44 s, $p < 0.005$).

In total, there were five errors (*i.e.* wrong medication or dosage) while recording medications with the traditional interface (10.7% of the registered medications) versus two errors with the vocal interfaces (2.6%). Even though the mock-up was not strictly identical to the anaesthesia record participants were used to, the selection of the medications was very similar to the original.

Finally, when used correctly, the opportunity to use speech input can also improve team situation awareness and mutual verification. There was indeed one example of a nurse registering by voice one medication, which was the wrong one; the error was immediately spotted by the doctor who could hear it.

4.4 Speech recognition accuracy

4.4.1 Keyword based strategy for the speech interface

The keyword based approach with speech recognition running permanently worked even better than expected. During the two hours and a half of cumulated time for sessions with speech recognition, no voice command was recognised by the system that was not targeted to the system. This ability of the system not to include non-intended speech is not trivial, since a speech recognition system will naturally tend to recognise possible words out of random speech or even noise. This result demonstrates the feasibility of using speech recognition without button activation even in noisy environment.

Another encouraging result was the flexibility of the keyword activation: if a user starts saying a command but aborts for any reason (*e.g.* hesitation, error), the user may simply begin once again. For instance, a user would say "Computer Propofol… uh… Computer Propofol bolus 60". This feature has been extensively used by the participants, in a very natural way and without experiencing any trouble.

As far as the video analyses have shown, starting each dictation targeted to the anaesthesia record by the keyword "Computer…" was sufficient to make it clear that what was being said was for the record and not for the other member of the medical team. There was no case of misunderstanding between the members of the medical team imputable to the vocal modality. This characteristic of the keyword based vocal interface would have been more difficult to achieve when using *e.g.* a speech input controlled by a button because in the absence of feedback, only the speaker typically knows when such a button is pressed.

4.4.2 Recognition rates

Even though this experiment was not aimed at measuring speech recognition rates, the data collected nevertheless yielded some statistics about the accuracy from novices using a minimally trained system for the first time.

The categorisation of the types of dictation errors, correct dictation and recognition rates is reported in Figure 6.

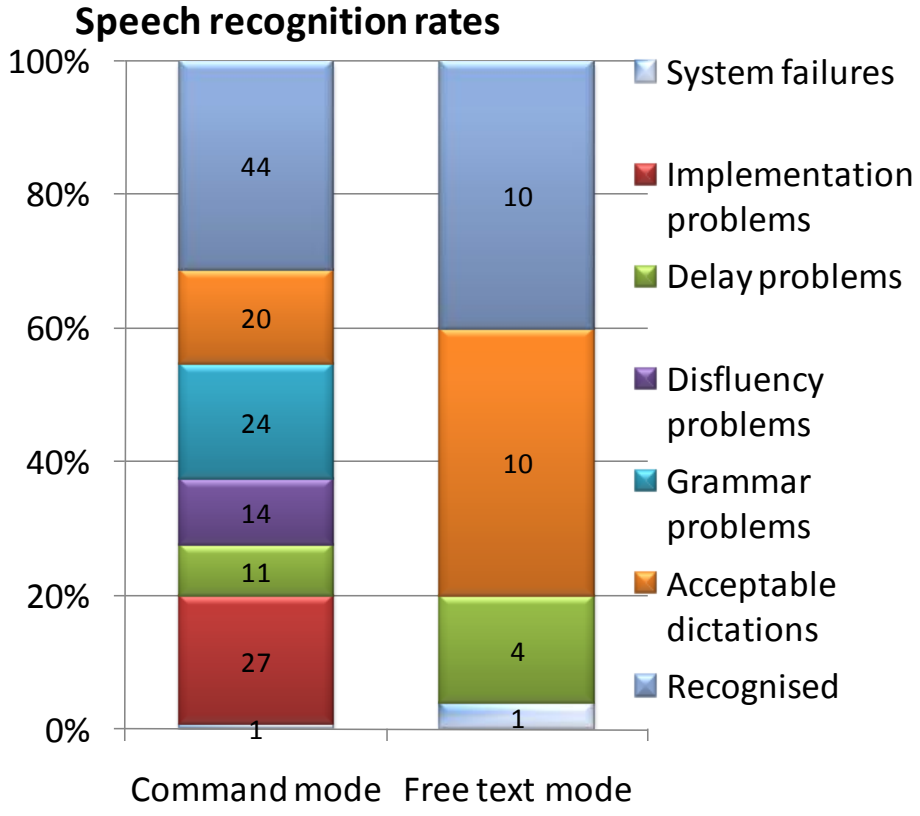


Figure 6: Categorisation of dictations and recognitions.

In command mode, the "non-acceptable" dictations (55%) were mainly due to implementation limitations (35% of them), *i.e.* features that would be added to the system if a new version was to be done. This includes missing abbreviation of medications, or the fact that the participants often dictated units when registering dosages, while the grammar expected only numbers. The second larger set of dictation problems is related to the lack of user compliance with the syntax (31%). Dysfluencies (*e.g.* "uh", repetitions) and delay problems (too long pauses) are responsible for 32% of the "non-acceptable" dictations.

When considering only the "acceptable dictations", the recognition rate was 69% in command mode, and 50% in free text mode. Within the "acceptable dictations", wrong recognitions are due to the speech recognition system limitations, but also to the speaker elocution that can be more or less suited to automatic speech recognition.

All attempts by participants to start the free text mode using the keywords "Computer remark…" succeeded. Then, in 20% of the cases, there was a delay problem due to the participants not waiting for the free text mode to be ready (~one second delay, sound feedback when ready) and speaking too early. The remaining types of non-successful dictations are too subjective to be classified.

There were two "system failures": the first in command mode where the system was not ready when the nurse did her first dictation, the second in free text mode where the program dedicated to free text crashed. In both cases, one dictation was missed.

The best recognition rates for one person were 86% recognition rate in command mode and 71% recognition rate in free text mode, and the worst rates for one person were, respectively, 57% and 20%. These recognition rates are still below those (98+%) that can be achieved by experienced speakers using a trained system [Happe *et al.* 2003].

4.4.3 Overall utility of speech interface

In the questionnaire, and as shown in Figure 7, the participants ranked the general usefulness of this speech interface to be 4 out of a maximum of 5, if recognition rates could reach satisfying levels.

During the operation, the speech interface reduces the delays in registrations, and it may therefore be assumed that it would help in producing more accurate and correct entries. In this regard, the average utility of the speech interface during operation was ranked 4.25/5.

Similarly, participants were asked to imagine a speech recognition system working with 100% recognition and rate this for its ability to improve the quality of the record in terms of completeness. The average response showed on average a ranking of 4 out 5.

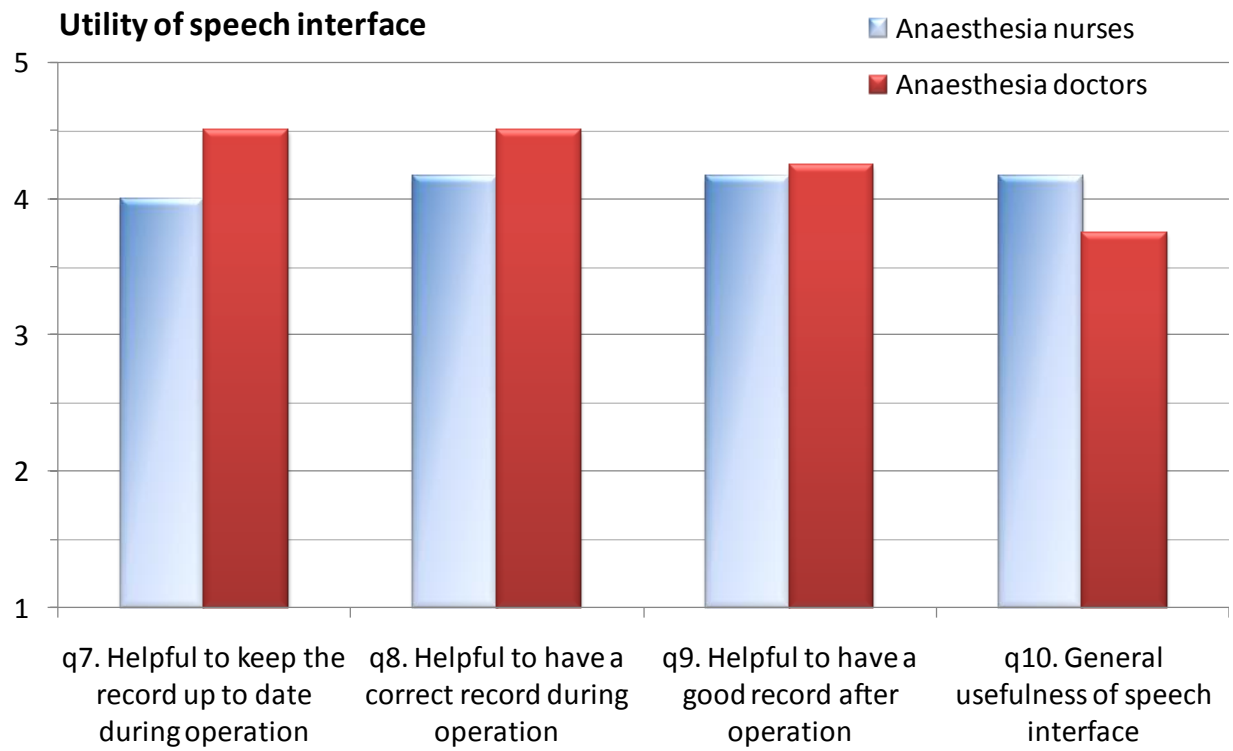


Figure 7: Questionnaire responses on the overall utility of the speech interface. (See Table 2 for the full questions).

Finally, in the free text section of the survey, some participants shared their views and concerns regarding a vocal interface. Six of the ten respondents reported that the vocal modality would be useful to have because it helps to produce more accurate and real-time data; five respondents said it would help in keeping hands free and a visual contact with the patient; one saw a possible improvement in hygiene. On the negative side, four of the respondents were concerned about having to learn a new tool and two about the increase in noise in the operation room.

5 Discussion

The proposed queue-based metric of the workload associated with delaying registrations is, the author suggests, a useful indicator of the mental workload related to the anaesthesia record. Measuring elements of performance in a secondary task is often needed in human factors research [Sauer 2000] and the author believes this metric to be an improvement over some other traditional indicators such as the time to completion, when it comes to handle queues of tasks and to allow an interruption of the scenario before all the tasks are completed. While queuing theory principles are used in simulations to model human performance [Liu 1997], they are apparently not commonly used so far to analyse real data, as does the queue-based metric suggest here.

While the supplemental vocal interface objectively allows a reduction of the queue of events waiting to be registered in the record, this experiment has not delivered data (and was not designed to do so) that show the gains in performance on the primary task. It may be expected that when users can concentrate on their primary task, their performance will benefit from this. However, there is the possibility that when events are quickly registered, this may have a potentially negative effect on situation awareness since the anaesthesiologist is no longer forced to keep registrations in mind. Perhaps this is similar to the potential loss of awareness of vital signs that happened when the transition from paper-based to electronics records took place. With the electronic record, it was no longer needed for the anaesthesiologist to write down vital sign trends, which were then automatically registered by the anaesthesia monitors.

As Table 4 shows, there were more events on average during sessions using speech recognition than during sessions with the traditional touch screen based interface. To a large extent, this is due to a difference in the way in which anaesthesiologists were registering events with the two interfaces. Thus, when participants used the traditional interface, there was a tendency for them to aggregate events together and then, when there was time for this, to register these events in combination when possible. For instance, when two bolus injections of a medication were made within a short time period, participants using the traditional interface were likely to record only a single event combining the sum of the two boluses, while they always detailed the two events when using voice input. Similarly, when using the traditional interface, practitioners would typically report only one event when they repeatedly modified the rate of an infusion within a short time period, while they tended to register each modification when registering with the vocal facilities. The same tendency was apparent when participants registered several acts of defibrillations or other actions.

It would have been desirable to have run the experiment with a much higher level of prior training of participants in using the speech interface; and similarly, it would have been desirable if participants had had prior familiarity with the anaesthesia simulator and the anaesthesia record mock-up. But this was unfortunately not possible due to time and resource constraints. In particular, if it had been possible to achieve recognition rates during the simulations comparable to those obtained with well-trained users operating mature systems, there would not have been a need of using the Wizard of Oz technique.

It should be emphasised that during crisis situations in real situations, the anaesthesia team typically calls for external assistance, and if some colleagues are available, a third person helps in handling the situation and in filling the anaesthesia record.

Conclusion

This paper has reported results of the evaluation of an anaesthesia record speech recognition interface that is permanently listening and becomes activated by keywords. The evaluation results show that a hands-free vocal interface may be used efficiently to register events while they are happening, thus avoiding an accumulation of events awaiting registration. The experiment has shown that speech based registration can be performed accurately even during emergencies and time critical scenarios, while providing some benefits for the team situation awareness.

The “average queue of events” metric introduced in this article appears to be a useful indicator of mental workload when users have to handle two or more simultaneous tasks.

Participants’ use of the speech recognition interface, arguably because of lack of training, did not yield a performance that would be satisfactory for daily use. In particular, the free text mode offered only poor recognition rates, especially when other people were speaking at the same time. However, the command mode performed better and was quite insensitive to background noise, reaching recognition rates around 70% when inputs complied with the grammar and the constraint of being dictated without pause. At the same time, the experiment also showed that the chosen speech recognition system will require an extensive training phase for each user, involving both time to train the individual voice profile on the machine, and also time to practice dictations so that commands are enunciated clearly and without hesitation.

More generally, the article provides some subjective and objective data that show some of the limits of the current touch screen based interface for the electronic anaesthesia record, and it has quantified some of the possible benefits that could be achieved by supplementing current interfaces with speech input facilities.

Acknowledgements

This work was supported by the Fifth European Community Research and Development Framework, Improving Human Potential Programme, within the ADVISES Research Training Network about "Analysis Design and Validation of Interactive Safety critical and Error-tolerant Systems". Thanks to Max Manus for providing the speech recognition software. Thanks to the Danish Institute for Medical Simulation of Herlev University Hospital, in particular Doctors Doris Østergaard, Ann Moller and Nini Vallebo for their help on the experimental design and during the simulations. Thanks to the participants from Køge Hospital, in particular Doctor Viggo Stryger for his expertise on the anaesthesia system, including during real anaesthesias, and critique of the prototype. Some credit should also be given to my supervisors Morten Hertzum (Roskilde University, DK) and Henning Boje Andersen (Risø National Laboratory) for organisation, inspiration, reviewing and substantial improvement of the writing style.

References

[Alapetite & Gauthereau 2005] Alexandre Alapetite & Vincent Gauthereau. Introducing vocal modality into electronic anaesthesia record systems: possible effects on work practices in the operating room. *Proceedings of EACE'2005 (Annual Conference of the European Association of Cognitive Ergonomics) 29 September - 1 October 2005, Chania, Crete, Greece; section II on Research and applications in the medical domain, 189–196. ACM International Conference Proceeding Series, vol. 132. University of Athens, 197-204,* ISBN:9-60254-656-5.

[Alapetite 2006] Alexandre Alapetite. Impact of noise and other factors on speech recognition in anaesthesia. *International Journal of Medical Informatics, available online in December 2006.* doi:10.1016/j.ijmedinf.2006.11.007

[Cowan 2000] Nelson Cowan, The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, Cambridge University Press, 2001, 24:87-114.* doi:10.1017/S0140525X01003922

[Detmer *et al.* 1995] William M. Detmer, Smadar Shiffman, Jeremy C. Wyatt, Charles P. Friedman, Christopher D. Lane, Lawrence M. Fagan. A Continuous-speech Interface to a Decision Support System: II. An Evaluation Using a Wizard-of-Oz Experimental Paradigm. *Journal of the American Medical Informatics Association, Jan–Feb 1995, 2(1):46–57.*

[Devos *et al.* 1991] Cathy B. DeVos, Martin D. Abel, John P. Abenstein. An evaluation of an automated anesthesia record keeping system. *Biomedical Sciences Instrumentation, 1991, 27:219-25.*

[Gaba *et al.* 1988] David M. Gaba, Abe DeAnda. A Comprehensive Anesthesia Simulation Environment: Re-creating the Operating Room for Research and Training. *Anesthesiology, 1988, 69:387-394.*

[Gröschel *et al.* 2004] J. Gröschel, F. Philipp, St. Skonetzki, H. Genzwürker, Th. Wetter, K. Ellinger. Automated speech recognition for time recording in out-of-hospital emergency medicine – an experimental approach. *Resuscitation, 2004, 60:205–212.* doi:10.1016/j.resuscitation.2003.10.006

[Happe *et al.* 2003] André Happe, Bruno Pouliquen, Anita Burgun, Marc Cuggia, Pierre Le Beux. Automatic concept extraction from spoken medical reports. *International Journal of Medical Informatics, 2003, 70:255-263*. doi:10.1016/S1386-5056(03)00055-8

[Jungk *et al.* 2000] Andreas Jungk, Bernhard Thull, Lutz Fehrle, Andreas Hoeft, Günter Rau. A case study in designing speech interaction with a patient monitor. *Journal of Clinical Monitoring and Computing, 2000, 16:295-307*. doi:10.1023/A:1011456205786

[Kozine 2007] Igor Kozine. Simulation of human performance in time-pressured scenarios, Proceedings of the Institution of Mechanical Engineers. *IMechE'2007, Vol. 221, Part O: Journal of Risk and Reliability, pp. 141–152*. doi:10.1243/1748006XJRR48

[Kushniruk & Patel 2004] Andre W. Kushniruk & Vimla L. Patel, Cognitive and usability engineering methods for the evaluation of clinical information systems. *Journal of Biomedical Informatics, 2004, 37(1):56–76*. doi:10.1016/j.jbi.2004.01.003

[Liu 1997] Yili Liu, Queueing Network Modeling of Human Performance of Concurrent Spatial and Verbal Tasks. *IEEE Transactions on Systems, Man and Cybernetics, Part A, March 1997, 27(2):195-207*. doi:10.1109/3468.554682

[Lovis *et al.* 2000] Christian Lovis, Robert H. Baud, Pierre Planche, Power of expression in the electronic patient record: structured data or narrative text? *International Journal of Medical Informatics, 2000, 58–59:101–110*. doi:10.1016/S1386-5056(00)00079-4

[Østergaard 2004] Doris Østergaard, National Medical Simulation training program in Denmark. *Critical Care Medicine, 32(2) (Supplement):S58-S60, February 2004*. doi:10.1097/01.CCM.0000110743.55038.94

[Sanjo *et al.* 1999] Yoshimitsu Sanjo, Tetsuo Yokoyama, Shigehito Sato, Kazuyuki Ikeda, Reiko Nakajima. Ergonomic automated anesthesia recordkeeper using a mobile touch screen with voice navigation. *Journal of Clinical Monitoring and Computing, 1999, 15:347-356*. doi:10.1023/A:1009972223750

[Sauer 2000] Juergen Sauer, Prospective memory: a secondary task with promise. *Applied Ergonomics, April 2000, 31(2):131-137*. doi:10.1016/S0003-6870(99)00042-3

[Sears *et al.* 2003] Andrew Sears, Jinjuan Feng, Kwesi Oseitutu, Clare-Marie Karat. Hands-Free, Speech-Based Navigation During Dictation: Difficulties, Consequences, and Solutions. *Human-computer interaction, 2003, 18:229–257*. doi:10.1207/S15327051HCI1803_2

[Shiffman *et al.* 1995] Smadar Shiffman, William M. Detmer, Christopher D. Lane, Lawrence M. Fagan. A continuous-speech interface to a decision support system: I. Techniques to accommodate for misrecognized input. *Journal of the American Medical Informatics Association, 1995, 2(1):36–45*.

Appendix

Appendix 1: Main speech recognition grammar (in Danish)

```
#JSGF V1.0;
grammar anaesthesia_commands {PspLanguage = 17;};

concept <command>;
concept <medication>;
concept <vaeskeind>;
concept <gas>;

declarations
{
        integer cmdid:<command>;
        integer medid:<medication>;
        string medName:<medication>;
        string medMode:<medication>;
        float medQuant:<medication>;
        string medAction:<medication>;
        string viName:<vaeskeind>;
        string viAction:<vaeskeind>;
        string gsName:<gas>;
        string gsAction:<gas>;
}

<commands> =
bemærk |
"akut indlæggelse" |
"alle kirurgiske procedurer afsluttet" |
allergi |
"allergi og faste ok" |
"anlæg af dræn" |
antibiotika |
anti-emetika |
"aspiration af ventrikelindhold til lunger" |
"assisteret ventilation" |
asystoli |
blodtryksfald |
cementering |
"dårlig oversigt" |
dræn |
duodenalsonde |
"ekstrem bradykardi" |
ekstubation |
endokrint |
engangskateter |
forbinding |
gastrointestinalt |
gennemlysning |
gipsning |
"I D ok" |
"i seng" |
induktion |
intubation |
ketalar |
"kirurg slut" |
"kirurg start" |
"klinisk hjertestop" |
koagulation |
"kontrolleret ventilation" |
"kortvarigt blodtryksfald" |
kramper |
kulderystelser |
kvalme |
lejring |
"må køre til stamafdeling" |
"malign hypertermi" |
metadon |
opvågningen |
"på lejet" |
"på stuen" |
"paravenøs injektion" |
"patient afleveret" |
"patient fastende" |
"patient fryser" |
"patient klar til operation" |
"patient modtaget på stuen" |
renalt |
respirationsstop |
respiratorisk |
revertering |
"se notat" |
"slut på anæstesi" |
"spinal tilfælde" |
"spontan respiration" |
"start af anæstesi" |
"stilet i tube" |
"stiv nakke" |
"store sekretmængder" |
"subkutan infusion" |
"subkutan injektion" |
sugning |
"svær intubation" |
"svær intubation med fiberscop" |
"svær intubation på spontan respiration" |
tandskade |
trendelenburg |
"udskrives til stamafdeling" |
"vandret leje" |
"vanskelig intravenøs-adgang" |
"varmt tæppe er givet" |
"venflon proppet" |
"venter på anæstesilæge" |
"venter på kirurg" |
"venter på portør" |
ventrikelaspiration |
ventrikelflimmer |
ventrikelsonde;

<command> = computer <commands>;

<medications> =
ACTRAPID {medName="ACTRAPID";} |
ADRENALYN {medName="ADRENALYN";} |
AMIODARON {medName="AMIODARON";} |
AMPICILLIN {medName="AMPICILLIN";} |
Atropin {medName="Atropin";} |
"ATROPIN NEOSTIGMIN" {medName="ATROPIN
NEOSTIGMIN";} |
Atrovent {medName="Atrovent";} |
BENZYL-PENICILLIN {medName="BENZYL-
PENICILLIN";} |
Bricanyl {medName="Bricanyl";} |
```

```
"BUPIVACAIN PLAIN" {medName="BUPIVACAIN PLAIN";} |
"BUPIVACAIN TUNG" {medName="BUPIVACAIN TUNG";} |
"CALCIUM CHLORID" {medName="CALCIUM CHLORID";} |
CEFUROXIM {medName="CEFUROXIM";} |
CLEMASTIN {medName="CLEMASTIN";} |
Combivent {medName="Combivent";} |
Cordarone {medName="Cordarone";} |
Cyklokapron {medName="Cyklokapron";} |
DEXAMETHASON {medName="DEXAMETHASON";} |
DIAZEPAM {medName="DIAZEPAM";} |
Diclocil {medName="Diclocil";} |
DICLOXACILLIN {medName="DICLOXACILLIN";} |
DIGOXIN {medName="DIGOXIN";} |
DOBUTAMIN {medName="DOBUTAMIN";} |
Dobutrex {medName="Dobutrex";} |
DOPAMIN {medName="DOPAMIN";} |
Dopram {medName="Dopram";} |
Dormicum {medName="Dormicum";} |
DOXAPRAM {medName="DOXAPRAM";} |
EFEDRIN {medName="EFEDRIN";} |
ESMERON {medName="ESMERON";} |
Fenemal {medName="Fenemal";} |
FENTANYL {medName="FENTANYL";} |
Fortecortin {medName="Fortecortin";} |
Furix {medName="Furix";} |
FUROSEMID {medName="FUROSEMID";} |
Garamycin {medName="Garamycin";} |
GENTAMICIN {medName="GENTAMICIN";} |
GLYCOPYRRON {medName="GLYCOPYRRON";} |
HYDROCORTISON {medName="HYDROCORTISON";} |
HYPNOMIDAT {medName="HYPNOMIDAT";} |
Ibuprofen {medName="Ibuprofen";} |
"INSULIN ACTRAPID" {medName="INSULIN ACTRAPID";} |
KETOGAN {medName="KETOGAN";} |
"Ketogan novum" {medName="Ketogan novum";} |
METAOXEDRIN {medName="METAOXEDRIN";} |
METHYLPREDNISOLON {medName="METHYLPREDNISOLON";} |
METOCLOPRAMID {medName="METOCLOPRAMID";} |
METRONIDAZOL {medName="METRONIDAZOL";} |
MIDAZOLAM {medName="MIDAZOLAM";} |
"MIDAZOLAM DORMICUM" {medName="MIDAZOLAM DORMICUM";} |
MIVACRON {medName="MIVACRON";} |
MORFIN {medName="MORFIN";} |
NALOXON {medName="NALOXON";} |
Narcanti {medName="Narcanti";} |
NAROPIN {medName="NAROPIN";} |
ONDANSETRON {medName="ONDANSETRON";} |
Oxycontin {medName="Oxycontin";} |
OXYNORM {medName="OXYNORM";} |
"Oxynorm Kapsel" {medName="Oxynorm Kapsel";} |
Paracetamol {medName="Paracetamol";} |
Petidin {medName="Petidin";} |
Primperan {medName="Primperan";} |
PROPOFOL {medName="PROPOFOL";} |
RAPIFEN {medName="RAPIFEN";} |
ROBINOL {medName="ROBINOL";} |
Robinul {medName="Robinul";} |
"ROBINUL NEOSTIGMIN" {medName="ROBINUL NEOSTIGMIN";} |
SALBUTAMOL {medName="SALBUTAMOL";} |
Salbuvent {medName="Salbuvent";} |
Solucortef {medName="Solucortef";} |
Solu-medrol {medName="Solu-medrol";} |
STESOLID {medName="STESOLID";} |
Sufenta {medName="Sufenta";} |
SUXAMETON {medName="SUXAMETON";} |
Tavegyl {medName="Tavegyl";} |
Teofyllin {medName="Teofyllin";} |
Teofylamin {medName="Teofylamin";} |
TERBUTALIN {medName="TERBUTALIN";} |
TIOMEBUMAL {medName="TIOMEBUMAL";} |
Toradol {medName="Toradol";} |
Tradolan {medName="Tradolan";} |
TRANEXAMSYRE {medName="TRANEXAMSYRE";} |
ULTIVA {medName="ULTIVA";} |
Ventoline {medName="Ventoline";} |
VERAPAMIL {medName="VERAPAMIL";} |
Voltaren {medName="Voltaren";} |
Zinacef {medName="Zinacef";} |
Zofran {medName="Zofran";};

<mMode> =
infusion {medMode="infusion";} |
bolus {medMode="bolus";};

<mQuant0> =
et {medQuant=0.1;} |
en {medQuant=0.1;} |
to {medQuant=0.2;} |
tre {medQuant=0.3;} |
fire {medQuant=0.4;} |
fem {medQuant=0.5;} |
seks {medQuant=0.6;} |
syv {medQuant=0.7;} |
otte {medQuant=0.8;} |
ni {medQuant=0.9;};

<mQuant> =
nul komma <mQuant0> |
et {medQuant=1;} |
en {medQuant=1;} |
to {medQuant=2;} |
tre {medQuant=3;} |
fire {medQuant=4;} |
fem {medQuant=5;} |
seks {medQuant=6;} |
syv {medQuant=7;} |
otte {medQuant=8;} |
ni {medQuant=9;} |
ti {medQuant=10;} |
femten {medQuant=15;} |
tyve {medQuant=20;} |
femogtyve {medQuant=25;} |
tredive {medQuant=30;} |
femogtredive {medQuant=35;} |
fyrre {medQuant=40;} |
femogfyrre {medQuant=45;} |
halvtreds {medQuant=50;} |
femoghalvtreds {medQuant=55;} |
tres {medQuant=60;} |
femogtres {medQuant=65;} |
halvfjerds {medQuant=70;} |
femoghalvfjerds {medQuant=75;} |
firs {medQuant=80;} |
femogfirs {medQuant=85;} |
halvfems {medQuant=90;} |
femoghalvfems {medQuant=95;} |
hundrede {medQuant=100;} |
tohundrede {medQuant=200;} |
trehundrede {medQuant=300;} |
firehundrede {medQuant=400;} |
femhundrede {medQuant=500;} |
sekshundrede {medQuant=600;} |
```

```
syvhundrede {medQuant=700;} |
ottehundrede {medQuant=800;} |
nihundrede {medQuant=900;} |
tusind {medQuant=1000;};

<mAction> =
slut {medAction="slut";};
/*stop {medAction="slut";};*/

<medication> = computer <medications> <mMode>
(<mQuant> | <mAction>);


<vAction> =
start {viAction="start";} |
slut {viAction="slut";};
/*stop {viAction="slut";};*/

<vaeskerind> =
"Glucose isotonisk" {viName="Glucose isotonisk";} |
"Voluven" {viName="Voluven 60 mg/ml";};
"Natrium klorid" {viName="NaCl";};

<vaeskeind> = computer <vaeskerind> <vAction>;


<gasser> =
SEVO {gsName="SEVOFLURANE";} |
SEVOFLORAN {gsName="SEVOFLURANE";} |
SEVOFLURANE {gsName="SEVOFLURANE";} |
ISOFLURANE {gsName="ISOFLURANE";} |
ENFLURANE {gsName="ENFLURANE";} |
Ilt {gsName="O2/ATM";};
O2 {gsName="O2";};
/*{gsName="O2/N20";};*/

<gAction> =
start {gsAction="start";} |
slut {gsAction="slut";};
/*stop {gsAction="slut";};*/

<gas> = computer <gasser> <gAction>;
```

```
/* start transcription */
computer {PHONETIC="k 6 m p j u: d 6;";};
bemærk {PHONETIC="b e m a 6 g;b e m E 6 g;";};
bolus {PHONETIC="b o l u s;";};
infusion {PHONETIC="e n f u s j o: n;";};
start {PHONETIC="s d A: d;";};
slut {PHONETIC="s l u d;";};
punktum {PHONETIC="p O N t O m;";};
komma {PHONETIC="k 6 m a;";};
nul {PHONETIC="n O l;";};
et {PHONETIC="e d;";};
en {PHONETIC="e: n;";};
to {PHONETIC="t o:;";};
tre {PHONETIC="t R E:;";};
fire {PHONETIC="f i: 6;";};
fem {PHONETIC="f E m;";};
seks {PHONETIC="s E g s;";};
syv {PHONETIC="s y w;";};
otte {PHONETIC="O: d @;";};
ni {PHONETIC="n i:;";};
ti {PHONETIC="t i:;t i;";};
femten {PHONETIC="f E m d =n;";};
tyve {PHONETIC="t y w @;";};
femogtyve {PHONETIC="f E m 6 t y: w @;";};
tredive {PHONETIC="t R a D v @;t R E D v @;";};
femogtredive {PHONETIC="f E m 6 t R E D v @;f E m 6 t R a D v @;";};
fyrre {PHONETIC="f 9: 6;";};
femogfyrre {PHONETIC="f E m 6 f 2 6;";};
halvtreds {PHONETIC="h a l t R E s;";};
femoghalvtreds {PHONETIC="f E m 6 h a l t R E s;";};
tres {PHONETIC="t R E s;";};
femogtres {PHONETIC="f E m 6 t R E s;";};
halvfjerds {PHONETIC="h a l f j a 6 s;h a l f j E 6 s;";};
femoghalvfjerds {PHONETIC="f E m 6 h a l f j a 6 s;f E m 6 h a l f j E
6 s;";};
firs {PHONETIC="f i 6 s;";};
femogfirs {PHONETIC="f E m 6 f i 6 s;";};
halvfems {PHONETIC="h a l f E m s;";};
femoghalvfems {PHONETIC="f E m 6 h a l f E m s;";};
hundrede {PHONETIC="h u n R 6 D @;";};
tohundrede {PHONETIC="t o h u n R 6 D @;";};
trehundrede {PHONETIC="t R E h u n R 6 D @;t R a h u n R 6 D @;t R A h
u n R 6 D @;";};
firehundrede {PHONETIC="f i 6 h u n R 6 D @;f i: 6 h u n R 6 D @;";};
femhundrede {PHONETIC="f e m h u n R 6 D @;f E m h u n R 6 D @;f e: m
h u n R 6 D @;f a m h u n R 6 D @;";};
sekshundrede {PHONETIC="s E g s h u n R 6 D @;s e g s h u n R 6 D
@;";};
syvhundrede {PHONETIC="s y w h u n R 6 D @;s y: w h u n R 6 D @;";};
ottehundrede {PHONETIC="O: d @ h u n R 6 D @;6 d @ h u n R 6 D @;6 d e
h u n R 6 D @;O d @ h u n R 6 D @;";};
nihundrede {PHONETIC="n i h u n R 6 D @;";};
tusind {PHONETIC="t u: s =n;";};
"akut indlæggelse" {PHONETIC="a k u d si e n l E g =l s @;";};
"alle kirurgiske procedurer afsluttet" {PHONETIC="a l @ si k i R u 6 w
i s g @ si p R o s @ d y: 6 si A w s l u d @ D;";};
allergi {PHONETIC="a l E 6 g i:;a l 6 g i:;";};
"allergi og faste ok" {PHONETIC="a l 6 g i: si 6 si f a s d @ si
6 g;a l E 6 g i: si 6 si f a s d @ si 6 g;a l 6 g i: si 6 w si f
a s d @ si 6 g;a l E 6 g i: si 6 w si f a s d @ si 6 g;a l 6 g
i: si o: w si f a s d @ si 6 g;a l E 6 g i: si o: w si f a s d @
si 6 g;a l 6 g i: si O: w si f a s d @ si 6 g;a l E 6 g i: si O:
w si f a s d @ si 6 g;";};
"anlæg af dræn" {PHONETIC="a n l E: g si a si d R E: n;";};
antibiotika {PHONETIC="a n t i b i o: t i k a;";};
anti-emetika {PHONETIC="a n t i si e m a t i k a;a n t i si E m
a t i k a;a n t i: si e m a t i k a;a n t i: si E m a t i k
a;";};
"aspiration af ventrikelindhold til lunger" {PHONETIC="a s b i A
s j o: n si a si v E n t R i g =l e n h 6 l si t e si l O N 6;a
s b i R A s j o: n si a si v E n t R i g =l e n h 6 l si t e si
l O N 6;a s b i A s j o: n si a si v E n t R i g =l e n h 6 l si
t e l si l O N 6;a s b i R A s j o: n si a si v E n t R i g =l e
n h 6 l si t e l si l O N 6;";};
"assisteret ventilation" {PHONETIC="a s i s d e: 6 D si v E n t
i l a s j o: n;";};
asystoli {PHONETIC="a s y s d o: l i;";};
blodtryksfald {PHONETIC="b l o D t R 9 g s f a l;";};
cementering {PHONETIC="s e m E n t e: e N;";};
"dårlig oversigt" {PHONETIC="d Q: l i si 6 w 6 s e g d;";};
dræn {PHONETIC="d R E: n;";};
duodenalsonde {PHONETIC="d u o d e n a: l s 6 n d @;";};
"ekstrem bradykardi" {PHONETIC="E g s d R E: m si b R A d y k A
d i:;";};
"ekstrem bradykardi" {PHONETIC="E g s d R E: m si b R A d y k A
d i:;";};
ekstubation {PHONETIC="E g s t u b a s j o: n;E g s d u b a s j
o: n;E g s t u a s j o: n;E g s u a s j o: n;";};
endokrint {PHONETIC="E n d o k R i: n d;";};
engangskateter {PHONETIC="e: n g A N s k a t e: d 6;e n g A N s
k a t e: d 6;e: n g A N s k a t e: d 6;";};
forbinding {PHONETIC="f 6 b e n e N;";};
gastrointestinalt {PHONETIC="g a s d R o e n t E s d i n a: l
d;";};
gennemlysning {PHONETIC="g E n =m l y: s n e N;";};
gipsning {PHONETIC="g i b s n e N;";};
"I D ok" {PHONETIC="i si d e: si 6 g;";};
"i seng" {PHONETIC="i si s E N;i si s e N;i si s A N;";};
induktion {PHONETIC="e n d u g s j o: n;";};
intubation {PHONETIC="e n t u b a s j o: n;";};
ketalar {PHONETIC="k a t a l A:;k e t a l A:;";};
"kirurg slut" {PHONETIC="k i R u 6 w si s l u d;";};
"kirurg start" {PHONETIC="k i R u 6 w si s d A: d;";};
"klinisk hjertestop" {PHONETIC="k l i: n i s g si j E 6 d @ s d
6 b;k l i: n i s g si j a 6 d @ s d 6 b;";};
koagulation {PHONETIC="k o a g u l a s j o: n;";};
"kontrolleret ventilation" {PHONETIC="k 6 n t R o l e: 6 D si v
E n t i l a s j o: n;";};
"kortvarigt blodtryksfald" {PHONETIC="k Q d v A: i d si b l o D
t R 9 g s f a l;";};
kramper {PHONETIC="k R A m b 6;";};
kulderystelser {PHONETIC="k u l @ R 2 s d =l s 6;";};
kvalme {PHONETIC="k v a l m @;";};
lejring {PHONETIC="l A j R E N;";};
```

"må køre til stamafdeling" {PHONETIC="m O: si k 2: 6 si t e si s d a m A w d e: l e N;m O: si k 2: 6 si t e l si s d a m A w d e: l e N;";};
"malign hypertermi" {PHONETIC="m a l i: n si h y b 6 d E 6 m i:;m a: l i n si h y b 6 t E 6 m i:;m a l i: n si h y b 6 t E 6 m i:;m a: l i n si h y b 6 t E 6 m i:;m a l i: n si h y: b 6 t E 6 m i:;m a: l i n si h y: b 6 t E 6 m i:;m a l i: n si h y b 6 d E 6 m i;m a: l i n si h y b 6 d E 6 m i;";};
metadon {PHONETIC="m e t a d o: n;";};
opvågningen {PHONETIC="6 b v O: w n e N =n;6 b v O w n e N =n;";};
"på lejet" {PHONETIC="p O si l A j @ D;";};
"på stuen" {PHONETIC="p O si s d u: =n;";};
"paravenøs injektion" {PHONETIC="p A A v e n 2: s si e n j E g s j o: n;";};
"patient afleveret" {PHONETIC="p a s j E n d si A w l e v e: 6 D;";};
"patient fastende" {PHONETIC="p a s j E n d si f a: s d =n @;";};
"patient fryser" {PHONETIC="p a s j E n d si f R y: s 6;";};
"patient klar til operation" {PHONETIC="p a s j E n d si k l A: si t e si o b @ R A s j o: n;p a s j E n d si k l A: si t e l si o b @ R A s j o: n;";};
"patient modtaget på stuen" {PHONETIC="p a s j E n d si m o D t a: @ D si p O si s d u: =n;";};
renalt {PHONETIC="R E n a: l d;";};
respirationsstop {PHONETIC="R E s b i R A s j o: n s d 6 b;";};
respiratorisk {PHONETIC="R E s b i A t o: i s g;R E s b i R A t o: i s g;";};
revertering {PHONETIC="R E v E 6 t e: e N;R a v E 6 t e: e N;R E 6 t e: e N;";};
"se notat" {PHONETIC="s e: si n o t a: d;";};
"slut på anæstesi" {PHONETIC="s l u d si p O si a n E s d e s i:;";};
"spinal tilfælde" {PHONETIC="s b i n a: l si t e l f E l @;";};
"spontan respiration" {PHONETIC="s b 6 n t a: n si R E s b i R A s j o: n;";};
"start af anæstesi" {PHONETIC="s d A: d si a si a n E s d e s i:;";};
"stilet i tube" {PHONETIC="s d i: l @ D si i si t u: b @;s d e l @ D si i si t u: b @;s d i l E d si i si t u: b @;s t e l @ D si i si t u: b @;s d i: l @ D si i si t u b @;s d e l @ D si i si t u b @;s d i l E d si i si t u b @;s t e l @ D si i si t u b @;";};
"stiv nakke" {PHONETIC="s d i w si n A g @;";};
"store sekretmængder" {PHONETIC="s d o: 6 si s e k R E d m E N d 6;s d o 6 si s e k R E d m E N d 6;s d o: 6 si s e k R E: d m E N d 6;s d o 6 si s e k R E: d m E N d 6;s d o: 6 si s e k R E m E N d 6;s d o 6 si s e k R E m E N d 6;";};
"subkutan infusion" {PHONETIC="s u b k u t a: n si e n f u s j o: n;";};
"subkutan injektion" {PHONETIC="s u b k u t a: n si e n j E g s j o: n;";};
sugning {PHONETIC="s u: n e N;";};
"svær intubation" {PHONETIC="s v E 6 si e n t u b a s j o: n;";};
"svær intubation med fiberscop" {PHONETIC="s v E 6 si e n t u b a s j o: n si m E D si f i: b 6 s g o: b;s v E: 6 si e n t u b a s j o: n si m E D si f i: b 6 s g o: b;s v a 6 si e n t u b a s j o: n si m E D si f i: b 6 s g o: b;s v E 6 si e n t u b a s j o: n si m e D si f i: b 6 s g o: b;s v E: 6 si e n t u b a s j o: n si m e D si f i: b 6 s g o: b;s v a 6 si e n t u b a s j o: n si m e D si f i: b 6 s g o: b;s v E 6 si e n t u b a s j o: n si m e: D si f i: b 6 s g o: b;s v E: 6 si e n t u b a s j o: n si m e: D si f i: b 6 s g o: b;";};
"svær intubation på spontan respiration" {PHONETIC="s v E 6 si e n t u b a s j o: n si p O si s b 6 n t a: n si R E s b i R A s j o: n;";};
tandskade {PHONETIC="t a n s g a: D @;t a n s g a: D @;d a n s g a: D @;t a: n s g a: D @;";};
trendelenburg {PHONETIC="t R E n d e: l =n b u 6;";};
"udskrives til stamafdeling" {PHONETIC="u D s g R i: w @ s si t e si s d a m A w d e: l e N;u D s g R i: w @ s si t e l si s d a m A w d e: l e N;";};
"vandret leje" {PHONETIC="v A n d R 6 D si l A j @;";};
"vanskelig intravenøs-adgang" {PHONETIC="v a n s g =l i si e n t R A v e n 2: s si a D g A N;";};
"varmt tæppe er givet" {PHONETIC="v A: m d si t E b @ si 6 si g i: w @ D;v A: m d si t E b @ si E 6 si g i: w @ D;v A: m d si t E b @ si 6 si g i: v @ D;v A: m d si t E b @ si E 6 si g i: v @ D;";};
"venflon proppet" {PHONETIC="v e n f l 6 n si p R 6 b @ D;v a n f l 6 n si p R 6 b @ D;v E n f l 6 n si p R 6 b @ D;v e n f l o: n si p R 6 b @ D;";};
"venter på anæstesilæge" {PHONETIC="v E n d 6 si p O: si a n E s d e s i l E: @;v e n d 6 si p O: si a n E s d e s i l E: @;v A n d 6 si p O: si a n E s d e s i l E: @;v E n d 6 si p 6 si a n E s d e s i l E: @;v e n d 6 si p 6 si a n E s d e s i l E: @;v A n d 6 si p 6 si a n E s d e s i l E: @;v E n d 6 si p O: si a n E s d e s i l E: j @;v e n d 6 si p O: si a n E s d e s i l E: j @;";};
"venter på kirurg" {PHONETIC="v E n d 6 si p O si k i R u 6 w;";};
"venter på portør" {PHONETIC="v E n d 6 si p O si p Q t 2 6;";};
ventrikelaspiration {PHONETIC="v E n t R i g =l a s b i R A s j o: n;v E n t R i g =l a s b i A s j o: n;";};
ventrikelflimmer {PHONETIC="v E n t R i g =l f l e m 6;";};
ventrikelsonde {PHONETIC="v E n t R i g =l s 6 n d @;";};
ACTRAPID {PHONETIC="a s d R A p i D;A g t R A p i D;A g t A p i D;";};
ADRENALYN {PHONETIC="a d R E n a l y: n;a d R E n a: l 9 n;a d R E n a l 9 n;";};
AMIODARON {PHONETIC="a m i o A R o: n;a m i o A R 6 N;a m i o d a R 6 N;a m i 6 d a R 6 N;";};
AMPICILLIN {PHONETIC="A m b i s i l i: n;A m p i s i l i: n;A m b e s i l i: n;";};
Atropin {PHONETIC="a t R o p i: n;";};
"ATROPIN NEOSTIGMIN" {PHONETIC="a t R o p i: n si n 6 s d i m i n;a t R o p i: n si n 6 s d i g m i n;a t R o p i: n si n 6 s d i g m i: n;a t R o p i: n si n e o s d i m i n;";};
Atrovent {PHONETIC="a t R O v E n d;a t R o v E n d;a t R o v A N;a 6 v E n d;";};
BENZYL-PENICILLIN {PHONETIC="b E n s y: l si p e n i s i l i: n;b e n s y: l si p e n i s i l i: n;";};
Bricanyl {PHONETIC="b R i k a n y: l;";};
"BUPIVACAIN PLAIN" {PHONETIC="b u b e v a k i: n si p l A j n;b u b e v a k i n si p l A j n;b u b e v a k a i: n si p l A j n;b u b e v A: s A j n si p l A j n;b u b e v a k i: n si p l a n;b u b e v a k i n si p l a n;b u b e v a k a i: n si p l a n;b u b e v A: s A j n si p l a n;";};
"BUPIVACAIN TUNG" {PHONETIC="b u b e v a k i: n si t O N;b u b e v a k i n si t O N;b u b e v a k a i: n si t O N;b u b e v A: s A j n si t O N;";};
"CALCIUM CHLORID" {PHONETIC="k a l s i O m si k l o R i D;k a l s j O m si k l o R i D;k a l s i O m si g l o R i D;k a l s j O m si g l o R i D;";};
CEFUROXIM {PHONETIC="s e f u 6 6 g s i m;s e f u 6 O g s i m;s e f u R 6 g s i m;s A f u: 6 O g s i m;";};
CLEMASTIN {PHONETIC="k l a s d i: n;k l A m a s d i: n;k l e: m a s d i: n;k l a s d i n;";};
Combivent {PHONETIC="k 6 m b i n d;k 6 m b i v E n d;k 6 m b i v a n d;k 6 m b i v A N;";};
Cordarone {PHONETIC="k Q: d a 6 n @;k Q: d A 6 n @;k Q: d a R 6 n @;s Q: d a 6 n @;";};
Cyklokapron {PHONETIC="s i g l o k a p R o: n;s y k l o k a p R o: n;s i g l o k a: b R 6 N;s y g l o k a p R o: n;";};
DEXAMETHASON {PHONETIC="d E g s a m e: d h a: s 6 n;d E g s a m e: d h a: s =n;d E g s a m e: d h a: s o: n;d A g s a m e: d h a: s 6 n;";};
DIAZEPAM {PHONETIC="d i a s E p A m;d i a s A p A m;d i a s @ p A m;d i a s e p A m;";};
Diclocil {PHONETIC="d i s l o s i l;d i s l o k i: l;d i s l o s i: l;d i s l o g i: l;";};
DICLOXACILLIN {PHONETIC="d i g l 6 g s a s i l i: n;d i s l o g a s i l i: n;d i s l o g s a s i l i: n;d i s l O w g a s i l i: n;";};
DIGOXIN {PHONETIC="d i 6 g s i: n;";};
DOBUTAMIN {PHONETIC="d 6 b t a m i: n;d 6 b u t a m i: n;";};
Dobutrex {PHONETIC="d 6 b t R E g;d 6 b t R a g;d 6 b t R 9 g;d 6 b t R e g s;";};
DOPAMIN {PHONETIC="d o: p a m i: n;";};
Dopram {PHONETIC="d 6 b R A m;d o p R A m;d O w p R A m;d o: p R A m;";};
Dormicum {PHONETIC="d Q: m i k O m;d 6 m i k O m;d o 6 m i k O m;";};
DOXAPRAM {PHONETIC="d 6 g s A b R A m;d 6 g s a p R A m;d 6 g s A b R A: m;";};
EFEDRIN {PHONETIC="f e R i: n;E f E R i: n;f e d R i: n;f E R i: n;";};
ESMERON {PHONETIC="E s m 6 6 n;E s m @ R o: n;e s m 6 6 n;a s m 6 6 n;";};
Fenemal {PHONETIC="f e n @ m a: l;f e n @ m a l;f e n m a: l;";};
FENTANYL {PHONETIC="f a n t a n y: l;f e n t a n y: l;f E n d a n y: l;f E n t a n y: l;";};
Fortecortin {PHONETIC="f 6 t A s o t i: n;f 6 t A s o 6 d i: n;f Q: t A s o t i: n;f 6 t A s o 6 d i n;";};
Furix {PHONETIC="f u: R i g s;f u: R i g;f u: i g;f u: i g s;";};
FUROSEMID {PHONETIC="f u R o: s @ m i D;";};
Garamycin {PHONETIC="g A R a m y s i: n;g A A m y s i: n;g A R A m y s i: n;g A R A: m y s i: n;";};
GENTAMICIN {PHONETIC="g E n t a m i s i: n;g E n t a m i k i: n;g E n t a m i: s i n;g E n t a m i: s i: n;";};
GLYCOPYRRON {PHONETIC="g l y k o p y R o: n;g l y k o p y R O n;g l y k o p y R R o: n;g l y k o p y: R o: n;";};
HYDROCORTISON {PHONETIC="h y d R o k Q t i s o: n;h y d R o s Q t i s o: n;";};
HYPNOMIDAT {PHONETIC="h y b n o m i: D 6;h y b n o: m i d a: d;h y b n o: m i: D 6;h y b n o m i d a: d;";};
Ibuprofen {PHONETIC="i b u p R o f =n;i b u p R o f E n;i b u p R o f e n;i: b u p R o f =n;";};
"INSULIN ACTRAPID" {PHONETIC="e n s u l i: n si a s d R A p i D;e n s u l i: n si A g t R A p i D;e n s u l i: n si A g t A p i D;";};
KETOGAN {PHONETIC="k E d 6 g A N;k E d 6 w =n;k e t 6 N =n;k a t o g a n;";};
"Ketogan novum" {PHONETIC="k E d 6 g A N si n o v O m;k E d 6 w =n si n o v O m;k e t 6 N =n si n o v O m;k a t o g a n si n o v O m;k E d 6 g A N si n o: v O m;k E d 6 w =n si n o: v O m;k e t 6 N =n si n o: v O m;k a t o g a n si n o: v O m;";};

METAOXEDRIN {PHONETIC="m e t a 6 g @ d R i: n;m e d a 6 g @ d R i: n;m e t a 6 g @ d R E N;m e d a 6 g @ d R E N;";};
METHYLPREDNISOLON {PHONETIC="m e t y l b R E: D n e s o: l 6 n;m e t y l b R E: D n i s o: l 6 n;m e t y l b R E: D n i s o l 6 N;m e t y l b R E: D n i s o l O: n;";};
METOCLOPRAMID {PHONETIC="m e t o s l o p R A m i: D;m e d o s l o p R A m i: D;m e t o s l o p R A m i D;m e t o k l o p R A m i: D;";};
METRONIDAZOL {PHONETIC="m e t R o n i d a: s 6 l;m e t R o n i d a: s o: l;";};
MIDAZOLAM {PHONETIC="m e d a s o l a m;m e d a s o l A m;m e d a: s 6 l A m;m e d a s o l A: m;";};
"MIDAZOLAM DORMICUM" {PHONETIC="m e d a s o l a m si d Q: m i k O m;m e d a s o l A m si d Q: m i k O m;m e d a: s 6 l A m si d Q: m i k O m;m e d a s o l A: m si d Q: m i k O m;m e d a s o l a m si d 6 m i k O m;m e d a s o l A m si d 6 m i k O m;m e d a: s 6 l A m si d 6 m i k O m;m e d a s o l A: m si d 6 m i k O m;";};
MIVACRON {PHONETIC="m i a k R o: n;m i A k R o: n;m v a k R o: n;";};
MORFIN {PHONETIC="m Q f i: n;";};
NALOXON {PHONETIC="n a: l 6 g =n;n a l 6 g =n;n a l o: g o: n;n a l O g s =n;";};
Narcanti {PHONETIC="n a 6 s A n t i:;n a 6 k a n t i:;n A k a n t i:;n a 6 s A n t i;";};
NAROPIN {PHONETIC="n a 6 o b i: n;n a 6 b e n;n A R o p i: n;n a 6 p i: n;";};
ONDANSETRON {PHONETIC="o n a n s @ t R o: n;o n a n s A t R o: n;o n a n s e R o: n;o n a n s @ t R O n;";};
Oxycontin {PHONETIC="6 g s y k 6 n t i: n;6 g s y k 6 n t i n;6 g s y g 6 n t i: n;";};
OXYNORM {PHONETIC="6 g s y n o 6 m;6 g s y n Q: m;";};
"Oxynorm Kapsel" {PHONETIC="6 g s y n o 6 m si k A p s =l;6 g s y n Q: m si k A p s =l;6 g s y n o 6 m si k A b s =l;6 g s y n Q: m si k A b s =l;";};
Paracetamol {PHONETIC="p A A s E t a m o: l;";};
Petidin {PHONETIC="p a t i: D i: n;p E d i d e;p E d i d i: n;p E d i d i n;";};
Primperan {PHONETIC="p R i m b @ R A n;p R i m b @ R A: n;p R i: m b @ R A n;p R i: m @ R A: n;";};
PROPOFOL {PHONETIC="p R 6 b 6 f 6 l;";};
RAPIFEN {PHONETIC="R A b i f =n;R A: b i f =n;R A p i: w =n;R A p i f =n;";};
ROBINOL {PHONETIC="R 6 b i n o: l;R 6 b e n o: l;R o b i n o: l;";};
Robinul {PHONETIC="R 6 b i n u l;R 6 b i n O l;R 6 b i n u: l;R 6 b e n O l;";};
"ROBINUL NEOSTIGMIN" {PHONETIC="R 6 b i n u l si n 6 s d i m i n;R 6 b i n O l si n 6 s d i m i n;R 6 b i n u: l si n 6 s d i m i n;R 6 b e n O l si n 6 s d i m i n;R 6 b i n u l si n 6 s d i g m i n;R 6 b i n O l si n 6 s d i g m i n;R 6 b i n u: l si n 6 s d i g m i n;R 6 b e n O l si n 6 s d i g m i n;";};
Salbuvent {PHONETIC="s a l b u: v E n d;s a l b u: v A N;";};
SALBUTAMOL {PHONETIC="s a l b u t a m o: l;s a l b u: t a m o: l;s a: l b u t a m o: l;s a: b u t a m o: l;";};
Solucortef {PHONETIC="s o l u s Q: d E: w;s o l u k o 6 d A f;s o l u k Q: t E f;s o l u k Q t E f;";};
Solu-medrol {PHONETIC="s o l u si m E D R 6 l;s o l u si m e: R 6 l;s o: l u si m E D R 6 l;s o: l u si m e: R 6 l;";};
STESOLID {PHONETIC="s E 6 l e D;s d E 6 l e D;s E 6 l i d;s d E s o l i D;";};
Sufenta {PHONETIC="s u f E n t a;s u f E n d a;s u f A n d a;s u f a n t a;";};
SUXAMETON {PHONETIC="s u g s a m e: d =n;s O g a m e: d =n;s O g A m e t o: n;s O g A m e t 6 N;";};
Tavegyl {PHONETIC="t a: v @ g y l;t a: w @ g y l;t A w e g 2 l;t A: w @ g y l;";};
Teofyllin {PHONETIC="d e o f y l e n;t e o f y l e n;d e o f y l i: n;t e o f y l i: n;";};
Teofylamin {PHONETIC="d e o f y l a m i: n;t e o f y l a m i: n;d e o f y l A m i: n;t e o f y l A m i: n;";};
TERBUTALIN {PHONETIC="t 6 b t a: l i n;t E 6 b u t a l i: n;d 6 b t a: l i n;t 6 b t a: l i: n;";};
TIOMEBUMAL {PHONETIC="t e o m @ b O m a: l;t e o m @ b m a: l;t e o m @ b u: m a: l;t e o m e b O m a: l;";};
Toradol {PHONETIC="t o R A d o: l;t o: R A d o: l;t o R A: d o: l;t o: R A: d o: l;";};
Tradolan {PHONETIC="t R A d o: l =n;t R A: d 6 l a: n;t R A: d 6 l a n;t R A d 6 l a: n;";};
TRANEXAMSYRE {PHONETIC="t R A n E g s a m s y: 6;t R A: n E g s a m s y: 6;";};
ULTIVA {PHONETIC="u l t i v a;u l t i a;u l t i: v a;u l t i: a;";};
Ventoline {PHONETIC="v E n d o l i: n @;v E n d o l i: n;v a n d o l i: n @;v a n d o l i: n;";};
VERAPAMIL {PHONETIC="v E R A p a m i: l;v e: R A p a m i: l;v E R A p A m i l;v E R A p a m i l;";};
Voltaren {PHONETIC="v 6 l t A: A n;v 6 l t a: 6 n;v 6 l t a: 6 6 n;v 6 l t a 6 n;";};
Zinacef {PHONETIC="s i n a s E f;s i: n a s E f;s i: n A s E f;s i n a: s a f;";};
Zofran {PHONETIC="s 6 f R A n;s 6 f R A: n;s 6 f R 6 n;s 6 f R a;";};
SEVO {PHONETIC="s e: 6;s e 6;s e: o;s e: v o;";};
SEVOFLORAN {PHONETIC="s v 6 w l 6 A n;s v 6 w l 6 A n;s f 6 f l o R a: n;s f 6 f l o R A: n;";};
SEVOFLURANE {PHONETIC="s v 6 w l u R A n @;s v 6 w l u R A n @;s v 6 w l u R A: n @;s v 6 w l u R A: n @;";};
ISOFLURANE {PHONETIC="e s o: f l u R A n @;i s o f l u R A n @;e s o: f l u R A: n @;i s o f l u R A: n @;";};
ENFLURANE {PHONETIC="e n f l u R A n @;e n f l u R A: n @;e n f l u 6 A n @;e n f l u R A n;";};
Ilt {PHONETIC="i: l d;i l d;e l d;";};
"Glucose isotonisk" {PHONETIC="g l u k o: s @ si i s o t o: n i s g;";};
Voluven {PHONETIC="v o l u: =n;v o l u: v =n;v o l u v E n;v o l u: @ n;";};
"Natrium klorid" {PHONETIC="n a t R i O m si k l o R i D;n a: t R i O m si k l o R i D;";};

Appendix 2: Detailed responses of the participants (in Danish)

Anaesthesia nurses

Generelt om elektronisk anæstesijournal

	q1. Hvor meget ligner Dräger-systemet, som I bruger på Køge Sygehus, og prototypen i Herlev hinanden mht. til de funktioner, der var nødvendige for registrering af medicin og bemærkninger?	q2. Hvor nyttigt mener du – ud fra din egen erfaring – det vil være at have en anæstesijournal, som er ajourført hele tiden under operationen?	q3. Hvor ofte bruger du anæstesijournalen som hjælp til at huske, hvad der er sket, eller som støtte til at tage nye beslutninger?
	1: Ingen lighed 5: Total lighed	1: Ikke nyttigt 5: Uundværligt	1: Aldrig 5: Hele tiden
Nurse 1	2	5	5
Nurse 2	4	4	3
Nurse 3	4	5	5
Nurse 4	3	5	5
Nurse 5	4	4	4
Nurse 6	3	4	3
Average	3.3	4.5	4.2

Om den traditionelle måde i den første session

	q4. I den første session uden talegenkendelse, hvor besværligt var det at udfylde anæstesijournalen under operationen?	q5. Har udfyldelse af anæstesijournalen på traditionel måde taget tid, som kunne have været brugt på patienten?	q6. Har udfyldelse af anæstesijournalen på traditionel måde forstyrret din primære opgave med patienten eller formindsket din koncentration?
	1: Umuligt 5: Let	1: Nej, ingen tid 5: Ja, for meget tid	1: Nej, ingen forstyrrelse 5: Ja, for meget forstyrrelse
Nurse 1	1	4	4
Nurse 2	3	1	1
Nurse 3	5	4	1
Nurse 4	3	3	3
Nurse 5	2	4	3
Nurse 6	2	4	3
Average	2.7	3.3	2.5

Om talegenkendelse som input i anden session

	q4b: I den anden session med talegenkendelse, hvor besværligt var det at udfylde anæstesijournalen under operationen?	q5b: Har udfyldelse af anæstesijournalen ved brug af talegenkendelse taget tid, som kunne have været brugt på patienten?	q6b: Har udfyldelse af anæstesijournalen med talegenkendelse forstyrret din primære opgave med patienten eller formindsket din koncentration?
	1: Umuligt 5: Let	1: Nej, ingen tid 5: Ja, for meget tid	1: Nej, ingen forstyrrelse 5: Ja, for meget forstyrrelse
Nurse 1	3	3	3
Nurse 2	2	5	5
Nurse 3	4	2	2
Nurse 4	4	2	2
Nurse 5	4	2	2
Nurse 6	3	3	3
Average	3.3	2.8	2.8

Om talegenkendelse som input i fremtiden

	Antag at den traditionelle måde at registrere på blev suppleret med et talegenkendelsessystem, som i det store og hele genkendte perfekt. Hvordan tror du den mest sandsynlige ændring vil være i de følgende krav til journalens kvalitet?			
	q7: At journalen er opdateret på et hvilket som helst tidspunkt under operationen	q8: At journalen er fuldstændig på et hvilket som helst tidspunkt under operationen	q9: At journalen er fuldstændig efter operationen	q10: Hvor nyttigt mener du det ville være at kunne bruge talegenkendelse som supplement til den nuværende måde med touchskærm og tastatur?
	1: Der vil ske en ændring i klart negativ retning 3: Der vil ikke ske nogen ændring 5: Der vil ske en ændring i klart positiv retning	1: Der vil ske en ændring i klart negativ retning 3: Der vil ikke ske nogen ændring 5: Der vil ske en ændring i klart positiv retning	1: Der vil ske en ændring i klart negativ retning 3: Der vil ikke ske nogen ændring 5: Der vil ske en ændring i klart positiv retning	1: Ikke nyttigt 5: Uundværligt
Nurse 1	3	4	3	3
Nurse 2	3	3	4	4
Nurse 3	5	5	5	5
Nurse 4	5	5	5	5
Nurse 5	4	4	4	4
Nurse 6	4	4	4	4
Average	4.0	4.2	4.2	4.2

Anaesthesia doctors

Generelt om elektronisk anæstesijournal

	q1. Hvor meget ligner Dräger-systemet, som I bruger på Køge Sygehus, og prototypen i Herlev hinanden mht. til de funktioner, der var nødvendige for registrering af medicin og bemærkninger?	q2. Hvor nyttigt mener du – ud fra din egen erfaring – det vil være at have en anæstesijournal, som er ajourført hele tiden under operationen?	q3. Hvor ofte bruger du anæstesijournalen som hjælp til at huske, hvad der er sket, eller som støtte til at tage nye beslutninger?
	1: Ingen lighed 5: Total lighed	1: Ikke nyttigt 5: Uundværligt	1: Aldrig 5: Hele tiden
Doctor 1	4	4	5
Doctor 2	3	4	5
Doctor 3	4	4	4
Doctor 4	4	4	3
Average	3.8	4.0	4.3

Om den traditionelle måde i den første session

	q4. I den første session uden talegenkendelse, hvor besværligt tror du det var for sygeplejersken at udfylde anæstesijournalen under operationen?	q5. Har udfyldelse af anæstesijournalen på traditionel måde for sygeplejersken taget tid, som kunne have været brugt på patienten?	q6. Har udfyldelse af anæstesijournalen på traditionel måde forstyrret sygeplejerskens primære opgave med patienten eller formindsket hendes koncentration?
	1: Umuligt 5: Let	1: Nej, ingen tid 5: Ja, for meget tid	1: Nej, ingen forstyrrelse 5: Ja, for meget forstyrrelse
Doctor 1	5	3	2
Doctor 2	3	4	4
Doctor 3	3	4	2
Doctor 4	1	1	2
Average	3.0	3.0	2.5

Om talegenkendelse som input i anden session

	q4b. I den anden session med talegenkendelse, hvor besværligt tror du det var for sygeplejersken at udfylde anæstesijournalen?	q5b. Har sygeplejerskens udfyldelse af anæstesijournalen ved brug af talegenkendelse taget tid, som kunne have været brugt på patienten?	q6b. Har sygeplejerskens udfyldelse af anæstesijournalen med talegenkendelse forstyrret hendes primære opgave med patienten eller formindsket hendes koncentration?
	1: Umuligt 5: Let	1: Nej, ingen tid 5: Ja, for meget tid	1: Nej, ingen forstyrrelse 5: Ja, for meget forstyrrelse
Doctor 1	4	2	2
Doctor 2	2	3	3
Doctor 3	3	2	2
Doctor 4	2	2	4
Average	2.8	2.3	2.8

Om talegenkendelse som input i fremtiden

	Antag at den traditionelle måde at registrere på blev suppleret med et talegenkendelsessystem, som i det store og hele genkendte perfekt. Hvordan tror du den mest sandsynlige ændring vil være i de følgende krav til journalens kvalitet?			
	q7: At journalen er opdateret på et hvilket som helst tidspunkt under operationen	q8: At journalen er fuldstændig på et hvilket som helst tidspunkt under operationen	q9: At journalen er fuldstændig efter operationen	q10: Hvor nyttigt mener du det ville være at kunne bruge talegenkendelse som supplement til den nuværende måde med touchskærm og tastatur?
	1: Der vil ske en ændring i klart negativ retning 3: Der vil ikke ske nogen ændring 5: Der vil ske en ændring i klart positiv retning	1: Der vil ske en ændring i klart negativ retning 3: Der vil ikke ske nogen ændring 5: Der vil ske en ændring i klart positiv retning	1: Der vil ske en ændring i klart negativ retning 3: Der vil ikke ske nogen ændring 5: Der vil ske en ændring i klart positiv retning	1: Ikke nyttigt 5: Uundværligt
Doctor 1	5	5	5	3
Doctor 2	4	4	4	4
Doctor 3	4	4	4	4
Doctor 4	5	5	4	4
Average	4.5	4.5	4.3	3.8

Transition 4

The methods involved and the analysis of the previous experiments [Alapetite 2007] form the main result of the thesis, and provide a last set of answers to the questions raised at the EACE'2005 conference (*cf.* Transition 2).

Another set of questions naturally emerged from experimenting with this prototype, in particular regarding the possible deployment of such a system.

Although it was not feasible to envisage larger scale experiments given the time and budget available for this project, it was possible to make a related survey with Vejle and Give hospital[1], to study some human factors issues influencing the acceptance and the success of the speech recognition system being introduced. This hospital was in the process of introducing the Max Manus speech recognition system into all its clinical departments in 2005 – 2006 to produce patient records. Previously, the work procedure consisted of physicians dictating record notes on tape or to an audio file, which was subsequently transcribed by medical secretaries. The new procedure involves physicians writing the medical records themselves directly on a computer (in Microsoft Word), normally with the help of an automatic speech recognition system (ASR). Using the ASR is not mandatory but strongly encouraged by the hospital and departments, but a few physicians apparently continue to use only a keyboard.

This study (January – December 2006) [Alapetite, Andersen, Hertzum 2007] was an opportunity to evaluate *via* electronic questionnaires and other indicators, the deployment, acceptance and success of a speech recognition system sharing technological similarities with the above mentioned prototype.

However, as often the case when deploying new technologies, the introduction of the new system coincides with deeper modifications of the work practices that are only partially due to this new system.

This fact was the occasion for discussing in the second paper [Alapetite & Gauthereau 2005] the possible consequences in the short and long term of a possible introduction of the voice enabled anaesthesia record prototype.

1 "Vejle and Give" is the name of one single hospital.

In the case of Vejle and Give Hospital, the major difference between the previous and the new work practice is not exclusively a matter of technology. Rather, it is the change to a new work system, involving speech technology as a strongly recommended option, which takes away from physicians the support from secretaries and makes physicians responsible for all efforts involved in producing and finalising documents for the patient records on the fly. Previously, the workload was shared between physicians and secretaries, who completed the transcription, sometimes caught lacunas or inconsistencies, and reminded physicians to check and approve the documents before they would be transferred to the patient medical record.

The fact that the introduction of speech recognition as a front end to the electronic medical record is accompanied, necessarily, with a change in work practices tends blur any measure of the success of the new technology alone. The ideal setting to study the success of the new speech recognition system would have been a comparison between a period where physicians would have used the new work procedures with only the keyboard to type the documents, and a period during which they could additionally have used the vocal modality.

Acceptance of Speech Recognition by Physicians: A Survey of Expectations, Experiences, and Social Influence

Alexandre Alapetite[1,2], Henning Boje Andersen[1], Morten Hertzum[2]

1. Systems Analysis Department; Risø National Laboratory; Technical University of Denmark; DK-4000 Roskilde; Denmark
2. Computer Science, Roskilde University, Universitetsvej 1; P.O. Box 260; DK-4000 Roskilde, Denmark

This section is the long version of the following article:

Alapetite, Andersen, Hertzum: Acceptance of Speech Recognition by Physicians: A Survey of Expectations, Experiences, and Social Influence. *Submitted to the International Journal of Human-Computer Studies, 2007.*

Abstract

Introduction: Speech recognition is being used more and more for medical applications, ranging from the production of pathology or X-ray reports to entries to the electronic medical record (EMR). The present study has surveyed physician views and attitudes before and after the introduction of speech technology as a front end to an electronic medical record. The survey was made in a hospital that recently (2006-2007) replaced traditional dictation-and-secretary transcription by speech technology as the preferred input mode to the electronic medical record for all physicians in clinical departments.

Objective: The aim of the survey was (i) to identify how attitudes and perceptions among physicians affected the acceptance and success of the speech recognition system and the new work procedures associated with it; and (ii) to assess the degree to which physicians' attitudes and expectations to the use of speech technology changed after actually using it.

Methods: The survey was based on two questionnaires. When they were about to begin training with the speech recognition system, physicians in three departments received an "expectations questionnaire" asking the physicians about their opinions and views about the use of the system. Subsequently, when they had had some experience with the system, physicians in six departments received an "experiences questionnaire" asking similar questions, eliciting respondents' retrospective perceptions of using the

speech recognition system and new work procedures. The survey data were supplemented with performance data from the speech recognition system.

Results: The surveyed physicians tended to report a more negative view of the system after having used it for some months than before. Retrospectively, physicians are approximately evenly divided between those who think it was a good idea to introduce speech recognition (33%), those who think it was not (31%) and those who are neutral (35%). Physicians who rated the traditional secretary-assisted system highly tended to be less in favour of introducing speech recognition. In particular, the physicians felt that they spent much more time producing medical records than before, including time correcting the speech recognition, and that the overall quality of records had declined.

Conclusion: Physicians tended to become somewhat more negative toward the use of the speech recognition system after having used it for some time. Nevertheless, workflow improvements and the possibility to access the records immediately after dictation were almost unanimously appreciated. Physicians' affinity with the system seems to be quite dependent on their perception of the associated new work procedures.

Keywords

Electronic medical record; Electronic patient record; speech recognition; transcription; technology acceptance

1 Introduction

Speech recognition has been refined and become more robust in recent years. The gradual maturation of the technology has been accompanied by adoptions of the technology in the medical domain, where it is used to enter comments into the electronic medical record (EMR), thus replacing the standard way of entering notes by physician dictation and subsequent transcription by medical secretaries or a dedicated service [Zafar *et al.* 1999]. At the same time as the technology has matured, speech recognition has been developed and implemented for languages spoken by much "smaller" populations, such as Danish (5.4 million speakers).

Vejle and Give Hospital, Denmark, has been one of the first hospitals to introduce speech recognition for all major specialties and departments. Having run a successful project on speech recognition in its radiology department since 2000, this regional hospital (349 beds, and 217 000 outpatients in 2006) began to implement plans for having all physicians in clinical departments use speech recognition to input physician notes and instructions into the EMR. The speech recognition system – software based on Philips Speech Magic, adapted to Danish and deployed by Max Manus A/S – was

rolled out in all clinical departments in 2005-2006, and has about 240 physician users as of 2007.

The main purpose of introducing speech recognition across all departments was to ensure a quicker completion of medical record entry and to achieve a higher quality of patient records. The old transcription system was known to sometimes produce backlogs of dictation tapes waiting to be transcribed, or transcriptions waiting to be checked and approved by physicians. Additionally, an expected consequence was to allow secretaries, who would no longer need to spend time on transcriptions, to take over other duties. It was hoped that the quality of medical records would be enhanced, since physicians would now be going to check and revise their written (speech recognised) record immediately while their intentions were still fresh in memory. While little is known so far about the impacts of speech recognition on the various stages of the writing process and on the quality of outcome [Honeycutt 2003], the above-mentioned goals fully match criteria such as those reported by [Mönnich & Wetter 2000].

The present study had two related objectives: First, to identify physicians' attitudes and expectations about speech recognition that might predict their subsequent level of satisfaction with actual use of the technology. Second, to assess possible changes between prior expectations to and subsequent experience with the technology as a replacement for the traditional mode of producing medical records.

2 Related work

Work about the acceptance of speech recognition falls into two main areas: speech recognition and technology acceptance. Studies of speech recognition have predominantly been devoted to recognition of spoken English. However, recognition rates of systems that recognise English are not necessarily transferable to a speech recognition system for Danish.

2.1 Speech recognition

For free-text dictation, speech recognition combines some characteristics of traditional dictation and of word processing [Leijten & Van Waes 2005]: on the one hand, quick and easy use of speech, and on the other, instantaneous graphical feedback and the possibility of jumping back and forth in the text. At the same time, speech recognition has its own advantages and drawbacks.

For transcription of text, state-of-the-art systems correctly recognize 72%-98% of the spoken words according to recent research [Alapetite 2006; Zafar *et al.* 2004; Al-Aynati & Chorneyko 2003; Kanal *et al.* 2001; Sears *et al.* 2001; Devine *et al.* 2000; Jungk *et al.* 2000; Ramaswamy *et al.* 2000; Zafar *et al.* 1999], while commercially reported recognition rates are generally above 95%. Several factors contribute to the differences in recognition rates across studies:

- Vocabulary affects speech recognition through its size and domain coverage. Large vocabularies with good domain coverage are attractive, simply because they enable recognition of more words. Conversely, the acoustic distinctiveness of words is larger in small vocabularies, increasing the likelihood of correct recognition. Small vocabularies are, however, mostly relevant for voice navigation. State-of-the-art systems for text transcription have vocabularies comprising tens of thousands of words and optional, add-on vocabularies for specific domains such as the medical domain.
- Speakers influence speech recognition by the clarity and consistency of pronunciation and the degree of fit between their pronunciation and the acoustic model of the system. Speaker-dependent systems achieve higher recognition rates than speaker-independent systems but require one or more training sessions – based on which the system adapts its acoustic model to the speaker – and may be more sensitive to variations of the background noise, microphone, and voice (*e.g.* due to a cold). Even after training, atypical speakers, including non-natives [Coniam 1999] as well as children and elderly [Wilpon & Jacobsen 1996], experience lower recognition rates than typical speakers.
- Noise affects speech recognition in two ways: (a) It distorts the speech signal, making it more difficult to discern the spoken words. (b) In the presence of noise, people alter their voice in an attempt to counter the distortion of the speech signal (the Lombard effect) [Lombard 1911]. Ambient noises, such as those heard in hospital wards or emergency rooms, are reported not to significantly affect speech recognition rates on average, especially when a suitable microphone is used [Alapetite 2006; Zafar *et al.* 1999]. However, in spite of numerous noise-cancellation techniques, loud noise and even moderate levels of noise may considerably degrade the performance of speech recognition systems [Gong 1995; Barker *et al.* 2005].
- Speech recognition systems are based on principles of statistical pattern matching [Young 1996]. However, in spite of this commonality, individual systems differ in their parameterization of the speech signal, the acoustic model of each phoneme, and the language model used in predicting the words most likely to follow the preceding words. Thus, different systems make different recognition errors, even when they achieve similar recognition rates. This difference can be used to improve recognition rates by fusing the outputs of multiple systems [Alapetite 2006; Fiscus 1997].

Studies of text transcription show that it takes more time for a person to produce a text by voice input followed by correction of the recognition errors than by dictation followed by proofreading after the text has been typed by a human typist whose time is not included in the comparison [Borowitz 2001; Al-Aynati & Chorneyko 2003]. Thus, the freeing of typist time for other tasks is achieved at the expense of spending more of the speaker's time. [Mohr *et al.* 2003] studied speech recognition as an aid for typists and found that editing a draft produced by speech recognition took longer than typing the audio-recorded text from scratch. The main time-related advantage of using speech recognition, as opposed to human typists, for text transcription appears to be a considerable reduction of the time from the production of the original dictation until the text is completed [Lai & Vergo 1997; Ramaswamy *et al.* 2000; Borowitz 2001]. It should be noted that previous experience with traditional dictation systems or word processing influences the use of speech recognition, in the sense that users tend to stick to their previous writing habits when they start using ASR [Leijten & Van Waes 2005].

[Zafar *et al.* 2004], who reviewed recognition errors made by speech recognition systems during text transcription, found that 9.4% of errors were nonsense errors and 1.6% critical errors. The presence of nonsense and critical errors complicates error correction. Attempts at easing error correction by utilising the confidence scores generated by speech recognition systems have yielded mixed results [Suhm *et al.* 2001; Feng & Sears 2004]. Error correction can be made by voice commands, making text production entirely hands-free, but this is inefficient compared to making the corrections by keyboard and mouse [Suhm *et al.* 2001]. Multimodal methods of text production are also recommended for ergonomic reasons [Juul-Kristensen *et al.* 2004].

2.2 Technology acceptance

Technology acceptance has been studied from many perspectives, including the theory of reasoned action (TRA) [Fishbein & Ajzen 1975], the theory of planned behaviour (TPB) [Ajzen 1985; Ajzen 1991], diffusion of innovations (DOI) [Rogers 2003], and the technology acceptance model (TAM) [Davis 1989; Davis 1993]. These perspectives generally agree that technology acceptance concerns the adoption processes through which individuals decide to acquire and deploy a technology for a specified purpose. They differ, however, in the factors considered to influence the adoption process. Recently, [Venkatesh *et al.* 2003] proposed a technology acceptance model that unified much of the previous work by encompassing an inclusive set of factors:

- *Performance expectancy:* "the degree to which an individual believes that using the system will help him or her attain gains in job performance" [Venkatesh *et al.* 2003:447]. Performance expectancy includes factors such as perceived usefulness (from TAM) and relative advantage (from DOI), which have been the strongest predictors of acceptance in previous studies. In the unified model performance expectancy was, likewise, a determinant of intention to use systems, and more so for men and younger employees.
- *Effort expectancy:* "the degree of ease associated with the use of the system" [Venkatesh *et al.* 2003:450]. Effort expectancy includes ease-of-use factors (from TAM and DOI), which have particularly been found to influence usage behaviour during early use of a system. In the unified model effort expectancy was, likewise, a determinant of intention to use, and more so for women, older employees, and with less experience using the system.
- *Social influence:* "the degree to which an individual perceives that important others believe he or she should use the new system" [Venkatesh *et al.* 2003:451]. Social influence includes subjective norm (from TRA and TPB) and image (from DOI). In the unified model, social influence was a determinant of intention to use, and more so for women, older employees, with less experience using the system, and when use was mandated.
- *Facilitating conditions:* "the degree to which an individual believes that an organizational and technical infrastructure exists to support use of the system" [Venkatesh *et al.* 2003:453]. Facilitating conditions include perceived behavioural control (from TPB) and compatibility (from DOI). In the unified model, the effect of facilitating conditions was subsumed by effort expectancy, except for an effect on usage for older employees with experience using the system.

In the study by [Venkatesh *et al.* 2003], the unified technology acceptance model explained 70% of the variance in individuals' intention to use systems. Many systems are however adopted in organizational contexts, which appear to be somewhat under-recognized in the unified model. Organization-level factors that affect the adoption of technologies include administrative intensity, centralization, external communication, functional differentiation, internal communication, managerial attitude toward change, professionalism, slack resources, specialization, and technical knowledge resources [Damanpour 1991].

Studies of adoption in organizational contexts often find that it is a two-stage process involving a formal decision to adopt a technology followed by actual deployment of the technology by users [Fichman 2000; Gallivan 2001]. This creates opportunities for lags between the formal, often organization-level decision and subsequent local deployment by individuals. One reason for these lags is that the formal decision to adopt a technology and the decisions about actual deployment are typically made by different people, who may disagree. Another reason may be that different considerations are salient to the formal decision and to actual deployment. Specifically, unrealistic expectations during the formal decision to adopt may lead to disappointment among the first employees that actually deploy a technology and these

disappointed expectations may, in turn, discourage and delay further deployment [Fichman & Kemerer 1999]. This way, unrealistic expectations produce a subtle combination of performance expectancy and social influence.

In the case of speech-input interfaces in the medical domain, some studies on technology acceptance have been reported, such as [Dillon & Norcio 1997] showing an effect of expertise and experience on the performance or the acceptance.

3 Survey method

A questionnaire was developed and deployed as a survey at Vejle and Give hospital (Denmark). The survey was divided into two phases, a prospective phase in which we surveyed physicians' expectations toward speech recognition and a subsequent retrospective phase where physicians' experiences with the technology were surveyed.

3.1 Participants

The survey participants were 186 anonymous physicians at Vejle and Give Hospital, about half of whom were introduced in 2005 to speech recognition to replace dictation-and-transcription, and the other half to be introduced as the study progressed during 2006. The departments involved were medicine, neurology, oncology, organ surgery, orthopaedic surgery, and otology.

3.2 Survey instrument

The survey instrument was a pair of related and overlapping questionnaires developed by the project group[1]: A prospective one asking respondents about their expectations and attitudes to the use of speech recognition technology as a front end to the EMR and a retrospective one asking them about their experiences with the technology. The two questionnaires partially overlap, asking respondents the same questions with only changes of tense. This allows us to compare answers before and after the introduction and use of the target technology. The expectations questionnaire contains 23 closed questions (Likert-type or Yes/no) and 1 open item, and the experiences questionnaire contains 19 closed questions (Likert-type) and 7 open items. The two questionnaires shared 10 closed question, differing only in tense (*cf.* Appendix A).

[1] The project group consisted of the authors and, from Vejle and Give Hospital, leaders of the speech recognition project, Aase Andreasen and Trine Ankjær. Useful input to the questionnaire was received from the company delivering and implementing the speech recognition technology, Max Manus A/S.

3.3 Procedure

The administration of the survey questionnaires followed the schedule for the introduction of the speech recognition system at the different hospital departments; see Table 1.

During 2006, the system was introduced successively into the otology, medicine, and oncology departments. About one month prior to their introduction to the system, physicians in each department received e-mails inviting them to answer the expectations questionnaire. When a department had been using speech recognition for about four months, physicians were once again invited by e-mail to participate in the second phase of the survey, this time answering the experiences questionnaire. Physicians in the three departments mentioned completed both the expectations questionnaire and the experiences questionnaire.

Three additional departments completed the experiences questionnaire only. During 2005, speech recognition had been introduced at the orthopaedic surgery, organ surgery, and neurology departments. The physicians in these departments received the experiences questionnaire after they had been using speech recognition for eight to twelve months. The additional data consolidate the analysis of the physicians' experiences using voice input.

The physicians were contacted *via* their professional e-mail address. The e-mail contained an introduction to the survey and the motivation for conducting it, explained how to participate and that participation was anonymous, and included a link to the questionnaire, which was Web-based. To lend the survey both practical relevance and scientific credibility, the e-mail was co-signed by the project manager of the speech recognition project at the hospital and the second author of this article. While participation in the survey was anonymous, each respondent had a unique identifier that enabled us to pair a respondent's expectations and experiences answers. For each of the questionnaires, two e-mail reminders were sent to non-respondents. Completion of each of the questionnaires was estimated to take fifteen minutes.

Table 1: Schedule for the survey questionnaires.

	Expectations questionnaire	Experiences questionnaire
Medicine	May 2006	September 2006
Neurology		August 2006
Oncology	August 2006	December 2006
Organ surgery		August 2006
Orthopaedic surgery		August 2006
Otology	March 2006	August 2006

3.4 Speech contribution rates

In addition to the data collected through the survey, the vendor of the speech-recognition system provided the average "speech contribution rate" (SCR) for each physician for each month of 2006. The speech contribution rate represents the percentage of words that remain unaltered when a physician reviews a document produced by speech recognition and performs any manual corrections and modifications deemed necessary. At Vejle and Give Hospital, approval is the responsibility of the physician who dictated the document to the speech-recognition system.

Thus, the speech contribution rate is similar to, but not identical with, a standard speech recognition rate. While a speech recognition rate compares the recognized text with the actual spoken text, the speech contribution rate compares the recognized text with the final text entered into the medical records. Thus, the speech contribution rate diverges from a speech recognition rate when a physician not only corrects the recognized text for misrecognitions but also revises it by adding, deleting, or changing formulations compared to the originally spoken text. Physicians may also differ in their willingness to correct inconsequential misrecognitions. Lacking the data required for computing the speech recognition rate, we find the speech contribution rate, which is calculated automatically by the speech-recognition system, a useful measure of the system's work-related quality.

3.5 Response rate

The survey data will be grouped in two ways during the analysis. First, one set of analyses will investigate the correlations between expectations and experiences. These analyses are based on the data from the 39 physicians who responded to both questionnaires (response rate: 39%). Second, another set of analyses will investigate the respondents' experiences using speech recognition. These analyses are based on the 98 responses (including the 39 above) to the experiences questionnaire, (response rate: 53%). A total of 112 questionnaires were received from the 186 physicians to whom invitations were distributed, yielding an overall response rate of 60%. Table 2 gives the response rates for the individual departments.

Table 2: Response rates for the survey.

Department	Physicians	Expectations questionnaire		Experiences questionnaire		Both expectations and experiences questionnaires	
		Respondents	Response rate	Respondents	Response rate	Respondents	Response rate
Medicine	60	29	48%	36	60%	23	38%
Neurology	24	-	-	10	42%	-	-
Oncology	23	11	48%	8	35%	6	26%
Organ surgery	20	-	-	11	55%	-	-
Orthopaedic surgery	42	-	-	20	48%	-	-
Otology	17 [1]	13	81%	13	76%	10	63%
Total	186 [2]	53	53%	98	53%	39	39%

[1] During the expectations survey only 16 physicians were employed in the Otology department
[2] A total of 112 physicians responded to either the expectations or the experiences questionnaire and 41 responded to only the experiences questionnaire

3.6 Differences between respondents and non-respondents

In order to characterise the sample of the population who answered at least one of the two questionnaires when compared to the non-respondents, we compared their respective speech contribution rates and average number of dictations, as reported in Table 3.

Table 3: Comparison of respondents and non-respondents based on the SCR statistics provided by the speech recognition system. In parentheses are the numbers of physicians for whom usage data were not available*.

Department	Survey respondents			Non-respondents		
	Physicians	Average dictations	Average SCR	Physicians	Average dictations	Average SCR
Medicine	41 (+1)	2609.6	84.6	13 (+5)	1550.1	84.7
Neurology	9 (+1)	2674.6	88.2	13 (+1)	2505.1	75.7
Oncology	8 (+5)	473.4	75.3	5 (+5)	554.0	80.6
Organ surgery	9 (+2)	3843.3	86.3	6 (+3)	3575.5	86.8
Orthopaedic surgery	17 (+3)	4423.5	89.4	19 (+3)	2243.7	80.4
Otology	16 (+0)	2855.8	88.3	1 (+0)	1077.0	89.0
Total	100 (+12)	2903.3	85.7	57 (+17)	2116.6	81.2

Note: SCR – speech contribution rate. * Data about their use of the speech recognition system were available for only 100 respondents (submitting at least one of the two questionnaires) out of 112 (89%) and 57 non-respondents out of 74 (77%).

Respondents produced, on average, significantly more dictations (+787 in average, $p < 0.005$, t-test, equality of variances not assumed) and achieved significantly higher speech contribution rates (+4.5 points, $p < 0.01$, t-test, equality of variances not assumed) than non-respondents. This result does not necessarily show that speech recognition worked better for the respondents than for the non-respondents but it indicates that respondents, on average, tended to leave more text unchanged than non-respondents did.

4 Results

4.1 Expectations versus experiences

Ten questions were included in both the expectations and the experiences questionnaire. The overall tendency was toward more negative experiences than expectations, and for six of the questions the physicians' experiences were significantly more negative than their expectations; see Table 4 and Appendix B. Several of the questions show a polarization effect in that the number of responses in the neutral middle category was reduced.

Table 4: Comparison of expectations and experiences, $N = 39$.

#	Questions (numbers refer to item numbers in Appendix A)		Expectations		Experiences	
			Positive	Negative	Positive	Negative
1	I think it is / was a good idea to introduce speech recognition for medical record keeping (Agree completely – Disagree completely)		44%	36%	34%	46%
5	My department head thinks it is / was a good idea to introduce speech recognition for medical record keeping (Agree completely – Disagree completely)		64%	13%	59%	8%
6	My colleagues think it is / was a good idea to introduce speech recognition for medical record keeping (Agree completely – Disagree completely)		43%	39%	41%	54%
8	After the introduction of speech recognition the quality of medical records will in general be / has in general (Improved a lot – Declined a lot)	**	13%	64%	3%	77%
9	Wrt. precision (*i.e.* that no superfluous information is included) medical records will / have turned out to (become more precise – become less precise)	*	23%	36%	10%	51%
10	Wrt. structure (*i.e.* that information is where it is supposed to be) medical records will / have turned out to (become more structured – become less structured)	*	23%	26%	13%	41%
11	Wrt. completeness (*i.e.* that all required information is included) medical records will / have turned out to (become more complete – become less complete)	**	10%	46%	5%	72%
12	Speech recognition will optimize / has optimized the process of keeping medical records (Agree completely – Disagree completely)	**	34%	39%	18%	67%
13	Speech recognition will produce / has produced appreciable time savings for the benefit of patient care (Agree completely – Disagree completely)		8%	82%	0%	90%
14	Due to speech recognition the amount of time I expect to spend / am spending on medical record keeping will be / has become (Much shorter – Much longer)	**	5%	85%	3%	95%

Note: 'Positive' gives the sum of positive responses (*e.g. agree completely* and *agree somewhat*). 'Negative' gives the sum of negative responses (*e.g. disagree completely* and *disagree somewhat*). Remaining responses are neutral or *Don't know*. Significant differences (Wilcoxon signed-rank test) between expectations and experiences are marked with asterisks: $^{*}\ p < 0.05$, $^{**}\ p < 0.01$.

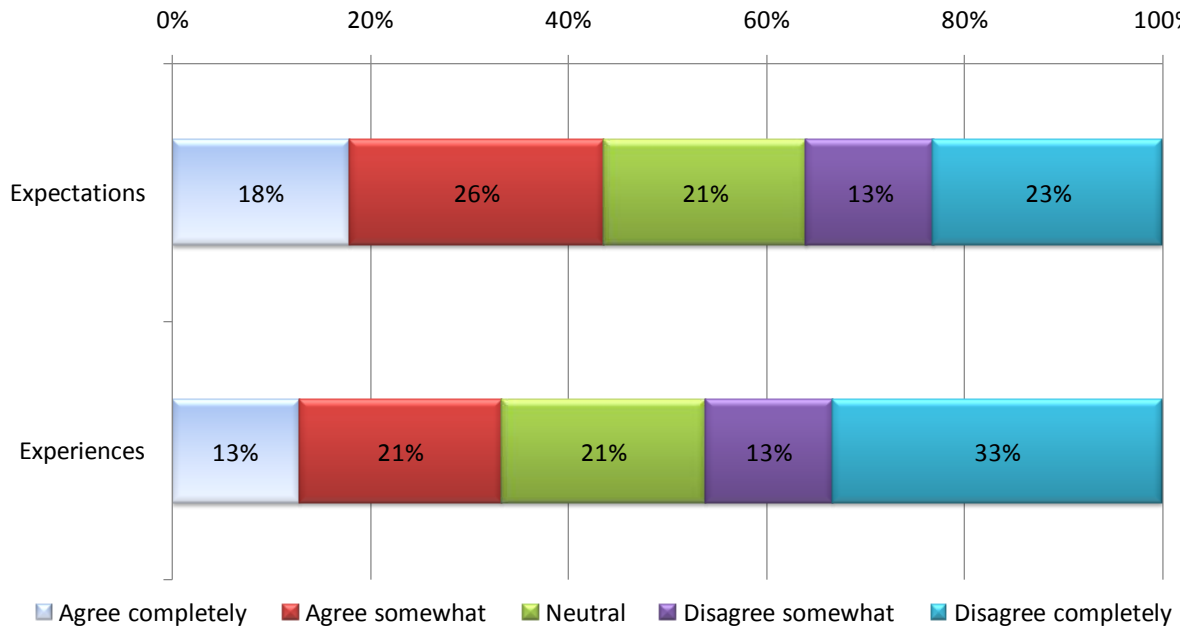


Figure 1: Physicians' responses to the question 1 "I think it is/was a good idea to introduce speech recognition for medical record keeping" in the expectations and experiences questionnaires, N = 39.

The physicians' overall assessment of whether it was a good idea to introduce speech recognition (Figure 1) shows a significant correlation of 0.71 between the expectations and experiences questionnaires ($p < 0.001$, Spearman's rho). That is, the variation in expectations predicted (r^2) 51% of the variation in experiences. At the same time, the difference in assessment before and after is slightly below the threshold of significance ($p = 0.051$, Mann-Whitney). For this question, four physicians (10%, $N = 39$) had more positive experiences than expectations, whereas 11 had more negative experiences than expectations (28%). Expectations varied across departments: physicians in the oncology department were significantly more negative in their overall assessment before they started using the system compared to physicians in the otology and medicine departments ($p < 0.005$, Kruskal Wallis).

4.2 Factors influencing overall assessment

In technology-acceptance research, factors that may influence people's acceptance of systems are typically correlated with (self-reported) usage of systems. Because use of the system that we investigated was mandatory, the items included in this study were instead correlated with physicians' overall assessment of whether it was a good idea to introduce speech recognition. Table 5 shows the correlations.

Table 5: Predictors of overall assessment of speech-recognition system, $N = 39$.

#	Items (numbers refer to the questions in Appendix A)	Overall assessment (expectations)	Overall assessment (experiences)
1	**Overall assessment**	1.00	0.71**
	Performance expectancy		
8	Quality of contents	0.50**	0.47**
12	Improved work process	0.48**	0.36*
	Effort expectancy		
3	Ease of learning	-0.01	-0.13
2	Ease of use	0.46**	0.30
14	Time spent	0.11	0.10
	Social influence		
5	Department head	0.37*	0.20
6	Colleagues	0.56**	0.35*
7	Medical secretaries	0.47*	0.34
	Facilitating conditions		
4	Transcription service provided by medical secretaries	-0.59**	-0.44**
29	Access to support during introduction +	0.31	0.18
30	Quality of support during introduction +	0.42**	0.30

Note: Items are single questions from the expectations questionnaire, except those marked with plusses, which are single questions from the experiences questionnaire. Significant correlations (Spearman's rho) are marked with asterisks: * $p < 0.05$, ** $p < 0.01$.

Each of the items concerning performance expectancy was significantly correlated with physicians' overall assessment of speech recognition before starting to use the system (expectations) as well as after having used the system for four months or more (experiences). Thus, expectations about improved quality of the contents of medical records and about improved work processes in the production of medical records were important predictors of the physicians' acceptance of speech recognition, predicting (r^2) 22% and 13%, respectively, of the variation in overall assessment after 4+ months of use.

Conversely, none of the three items concerning effort expectancy was significantly correlated with physicians' overall assessment after having gained experience with the system. Ease of use was however significantly correlated with overall assessment before physicians started using the system, suggesting that this item affected physicians' expectations but lost importance as physicians gained experience with the system.

Before they started using the speech-recognition system, physicians' overall assessment correlated significantly with their perception of whether their department head, their colleagues, and the medical secretaries were in favour of the introduction of speech recognition. These three social-influence items predicted (r^2) 14%, 31%, and 22%, respectively, of the variation in overall assessment before starting to use the system. After having gained personal experience with the system, colleagues was the only one of the three social-influence items that still correlated significantly with overall assessment. Physicians' perception of their colleagues' assessment of the system explained as much of the physicians' overall assessment as their performance expectancy. Conversely, the social influence of department heads and medical secretaries appeared to fade away when the physicians started using the system.

Among the facilitating conditions, the transcription service provided by the medical secretaries was significantly negatively correlated with physicians' overall assessment of speech recognition, explaining 35% of the variation in overall assessment before physicians started using the systems and 19% of the variation in overall assessment after they had gained experience using it.

4.3 Experience of speech recognition

Physicians' experiences with the speech-recognition system were collected when they had used the system for four months or more. Table 6 shows their responses and the correlation between individual responses and the speech contribution rate, *i.e.* the extent to which the user accepts the system produced text (see Section 3.4).

The overall pattern of responses from the 98 physicians responding to the experiences questionnaire was similar (*cf.* Tables 4 and 6) to that of the sub-group of 39 physicians responding to both questionnaires and whose data we have discussed in sections 4.2 and 4.3. However, the physicians who answered only the experiences questionnaire ($N = 59$) tended to be somewhat more positive than the 39 answering both questionnaires ($p < 0.08$, Mann-Whitney).

Table 6: Experience of speech recognition, $N = 98$.

#	Question (numbers refer to item numbers in Appendix A)	Experiences		Correlation with SCR
		Positive	Negative	
1	I think it was a good idea to introduce speech recognition for medical record keeping (Agree completely – Disagree completely)	33%	31%	0.11
5	My department head thinks it was a good idea to introduce speech recognition for medical record keeping (Agree completely – Disagree completely)	70%	6%	0.01
6	My colleagues think it was a good idea to introduce speech recognition for medical record keeping (Agree completely – Disagree completely)	14%	46%	0.10
8	After the introduction of speech recognition the quality of medical records has in general (Improved a lot – Declined a lot)	15%	62%	0.25*
9	Wrt. precision (*i.e.* that no superfluous information is included) medical records have turned out to (become more precise – become less precise)	16%	43%	0.23*
10	Wrt. structure (*i.e.* that information is where it is supposed to be) medical records have turned out to (become more structured – become less structured)	20%	23%	0.16
11	Wrt. completeness (*i.e.* that all required information is included) medical records have turned out to (become more complete – become less complete)	11%	60%	0.28**
12	Speech recognition has optimized the process of keeping medical records (Agree completely – Disagree completely)	29%	58%	0.18
13	Speech recognition has produced appreciable time savings for the benefit of patient care (Agree completely – Disagree completely)	7%	83%	0.18
14	Due to speech recognition the amount of time I am spending on medical record keeping has become (Much shorter – Much longer)	3%	94%	0.20
24	Today the number of recognition errors is at an acceptable level (Agree completely – Disagree completely)	22%	69%	0.33**
25	The time and effort I spend correcting recognition errors is at an acceptable level (Agree completely – Disagree completely)	17%	76%	0.26*
26	I know how the system can learn from my corrections of recognition errors (Agree completely – Disagree completely)	55%	18%	0.08
27	The system becomes gradually better at recognizing my speech when I mark recognition errors (Agree completely – Disagree completely)	36%	44%	0.22*

Note: 'Positive' gives the percentage of positive responses (*e.g. agree completely* and *agree somewhat*). 'Negative' gives the percentage of negative responses (*e.g. disagree completely* and *disagree somewhat*). Remaining responses are neutral or *Don't know*. 'Correlation with SCR' gives the correlation (Spearman's rho) between physicians' responses and their speech contribution rate; significant correlations are marked with asterisks: * $p < 0.05$, ** $p < 0.01$.

Concerning their overall assessment of whether it was a good idea to introduce speech recognition, respondents were distributed about equally across positive (33%), neutral (35%), and negative responses (31%). Notably, overall assessment was not significantly correlated with speech contribution rate. Several other questions indicate that the technical performance of the system was unsatisfactory. Particularly, 69% of physicians disagreed that the number of recognition errors was at an acceptable level, and 76% disagreed that the time and effort they spent correcting recognition errors was at an acceptable level. Unsurprisingly, disagreeing on these questions correlated significantly, though weakly, with low speech contribution rates.

It appears that the introduction of the speech-recognition system has affected medical record keeping negatively in two important ways. First, the time and effort involved in producing medical records is perceived to have increased. Indeed, 94% of physicians found that they now spent more time on medical record keeping, and 83% disagreed that speech recognition had produced timesaving for the benefit of patient care. Second, the quality of the records is perceived to have suffered. Thus, 62% indicate that the general quality of records has declined and 60% that medical records have become less complete. For these two items, there was a significant, though weak, correlation with speech contribution rate, indicating that physicians who experienced a decrease in quality and completeness made more changes to the recognized text compared to physicians who experienced an increase in quality and completeness. This indicates that physicians attempted to compensate for the perceived inadequacies of the speech-recognition system. With respect to precision and structure – two other quality attributes – responses were more mixed, but few physicians experienced an improvement (16% and 20%, respectively).

Physicians perceived their department heads as being in favour of the speech-recognition system (48% completely agreed to this item). This suggests strongly that department heads have provided the managerial support necessary to carry through the introduction of the system. Interestingly, physicians perceived their colleagues to be somewhat more negative toward the introduction of the speech-recognition system than their colleagues were in their own overall assessment of the introduction of the system (*cf.* the first and third items in Table 6). This may suggest that when talking with each other about the system the physicians have highlighted negative aspects. One positive aspect was that 55% of physicians agreed that they knew how the system could learn from their correction of recognition errors. As described below, this led to gradual performance improvements.

4.4 Evolution of speech contribution rates

During their first month of using the speech-recognition system, the physicians made an average of 130 dictations. From their second through to their eleventh month of using the system the average number of monthly dictations made by a physician was in the range 320 to 417. This indicates that the system was widely used and that the physicians gained considerable experience. The average duration of a dictation was 17.5 seconds.

Figure 2 shows a steady improvement in the speech contribution rate for the survey respondents as they gained experience using the system. During their first month of use they achieved an average speech contribution rate of 79%, but after eleven months of usage this had increased to 94%, an average monthly increase of 1.4 percentage points (least square linear trend, coefficient of determination R^2 = 0.94). It should, however, be kept in mind that fewer physicians have used the system for eleven months than for one month. This is not an indication that physicians are discontinuing their use of the system but that different departments started using it at different points in time. The highest speech contribution rate is achieved by two physicians with eleven months of experience, who have an average of over 640 dictations per month.

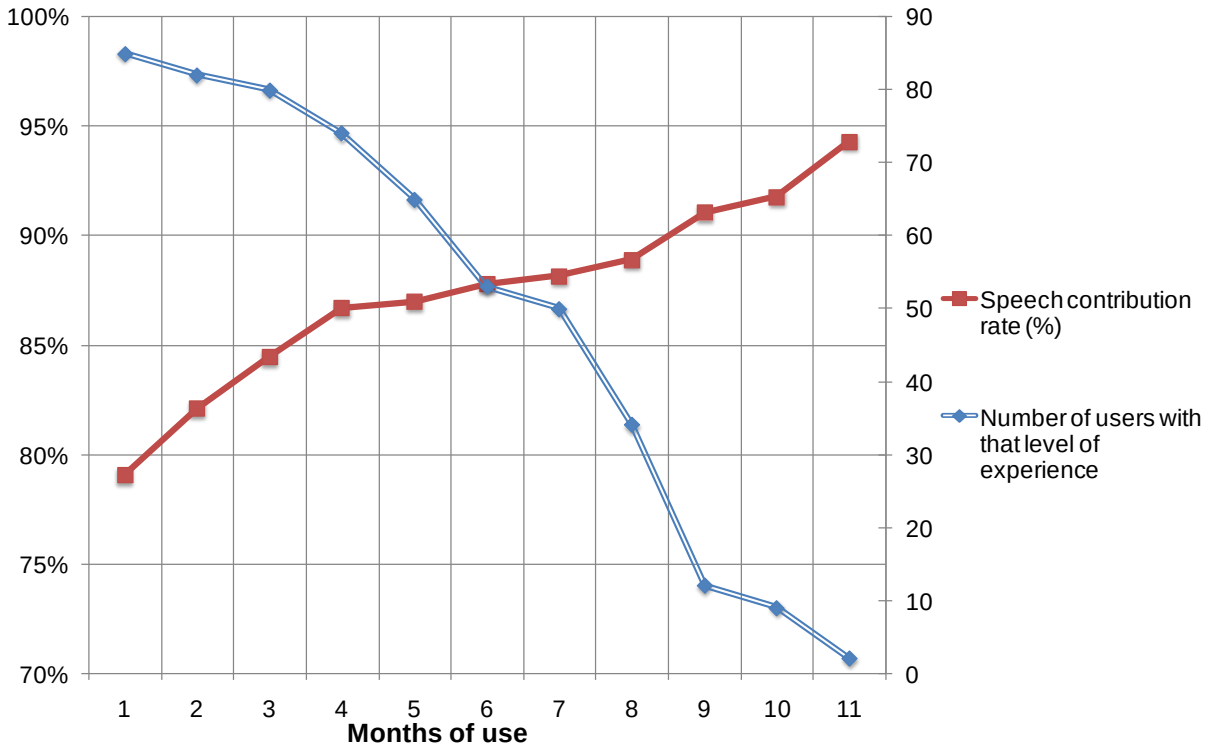


Figure 2: Speech contribution rate as a function of months of experience using the system, and number of physicians at the different levels of experience.

Speech contribution rates varied considerably across physicians. Figure 3 shows that large variation existed even for physicians with the same level of experience with the system. As an example, the bottommost curve shows that after using the system for one month three physicians had speech contribution rates below 52%, three above 94%, and the remaining 79 physicians between 60% and 92%. With increasing levels of experience the variation across physicians decreased (standard deviation σ = 11.9, 6.2, and 3.9 percentage points for 1, 5, and 10 months of experience, respectively). Most of the improvement in average speech contribution rate with increasing levels of experience consisted of physicians with low initial speech contribution rates catching up with the other physicians.

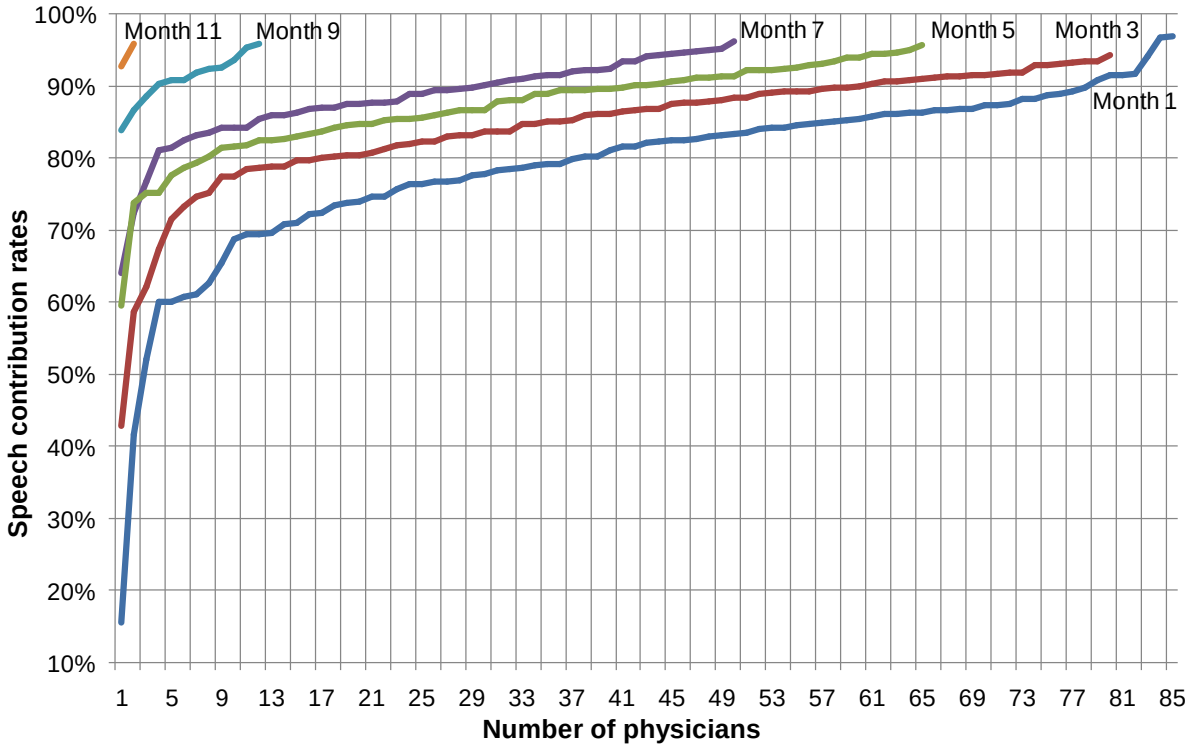


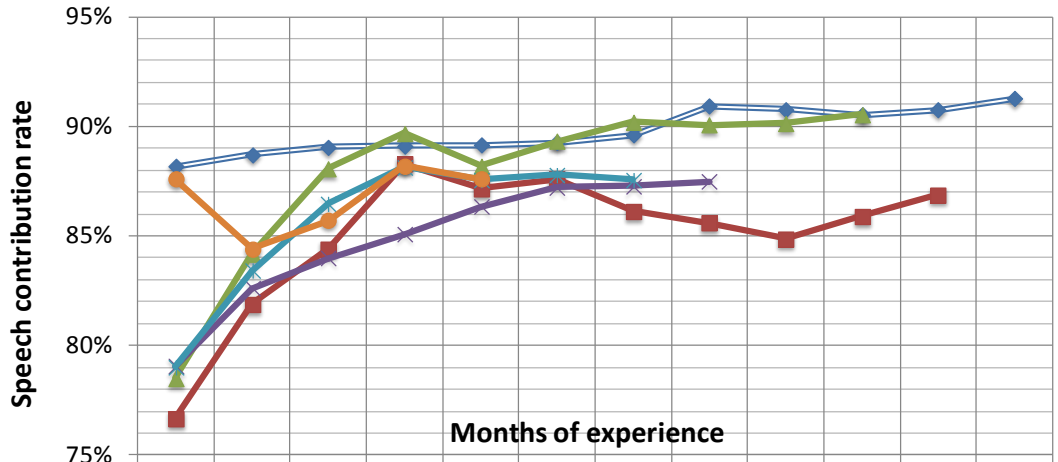
Figure 3: Accumulated distribution of speech contribution rates at different levels of experience (*i.e.* months of using the speech-recognition system).

4.5 Focus on groups of users

The previous section has shown that there is a high variability between subjects, especially when they start using the speech recognition system. Therefore, in order to reduce the effect caused by the evolution of the population, we will now follow stable groups of users, grouped by their total seniority as of January 2007. Groups with fewer than five users are excluded. Here again, only the respondents to the survey are taken into account.

In accordance with Figure 3, Figure 4 shows a high variability between groups, and similarly to Figure 2, Figure 4 reports a positive evolution of the speech contribution rate for all the groups. The highest gain is quite logically for the groups starting with the lowest speech contribution.

The new information provided by Figure 4 is that whatsoever the variability among groups, the speech contribution rate increases in average for all of them during the covered period. Another observation is that groups mostly tend to keep their rank when sorted by their speech contribution rates.



	1	2	3	4	5	6	7	8	9	10	11	12
21 users, 12 months	88,2%	88,7%	89,0%	89,1%	89,1%	89,3%	89,6%	90,9%	90,8%	90,5%	90,8%	91,3%
6 users, 11 months	76,7%	81,9%	84,4%	88,3%	87,2%	87,6%	86,1%	85,6%	84,9%	85,9%	86,9%	
10users , 10 months	78,5%	84,2%	88,1%	89,7%	88,2%	89,3%	90,2%	90,1%	90,1%	90,5%		
24 users, 8 months	79,0%	82,6%	84,0%	85,1%	86,3%	87,2%	87,3%	87,5%				
14 users, 7 months	79,0%	83,4%	86,5%	88,1%	87,6%	87,8%	87,6%					
8 users, 5 months	87,6%	84,4%	85,7%	88,2%	87,6%							

Figure 4: Evolution of speech contribution rates for various groups of users.

4.6 Differences between groups of respondents

Some analyses were conveyed to highlight some possible differences among groups of respondents.

4.6.1 Differences within the "expectations" questionnaire

In the "expectations" questionnaire (N = 41), there was no statistically significant difference between departments for most questions. However, each of the three departments had at least one question significantly differing from the two other departments ($p < 0.005$; Kruskal Wallis), as detailed below.

As reported by the questions 1 and 2, the oncology department was clearly against the introduction of the new system (mean of 4.3 on a scale where 1 is very positive, 3 is neutral, and 5 is very negative) and expected some difficulties (4.3/5 where 5 is very difficult), while the two other departments, *i.e.* otology and medicine, were somehow positive (means of 2.9 and 2.7/5) and did not expect much difficulties (2.5 and 2.8/5).

The medicine department perceived their chef as very positive towards the introduction of the new system (question 5, mean 1.5/5) while the two other departments were only slightly positive (means of 2.8 for otology and 2.5 for oncology).

The last clearly significant difference ($p < 0.004$) is between the otology department that reports not to be often sought for help by colleagues regarding IT issues (question 23, mean of 3.8 where 1 is positive, 3 is neutral and 5 is negative) while the two other departments were somehow positive (means of 2.2 and 2.7/5).

4.6.2 Differences within the "retrospective" questionnaire

In the "retrospective" ($N = 41$) questionnaire, there was no statistically significant difference for most questions between the three departments of orthopaedic surgery, organ surgery and neurology. The only question with a very significant difference ($p < 0.01$; Kruskal Wallis) is the one regarding the perceived opinion of the colleagues of the general merit of introducing the new system, for which the orthopaedic surgery responded positively (mean of 2.6), the neurology department was neutral (mean of 3.2) and the organ surgery somehow negative (mean of 3.6).

4.6.3 Differences within the "after" questionnaire

In the "after" questionnaire ($N = 57$), none of the question was answered significantly differently between the three departments of otology, medicine and oncology.

4.6.4 Differences between the "after" and "retrospective" questionnaires

Globally, there is a significant difference (Kruskal Wallis, $p < 0.03$) between the responses given by the respondents to the "after" ($N = 57$) and "retrospective" ($N = 41$) surveys. Many of the questions were indeed answered significantly differently between the "after" and the "retrospective" groups (7 questions with $p \leq 0.02$). However, the difference seems to be mainly due to differences among departments.

4.6.5 Differences of "speech contribution rate"

In this section, we study the variability of speech contribution rates between the various departments, for all the respondents of the "after" and "retrospective" surveys, for whom we have individual speech contribution rates. The limitations of this "speech contribution rate" parameter, acknowledged earlier in the article, should be kept in mind.

The Levene test of homogeneity of variances is not satisfied (3.7, $p < 0.005$) therefore One-way ANOVA cannot be used directly, and a non-parametric test (Kruskal Wallis) is used instead.

The Kruskal Wallis test on all the departments ("after" and "retrospective" surveys) shows a significant difference between their speech contribution rates ($p < 0.02$). This could however be due to only a couple of particular departments, and therefore a closer look is taken, by studying separately the departments of the "retrospective" and then "after" surveys.

There is no significant difference of speech contribution rates between departments of the "retrospective" survey (Kruskal Wallis, $p > 0.53$). Their differences are therefore mostly due to the physicians' individual variability, rather than to the departments' respective influence.

On the contrary, there is a significant difference of speech contribution rates between the departments of the "after" survey (Kruskal Wallis, $p < 0.022$). The fact that the mean rank of the medicine department (24.3) is below the oncology department (27.2) while their mean is respectively 84.5 and 79.8 (opposite order) illustrates the extreme individual variability inside the oncology department (standard deviation 19.5 points).

We can conclude that there is a significant difference of speech contribution rate between departments, which is mainly due to the otology department reaching a significantly higher score and the oncology department being significantly lower, while the difference among the other departments is only a little significant ($p = 0.051$). Otherwise, the individual variability generally overcomes the departments' influence.

4.7 Typical comments from the respondents

Out of 98 respondents to the experiences questionnaire, 94 expressed comments in at least one of the 7 open questions. The main trends revealed by this questionnaire were supported by the free comments, which additionally covered points not addressed by the questionnaire. For instance, 33 respondents expressed negative feelings about doing a "secretary's job".

"*Why use a high-salary highly qualified physician, who can type with only two fingers, to do secretarial tasks that could be done better and more cheaply by a secretary mastering ten-finger typing?*" (Translated from Danish to English)

On the other hand, 14 respondents indicated that the reduced involvement of the secretaries provides independence and removes some errors.

"*The record is just like I want it to be". "There is no more .....* [*sic*; signs used by the secretaries when they cannot understand what has been said on the tape] *notes from the secretaries. There are no more secretary errors altering the meaning.*"

A contrary view was expressed by 3 physicians concerned that secretaries are no longer there to capture errors or inadequacies, especially checking reference codes and related documents.

"*The control function usually provided by the good secretaries is lost, for instance on checking the* [...] *codes. Another example is if one dictates a need to refer the patient to another clinician and forgets to actually make this referral document.*"

Most respondents found that the new work procedure is optimising the workflow; as much as 77 respondents offered comments expressing this view: *e.g.,* "*Records are done on the fly*"; "*Records are immediately available for further use*". However, 83 physicians also indicated that the use of speech recognition takes too much time. Comments are mentioning from 20% to 300% more time than the previous procedure. It should be noted that a substantial number of respondents asked (17 explicitly) for more information during the introduction, to spend more time on the integration of the new system, and to re-assign some resources consequently.

"*Under the introduction of the new system, it is of major importance to recognise that it takes more time* [for the physicians]*, and to take that into account when informing the co-workers and do some planning accordingly.*"

Many critical comments concern the integration of the speech recognition system into the existing EMR system (12 comments) and towards the user interface, which is seen as too slow (to start, to react) and requiring too much mouse interaction (29 comments). Six respondents state that they often avoid using the speech recognition system and use the keyboard instead.

"*For short documents, it is less time consuming to write directly in the patient record* [with the keyboard]*, since speech recognition is too slow.*"

Regarding the pure speech recognition capabilities, 61 comments call for improvements. In particular, respondents complain that the system does not appear to learn from the corrections they enter.

"*One feels like Sisyphus, correcting the same things again and again.*"

27 comments report some difficulties with complex sentences.

"*More telegram-style, resulting in less descriptive and less nuanced texts.*"

18 respondents consider the types of errors produced by speech recognition more difficult to spot than previous transcription errors and potentially more harmful (10 comments). These observations are corroborated by the findings of a previous study [Honeycutt 2003].

"*The system does not perform well in recognising the small words, which often have a crucial impact on the meaning (e.g. 'and', 'not', etc.)*"

Some respondents suggest using speech recognition only in some specific areas (8 comments), in particular for urgent, medium-to-long documents, with short typical sentences, and only in a department with low background noise.

Overall, 7 respondents express enthusiasm toward the technology, while 26 report that they experience an increase in stress or a decrease in work satisfaction. Finally, a few comments were made by respondents who expressed surprise and worry that their speech contribution rates are monitored by their superior; not least because they have doubts regarding its accuracy and relevance (those limitations have been explained in the article).

"I believe however that I have to correct more than reported in the speech contribution rate. [...] *I was astonished that one monitors physicians' speech contribution rates! They can indeed be manipulated in various ways, such as: 1) Avoid correcting one's own errors; 2) Stick to short and simple formulations; 3) Use standard phrases."*

The opinions reported by the physicians in the free comments are to some extent similar to other experiments reported in the literature [Lai & Vergo 1997].

5 Discussion

5.1 Expectations, experiences and social influence

The two major objectives behind the introduction of speech technology as a front end to the EMR system were to achieve a more rational workflow and thus a quicker completion of records and to enhance the quality of medical records. The first objective has been achieved as judged by the physicians themselves, and information from the hospital corroborates this entirely. Respondents indicate, and express their appreciation, that workflow has improved, and that records are now accessible immediately after dictation. Still, physicians' experiences were more negative than their expectations, particularly with respect to the quality of medical records and the time spent producing them. The technical performance of the system was experienced as unsatisfactory, particularly with respect to the number of recognition errors and the time and effort required to correct them. Respondents almost unanimously reported that the time they personally spent producing medical records has increased, and they also agreed that speech recognition had not led to overall time savings for the benefit of patient care. Physicians experienced that the quality of medical records had declined in general and particularly with respect to record completeness. Finally, respondents are approximately equally divided between those who, in retrospect, think it was and those who think it was not a good idea to introduce speech recognition.

With respect to predictors of the physicians' acceptance of the system, our results indicate that their overall assessment of speech recognition prior to using it was the strongest among the predictors we have tested. This suggests that asking prospective users for their assessment of whether the introduction of a system is a good idea can be used as an early, cheap, and rather reliable indicator of whether they will approve of

the system after having used it for some time. This finding discords, however, with [Root & Draper 1983] who found little correlation between people's assessments before and after they had experience with a system.

For the predictors identified in previous technology-acceptance studies, we find that performance expectancy and social influence were moderate predictors of our respondents' overall assessment of speech recognition before they began to use it. After having gained experience with the system, performance expectancy and perception of colleagues' overall assessment of speech recognition still provided some prediction of overall assessment. Effort expectancy in terms of perceived ease of use was a moderate predictor of overall assessment before starting to use the system but not after months of use. These results are in agreement with previous technology acceptance studies with respect to the presence of significant correlations, the general magnitude of correlations, as well as the effect of experience with the system [Davis 1989; Adams *et al.* 1992; Davis 1993; Venkatesh *et al.* 2003]. It should be noted that Venkatesh et al. (2003) find that social influence is mainly a predictor of technology acceptance when use of a technology is mandatory, as was the case in our study.

As in previous studies (*e.g.* [Venkatesh *et al.* 2003]) facilitating conditions were perceived rather similarly to effort expectancy, except for the moderate and lasting negative influence of physicians' perception of the transcription service previously provided by medical secretaries. While general dissatisfaction with a previous solution may have an only temporary, and supposedly positive, effect on people's assessment of a new technology, our study suggests that a long-lasting and generally well-liked previous solution has a lasting negative effect on people's assessment of a new technology.

With respect to physicians' performance with the speech-recognition system, their speech contribution rate correlated only weakly with their assessments of the system. Weak correlations between assessments and performance measures have also been found in previous studies [Frøkjær *et al.* 2000; Hornbæk & Law 2007]. Physicians' speech contribution rate improved over time, particularly for physicians with low initial rates, and after nine months of use, physicians had an average speech contribution rate of 91%. Thus, having used the system for dictating several thousand EMR entries physicians still experienced that they had to revise one in every eleven words of the text produced by the speech-recognition system. This was perceived as unsatisfactory and time consuming, especially because many physicians felt that they were correcting the same errors repeatedly, as also strongly reflected in respondents' free-text comments.

The system vendor emphasizes that this is the first generation of their system for recognition of Danish medical speech and they maintain that the second generation, currently under deployment, is faster and has a higher recognition accuracy. Anecdotal evidence seems to support this. Moreover, recognition rates typically reported for recognition of English speech (see Section 2) lends credibility to this assertion.

5.2 Limitations of the survey

This study has a number of limitations that should be taken into account in interpreting the results. First, the speech-recognition system was introduced simultaneously with new work procedures as secretary efforts were being replaced by physician efforts. This makes it difficult or perhaps impossible to distinguish effects of using the speech recognition system from effects of the new work procedures, which changed roles and responsibilities in the production of the medical records. Second, while a response rate of 60% is comparable with other surveys of technology acceptance (*e.g.* [Adams *et al.* 1992; Hebert & Benbasat 1994; Fichman & Kemerer 1999]) it calls for caution in interpreting the results. The slightly higher speech contribution rate of respondents compared to non-respondents suggests that the respondents' somewhat half-hearted responses may well be an upper bound on the enthusiasm of the total population of physicians toward the speech-recognition system. Third, physicians who answered both questionnaires received the experiences questionnaire after having used the speech-recognition system for about four months, and speech contribution rates were studied over the first eleven months of use. While this entails that the physicians had considerable experience with the system, it remains unknown whether their assessment of the system had stabilized and it appears that their performance was still improving. Fourth, the speech contribution rate used in this survey is different from a standard speech recognition rate. As explained, the speech contribution rate provides a measure of the work-related quality of the speech-recognition system.

Conclusion

Speech-recognition technology is continuously being refined and is gradually becoming adopted as an alternative to typing text or to dictation and subsequent transcription by secretaries. This study reports the results of a survey of the first hospital to introduce speech recognition in Danish for all clinical specialties and departments. We have found that:

- Physicians' expectations tended to be more positive than their experiences. It is seen as a valuable benefit of the technology that it makes it possible to access records right after their dictation is completed. Yet, the physicians felt that they spent much more time producing medical records with the new system and associated work procedures, that the overall quality of records had declined, and

that the performance of the system in terms of recognizing speech was unsatisfactory.

- Performance expectancy, effort expectancy (especially ease of use), social influence, and facilitating conditions were all moderate predictors of physicians' overall assessment of the speech-recognition system before they started using it. While the performance-expectancy items – quality of contents and improved work process – remained significant predictors also after physicians had gained experience with the system, the only other significant items were colleagues (a social influence) and the transcription service previously provided by the medical secretaries (a facilitating condition).
- The percentage of words that remained unaltered when physicians proofread their medical records (the speech contribution rate) increased as physicians gained experience with the system. While this indicates a gradual performance improvement, the average speech contribution rate after nine months of use was only 91%. Physicians' speech contribution rates correlated only weakly with their assessment of the system.
- Physicians are approximately equally divided among those who think, in retrospect, that the introduction of speech recognition was a good idea, that it was not, and those who are neutral.

While acknowledging that most physicians in the present study have shown a less than enthusiastic reception of speech recognition technology, it should not be overlooked that one third of physicians were positive in their overall assessment of the speech-recognition system after they had gained experience with it. This provides some basis for further efforts to improve speech recognition in Danish and other "relatively small" languages and introduce it for medical record keeping. It needs to be documented to which extent longer periods of practice as well as more mature generations of the technology will lead to higher levels of satisfaction among physician users.

References

[Adams *et al.* 1992] D.A. Adams, R.R. Nelson, P.A. Todd. Perceived usefulness, ease of use, and usage of information technology: A replication. *MIS Quarterly, 1992, 16(2):227-247.* doi:10.2307/249577

[Ajzen 1985] Icek Ajzen. From intentions to actions: A theory of planned behavior. In J. Kuhl and J. Beckmann (eds.), Action Control: From Cognition to Behavior. *Springer, New York, 1985, pp. 11-39.*

[Ajzen 1991] Icek Ajzen. The theory of planned behaviour. *Organizational Behavior and Human Decision Processes, 1991, 50(2): 179-211.* doi:10.1016/0749-5978(91)90020-T

[Alapetite 2006] Alexandre Alapetite. Impact of noise and other factors on speech recognition in anaesthesia. *International Journal of Medical Informatics, available online in December 2006.* doi:10.1016/j.ijmedinf.2006.11.007

[Al-Aynati & Chorneyko 2003] Maamoun M. Al-Aynati & Katherine A. Chorneyko. Comparison of Voice-Automated Transcription and Human Transcription in Generating Pathology Reports. *Archives of Pathology and Laboratory Medicine, 2003, 127(6):721–725*.

[Barker *et al.* 2005] J.P. Barker, M.P. Cooke, D.P.W. Ellis. Decoding speech in the presence of other sources. *Speech Communication, 2005, 45(1):5-25* . doi:10.1016/j.specom.2004.05.002

[Borowitz 2001] Stephen M. Borowitz. Computer-based speech recognition as an alternative to medical transcription. *Journal of the American Medical Informatics Association, 2001, 8(1):101-102*.

[Coniam 1999] D. Coniam. Voice recognition software accuracy with second language speakers of English. *System, 1999, 27(1):49-64*. doi:10.1016/S0346-251X(98)00049-9

[Damanpour 1991] Fariborz Damanpour. Organizational innovation: A meta-analysis of effects of determinants and moderators. *Academy of Management Journal, 1991, 34(3):555-590*. doi:10.2307/256406

[Davis 1989] F.D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 1989, 13(3):319-340*. doi:10.2307/249008

[Davis 1993] Fred D. Davis. User acceptance of information technology: systems characteristics, user perceptions and behavioral impacts. *International Journal of Man-Machine Studies, 1993, 38(3):475-487*. doi:10.1006/imms.1993.1022

[Devine *et al.* 2000] Eric G. Devine, Stephan A. Gaehde, Arthur C. Curtis. Comparative Evaluation of Three Continuous Speech Recognition Software Packages in the Generation of Medical Reports. *Journal of the American Medical Informatics Association. 2000,7(5):462–468*.

[Dillon & Norcio 1997] Thomas W. Dillon & A. F. Norcio - User performance and acceptance of a speech-input interface in a health assessment task. International Journal of Human-Computer Studies (1997) 47:591-602

[Feng & Sears 2004] Jinjuan Feng & Andrew Sears. Using confidence scores to improve hands-free speech based navigation in continuous dictation systems. *ACM Transactions on Computer-Human Interaction, 2004, 11(4):329-356*. doi:10.1145/1035575.1035576

[Fichman & Kemerer 1999] Robert G. Fichman & Chris F. Kemerer. The illusory diffusion of innovation: An examination of assimilation gaps. *Information Systems Research, 1999, 10(3):255-275*.

[Fichman 2000] Robert G. Fichman. The diffusion and assimilation of information technology innovations. In R.W. Zmud (ed.), Framing the Domains of IT Management: Projecting the Future through the Past. *Pinnaflex Educational Resources, 2000, Cincinnati (OH, USA), pp. 105-127*.

[Fiscus 1997] Jonathan G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). *In: Proceedings of IEEE'1997 Workshop Automatic Speech Recognition and Understanding*. doi:10.1109/ASRU.1997.659110

[Fishbein & Ajzen 1975] M. Fishbein, Icek Ajzen. Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research. *Addison-Wesley, 1975, Reading (MA, USA)*.

[Frøkjær *et al.* 2000] Erik Frøkjær, Morten Hertzum, Kasper Hornbæk. Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? *In Proceedings of the CHI'2000 Conference on Human Factors in Computing Systems. ACM Press, 2000, New York, pp. 345-352*. ISBN:1-58113-216-6

[Gallivan 2001] Michael J. Gallivan. Organizational adoption and assimilation of complex technological innovations: Development and application of a new model. *ACM SIGMIS Database, 2001, 32(3):51-85*. ISSN:0095-0033

[Gong 1995] Yifan Gong. Speech recognition in noisy environments: a survey. *Speech Communication, 1995, 16(3):261-291*. doi:10.1016/0167-6393(94)00059-J

[Hebert & Benbasat 1994] M. Hebert, I. Benbasat. Adopting information technology in hospitals: The relationship between attitudes/expectations and behaviour. *Hospital & Health Services Administration, 1994, 39(3):369-383.*

[Honeycutt 2003] Lee Honeycutt. Researching the use of voice recognition writing software. Computers and Composition (2003), 20:77–95. doi:10.1016/S8755-4615(02)00174-3

[Hornbæk & Law 2007] Kasper. Hornbæk & Effie Lai-Chong Law. Meta-analysis of correlations among usability measures. *In: Proceedings of the CHI'2007 Conference on Human Factors in Computing Systems, ACM Press, 2007, New York, pp. 617-626*. doi:10.1145/1240624.1240722

[Jungk *et al.* 2000] Andreas Jungk, Bernhard Thull, Lutz Fehrle, Andreas Hoeft, Günter Rau. A case study in designing speech interaction with a patient monitor. *Journal of Clinical Monitoring and Computing, 2000, 16:295-307*. doi:10.1023/A:1011456205786

[Juul-Kristensen *et al.* 2004] B. Juul-Kristensen, B. Laursen, M. Pilegaard, B.R. Jensen. *Physical workload during use of speech recognition and traditional computer input devices. Ergonomics, 2004, 47(2):119-133*. doi:10.1080/00140130310001617912

[Kanal *et al.* 2001] K.M. Kanal, N.J. Hangiandreou, A.M. Sykes, H.E. Eklund, P.A. Araoz, J.A. Leon, B.J. Erickson. Initial evaluation of a continuous speech recognition program for radiology. *Journal of Digital Imaging, 2001, 14(1):30-37*. doi:10.1007/s10278-001-0022-z

[Lai & Vergo 1997] Jennifer Ceil Lai & John George Vergo. MedSpeak: Report creation with continuous speech recognition. *In: Proceedings of the CHI'1997 Conference on Human Factors in Computing Systems. ACM Press, 1997, New York, pp. 431-438*. ISBN:0-89791-802-9

[Leijten & Van Waes 2005] Mariëlle Leijten, Luuk Van Waes. Writing with speech recognition: The adaptation process of professional writers with and without dictating experience. Interacting with Computers (2005) 17:736–772. doi:10.1016/j.intcom.2005.01.005

[Lombard 1911] E. Lombard. Le signe de l'élévation de la voix. *Annales Maladies Oreille, Larynx, Nez, Pharynx, 1911, 31:101–119*.

[Mohr *et al.* 2003] David N. Mohr, David W. Turner, Gregory R. Pond, Joseph S. Kamath, Kathy B. De Vos, Paul C. Carpenter. Speech Recognition as a Transcription Aid: A Randomized Comparison With Standard Transcription. *Journal of the American Medical Informatics Association, 2003, 10(1):85–93*. doi:10.1197/jamia.M1130

[Mönnich & Wetter 2000] G. Mönnich & T. Wetter. Requirements for speech recognition to support medical documentation. Methods of Information in Medicine (2000), 39(1):63-9.

[Ramaswamy *et al.* 2000] Mohan R. Ramaswamy, Gregory Chaljub, Oliver Esch, Donald D. Fanning, Eric van Sonnenberg. Continuous speech recognition in MR imaging reporting: Advantages, disadvantages, and impact. *American Journal of Roentgenology, 2000, 174(3):617-622*.

[Rogers 2003] Everett M. Rogers. Diffusion of Innovations, Fifth Edition. Free Press, 2003, New York. ISBN:0743222091

[Root & Draper 1983] R.W. Root, S. Draper. Questionnaires as a software evaluation tool. In Proceedings of the CHI'1983 Conference on Human Factors in Computing Systems. *ACM Press, 1983, New York, pp. 83-87*. doi:10.1145/800045.801586

[Sears *et al.* 2001] Andrew Sears, Clare-Marie Karat, Kwesi Oseitutu, Azfar Karimullah, Jinjuan Feng. Productivity, satisfaction, and interaction strategies of individuals with spinal cord injuries and traditional users interacting with speech recognition software. *Universal Access in the Information Society, 2001, 1(1):4-15*. doi:10.1007/s102090100001

[Suhm *et al.* 2001] Bernhard Suhm, Brad Myers, Alex Waibel. Multimodal error correction for speech user interfaces. *ACM Transactions on Computer-Human Interaction, 2001, 8(1):60-98*. doi:10.1145/371127.371166

[Venkatesh *et al.* 2003] V. Venkatesh, M.G. Morris, G.B. Davis, F.D. Davis. User acceptance of information technology: Toward a unified view. *MIS Quarterly, 2003, 27(3):425-478*.

[Wilpon & Jacobsen 1996] J.G. Wilpon, C.N. Jacobsen. A study of speech recognition for children and the elderly. *Proceedings of ICASSP'1996, the International Conference on Acoustics, Speech, and Signal Processing, Vol. I. IEEE, 1996, Los Alamitos, CA, pp. 349-352*. doi:10.1109/ICASSP.1996.541104

[Young 1996] Steve Young. Review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine, 1996, 13(5):45-57*. doi:10.1109/79.536824

[Zafar *et al.* 1999] Atif Zafar, J. Marc Overhage, Clement J. McDonald. Continuous speech recognition for clinicians. *Journal of the American Medical Informatics Association, 1999, 6(3):195–204*.

[Zafar *et al.* 2004] Atif Zafar, Burke Mamlin, Susan Perkins, Anne M. Belsito, J. Marc Overhage, Clement J. McDonald. A simple error classification system for understanding sources of error in automatic speech recognition and human transcription. *International Journal of Medical Informatics, 2004, 73:719-730*. doi:10.1016/j.ijmedinf.2004.05.008

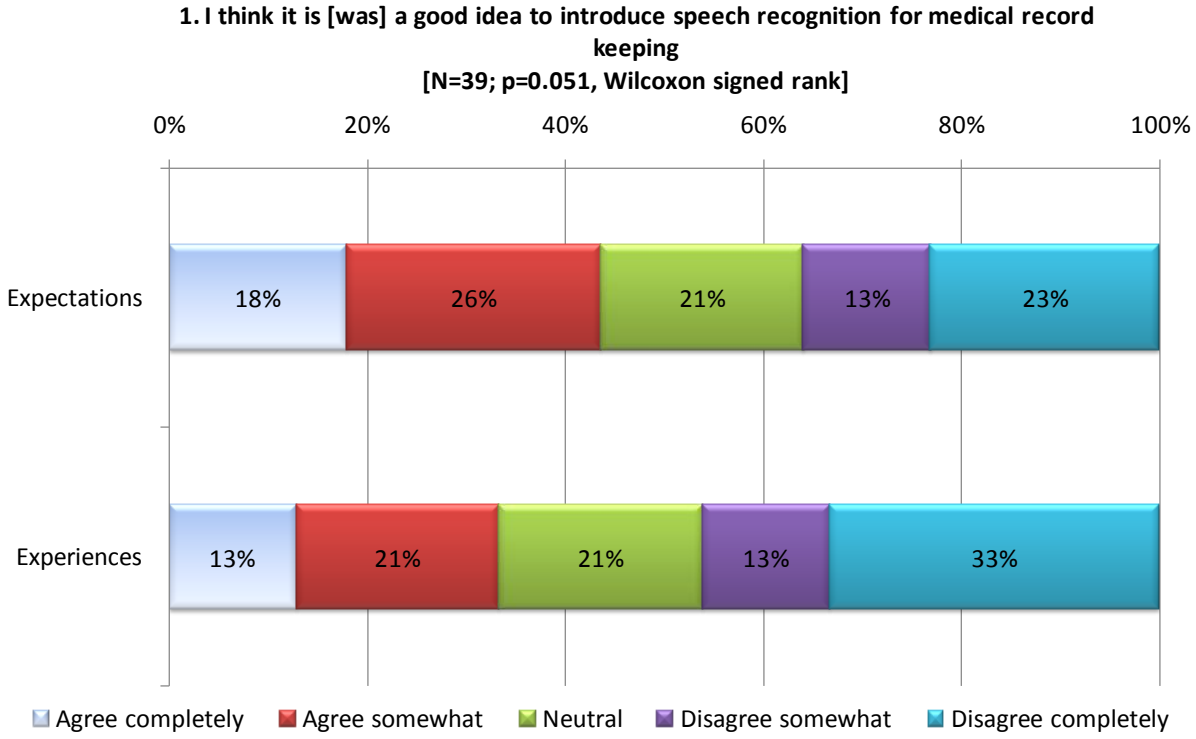
Appendix A: Expectations and experiences questionnaires

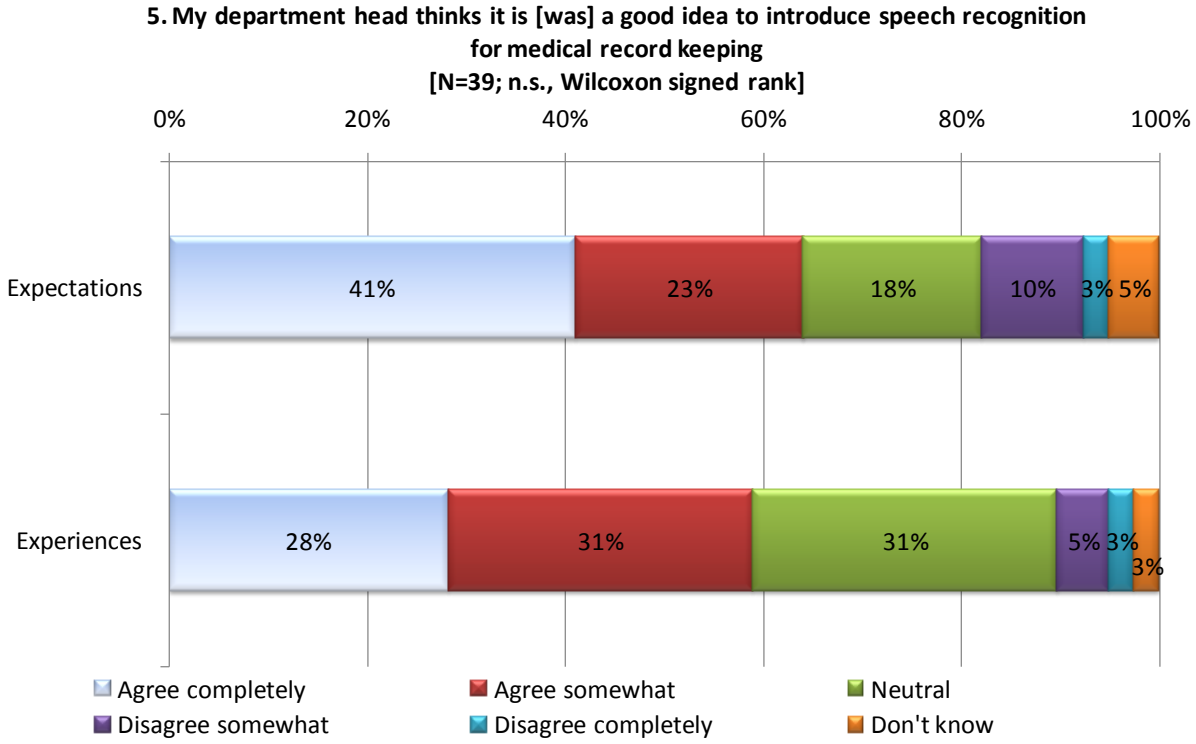
Questions included in the expectations questionnaire are indicated with plusses in the column B (before). Questions included in the experiences questionnaire are indicated with plusses in the column A (after), and variations in their wording compared to expectations questions are in italics. All questions have an additional "Don't know" option. Open questions have been left out.

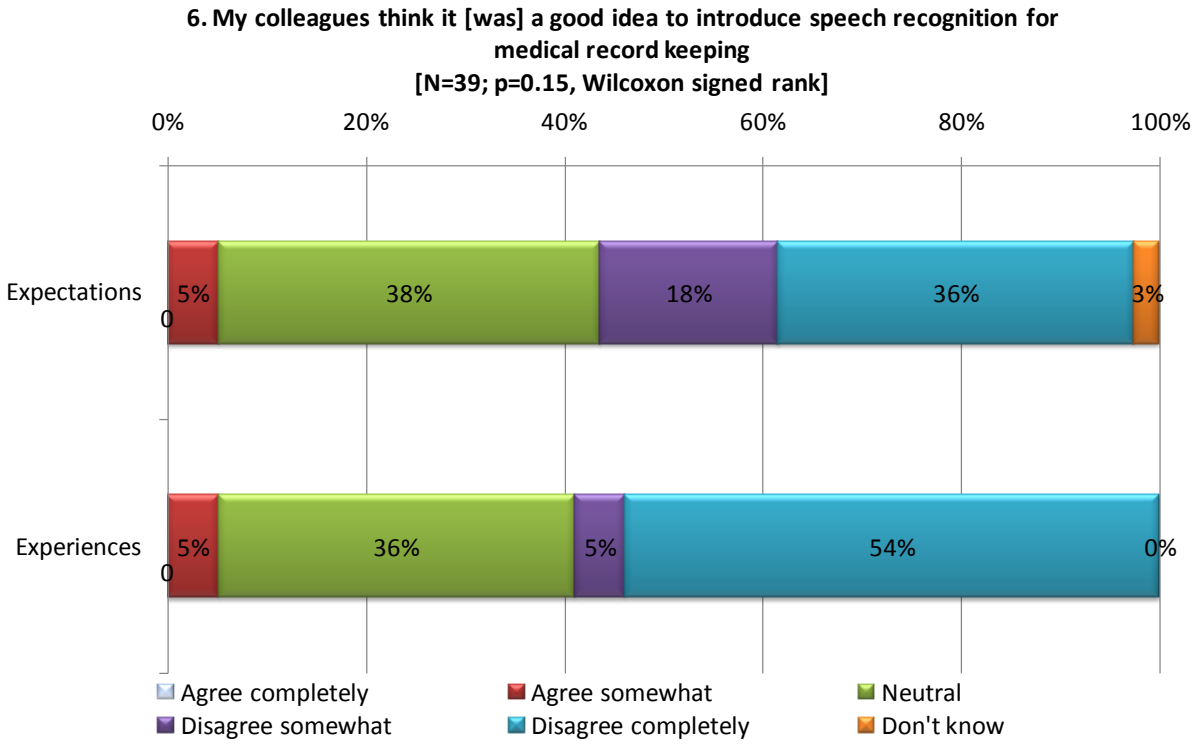
B	A	**Question items**
+	+	1. I think it is [*was*] a good idea to introduce speech recognition for medical record keeping. (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
+		2. I expect it to be easy to use speech recognition once I have become used to it. (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
+		3. I expect to have to spend much effort to become used to working with speech recognition. (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
+		4. The service provided by our secretarial staff is of such high standard that speech recognition will hardly be able to match it. (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
+	+	5. My department head thinks it is [*was*] a good idea to introduce speech recognition for medical record keeping. (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
+	+	6. My colleagues think it [*was*] a good idea to introduce speech recognition for medical record keeping. (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
+		7. Our secretaries think it is good idea to introduce speech recognition for medical record keeping. (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
+	+	8. After the introduction of speech recognition the quality of medical records will in general [*has in general turned out to*] … (Improve a lot, Improve somewhat, Remain the same, Decline somewhat, Decline a lot)
+	+	9. With respect to precision (*i.e.* that no superfluous information is included), medical records will [*have turned out to*] … (Become more precise, Remain at the same level, Become less precise)
+	+	10. With respect to structure (*i.e.* that information is where it is supposed to be), medical records will [*have turned out to*] … (Become more structured, Remain at the same level, Become less structured)
+	+	11. With respect to completeness (*i.e.* that all required information is included) medical records will [*have turned out to*] … (Become more complete, Remain at the same level, Become less complete)
+	+	12. I expect that speech recognition will optimize [*Speech recognition has optimized*] the process of keeping the medical record. (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
+	+	13. I expect speech recognition will produce [*Speech recognition has produced*] appreciable time savings for the benefit of patient care. (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
+	+	14. I expect that the time I spend on producing medical records in the long run will become [*The time I spend on producing medical records has become*] … (a lot shorter, shorter, the same, longer, a lot longer)
+		15. Have you talked with colleagues about their experience with speech recognition? (Yes, No)
+		16. If yes: How was their experience? (Largely positive, Both positive and negative, Largely negative)
+		17. I like to try out new technology. (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)

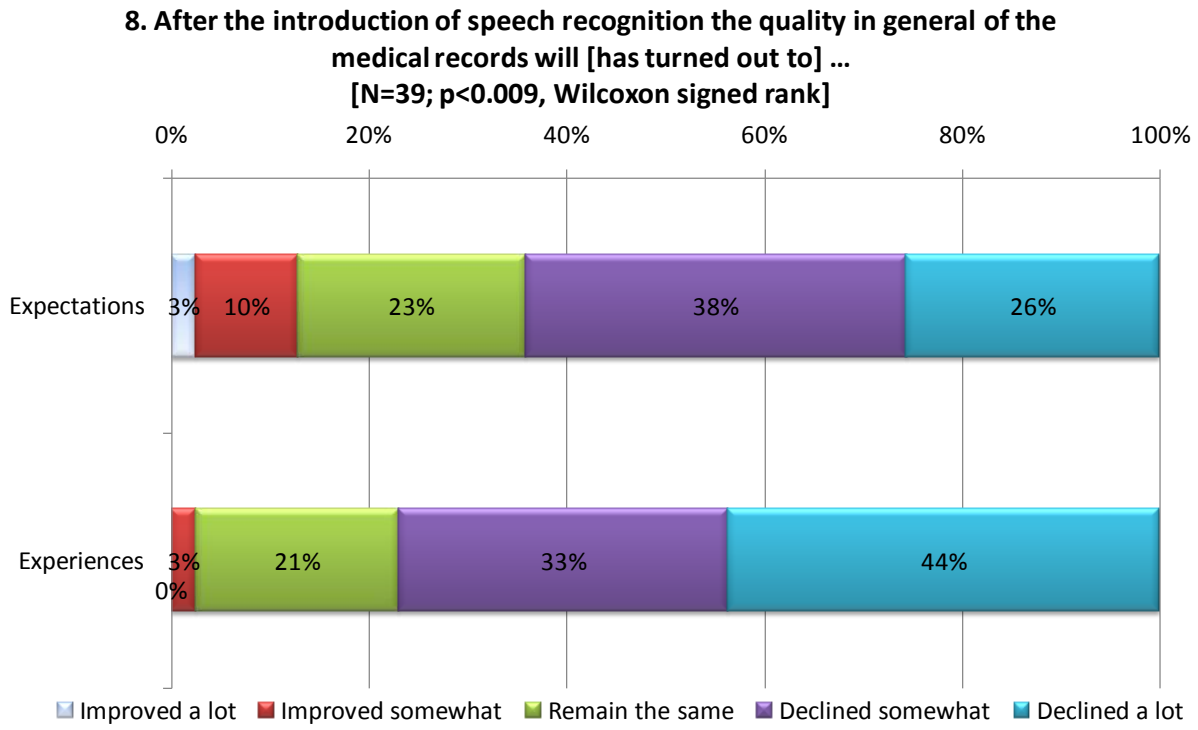
+ 18. I am not comfortable when I have to use a new IT system. (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
+ 19. The use of IT during the clinical work will often raise my level of stress. (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
+ 20. The use of IT will in general lead staff to be more efficient in their clinical work. (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
+ 21. The use of IT will in general make it easier for staff to complete their clinical work. (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
+ 22. When new IT is introduced in our departments/wards, it usually leads to benefits for patients. (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
+ 23. I am often asked for advice about our IT systems by my colleagues. (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
  + 24. *Today, the number of recognition errors is at an acceptable level.* (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
  + 25. *The time and effort I spend correcting recognition errors is at an acceptable level.* (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
  + 26. *I know how the system can learn from my corrections of recognition errors.* (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
  + 27. *The system is gradually becoming better at recognizing my speech when I mark recognition errors.* (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
  + 29. *During the introduction of speech recognition the access to support was satisfactory.* (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
  + 30. *During the introduction of speech recognition, the quality of support was satisfactory.* (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
  + 31. *Today, the access to support is satisfactory.* (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)
  + 32. *Today, the quality of support is satisfactory.* (Agree completely, Agree somewhat, Yes-and-no, Disagree somewhat, Disagree completely)

Appendix B: Distribution of the responses

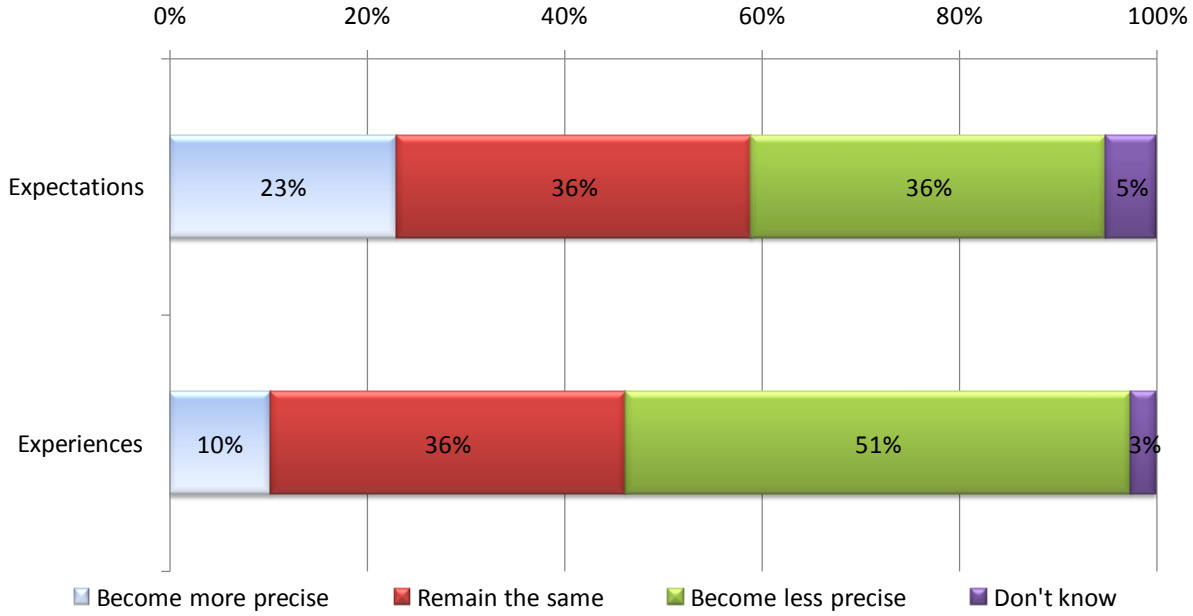




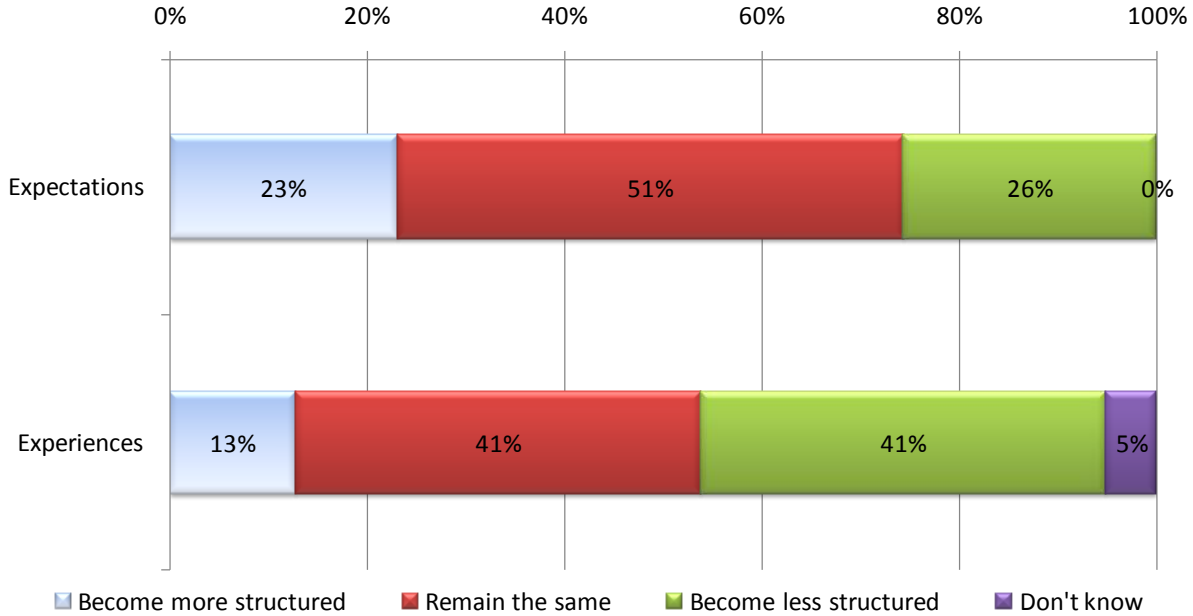

6. My colleagues think it [was] a good idea to introduce speech recognition for medical record keeping
[N=39; p=0.15, Wilcoxon signed rank]
0%
20%
40%
60%
80%
100%
Expectations
0
5%
38%
18%
36%
3%
Experiences
0
5%
36%
5%
54%
0%
Agree completely
Agree somewhat
Neutral
Disagree somewhat
Disagree completely
Don't know


8. After the introduction of speech recognition the quality in general of the medical records will [has turned out to] …
[N=39; p<0.009, Wilcoxon signed rank]
0%
20%
40%
60%
80%
100%
Expectations
3%
10%
23%
38%
26%
Experiences
0%
3%
21%
33%
44%
Improved a lot
Improved somewhat
Remain the same
Declined somewhat
Declined a lot

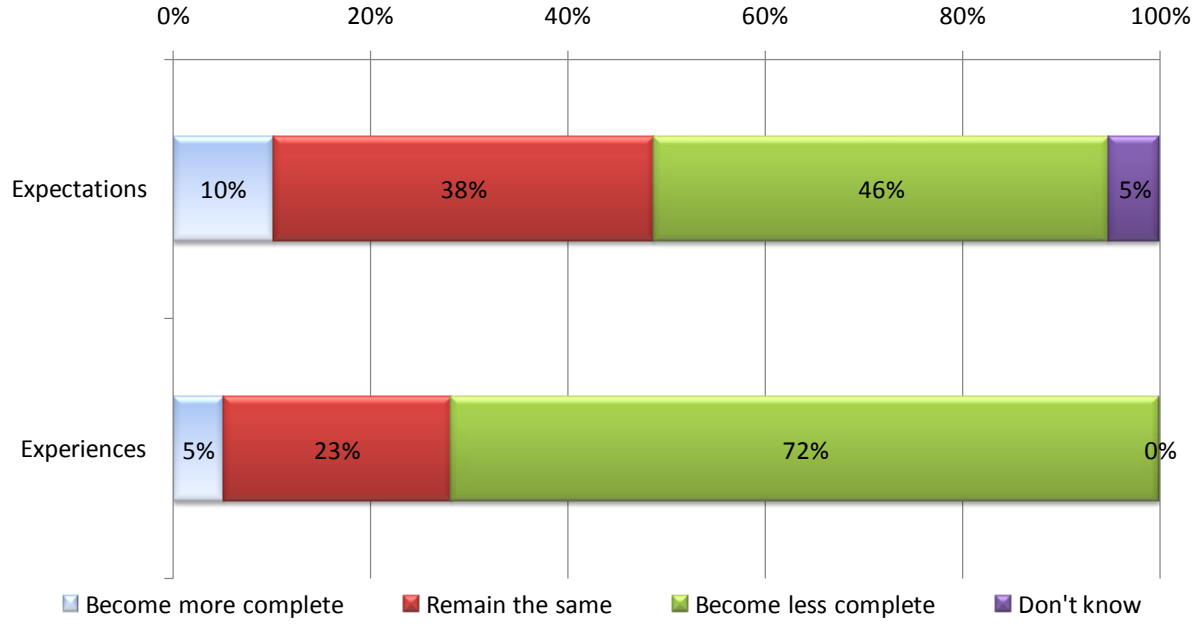
9. With respect to precision (i.e., that no superfluous information is included) I expect medical records to [medical records have turned out to] … [N=39; p<0.05, Wilcoxon signed rank]



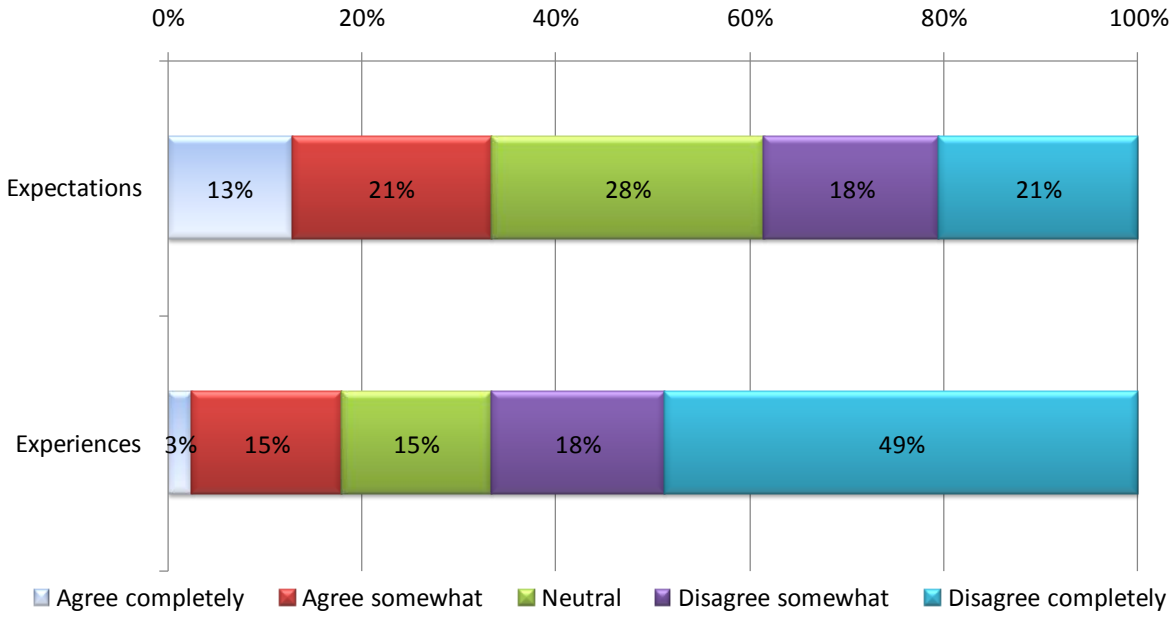
10. With respect to structure (i.e., that information is where it is supposed to be) I expect medical records to [medical records have turned out to] … [N=39, p<0.04, Wilcoxon signed rank]



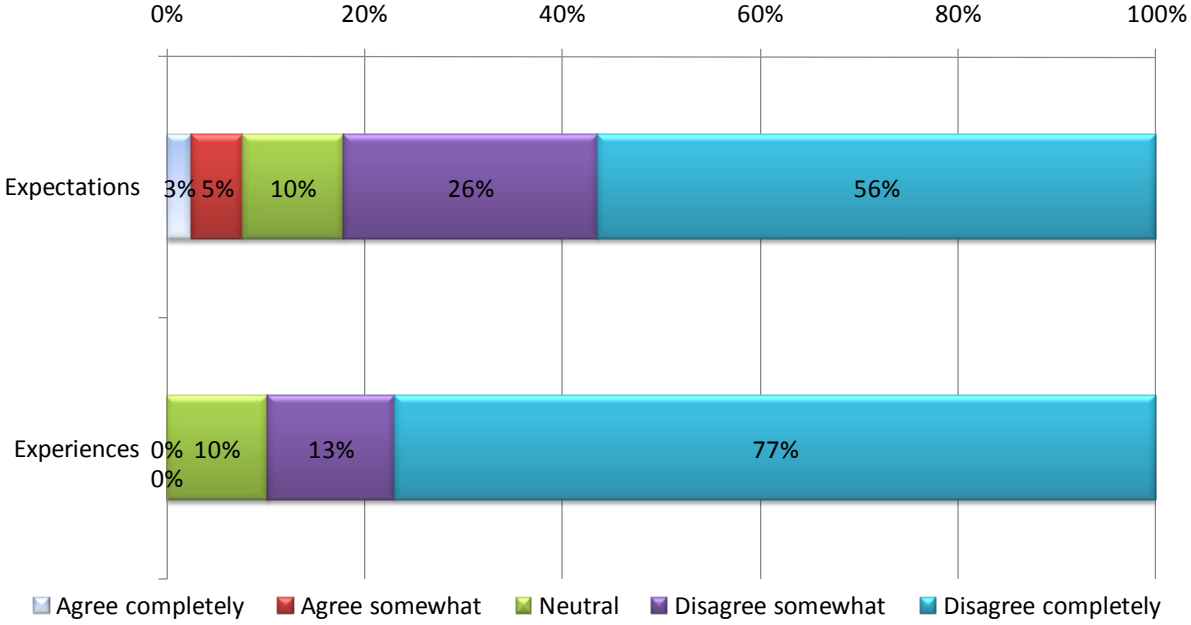
11. With respect to completeness (i.e., that all required information is included) medical records will [have turned out to] …
[N=39; p<0.01, Wilcoxon signed rank]



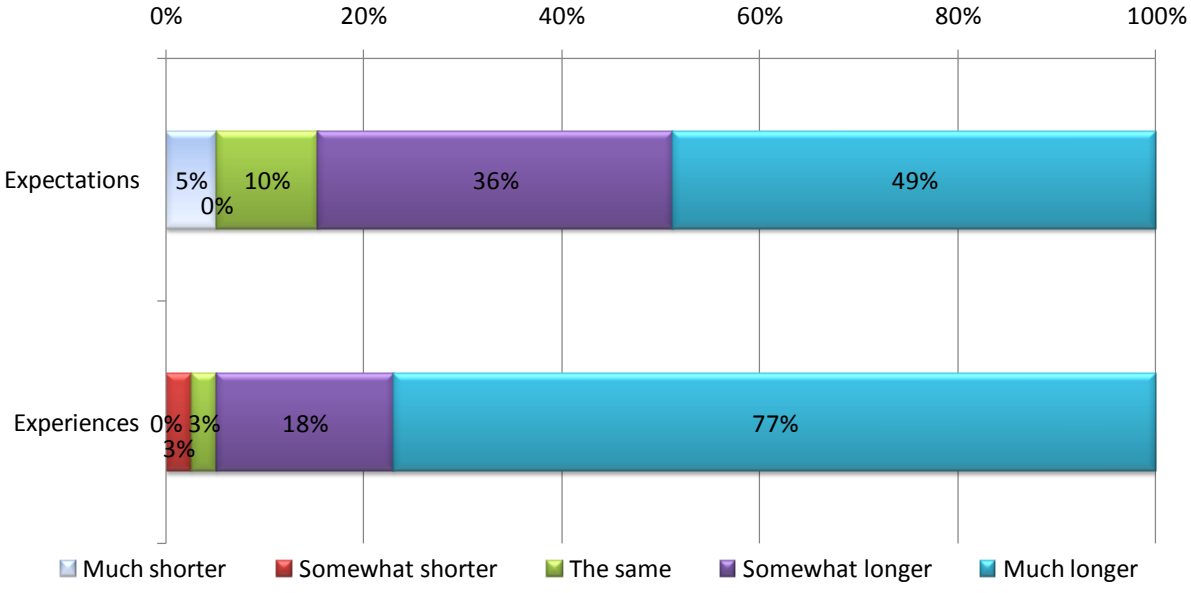
12. I expect that speech recognition will optimize [Speech recognition has optimized] the process of keeping the medical record
[N=39, p<0.005, Wilcoxon signed rank]



13. I expect speech recognition to produce [Speech recognition has produced] appreciable time savings for the benefit of patient care [N=39; p<0.06, Wilcoxon signed rank]



14. I expect that the time I spend on producing medical records in the long run will become [The time I spend on producing medical records has become] … [N=39; p<0.01, Wilcoxon signed rank]



Transition 5

1 Blurring effect of the new work procedures

In the previous paper [Alapetite, Andersen, Hertzum 2007], the fact that the speech recognition system was introduced simultaneously with new work procedures made it difficult to distinguish the qualities and drawbacks of the speech recognition engine, from the effect of the new work procedures that have changed some roles and responsibilities in the workflow of the production of medical records.

From a survey point of view, it would have been more interesting to either first change the work procedures before introducing the new speech recognition system, or *vice versa*. However, this approach is unfortunately impractical.

In particular, trying to introduce a speech recognition system before changing the work procedures, that is to say while keeping the traditional workflow (using a Dictaphone followed by a transcription done by secretaries), has repetitively been reported as a wrong approach, with examples of failure [Hvidberg 2003]. The main reason is the fact that current speech recognition systems still need users to comply with a somewhat artificial way of speaking if speakers want to achieve high recognition rates; hence, if users receive feedback and correct the misrecognitions themselves, both users and the speech recognition system may improve their interaction [Mohr *et al.* 2003]. Another reason is that it is sometimes more time consuming for secretaries to correct the text output from a speech recognition engine than to transcribe the whole text from an audio record.

For the experiment, the ideal case would probably have been to introduce the speech recognition system after having changed the work procedures, which means to compare the performance of the new speech recognition system with a system using a similar workflow, such as physicians typing the medical records with a keyboard. It was however unrealistic to ask the physicians to have a phase with keyboard only just to get a reference to evaluate the speech recognition system. Nevertheless, it should be noted that a few physicians (possibly computer skilled) reported not taking advantage of the speech recognition facilities by using only the keyboard.

2 Recommendations for deploying similar systems in the future

As this survey has shown, only about a third of the physicians in this hospital have perceived the deployment of this speech recognition system positively. Furthermore, it appeared difficult to measure the objective effect of speech recognition, as the speech contribution rates provided are not ideal, responses to the questionnaires are subjective, and the introduction of speech recognition coincided with a modification of the work procedures.

The lack of objective and robust data is problematic, not only for the managers, but also for the physicians – the users of the system – as they have shown in their free comments to be receptive to seriously established facts. Those facts in turn would influence their perceived usefulness of the system and likely the performances of the speech recognition, which is known to perform better with users that are positive towards this technology.

Therefore, I would recommend to managers to have an intermediary test step during the deployment of the system to build clear and objective data, including financial considerations and benefits for the patients. This period should last more than 4 months, as speech contribution rates are improving quickly at the beginning, and for the same reason, the first 2 months should not be taken into consideration. At least 10 participants (the statistical power should be evaluated) have to be involved in order to obtain reliable results, due to a high variability between users, especially during the first months.

One solution would be to use some secretary resources during a couple of months to make a quality survey: based on the audio recording of the dictation as an additional source of information, they could review the quality of the documents produced by the physicians with the speech recognition system.

3 Improvement ideas

Inspired by the literature (*e.g.* [Lai & Vergo 1997]), the results of the previous experiments and the feedback received in the above survey, a few improvement ideas come to mind.

The first one is a suggestion, yet to be experimented, to use "XXX" signs just as human transcriptionists do when they cannot understand what has been said with a high enough confidence. This could be supplemented by drop-down lists of alternatives. This would be an alternative to the current typical behaviour of speech recognition systems in free text mode which is indeed to guess as soon there is something audible, thus resulting in writing potentially dangerous text, with errors harder to spot and correct. More advanced feedback and user interfaces to this specific problem are proposed in e.g. [Lai & Vergo 1999].

The second suggestion is to propose a redundant multimodal user interface with enough vocal commands to give the user the possibility to choose between voice, keyboard and mouse for most actions besides dictations (*e.g.* corrections, file handling, start and stop dictation). This has already been partially proposed by most speech input systems. Although some studies have reported that it typically takes longer time to perform those aside actions with voice commands than with mouse and keyboard [Sears *et al.* 2003], this does not take into account the annoyance of switching to mouse and keyboard when the user does not want or cannot do so. Furthermore, it has been observed that given a redundant multimodal interface, users tend to "find a pattern of modality usage and stick with it" [Lai & Vergo 1997] for a given scenario. This pattern of usage is likely to be different from user to user, thus offering a redundant multimodal interface may therefore satisfy a wider population.

4 Natural languages

Regarding the reported poor speech recognition rates for specific parts of the natural language (*e.g.* negations), while this depends of course on the quality of the speech recognition engine, a major factor is actually the type of natural language to be recognised, as detailed in [Alapetite 2006]. For instance, the negation in a Danish sentence is typically expressed by one word "ikke" /ˈegə/ (similar to "not" in English) often shortened to /ˈeg/, while it is typically expressed with two words in French at two different locations in a sentence "ne … pas" /nə … pa/ thus providing redundancy on the crucial part and a higher likelihood to avoid this type of misunderstandings. This was to exemplify that each natural language is subject to different issues with speech recognition and therefore some might by essence perform better than others, in a given context.

5 Follow up survey

After the mitigated acceptance of the speech recognition system in Vejle hospital reported in the previous paper [Alapetite, Andersen, Hertzum 2007], there were some doubts that the physicians responding to our survey were not entirely objective in estimating the impact of the new work procedure involving speech recognition on the quality of the produced medical records.

Therefore, a dedicated quality survey was needed [Andersen, Alapetite *et al.* 2007], to blindly compare "within subjects" the quality of the documents produced with the traditional system using Dictaphones and transcriptions by secretaries, and documents produced with the new work procedure where the physicians are producing the documents themselves directly on a computer, with the possible help of speech recognition.

References

[Hvidberg 2003] Jens Hvidberg. Talegenkendelse - muligheder og barrierer for anvendelse til klinisk dokumentation (In Danish). *Master's thesis, Aalborg University, May 2003*.

[Lai & Vergo 1997] Jennifer Ceil Lai & John George Vergo. MedSpeak: Report creation with continuous speech recognition. *In: Proceedings of the CHI'1997 Conference on Human Factors in Computing Systems. ACM Press, 1997, New York, pp. 431-438*. ISBN:0-89791-802-9

[Lai & Vergo 1999] Jennifer Ceil Lai & John George Vergo. Speech recognition confidence level display. *US Patent 6,006,183, 1999*.

[Mohr *et al.* 2003] David N. Mohr, David W. Turner, Gregory R. Pond, Joseph S. Kamath, Kathy B. De Vos, Paul C. Carpenter. Speech Recognition as a Transcription Aid: A Randomized Comparison With Standard Transcription. *Journal of the American Medical Informatics Association, 2003, 10(1):85–93*. doi:10.1197/jamia.M1130

[Sears *et al.* 2003] Andrew Sears, Jinjuan Feng, Kwesi Oseitutu, Clare-Marie Karat. Hands-Free, Speech-Based Navigation During Dictation: Difficulties, Consequences, and Solutions. *Human-computer interaction, 2003, 18:229–257*. doi:10.1207/S15327051HCI1803_2

Blinded comparison of quality of medical records produced with speech recognition or traditional dictation and transcription

Henning Boje Andersen[1], Alexandre Alapetite[1,2], Peter Ivan Andersen[3], Aase Andreasen[3], Per Hølmer[4], Stig Jørring[4], Claus Varnum[3]

1. Systems Analysis Department; Risø National Laboratory; Technical University of Denmark; DK-4000 Roskilde; Denmark
2. Computer Science, Roskilde University, Universitetsvej 1; P.O. Box 260; DK-4000 Roskilde, Denmark
3. Vejle and Give Hospital; Kabbeltoft 25; DK-7100 Vejle; Denmark
4. Orthopaedic surgery; Hillerød Hospital; Helsevej 2; DK-3400 Hillerød, Denmark

This section is a draft of the following journal article:

Henning Boje Andersen, Alexandre Alapetite, Peter Ivan Andersen, Aase Andreasen, Per Hølmer, Stig Jørring, Claus Varnum. Blinded comparison of quality of medical records produced with speech recognition or traditional dictation and transcription. *To be submitted, 2007.*

Keywords

Electronic medical record; Electronic patient record; speech recognition; transcription; quality

1 Introduction

The task of creating and maintaining patient medical records is crucial to patient care. At the same time, it is a time consuming and largely uninteresting chore for physicians. In many and perhaps most countries and hospitals the entering of notes into the medical record is traditionally carried out as a sequence beginning with a physician's dictation (into a sound recording device) of the note to be entered, then a transcription of this by a secretary or some specialised transcription service, and finally the approval by the physician of the transcribed text. The transcription service can range from specialized internal medical secretaries to general purpose external companies. In order to reduce the costs and turnaround time, speech recognition has been recommended has a potential alternative (e.g. [Zick & Olsen 2001]).

Vejle and Give Hospital in Denmark has been the first Danish hospital to introduce speech recognition for all major specialties and departments. Since 2000, the hospital (349 beds and 217 000 outpatients in 2006) has run a successful project on speech recognition technology in its radiology department, and in 2004 it laid out plans for having all physicians in all clinical departments use speech recognition to input physician notes and instructions into the electronic medical record (EMR). The speech recognition system (software based on Philips Speech Magic, adapted to Danish and deployed by Max Manus A/S) was rolled out in all clinical departments in 2005-2006. It currently (2007) has about 240 physician users, including non-clinical departments

The traditional work practices involved physicians dictating to an audio recorder their notes for the EMR, handing over the tapes or files to medical secretaries in each of their respective departments. In contrast, the new work procedures require physicians to generate on a computer terminal the EMR notes themselves with the (potential) help of speech recognition system. The intention behind the introduction of speech recognition was primarily to reduce the need of secretaries for transcription, to optimize workflows by reducing the time it takes for producing EMR notes. However, it was also argued that the new work process would be expected to improve the quality of the generated documents by requiring physicians to correct and approve their own dictations when they still have the case in mind.

A previous survey [Alapetite, Andersen, Hertzum 2007] – involving some of the authors of this article – has taken the chance to follow the deployment of speech recognition at Vejle and Give Hospital in 2006, to find some factors affecting the acceptation and the success of the integration of the speech recognition system and associated new work procedures. The survey consisted of two questionnaires sent to the physicians of the departments moving to speech recognition, a few weeks before and then about 4 months after the actual introduction of speech recognition.

The results of the survey have shown that *a posteriori*, physicians (N = 98) were equally distributed between those who think it was a good idea (33%), neutral, and those who do not think it was a good idea (31%) to introduce speech recognition in their department. The survey also asked about the subjective opinion of the physicians on the evolution of the quality of the patient records:

- 15% reported an improvement of the general quality while 62% reported a deterioration;
- 16% agreed that the precision has increased versus 43% who rather agreed that the precision has decreased;
- 20% thought the records have become better structured while 22% thought they have become less structured;

- 11% believe the records have increased in precision versus 60% who believe the prevision has decreased.

Since this subjective data suggest a reduction in the quality of the records, in opposition with the initial hypothesis envisaging an improvement, a follow up survey was designed to collect objective data. This is what is reported in the current article. Due to the limited resources to select and extract the records, and even more the limited available time for the two expert judges with have used to assess the quality, the study limits itself to seven physicians in one department, namely orthopaedic surgery, for a total of 74 records produced with traditional transcription and the same number produced by speech recognition.

2 Related work

A number of studies have been published that focus on comparing the speech recognition rates between various speech recognition engines [Devine *et al.* 2000], word error rates between human transcription and speech recognition [Zafar *et al.* 2004], or total time spent by physicians when using traditional transcription or speech recognition. For instance, [Zick & Olsen 2001] found on the one hand similar error rates in English between speech recognition (Dragon NaturallySpeaking Medical suite version 4) and their traditional transcription service (external, contacted by telephone) with an "accuracy" of respectively 98.5% and 99.7%, and an average of 2.5 "corrections" needed per chart produced by speech recognition versus 1.2 for the traditional transcription. On the other hand, [Zick & Olsen 2001] found a much shorter turnaround time for speech recognition (3.7 minutes in average) than for traditional transcription (39.6 minutes). Similarly, [Borowitz 2001] found that the time spent by a physician dictating and editing notes was 16% higher using speech recognition (IBM ViaVoice Millennium) than traditional transcription. Other studies [Lai & Vergo 1997] suggest that this additional time may be accepted by physicians when the reduction in report turnaround time is significant and interesting, and when work procedures and interface design are adapted to compensate the inherent imperfections of current speech recognition systems.

However, no studies, apparently, have been reported that compare the respective final quality after correction of the documents produced by the traditional dictation system versus the ones entirely produced by the physicians with a possible use of speech recognition. Finally, due to the specificities of the Danish language and the relative youthfulness of its implementation in speech recognition systems, it has been argued that results for the English language may not be directly transferable [Alapetite 2006].

3 Method and materials

The overall design of this study involved a blinded review of medical records that were pair-wise selected so that each pair consisted of two records dictated by the same physician, one record produced with automatic speech recognition (ASR) and the other with traditional secretarial transcription, and so that the records were comparable (length, same type of record, diagnosis or patient visit). Records were to be presented in random order to two independent reviewers who, first, should review all records independently, and second, in a consensus review discus their findings and their ratings of these. Based on pilot trial it was determined that quality would be judged a "dictation error" scale and a "transcription error" scale, each scale being ordinal and going from 0 to 5.

3.1 Selection of the medical record

Medical record dictations were drawn from Vejle and Give hospital's department of orthopaedic surgery (12 senior physicians and, on average, 5-6 physicians in training). A random, balanced selection of dictations was made from medical records on a pair-wise basis. The selection ensured that each pair of records was comparable, as just described, and so that the only systematic difference was the production method (speech recognition [ASR] versus traditional transcription performed with the "Dicom" system, a recording and playback dictation tool). The quality of the pairing process was partially assessed by doing some correlations between the error levels of Dicom and ASR for records in free format, fixed format, and the two combined.

Another distinction applied in the selection was the distribution of records into two types: records that followed a more or less "fixed format" (description of operation, knee, hip, etc.) and records that concerned a "free format", having a less repetitive or fixed pattern (description of patient complaints, anamnesis, etc.) The reason for making a deliberate distinction and selection of both "fixed-format" and "free-format" records was a hypothesis that the former, following a more recurrent pattern of expressions, might require fewer corrections of speech recognition errors. See in Appendix for an example of record.

Table 1: Distribution of the selected records.

	Dicom	ASR	Total
Free format	4 physicians × 9 records = 36	4 physicians × 9 records = 36	2 modes × 4 physicians × 9 records = 72
Fixed format	3 physicians × 12 records = 36	3 physicians × 12 records = 36	2 modes × 3 physicians × 12 records = 72
Total	72	72	144

To compute the number of pairs of records required for each of the two types (free and fixed format), a power analysis was carried out with respect to a subsequent test of significance of possible differences with the Wilcoxon paired rank sum test. The power analysis (assumptions: alpha level of 0.05; approximately normal distributed data on a normalised rank-based scale; the smallest difference to be detected 1/2 SD; 80% chance of detecting a difference) showed that less than 30 pairs would be required on the assumptions stated. Using a more conservative choice, it was determined that 36 pairs should be reviewed for each type of record.

3.2 Quality rating

Once the medical records selected and the pairing done between Dicom and ASR records, a code was assigned to each record, and they were then randomized. Any sign that could tell if the record was produced thru Dicom or ASR was erased.

Two reviewers were assigned the mission to review and rate independently the 144 records. The reviewers (senior consultants, authors 5 and 6) belong to different hospital and have no involvement with the speech recognition at Vejle and Give Hospital. Having reviewed and rated the record independently, the two reviewers and a moderator would meet to perform a consensus review.[1]

Two scales, a transcription scale and a dictation scale, were defined to capture reviewers' ratings, each scale going from zero (no error) to five (serious error) as detailed in Table 2. Points above zero on the transcription scale were meant to reflect errors made by the secretary or the speech recognition system. Similarly, points above zero on the dictation scale were meant to reflect errors made by the physician during dictation. The distinction between the two scales is clear in principle: a given error or inadequacy is a transcription error if it is more likely than unlikely that what the record says is not what the physician has said during dictation. Conversely, it is a dictation error if it is more likely than unlikely that what the record says is what the physician has said. Due to the fact that the original audio dictations were not available, it was not possible to use a precise error classification system such as proposed in [Zafar *et al.* 2004]. The classification in "dictation" or "transcription" error was therefore slightly subjective but in most cases simple to assess for the reviewers.

[1] The individual ratings of reviewers and the analysis (kappa statistics) is not discussed in the draft of this paper.

Table 2: Description of error scales

Rating	Transcription scale (meaning of rating):	Dictation scale (meaning of rating):
0	No error / inadequacy	No error or inadequacy
1	Small, inconsequential error, including linguistic error (plural/sing.), anything that slows down reading (commas separating phrases)	Small, inconsequential error, typically sloppy dictation, omitting left/right, naming joint, but where context makes it clear what is meant
2	Somewhat larger error (or three small errors)	Somewhat larger error (or three small errors)
3	Larger error - but it is still easy to guess what was said	Larger error - but it is still easy to guess what was meant. This is an error which any physician would correct if he/she spots it and has time and opportunity to do so
4	Serious error, though still possible with some uncertainty to guess or infer what the dictating physician has said	Serious error, though still possible with some uncertainty to guess or infer what the dictating physician wanted to say
5	Serious error: the phrase or the lacuna is meaningless and it is not possible to guess what the dictating physician has said	Serious error: the phrase or the lacuna is meaningless and it is not possible to guess what the dictating physician wanted to say

3.3 Statistical analysis

In order to statistically compare the respective error levels of the records produced by the traditional Dicom and the newer ASR, the Wilcoxon matched pairs – a non-parametric test – was used, taking advantage of the pairing process described above. This test may tell if there is any statistically significant difference between the error levels of Dicom and ASR.

Another set of analysis was done without keeping the pairs and therefore using another independent-samples non-parametric test, namely Man-Whitney U. This test was used to corroborate the findings of the Wilcoxon matched pairs, when the correlation between the two parts of the pairs (Dicom and ASR) was week. This test was also used for statistics outside the pairs, regarding differences of error levels between free and fixed formats.

4 Results

Table 3 reports the number of records in each error level (from level zero with no error, to level five with severe errors) in the dictation and transcription parts, for Dicom and ASR modes.

Table 3: Error level comparison between Dicom and ASR.

			Number of records with each error level								Wilcoxon matched pairs
			0's	1's	2's	3's	4's	5's	mean	σ	
All formats (N=144)	Dictation (ρ = 0.4; p > 0.75)	Dicom	63	3	2	2	2	0	0.29	0.88	Z = 0.33 p > 0.74
		ASR	61	6	5	0	0	0	0.22	0.56	
	Transcription (ρ = 0.26; p* < 0.03)	Dicom	48	16	4	2	1	1	0.54	0.99	Z = -2.43 p* < 0.02
		ASR	37	20	4	1	3	7	1.08	1.61	

ρ is the Pearson correlation between the two categories (Dicom and ASR).
σ is the standard deviation. p is the statistical significance; * at 0.05 level; ** at 0.01 level.

As shown in Table 3, there is no significant difference between the quality of the dictations with Dicom and ASR ($p > 0.74$, Wilcoxon matched pairs). As there is a low paired samples correlation between the ratings of Dicom and speech recognition ($\rho = 0.4$, $p > 0.75$, Pearson), it is prudent to verify the results with an independent-samples test, which actually provides a nearly identical outcome ($p > 0.73$, Mann-Whitney U).

In contrast, there is a significant difference in the quality of the transcriptions in favour of Dicom over ASR ($p < 0.02$, Wilcoxon matched pairs). Since there is a good paired samples correlation between the ratings of Dicom and speech recognition ($\rho = 0.26$, $p < 0.03$, Pearson), the pairing is therefore satisfying for the matched pairs statistics.

The next table, Table 4, provides the same analysis as Table 3, but detailed for the free format and fixed formats. Similarly to the first part, the Wilcoxon matched pairs test was verified by a Mann-Whitney U test when the paired samples correlation between Dicom and ASR was too weak. The results between the two tests were always showing the same trends.

As visible in Table 4, there is a significant difference in the quality of the transcriptions in favour of Dicom over ASR ($p < 0.01$, Wilcoxon matched pairs) for the free format, but not significant for the fixed format. Differences of error levels for dictation are not significant for both free format and fixed format.

Table 4: Error level comparison between Dicom and ASR detailed for free and fixed formats.

			Number of records with each error level								Wilcoxon matched pairs
			0's	1's	2's	3's	4's	5's	mean	σ	
Free format (N=72)	Dictation (ρ = 0.37; p* < 0.03)	Dicom	31	2	2	1	0	0	0.25	0.70	Z = 0.72 p > 0.47
		ASR	32	2	2	0	0	0	0.17	0.51	
	Transcription (ρ = -0.10; p > 0.57)	Dicom	31	3	1	1	0	0	0.22	0.64	Z = -2.59 p** < 0.01
		ASR	18	13	2	0	2	1	0.83	1.23	
Fixed format (N=72)	Dictation (ρ = -0.15; p > 0.38)	Dicom	32	1	0	1	2	0	0.33	1.04	Z = 0 p = 1
		ASR	29	4	3	0	0	0	0.28	0.62	
	Transcription (ρ = 0.34; p* < 0.042)	Dicom	17	13	3	1	1	1	0.86	1.18	Z = -0.96 p > 0.33
		ASR	19	7	2	1	1	6	1.33	1.90	

Since a difference has been identified between free and fixed formats, a closer look is taken and Table 5 shows the differences between the two formats.

Table 5: Error level comparison between free and fixed formats.

			mean	σ	Man-Whitney U
All modes (Dicom + ASR) (N=144)	Dictation	Free format	0.21	0.60	Z = -0.51 p > 0.60
		Fixed format	0.31	0.85	
	Transcription	Free format	0.53	1.02	Z = -2.45 p* < 0.015
		Fixed format	1.10	1.59	

While there is still no significant difference between the error levels of the dictations in free or fixed format, the error level of the transcriptions is significantly higher for the fixed format than the free format ($p < 0.0015$, Man-Whitney U).

In order to tell if this difference of error levels in the transcriptions is due to a variation of performance of Dicom and/or ASR between free format and fixed format, a detailed analysis is reported in Table 6. This show that the only significant difference in the error levels between free and fixed format are for the transcriptions produced by Dicom ($p < 0.001$, Man-Whitney U).

Table 6: Error level comparison between free and fixed formats, detailed for Dicom and ASR.

			mean	σ	Man-Whitney U
Dicom (N=72)	Dictation	Free format	0.25	0.70	Z = -0.25
		Fixed format	0.33	1.04	p > 0.80
	Transcription	Free format	0.22	0.64	Z = -3.37
		Fixed format	0.86	1.18	p** < 0.001
ASR (N=72)	Dictation	Free format	0.17	0.51	Z = -0.95
		Fixed format	0.28	0.62	p > 0.33
	Transcription	Free format	0.83	1.23	Z = -0.47
		Fixed format	1.33	1.90	p > 0.63

5 Discussion

The result show a general tendency of a reduced transcription quality when using speech recognition (ASR) than using the traditional method based on audio recording of a dictation followed by secretary-based transcription (Dicom). The transcription method, *i.e.* Dicom or ASR, does not seem to have a significant impact on the dictation error level.

The transcription errors identified by reviewers were in most cases trivial and the intended meaning easy to guess. However, a small number of meaningless expressions (transcription errors) were included. For instance, in a couple of cases a meaningless word was added to the end of a sentence (the word “fremragende”, “excellent” in English, appeared unexpectedly).

It has also turned out that the higher number of errors of speech recognition records is restricted to the free format records, that is to say a type of dictations with *ad hoc*, variable and possibly long phrases with a rich vocabulary. Indeed, in the case of fixed format dictations, even if the speech recognition seem to produce slightly more errors than the traditional transcription, the difference is not significant. The difference of performance between Dicom and ASR in fixed format seems mainly due to secretaries making more transcription errors for fixed format, while there is no significant difference between fixed and free format for speech recognition.

It is somehow surprising to find that fixed format has a higher average error level than free format, although the difference is only significant for Dicom.

Conclusion

The results show that the error level is overall significantly higher for transcriptions produced by speech recognition than secretaries, hence confirming the results from a previous survey [Alapetite, Andersen, Hertzum 2007] made in the same hospital.

However, the quality of the transcriptions produced by speech recognition being lower than those produced by secretaries in the case of the free format, but comparable in the case of fixed format suggest that in this hospital setup, speech recognition should be privileged for some precise tasks.

This is in agreement with some comments from physicians from the same hospital who answered the survey [Alapetite, Andersen, Hertzum 2007] and argued for the use speech recognition in some specific cases, that is for urgent, medium to long documents, with short typical sentences, in a department with limited background noise.

References

[Alapetite 2006] Alexandre Alapetite. Impact of noise and other factors on speech recognition in anaesthesia. *International Journal of Medical Informatics, available online in December 2006*. doi:10.1016/j.ijmedinf.2006.11.007

[Alapetite, Andersen, Hertzum 2007] Alexandre Alapetite, Henning Boje Andersen, Morten Hertzum. Acceptance of Speech Recognition by Physicians: A Survey of Expectations, Experiences, and Social Influence. *Submitted to the International Journal of Human-Computer Studies, 2007*.

[Al-Aynati & Chorneyko 2003] Maamoun M. Al-Aynati & Katherine A. Chorneyko. Comparison of Voice-Automated Transcription and Human Transcription in Generating Pathology Reports. *Archives of Pathology and Laboratory Medicine, 2003, 127(6):721–725*.

[Borowitz 2001] Stephen M. Borowitz. Computer-based speech recognition as an alternative to medical transcription. *Journal of the American Medical Informatics Association, 2001, 8(1):101-102*.

[Devine *et al.* 2000] Eric G. Devine, Stephan A. Gaehde, Arthur C. Curtis. Comparative Evaluation of Three Continuous Speech Recognition Software Packages in the Generation of Medical Reports. *Journal of the American Medical Informatics Association. 2000,7(5):462–468*.

[Henricks *et al.* 2002] Walter H. Henricks, Kavous Roumina, Bradley E. Skilton, Debra J. Ozan, Gwendolyn R. Goss. The Utility and Cost Effectiveness of Voice Recognition Technology in Surgical Pathology. *Modern Pathology, May 2002, 15(5):565-71*. doi:10.1038/modpathol.3880564

[Lai & Vergo 1997] Jennifer Ceil Lai & John George Vergo. MedSpeak: Report creation with continuous speech recognition. *In: Proceedings of the CHI'1997 Conference on Human Factors in Computing Systems. ACM Press, 1997, New York, pp. 431-438*. ISBN:0-89791-802-9

[Zafar *et al.* 2004] Atif Zafar, Burke Mamlin, Susan Perkins, Anne M. Belsito, J. Marc Overhage, Clement J. McDonald. A simple error classification system for understanding sources of error in automatic speech recognition and human transcription. *International Journal of Medical Informatics, 2004, 73:719-730*. doi:10.1016/j.ijmedinf.2004.05.008

[Zick & Olsen 2001] Robert G. Zick, Jon Olsen. Voice Recognition Software Versus a Traditional Transcription Service for Physician Charting in the ED. *American Journal of Emergency Medicine, July 2001, 19(4):295-298*. doi:10.1053/ajem.2001.24487

Appendix

Example of record in free format from Vejle and Give hospital, produced by speech recognition:

NOTAT 271106 11.09
1(1)

Ortopædkirurgisk sektion
Ortopædkirurgisk afd
Kabbeltoft 25/Tykhøjetvej 4-6
7100 Vejle/Give

124153-1964

N 12

OP/UNDERSØGELSE

[redacted]

INDIKATION Manglende heling i midtskafts højresidig humerusfraktur.
OP.STUENR. Operationsstue 22.
OP.TIDSP. START/SLUT
Start kl. 10.55 slut kl. 12.20.
OPERATØR/ASSISTENT
Operatø[redacted] assistent [redacted]
ANÆSTESI UA.
OP/US.(KODE) KNBJ51 Intern fiksation m. marvsøm af humerusfraktur
PATOLOGI Midtskaft humerusfraktur med manglende heling, men antydning af hypertrofi og callusdannelse specielt omkring distale fragment.
PROCEDURE Bugleje. Incision fra olecranon spidsen og 8 cm proksimalt. Tricepssenen deles med diatermi og fossa olecrani lægges fri. Ved hjælp af sigteapparat åbnes ind til knoglemarhulen og åbningen udvides med konisk bor og hulafbider. Distale fragment reames op med håndreamer. Man forsøger at få leder igennem til proksimale fragment. Dette lykkes ikke, efter en del manipulation lykkes det at få åbnet til proksimale fragment med reamer, herefter reames op i proksimale fragment. Herefter indføres 9 mm tykt 260 mm lang T2 humerus søm. Kommer glat igennem til lige under caput humeri.
Ved hjælp af sigteapparatur låses distalt med 2 statiske skruer. Vejledt af gennemlysning 2 statiske låseskruer proksimalt og der er herefter fuld stabilitet og rimelig kontakt. Inden indføring af sømmet har man lagt en del knoglesmuld op omkring frakturen knoglesmuld at høste i forbindelse med åbning af knoglen. Til slut låse lukkes sømmet med end cap.
Låseskruer indført gennem separate stikincisioner.
Incision lige over albuen lukkes med Vicryl i tricepssenen Vicryl i subcutis metalklips i hud.
Stikincisioner lukkes med metalklips.
PLAN Røntgenkontrol er bestilt Kan mobiliseres med fri bevægelighed fra i morgen. Skal nok anvende collar and cuff de første 10-14 dage. Fjernelse af metalklips hos egen læge om 2 uger, skal inden udskrivelsen instrueres af fysioterapeut i ubelastede bevægeøvelser. Ambulant røntgenkontrol hos operatøren om ca. 4 uger kan der begynde at bevæge yderligere mod modstand og muligvis belaste. Røntgenkontrol efter 112 uger.
Der er i tilslutning til operationen givet Zinacef 1,5 g intravenøst, der skal ikke gives yderligere antibiotika.

General conclusion

In the domain of human–computer interaction, thanks to improvements in processing power and algorithms since the 1950's, speech recognition performance has been increasing spectacularly, nowadays offering accurate recognition in many languages and for many applications, including safety critical systems.

During the same period, in the medical domain, anaesthesia has made tremendous progress in terms of safety and quality of health care. To a large extent, this is due to improvements in the monitoring devices and work procedures. Recently, the introduction of electronic anaesthesia records has further contributed to this evolution, by solving many of the issues associated with the former paper based generation, although not without creating a smaller set of new problems such as ergonomic issues.

During discussions at Risø at the crossing of the two above-mentioned domains germinated the seminal idea of using speech recognition to enhance electronic anaesthesia records.

The aim of the thesis was thus to experiment with speech recognition in the anaesthesia domain to try to tell if this is a relevant way to go, and if yes, under which conditions.

1 The research question

The establishment of the precise research question was an iterative process that was done during the first months of the PhD studies, inspired by substantial literature reviewing, in interaction with some partners of this project, and further refined by interviews with various people. The exact topic only reached a stable status after the first conference, EACE'2005 (*cf.* Transition 2), then allowing a progression in a quite linear and logical manner.

The research sub-questions are taken from calls published by authors of related attempts. In particular, one goal was to study human behaviour when using speech recognition during real time experiments [Smith *et al.* 1990]. A related goal was to identify in which task areas speech recognition could be most useful, in particular when considering ergonomic concerns [Jungk *et al.* 2000], vigilance and contact with the patient [Sanjo *et al.* 1999].

In order to focus and to limit the parameter space, a hypothesis is made, namely that speech recognition may provide valuable gains in the case of busy anaesthesia situations, with regard to the quality of the recording and contact with the patient.

2 Recapitulation

The first paper [Alapetite & Gauthereau 2005] provides a reflexion, from a sociological viewpoint, on the possible short-term and long-term consequences – both positive and negative – of introducing speech recognition in the anaesthesia theatre.

The second paper [Alapetite 2006] attempts to identify the boundaries of speech recognition during anaesthesia, with respect to the background noise and its direct impact on the audio channel and indirect impact on speech recognition by altering human behaviour. In particular, it shows that background noises can be overcome by properly choosing the type of microphone and phraseology.

The third paper [Alapetite 2007] reports findings from a simulation experiment aimed at validating the above-mentioned hypothesis and at quantifying the gains offered by speech recognition during a busy anaesthesia scenario. Beside, the "average queue of events" metric is proposed as a convenient way of measuring the workload in the case of simultaneous tasks. Although not mature enough to envision a real use outside a simulator, the prototype that has been developed and tested confirms the potential of speech recognition during busy anaesthesia situations, by reducing significantly mental workload, delays before registration, inaccuracies and lacks, and by increasing visual contact with the patient.

The fourth paper [Alapetite, Andersen, Hertzum 2007] takes the opportunity to study the deployment, acceptance and success of a speech recognition solution – technologically similar to the one used in the former experiments – in a hospital, with a focus on the human factors. One of the main findings is that the general *a priori* opinion of the users about speech recognition is positively correlated with their opinion *a posteriori*. Another result is that the relatively low user acceptance was to a large extent due to the new work procedures that were not entirely adapted adequately to the new workflow involving speech recognition.

The fifth paper [Andersen, Alapetite *et al.* 2007] finally quantifies the impact on the quality of some medical records of new work procedures relying on a speech recognition system when compared with the traditional secretary-based transcriptions. While the general quality tends to be lower with speech recognition, this depends very much on the type of record.

3 Conclusion

The conclusions of the last article [Andersen, Alapetite *et al.* 2007] are consistent with the previous article [Alapetite, Andersen, Hertzum 2007] and to some extent similar to the conclusions of the whole thesis. Even after some years of improvement, in office environment and for transcription tasks, speech recognition remains best suited for non-typists [Grasso 1995], at least considering the specific system in Danish used during the various experiments. Speech recognition has a number of clear advantages in some specific areas and situations, in particular when near real time information must be entered into a computer, when entering this data is a secondary task, or when competing input devices are not well suited such as when the user is moving around or needs free hands to *e.g.* take care of a patient.

The thesis addressed a substantial part of the original research question, in particular with some clear results about the gains provided by speech recognition during busy anaesthesia situations in terms of anaesthesia record quality and time available for the patient. More generally, the thesis provides some information that may help to tell when speech recognition is a good choice or not, and when it is likely to succeed or not.

Regarding the larger research question of the network in which this PhD took place, namely ADVISES, the European research training network about "Analysis Design and Validation of Interactive Safety-critical and Error-tolerant Systems", this thesis fulfilled some of the expectations by investigating in a safety-critical domain, that is anaesthesia, some HCI solutions to improve patient safety. Furthermore, during the experiments reported in [Alapetite 2006], promising results were obtained using a redundant speech recognition architecture to both increase speech recognition accuracy and fault tolerance.

4 Future work

Indisputably, there is a need to further define the role of speech recognition and to improve its integration into existing systems. An eye must be kept on the progress of speech recognition engines, since better accuracy will lead to new perspectives and offer new integration possibilities. Additional work would be needed to properly integrate human factors in the development of speech-enabled interfaces.

I hope the individual articles together with the thesis as a whole will be valuable to the scientific community, hospitals and industrial entrepreneurs interested in providing improved anaesthesia equipments.

Bibliography

[Adams *et al.* 1992] D.A. Adams, R.R. Nelson, P.A. Todd. Perceived usefulness, ease of use, and usage of information technology: A replication. *MIS Quarterly, 1992, 16(2):227-247.* doi:10.2307/249577

[Ajzen 1985] Icek Ajzen. From intentions to actions: A theory of planned behavior. In J. Kuhl and J. Beckmann (eds.), Action Control: From Cognition to Behavior. *Springer, New York, 1985, pp. 11-39.*

[Ajzen 1991] Icek Ajzen. The theory of planned behaviour. *Organizational Behavior and Human Decision Processes, 1991, 50(2): 179-211.* doi:10.1016/0749-5978(91)90020-T

[Alapetite 2004] Alexandre Alapetite. XML en PHP5 avec la bibliothèque interne DOM (XML in PHP5 with the DOM internal library). *Direction|PHP, special issue 1 Tout sur PHP5 (All about PHP5), September 2004, Nexen Services SA (France).* ISSN:1765-2634. http://www.directionphp.biz/a_la_une.php?mois=2004-h1

[Alapetite 2005.a] Alexandre Alapetite. Voice recognition in multimodal systems: the case of anaesthesia patient journal. *In: Proceedings of the first ADVISES Young Researchers Workshop. Hans H.K. Andersen, Asmatullah Nayebkheil (eds.), Risø National Laboratory (DK), Systems Analysis Department. Risø-R-1516(EN),* ISBN:87-550-3443-8*, p. 5-9, April 2005.*

[Alapetite 2005.b] Alexandre Alapetite. Content accessibility of Web documents: Overview of concepts and needed standards. *In: COGAIN (European Network of Excellence) deliverable D6.1 "State of the art report of evaluation methodology", pages 28-34, September 2005. Long version in Risø-R-1576(EN),* ISBN:87-550-3546-9*, Risø National Laboratory, October 2006.*

[Alapetite & Gauthereau 2005] Alexandre Alapetite & Vincent Gauthereau. Introducing vocal modality into electronic anaesthesia record systems: possible effects on work practices in the operating room. *Proceedings of EACE'2005 (Annual Conference of the European Association of Cognitive Ergonomics) 29 September - 1 October 2005, Chania, Crete, Greece; section II on Research and applications in the medical domain, 189–196. ACM International Conference Proceeding Series, vol. 132. University of Athens, 197-204,* ISBN:9-60254-656-5.

[Alapetite 2006] Alexandre Alapetite. Impact of noise and other factors on speech recognition in anaesthesia. *International Journal of Medical Informatics, available online in December 2006.* doi:10.1016/j.ijmedinf.2006.11.007

[Alapetite 2007] Alexandre Alapetite. Speech recognition for the anaesthesia record during crisis scenarios. *International Journal of Medical Informatics, accepted August 2007.* doi:10.1016/j.ijmedinf.2007.08.007

[Alapetite, Andersen, Hertzum 2007] Alexandre Alapetite, Henning Boje Andersen, Morten Hertzum. Acceptance of Speech Recognition by Physicians: A Survey of Expectations, Experiences, and Social Influence. *Submitted to the International Journal of Human-Computer Studies, 2007.*

[Al-Aynati & Chorneyko 2003] Maamoun M. Al-Aynati & Katherine A. Chorneyko. Comparison of Voice-Automated Transcription and Human Transcription in Generating Pathology Reports. *Archives of Pathology and Laboratory Medicine, 2003, 127(6):721–725.*

[Andersen & Hansen 1995] Henning Boje Andersen, John Paulin Hansen. Multi-modal recording and analysis of interaction among operators and work systems. *Proceedings of HMI-AI-AS'1995, the fifth international conference on human-machine interaction and artificial intelligence in aerospace, Toulouse, France, September 27-29 1995. Expended version in Risø-R-939(EN), Risø National Laboratory, Denmark, December 1996*. ISBN:0106-2840.

[Andersen *et al.* 2000] Henning Bøje Andersen, C. R. Pedersen, Hans H. K. Andersen. Using eye tracking data to indicate team situation awareness. *In: Usability evaluation and interface design: Cognitive engineering, intelligent agents and virtual reality. Proceedings of HCI International'2001, International conference on human-computer interaction, 1(9): 1318-1322, New Orleans (LA, USA), 5-10 August 2001. D. Harris, M. J. Smith, G. Salvendy, R. J. Koubek (eds.), Lawrence Erlbaum Associates, Inc., Mahwah (NJ, USA), 2001.* ISBN:0-8058-3609-8

[Andersen *et al.* 2002] Henning Boje Andersen, Marlene Dyrløv Madsen, Niels Hermann, Thomas Schiøler, Doris Østergaard. Reporting adverse events in hospitals: A survey of the views of doctors and nurses on reporting practices and models of reporting. *In: Investigation and reporting of incidents and accidents. Workshop (IRIA 2002), Glasgow (GB), 17-20 June 2002. Chris Johnson (ed.), (University of Glasgow, Department of Computing Science, Glasgow, 2002) (GISTTechnical Report, G2002-2) p. 127-136.*

[Andersen & Andersen 2003] Hans H. K. Andersen & Verner Andersen. Establishing user requirements in HCI - A case-study in medical informatics. *Proceedings of HCI International'2003, International conference on human-computer interaction, Vol. 1, Theory and practice, Part 1, Crete (GR), 22-27 Jun 2003, J. Jacko & C. Stephanidis (eds.), (Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 2003) p. 611-615.*

[Andersen & Alapetite 2006] Henning Boje Andersen & Alexandre Alapetite. See – Sight Effectiveness Enhancement, Results of the Aeronautical Evaluation. *See (European project) deliverable D6.3, December 2005. Risø-R-1573(EN), Risø National Laboratory, November 2006*. ISBN:87-550-3541-8

[Andersen, Alapetite *et al.* 2007] Henning Boje Andersen, Alexandre Alapetite, Peter Ivan Andersen, Aase Andreasen, Per Hølmer, Stig Jørring, Claus Varnum. Blinded comparison of quality of medical records produced with speech recognition or traditional dictation and transcription. *To be submitted, 2007.*

[Barker *et al.* 2005] J.P. Barker, M.P. Cooke, D.P.W. Ellis. Decoding speech in the presence of other sources. *Speech Communication, 2005, 45(1):5-25* . doi:10.1016/j.specom.2004.05.002

[Baskerville 1999] Richard Baskerville. Investigation information systems with action research. *Communications of the Association for Information Systems, Volume2, Article 19, October 1999.*

[Beuscart-Zéphir *et al.* 2001] M.C. Beuscart-Zéphir, F. Anceaux, V. Crinquette, J.M. Renard. Integrating user's activity modelling in the design and assessment of hospital electronic patient records: the example of anesthesia. *International Journal of Medical Informatics*, 2001, 64:157-171. doi:10.1016/S1386-5056(01)00210-6

[Bolt 1980] Richard A. Bolt. "Put-that-there": Voice and gesture at the graphics interface. *Proceedings of the 7th annual conference on Computer graphics and interactive techniques, Seattle, Washington, USA, pp. 262-270*. doi:10.1145/800250.807503

[Borowitz 2001] Stephen M. Borowitz. Computer-based speech recognition as an alternative to medical transcription. *Journal of the American Medical Informatics Association, 2001, 8(1):101-102.*

[Bouchet & Nigay 2004] Jullien Bouchet & Laurence Nigay. ICARE: a component-based approach for the design and development of multimodal interfaces. *CHI'2004 Conference on Human Factors in Computing Systems, pp 1325-1328, Vienna, Austria*. doi:10.1145/985921.986055

[Chopra *et al.* 1992] V. Chopra, J.G. Bovill, J. Spierdijk, Floor Koornneef. Reported significant observations during anaesthesia: a prospective analysis over an 18-month period. *British Journal of Anaesthesia, 1992, 68:13-17.*

[Coniam 1999] D. Coniam. Voice recognition software accuracy with second language speakers of English. *System, 1999, 27(1):49-64*. doi:10.1016/S0346-251X(98)00049-9

[Cowan 2000] Nelson Cowan, The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, Cambridge University Press, 2001, 24:87-114*. doi:10.1017/S0140525X01003922

[Damanpour 1991] Fariborz Damanpour. Organizational innovation: A meta-analysis of effects of determinants and moderators. *Academy of Management Journal, 1991, 34(3):555-590*. doi:10.2307/256406

[Danis & Karat 1995] Catalina Danis & John Karat. Technology-driven design of speech recognition systems. Proceedings of DIS'1995, Symposium on Designing Interactive Systems, pp. 17-24. doi:10.1145/225434.225437

[Davis *et al.* 1952] K. H. Davis, R. Biddulph, S. Balashek. Automatic Recognition of Spoken Digits. Journal of the Acoustical Society of America, November 1952, 24(6):637-642. doi:10.1121/1.1906946

[Davis 1989] F.D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 1989, 13(3):319-340*. doi:10.2307/249008

[Davis 1993] Fred D. Davis. User acceptance of information technology: systems characteristics, user perceptions and behavioral impacts. *International Journal of Man-Machine Studies, 1993, 38(3):475-487*. doi:10.1006/imms.1993.1022

[Detmer *et al.* 1995] William M. Detmer, Smadar Shiffman, Jeremy C. Wyatt, Charles P. Friedman, Christopher D. Lane, Lawrence M. Fagan. A Continuous-speech Interface to a Decision Support System: II. An Evaluation Using a Wizard-of-Oz Experimental Paradigm. *Journal of the American Medical Informatics Association, Jan–Feb 1995, 2(1):46–57.*

[Devine *et al.* 2000] Eric G. Devine, Stephan A. Gaehde, Arthur C. Curtis. Comparative Evaluation of Three Continuous Speech Recognition Software Packages in the Generation of Medical Reports. *Journal of the American Medical Informatics Association. 2000,7(5):462–468.*

[Devos *et al.* 1991] Cathy B. DeVos, Martin D. Abel, John P. Abenstein. An evaluation of an automated anesthesia record keeping system. *Biomedical Sciences Instrumentation, 1991, 27:219-25.*

[Dillon & Norcio 1997] Thomas W. Dillon & A. F. Norcio - User performance and acceptance of a speech-input interface in a health assessment task. International Journal of Human-Computer Studies (1997) 47:591-602

[Dokas & Alapetite 2006] Ioannis Dokas & Alexandre Alapetite. A view on the Web engineering nature of Web based expert systems. *Poster paper in the proceedings of ICSOFT'2006, the 1st International Conference on Software and Data Technologies, 11-14 September 2006, Setubal, Portugal, pages 280-283. A development process meta-model for Web based expert systems: the Web engineering point of view. Long version in Risø-R-1570(EN),* ISBN:87-550-3536-1*, Risø National Laboratory, October 2006*.

[Dybkjær & Dybkjær 2002] Hans Dybkjær & Laila Dybkjær. Experiences from a Danish Spoken Dialogue System. *Proceedings of the 2nd Danish HCI Research Symposium, 7 November 2002, DIKU technical report 02/19, pp. 15-18, Erik Frøkjær & Kasper Hornbæk (Eds.), University of Copenhagen, Denmark,* ISSN:0107-8283.

[Ericsson & Simon 1984] K.A. Ericsson & H.A. Simon. Protocol analysis: Effects of verbalisation. *MIT Press (MA, USA) 1984*.

[Feng & Sears 2004] Jinjuan Feng & Andrew Sears. Using confidence scores to improve hands-free speech based navigation in continuous dictation systems. *ACM Transactions on Computer-Human Interaction, 2004, 11(4):329-356*. doi:10.1145/1035575.1035576

[Fichman & Kemerer 1999] Robert G. Fichman & Chris F. Kemerer. The illusory diffusion of innovation: An examination of assimilation gaps. *Information Systems Research, 1999, 10(3):255-275*.

[Fichman 2000] Robert G. Fichman. The diffusion and assimilation of information technology innovations. In R.W. Zmud (ed.), Framing the Domains of IT Management: Projecting the Future through the Past. *Pinnaflex Educational Resources, 2000, Cincinnati (OH, USA), pp. 105-127*.

[Fiscus 1997] Jonathan G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). *In: Proceedings of IEEE'1997 Workshop Automatic Speech Recognition and Understanding*. doi:10.1109/ASRU.1997.659110

[Fishbein & Ajzen 1975] M. Fishbein, Icek Ajzen. Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research. *Addison-Wesley, 1975, Reading (MA, USA)*.

[Frøkjær *et al.* 2000] Erik Frøkjær, Morten Hertzum, Kasper Hornbæk. Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? *In Proceedings of the CHI'2000 Conference on Human Factors in Computing Systems. ACM Press, 2000, New York, pp. 345-352*. ISBN:1-58113-216-6

[Gaba *et al.* 1988] David M. Gaba, Abe DeAnda. A Comprehensive Anesthesia Simulation Environment: Re-creating the Operating Room for Research and Training. *Anesthesiology, 1988, 69:387-394*.

[Gallivan 2001] Michael J. Gallivan. Organizational adoption and assimilation of complex technological innovations: Development and application of a new model. *ACM SIGMIS Database, 2001, 32(3):51-85*. ISSN:0095-0033

[Gauthereau 2004] Vincent Gauthereau. Emergent Structures in Drug Dispensing to inpatients: implications for patient safety. *Cognition, Technology and Work, 2004, 6(4):223-238.* doi:10.1007/s10111-004-0152-4

[Gelfand & Silman 1979] Stanley A. Gelfand & Shlomo Silman. Effects of small room reverberation upon the recognition of some consonant features. *The Journal of the Acoustical Society of America, July 1979, 66(1):22-29.* doi:10.1121/1.383075

[Giorgino *et al.* 2005] Toni Giorgino, Ivano Azzini, Carla Rognoni, Silvana Quaglini, Mario Stefanelli, Roberto Gretter, Daniele Falavigna. Automated Spoken Dialog System for Hypertensive Patient Home Management. *International Journal of Medical Informatics, 2005, 74(2): 159-167.* doi:10.1016/j.ijmedinf.2004.04.026

[Gong 1995] Yifan Gong. Speech recognition in noisy environments: a survey. *Speech Communication, 1995, 16(3):261-291.* doi:10.1016/0167-6393(94)00059-J

[Grasso 1995] Michael A. Grasso. Automated Speech Recognition in Medical Applications. *M.D. Computing, 1995, 12(1):16-23.*

[Gravenstein 1989] J.S. Gravenstein. The uses of the anesthesia record. *Journal of Clinical Monitoring, 1989, 5:256-265.* doi:10.1007/BF01618258

[de Graaf *et al.* 1997] P.M. de Graaf, G.C. van den Eijkel, H.J. Vullings, B.A. de Mol. A decision-driven design of a decision support system in anesthesia. *Artificial Intelligence in Medicine, October 1997, 11(2):141-53.*

[Gröschel *et al.* 2004] J. Gröschel, F. Philipp, St. Skonetzki, H. Genzwürker, Th. Wetter, K. Ellinger. Automated speech recognition for time recording in out-of-hospital emergency medicine – an experimental approach. *Resuscitation, 2004, 60:205–212.* doi:10.1016/j.resuscitation.2003.10.006

[Hamilton 1990] William K. Hamilton. Will we see automated record keeping systems in common use in anaesthesia during our lifetime? *Journal of Clinical Monitoring, 1990, 6(4):333-334.*

[Hansen 1996] John H.L. Hansen. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Communication, 1996, 20:151–173.* doi:10.1016/S0167-6393(96)00050-7

[Happe *et al.* 2003] André Happe, Bruno Pouliquen, Anita Burgun, Marc Cuggia, Pierre Le Beux. Automatic concept extraction from spoken medical reports. *International Journal of Medical Informatics, 2003, 70:255-263.* doi:10.1016/S1386-5056(03)00055-8

[Hebert & Benbasat 1994] M. Hebert, I. Benbasat. Adopting information technology in hospitals: The relationship between attitudes/expectations and behaviour. *Hospital & Health Services Administration, 1994, 39(3):369-383.*

[Henricks *et al.* 2002] Walter H. Henricks, Kavous Roumina, Bradley E. Skilton, Debra J. Ozan, Gwendolyn R. Goss. The Utility and Cost Effectiveness of Voice Recognition Technology in Surgical Pathology. *Modern Pathology, May 2002, 15(5):565-71.* doi:10.1038/modpathol.3880564

[Hirsch & Pearce 2000] Hans-Günter Hirsch & David Pearce. The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions. *In: Proceedings of ISCA ITRW ASR'2000 "Automatic Speech Recognition: Challenges for the Next Millennium", September 18-20 2000, Paris, France*.

[Hornbæk & Law 2007] Kasper. Hornbæk & Effie Lai-Chong Law. Meta-analysis of correlations among usability measures. *In: Proceedings of the CHI'2007 Conference on Human Factors in Computing Systems, ACM Press, 2007, New York, pp. 617-626*. doi:10.1145/1240624.1240722

[Honeycutt 2003] Lee Honeycutt. Researching the use of voice recognition writing software. Computers and Composition (2003), 20:77–95. doi:10.1016/S8755-4615(02)00174-3

[Hvidberg 2003] Jens Hvidberg. Talegenkendelse - muligheder og barrierer for anvendelse til klinisk dokumentation (In Danish). *Master's thesis, Aalborg University, May 2003*.

[Jungk *et al.* 2000] Andreas Jungk, Bernhard Thull, Lutz Fehrle, Andreas Hoeft, Günter Rau. A case study in designing speech interaction with a patient monitor. *Journal of Clinical Monitoring and Computing, 2000, 16:295-307*. doi:10.1023/A:1011456205786

[Juul-Kristensen *et al.* 2004] B. Juul-Kristensen, B. Laursen, M. Pilegaard, B.R. Jensen. *Physical workload during use of speech recognition and traditional computer input devices. Ergonomics, 2004, 47(2):119-133*. doi:10.1080/00140130310001617912

[Kanal *et al.* 2001] K.M. Kanal, N.J. Hangiandreou, A.M. Sykes, H.E. Eklund, P.A. Araoz, J.A. Leon, B.J. Erickson. Initial evaluation of a continuous speech recognition program for radiology. *Journal of Digital Imaging, 2001, 14(1):30-37*. doi:10.1007/s10278-001-0022-z

[de Keyser & Nyssen 1993] V. de Keyser & A.S. Nyssen. Human errors in anesthesia. *Le Travail Humain, 1993, 56(2-3):243-266*.

[Kozine 2007] Igor Kozine. Simulation of human performance in time-pressured scenarios, Proceedings of the Institution of Mechanical Engineers. *IMechE'2007, Vol. 221, Part O: Journal of Risk and Reliability, pp. 141–152*. doi:10.1243/1748006XJRR48

[Kushniruk & Patel 2004] Andre W. Kushniruk & Vimla L. Patel, Cognitive and usability engineering methods for the evaluation of clinical information systems. *Journal of Biomedical Informatics, 2004, 37(1):56–76*. doi:10.1016/j.jbi.2004.01.003

[Lai & Aarabi 2004] Calvin Yiu-Kit Lai, Parham Aarabi. Multiple-microphone time-varying filters for robust speech recognition. *In: Proceedings of ICASSP'2004, International Conference on Acoustics, Speech, and Signal Processing*.

[Lai & Vergo 1997] Jennifer Ceil Lai & John George Vergo. MedSpeak: Report creation with continuous speech recognition. *In: Proceedings of the CHI'1997 Conference on Human Factors in Computing Systems. ACM Press, 1997, New York, pp. 431-438*. ISBN:0-89791-802-9

[Lai & Vergo 1999] Jennifer Ceil Lai & John George Vergo. Speech recognition confidence level display. *US Patent 6,006,183, 1999*.

[Lane *et al.* 1961] H. L. Lane, A. C. Catania, S. S. Stevens. Voice Level: Autophonic Scale, Perceived Loudness, and Effects of Sidetone. *Acoustical Society of America, 1961*. doi:10.1121/1.1908608

[Lave 1993] Jean Lave. The practice of learning. *S. Chaiklin & J. Lave. Understanding practice: Perspectives on activity and context, 3-32. Cambridge (UK) University Press 1993.* ISBN:0521558514.

[Lechner *et al.* 2002] Alicia Lechner, Kevin Ecker, Patrick Mattson. Voice recognition – Software solutions in realtime ATC workstations. *Aerospace and Electronic Systems Magazine, IEEE, 2002, 17(11):11-16.* doi:10.1109/MAES.2002.1047373

[Leijten & Van Waes 2005] Mariëlle Leijten, Luuk Van Waes. Writing with speech recognition: The adaptation process of professional writers with and without dictating experience. Interacting with Computers (2005) 17:736–772. doi:10.1016/j.intcom.2005.01.005

[Levenshtein 1965] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR, 1965, 163(4):845-848, (Russian). English translation in Soviet Physics Doklady, 1966, 10(8):707-710.*

[Liu 1997] Yili Liu, Queueing Network Modeling of Human Performance of Concurrent Spatial and Verbal Tasks. *IEEE Transactions on Systems, Man and Cybernetics, Part A, March 1997, 27(2):195-207.* doi:10.1109/3468.554682

[Lombard 1911] E. Lombard. Le signe de l'élévation de la voix. *Annales Maladies Oreille, Larynx, Nez, Pharynx, 1911, 31:101–119.*

[Lovis *et al.* 2000] Christian Lovis, Robert H. Baud, Pierre Planche, Power of expression in the electronic patient record: structured data or narrative text? *International Journal of Medical Informatics, 2000, 58–59:101–110.* doi:10.1016/S1386-5056(00)00079-4

[Luketich *et al.* 2002] J.D. Luketich, H.C. Fernando, P.O. Buenaventura, N.A. Christie, S.C. Grondin, P.R. Schauer. Results of a randomized trial of HERMES-assisted versus non-HERMES-assisted laparoscopic antireflux surgery. *Surgical Endoscopy, 2002, 16:1264-1266.* doi:10.1007/s00464-001-8222-7

[Matsushita 2003] Masahiko Matsushita, Hiromitsu Nishizaki, Takehito Utsuro, Yasuhiro Kodama, Seiichi Nakagawa. Evaluating Multiple LVCSR Model Combination in NTCIR-3 Speech-Driven Web Retrieval Task. *Proceeding of Eurospeech'2003, the 8th European Conference on Speech Communication and Technology, pp. 1205-1208.*

[McCowan *et al.* 2005] Iain McCowan, Darren Moore, John Dines, Daniel Gatica-Perez, Mike Flynn, Pierre Wellner, Hervé Bourlard. On the Use of Information Retrieval Measures for Speech Recognition Evaluation. *IDIAP (Institut Dalle Molle d'Intelligence Artificielle Perceptive), Martigny, Switzerland. IDIAP Research Report IDIAP-RR 04-73, March 2005.* http://www.idiap.ch/ftp/reports/2004/rr04-73.pdf

[Mohr *et al.* 2003] David N. Mohr, David W. Turner, Gregory R. Pond, Joseph S. Kamath, Kathy B. De Vos, Paul C. Carpenter. Speech Recognition as a Transcription Aid: A Randomized Comparison With Standard Transcription. *Journal of the American Medical Informatics Association, 2003, 10(1):85–93.* doi:10.1197/jamia.M1130

[Mönnich & Wetter 2000] G. Mönnich & T. Wetter. Requirements for speech recognition to support medical documentation. Methods of Information in Medicine (2000), 39(1):63-9.

[Nardi 1996] B.A. Nardi. Context and Consciousness: Activity Theory and Human-Computer Interaction. *Cambridge MA, London: MIT Press 1996.*

[Østergaard 2004] Doris Østergaard, National Medical Simulation training program in Denmark. *Critical Care Medicine, 32(2) (Supplement):S58-S60, February 2004.* doi:10.1097/01.CCM.0000110743.55038.94

[Pallett 1985] David S. Pallett. Performance Assessment of Automatic Speech Recognizers. *Journal of Research of the National Bureau of Standards, September-October 1985, 90(5).* ISSN:0160-1741

[Paternò 2004] Fabio Paternò. Multimodality and Multiplatform Interactive Systems. *Proceedings of WCC'2004, the 18th IFIP (International Federation for Information Processing) World Computer Congress, Toulouse, August 2004, Kluwer Academic Publishers, René Jacquart (ed.), "Building the Information Society", pp. 421-426,* ISBN:1-4020-8156-1

[Pronovost *et al.* 2003] Peter Pronovost, Brad Weast, Mandalyn Schwarz, Rhonda M. Wyskiel, Donna Prow, Shelley N. Milanovich, Sean Berenholtz, Todd Dorman Pamela Lipsett. Medication reconciliation: a practical tool to reduce the risk of medication errors. *Journal of Critical Care, 2003, 18(4):201-205.* doi:10.1016/j.jcrc.2003.10.001

[Ramaswamy *et al.* 2000] Mohan R. Ramaswamy, Gregory Chaljub, Oliver Esch, Donald D. Fanning, Eric van Sonnenberg. Continuous speech recognition in MR imaging reporting: Advantages, disadvantages, and impact. *American Journal of Roentgenology, 2000, 174(3):617-622.*

[Rasmussen 1986] Jens Rasmussen. Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering. *Book published by Elsevier Science Inc., 1986.* ISBN:0444009876

[Robinson 1957] D. W. Robinson. The subjective loudness scale. *Acustica, 1957, Vol. 7.*

[Rogers 2003] Everett M. Rogers. Diffusion of Innovations, Fifth Edition. Free Press, 2003, New York. ISBN:0743222091

[Root & Draper 1983] R.W. Root, S. Draper. Questionnaires as a software evaluation tool. In Proceedings of the CHI'1983 Conference on Human Factors in Computing Systems. *ACM Press, 1983, New York, pp. 83-87.* doi:10.1145/800045.801586

[Saastamoinen 2005] Juhani Saastamoinen, Zdenek Fiedler, Tomi Kinnunen, Pasi Fränti. On factors affecting MFCC-based speaker recognition accuracy. *International Conference on Speech and Computer (SPECOM'2005), Patras, Greece, pp. 503-506, October 2005.*

[Sanjo *et al.* 1999] Yoshimitsu Sanjo, Tetsuo Yokoyama, Shigehito Sato, Kazuyuki Ikeda, Reiko Nakajima. Ergonomic automated anesthesia recordkeeper using a mobile touch screen with voice navigation. *Journal of Clinical Monitoring and Computing, 1999, 15:347-356.* doi:10.1023/A:1009972223750

[Sauer 2000] Juergen Sauer, Prospective memory: a secondary task with promise. *Applied Ergonomics, April 2000, 31(2):131-137.* doi:10.1016/S0003-6870(99)00042-3

[Schapira & Sharma 2001] Emilio Schapira & Rajeev Sharma 2001. Experimental Evaluation of Vision and Speech based Multimodal Interfaces. *In: Proceedings of the 2001 Workshop on Perceptive User interfaces (Orlando, Florida, USA). PUI'2001, vol. 15. ACM Press, New York, 1-9.* doi:10.1145/971478.971481

[Schmitz & Weiss 2004] Achim Schmitz & M. Weiss. Are we going to talk with our anaesthesia monitors in the future? *Acta Anaesthesiologica Scandinavica, 2004, 48(2):255-256.* doi:10.1111/j.0001-5172.2004.00295b.x

[Sears *et al.* 2001] Andrew Sears, Clare-Marie Karat, Kwesi Oseitutu, Azfar Karimullah, Jinjuan Feng. Productivity, satisfaction, and interaction strategies of individuals with spinal cord injuries and traditional users interacting with speech recognition software. *Universal Access in the Information Society, 2001, 1(1):4-15.* doi:10.1007/s102090100001

[Sears *et al.* 2003] Andrew Sears, Jinjuan Feng, Kwesi Oseitutu, Clare-Marie Karat. Hands-Free, Speech-Based Navigation During Dictation: Difficulties, Consequences, and Solutions. *Human-computer interaction, 2003, 18:229–257.* doi:10.1207/S15327051HCI1803_2

[Shiffman *et al.* 1995] Smadar Shiffman, William M. Detmer, Christopher D. Lane, Lawrence M. Fagan. A continuous-speech interface to a decision support system: I. Techniques to accommodate for misrecognized input. *Journal of the American Medical Informatics Association, 1995, 2(1):36–45.*

[Smith *et al.* 1987] N. Ty Smith, M. L. Quinn, A.J. Sarnat. Speech Recognition for the Automated Anesthesia Record. *In: The Automated Anesthesia Record and Alarm Systems, Chapter 11, Jonathan S. Gravenstein, Ronald S. Newbower, Allen K. Ream, N. Ty Smith (eds.), Butterworths, 1987, pp. 115-134.*

[Smith *et al.* 1990] N. Ty Smith, Robin A. Brien, Daniel C. Pettus, Brian R. Jones, Michael L. Quinn, Andrew Sarnat. Recognition accuracy with a voice-recognition system designed for anaesthesia record keeping. *Journal of Clinical Monitoring, 1990, 6(4):299-306.*

[Sobel 1981] Carolyn Panzer Sobel. A generative phonology of Danish. *Ph.D. Thesis 1981, City University of New York.*

[Steinlen & Bohn 1999] Anja K. Steinlen & Ocke-Schwen Bohn. Acoustic studies comparing Danish vowels, British English vowels and Danish-accented British English vowels. *Collected Papers (CD-ROM) of the 137th Meeting of the Acoustical Society of America and the 2nd Convention of the European Acoustics Association, Forum Acousticum, Paper 2pSCb21, Technical University of Berlin, Germany. Abstract in the Journal of the Acoustical Society of America, 1999, 105(2):1097.* doi:10.1121/1.425143

[Suhm *et al.* 2001] Bernhard Suhm, Brad Myers, Alex Waibel. Multimodal error correction for speech user interfaces. *ACM Transactions on Computer-Human Interaction, 2001, 8(1):60-98.* doi:10.1145/371127.371166

[Venkatesh *et al.* 2003] V. Venkatesh, M.G. Morris, G.B. Davis, F.D. Davis. User acceptance of information technology: Toward a unified view. *MIS Quarterly, 2003, 27(3):425-478.*

[Weber *et al.* 2005] Steen Weber, Jette Lundtang Paulsen, Henning Boje Andersen. Comparing of training effects in real and virtual environments. *In: Tagungsband: Virtual reality und augmented reality zum Planen, Testen und Betreiben technischer Systeme. 8. IFF-Wissenschaftstage, Magdeburg (DE), 22-24 June 2005. M. Schenk (ed.), (Fraunhofer-Institut für Fabrikbetrieb und -automatisierung IFF, Magdeburg) pp. 145-154.*

[Wilpon & Jacobsen 1996] J.G. Wilpon, C.N. Jacobsen. A study of speech recognition for children and the elderly. *Proceedings of ICASSP'1996, the International Conference on Acoustics, Speech, and Signal Processing, Vol. I. IEEE, 1996, Los Alamitos, CA, pp. 349-352.* doi:10.1109/ICASSP.1996.541104

[Young 1996] Steve Young. Review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine, 1996, 13(5):45-57*. doi:10.1109/79.536824

[Zafar *et al.* 1999] Atif Zafar, J. Marc Overhage, Clement J. McDonald. Continuous speech recognition for clinicians. *Journal of the American Medical Informatics Association, 1999, 6(3):195–204*.

[Zafar *et al.* 2004] Atif Zafar, Burke Mamlin, Susan Perkins, Anne M. Belsito, J. Marc Overhage, Clement J. McDonald. A simple error classification system for understanding sources of error in automatic speech recognition and human transcription. *International Journal of Medical Informatics, 2004, 73:719-730*. doi:10.1016/j.ijmedinf.2004.05.008

[Zick & Olsen 2001] Robert G. Zick, Jon Olsen. Voice Recognition Software Versus a Traditional Transcription Service for Physician Charting in the ED. *American Journal of Emergency Medicine, July 2001, 19(4):295-298*. doi:10.1053/ajem.2001.24487

RECENT RESEARCH REPORTS

#118 Paolo Bouquet, editor. *CONTEXT'07 Doctoral Consortium Proceedings*, Roskilde, Denmark, October 2007.

#117 Kim S. Henriksen. *A Logic Programming Based Approach to Applying Abstract Interpretation to Embedded Software*. PhD thesis, Roskilde, Denmark, October 2007.

#116 Marco Baroni, Alessandro Lenci, and Magnus Sahlgren, editors. *Proceedings of the 2007 Workshop on Contextual Information in Semantic Space Models: Beyond Words and Documents*, Roskilde, Denmark, August 2007.

#115 Paolo Bouquet, Jérôme Euzenat, Chiara Ghidini, Deborah L. McGuinness, Valeria de Paiva, Luciano Serafini, Pavel Shvaiko, and Holger Wache, editors. *Proceedings of the 2007 workshop on Contexts and Ontologies Representation and Reasoning (C&O:RR-2007)*, Roskilde, Denmark, August 2007.

#114 Bich-Liên Doan, Joemon Jose, and Massimo Melucci, editors. *Proceedings of the 2nd International Workshop on Context-Based Information Retrieval*, Roskilde, Denmark, August 2007.

#113 Henning Christiansen and Jørgen Villadsen, editors. *Proceedings of the 4th International Workshop on Constraints and Language Processing (CSLP 2007)*, Roskilde, Denmark, August 2007.

#112 Anders Kofod-Petersen, Jörg Cassens, David B. Leake, and Stefan Schulz, editors. *Proceedings of the 4th International Workshop on Modeling and Reasoning in Context (MRC 2007) with Special Session on the Role of Contextualization in Human Tasks (CHUT)*, Roskilde, Denmark, August 2007.

#111 Ioannis Hatzilygeroudis, Alvaro Ortigosa, and Maria D. Rodriguez-Moreno, editors. *Proceedings of the 2007 workshop on REpresentation models and Techniques for Improving e-Learning: Bringing Context into the Web-based Education (ReTIeL'07)*, Roskilde, Denmark, August 2007.

#110 Markus Rohde. *Integrated Organization and Technology Development (OTD) and the Impact of Socio-Cultural Concepts — A CSCW Perspective*. PhD thesis, Roskilde University, Roskilde, Denmark, 2007.

#109 Keld Helsgaun. An effective implementation of *k*-opt moves for the Lin-Kernighan TSP heuristic. 2006, Roskilde University, Roskilde, Denmark.

#108 Pernille Bjørn. *Virtual Project Teams — Distant Collaborative Practice and Groupware Adaptation*. PhD thesis, Roskilde University, Roskilde, Denmark, 2006.

#107 Henrik Bulskov Styltsvig. *Ontology-based Information Retrieval*. PhD thesis, Roskilde University, Roskilde, Denmark, 2006.