journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Speech recognition for the anaesthesia record during crisis scenarios

Alexandre Alapetite[*]

Systems Analysis Department, Risø National Laboratory, Technical University of Denmark, DK-4000 Roskilde, Denmark

ARTICLE INFO

Article history:
Received 27 January 2007
Received in revised form 1 July 2007
Accepted 21 August 2007

Keywords:
Anaesthesia record
Quality
Vocal interface
Speech recognition
Workload
Secondary task

ABSTRACT

Introduction: This article describes the evaluation of a prototype speech-input interface to an anaesthesia patient record, 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.

[*] Tel.: +45 46775182; fax: +45 46775199.
E-mail address: alexandre@alapetite.net.

1386-5056/$ – see front matter © 2007 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2007.08.007

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 [1]. 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 [8], 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 [3]. 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 Human Computer Interaction (HCI) 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 [9]: 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. 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.

2.1. *The electronic anaesthesia record*

The anaesthesia information management system in use at participants' home hospital, Recall AIMS from Dräger Medical, includes an anaesthesia record component with a touch-screen and a keyboard. 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. This system was used as a reference for the design of the mock-up.

2.2. *Speech recognition software*

For voice dictation in free speech mode, or "natural language", the speech recognition system Philips 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. 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 min during which the speech recognition system is trained on the user's voice.

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 [2]. 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.

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 s. If, perhaps through hesitation, the user has not completed the intended sentence before the 2-s 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 literature: Detmer et al. used a button to activate the speech recognition system [5], Sanjo et al. used a touch-screen to initiate the dialog [6], Jungk et al. did the dictations separately after the operations [3].

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 [2]. 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" [7].

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 [10].

2.3.1. Speech grammar

The main principles of the syntax to follow when dictating commands to the system were discussed with an anaesthesiologist from Køge Hospital. 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 [11]. 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 min to learn how to address the system. In spite of its simplicity, the grammar was aimed to cover the main user needs.

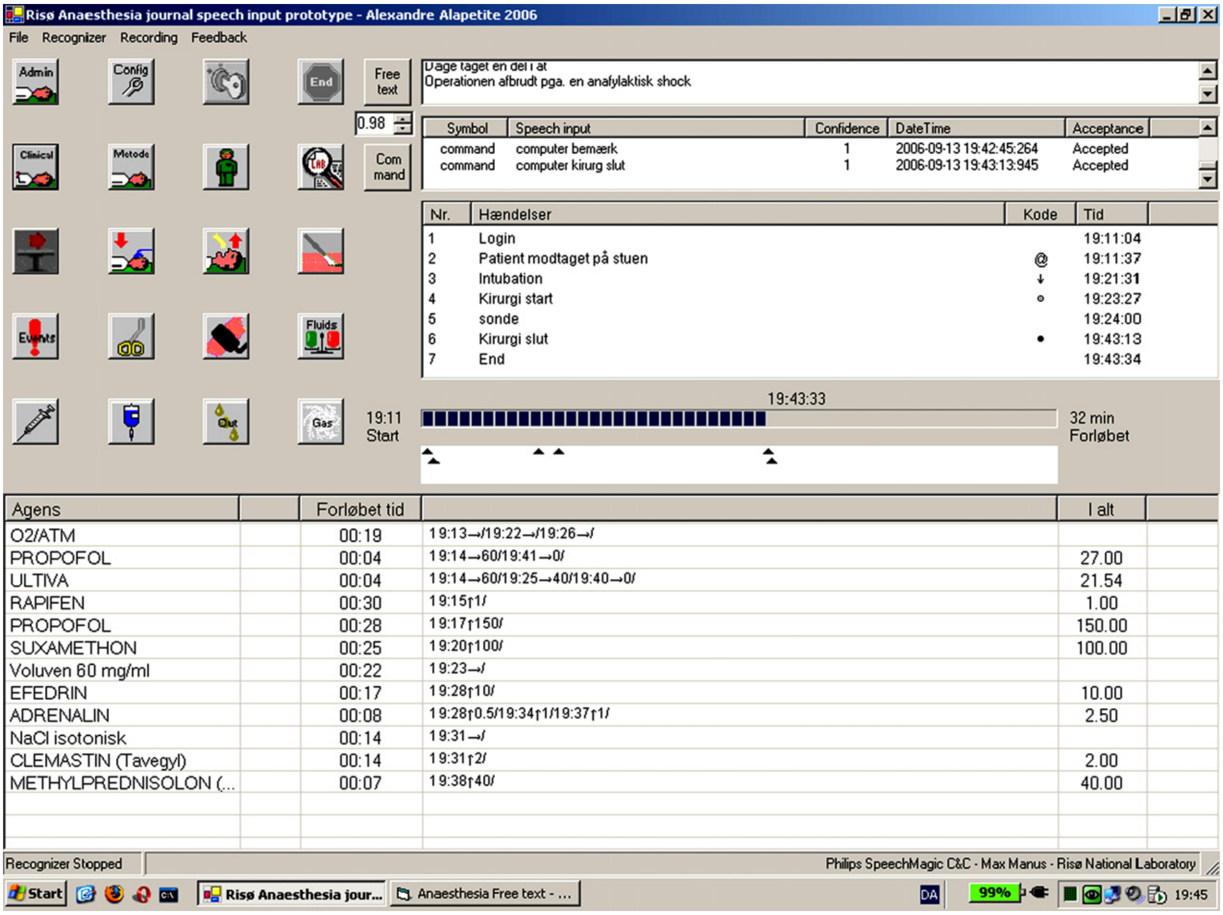
As reported in Table 1, there are five types of speech commands:

(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 2 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 1000 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 s.

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
COMPUTER <medication> INFUSION (<dosage> | STOP)	COMPUTER Propofol INFUSION 60	50 dosages
COMPUTER <liquid or gas> (START | STOP)	COMPUTER Oxygen START	3 liquids, 5 gases
COMPUTER REMARK {wait 1s} <free text> {wait 2s}	COMPUTER REMARK . . . Patient has fever between 38 and 39 °C . . .	Unlimited



Dage taget en del i at
Operationen afbrudt pga. en anafylaktisk shock

Symbol	Speech input	Confidence	DateTime	Acceptance
command	computer bemærk	1	2006-09-13 19:42:45:264	Accepted
command	computer kirurg slut	1	2006-09-13 19:43:13:945	Accepted

Nr.	Hændelser	Kode	Tid
1	Login		19:11:04
2	Patient modtaget på stuen	@	19:11:37
3	Intubation	↓	19:21:31
4	Kirurgi start	○	19:23:27
5	sonde		19:24:00
6	Kirurgi slut	●	19:43:13
7	End		19:43:34

Agens		Forløbet tid		I alt
O2/ATM		00:19	19:13→/19:22→/19:26→/	
PROPOFOL		00:04	19:14→60/19:41→0/	27.00
ULTIVA		00:04	19:14→60/19:25→40/19:40→0/	21.54
RAPIFEN		00:30	19:15↑1/	1.00
PROPOFOL		00:28	19:17↑150/	150.00
SUXAMETHON		00:25	19:20↑100/	100.00
Voluven 60 mg/ml		00:22	19:23→/	
EFEDRIN		00:17	19:28↑10/	10.00
ADRENALIN		00:08	19:28↑0.5/19:34↑1/19:37↑1/	2.50
NaCl isotonisk		00:14	19:31→/	
CLEMASTIN (Tavegyl)		00:14	19:31↑2/	2.00
METHYLPREDNISOLON (...		00:07	19:38↑40/	40.00

Fig. 1 – Mock-up of the anaesthesia record with speech commands.

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 (1.6 GHz, 768 MB of memory, Windows XP) linked to a 17″ touch-screen and to a head-set microphone (∼2.5 cm from the mouth) model PC145-USB 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 (Fig. 1), was developed with the programming framework Microsoft C# .NET 2.0. 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.

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. 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.

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 doctor will make the decisions regarding the strategy to follow, but doctor and 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 Ref. [1]. 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, Herlev University Hospital (Copenhagen region, Denmark) in one of the institute's full-scale simulators used for training anaesthesiologists [12], following principles similar to but newer than those reported in Ref. [13]. 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, 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.

For this experiment, the normal anaesthesia simulator setting was supplemented with the prototype electronic anaesthesia record system with speech input, with the touchscreen 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. The four video signals were mixed to produce a single picture divided into four areas (Fig. 2), thus avoiding all problems of synchronisation. A stereo microphone was placed in the middle of the operating room.

3.2.3. *Participants*

The 12 participants were volunteers from Køge Hospital. Their department was chosen because they had been using an

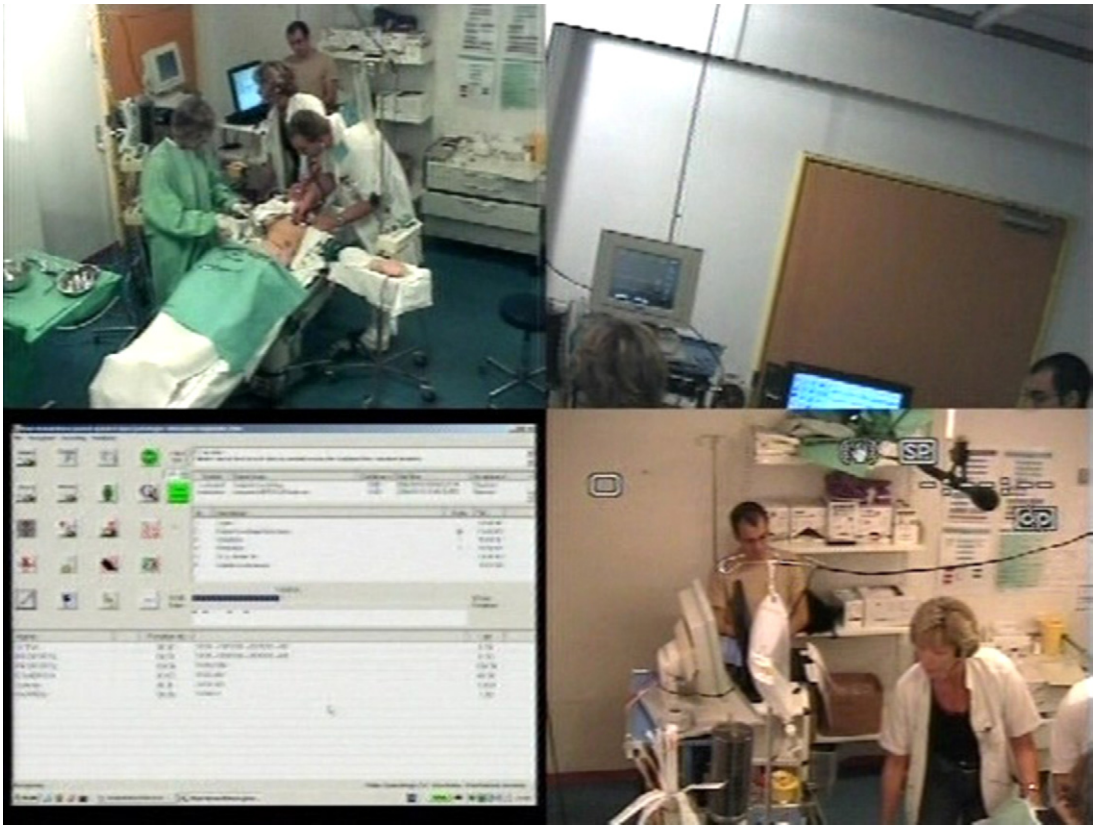


Fig. 2 – Recording sound and four videos at a time.

electronic anaesthesia record for some years. There were six 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.

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 [5]. 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 input recognised by the software. It was originally planned to do the Wizard-of-Oz remotely, 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.

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 min: 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 3 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 1 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. There are indeed many events that are not registered during

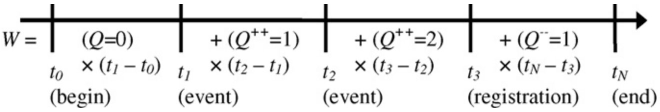


Fig. 3 – Example of workload calculation using the proposed approach based on queuing theory.

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 [9,14]. 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.

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

where W is the averaged queue of events to be registered (workload), t_n the time in seconds of an event or a registration, Q_n 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.

In the example shown in Fig. 3, lasting 40 s where each interval is 10 s, with two events and then one registration, the average queue size is:

$$W = \frac{[(0 \times 10\,\mathrm{s}) + (1 \times 10\,\mathrm{s}) + (2 \times 10\,\mathrm{s}) + (1 \times 10\,\mathrm{s})]}{40\,\mathrm{s}} = 1$$

3.3.1. *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 2. This type of video analysis is common in HCI studies [15]. 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.

3.3.2. *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" [2], 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 inter-

Table 2 – Short excerpt from a transcript of session 12, translated into English

	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)

The code "ASR" stands for "Automatic speech recognition".

face 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 1–5, where 1 is full disagreement, 3 is neutral and 5 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 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).

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 in Fig. 4 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 Fig. 4), partly due to small samples. The small differences could at least be partially explained by the fact the participants were accustomed to the traditional interface, but tried the speech interface for the first time.

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.

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.

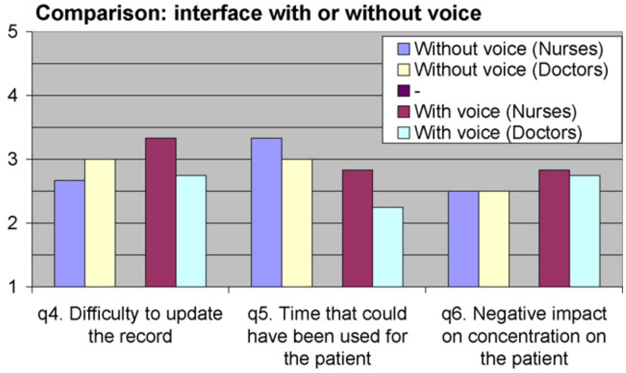


Fig. 4 – Questionnaire responses on time and difficulty to keep the anaesthesia record updated during the scenario, with or without voice.

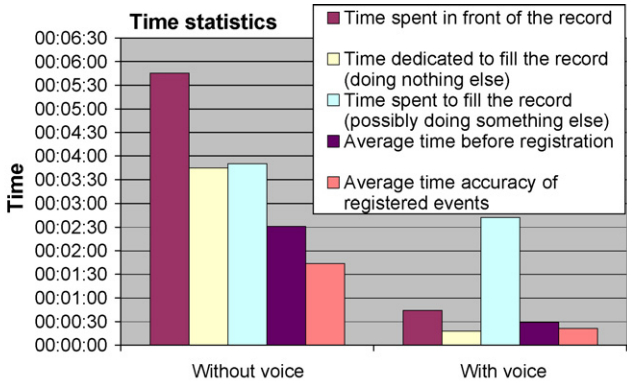


Fig. 5 – Measurements of delays and time used to fill the anaesthesia record, with or without voice.

The sessions lasted on average 31 min without voice and 26 min with voice, but the differences are not significant ($p = 0.14$, *t*-test). As shown in Fig. 5, 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 3). This means it took on average 17 s per event registration with the traditional interface, and almost three times less with speech recognition, down to 6 s per registration ($p < 0.002$).

In Fig. 5, 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 (Fig. 5). 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 five 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.

Table 3 – Statistics measures from video analysis, with or without voice, each condition averaged across six sessions

	Without voice	With voice	*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	$p < 0.005$

With the traditional interface, the long delay before registration leads to queues of events that accumulate, as reported in Fig. 6, and the queue increases all along the anaesthesia scenario. In contrast, the queue is kept small with the speech interface. As shown in Table 3, 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.

In Fig. 5, 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 3, 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.

In Table 3, 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

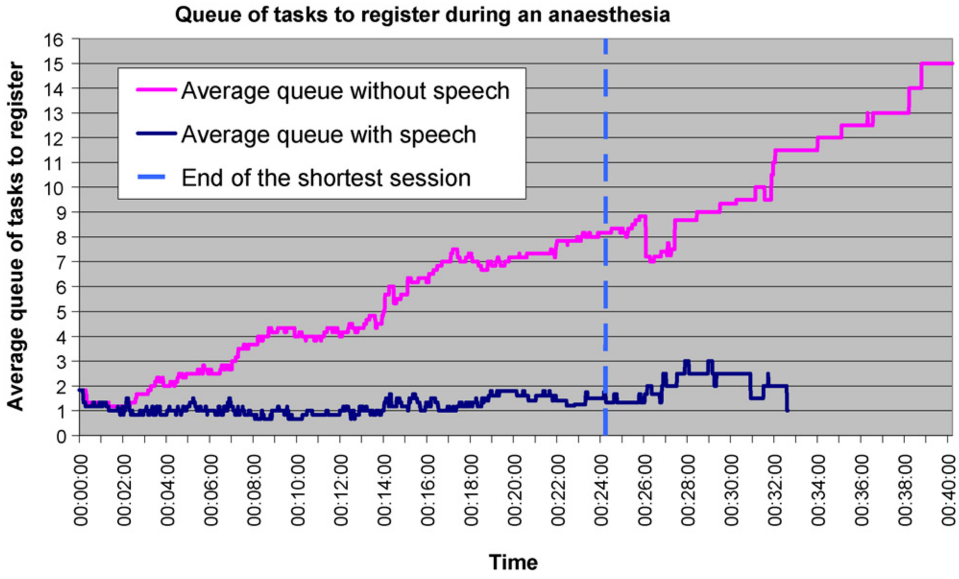


Fig. 6 – Evolution of the averaged queue of events to register during the anaesthesia scenarios, with or without voice.

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 Fig. 5, the time accuracy of the registered events was almost five times higher with the vocal interface (21 s 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 2 h 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.

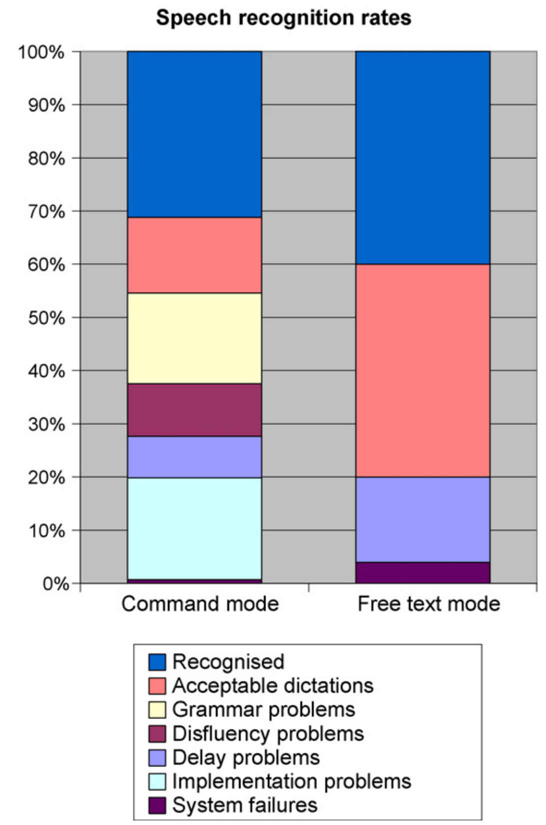


Fig. 7 – Categorisation of dictations and recognitions.

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 Fig. 7.

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 (∼1 s delay, sound feedback when ready) and speaking too early. The remaining types of non-successful dictations are too subjective to be classified.

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 [16].

4.4.3. *Overall utility of speech interface*

In the questionnaire, 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 a 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.

Finally, in the free text section of the survey, some participants shared their views and concerns regarding a vocal interface. Six of the 10 respondents reported that the vocal modality would be useful to have because it helps to produce more accurate and real-time data; 5 respondents said it would help in keeping hands free and a visual contact with the patient; 1 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 [17] 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 [14], 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 secondary or 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 3 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 speech 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 speech 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.

6. 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.

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 which 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 [1].
- Background noise and stress are among the factors having a negative effect on speech recognition rates [2].
- Some experiments have been done to investigate the potential of speech recognition in anaesthesia, mainly during calm situations and not entirely realistic anaesthesia scenarios [3]. Questionnaire surveys [4] and simulations [5] 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 [6].

What the study has added to the body of knowledge?

- 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 [7].

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

[1] A. Alapetite, V. Gauthereau, Introducing vocal modality into electronic anaesthesia record systems: possible effects on work practices in the operating room, in: 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, 2005, pp. 197–204.

[2] A. Alapetite, Impact of noise and other factors on speech recognition in anaesthesia, Int. J. Med. Inf. (available online December 2006, doi:10.1016/j.ijmedinf.2006.11.007).

[3] A. Jungk, B. Thull, L. Fehrle, A. Hoeft, G. Rau, A case study in designing speech interaction with a patient monitor, J. Clin. Monit. Comput. 16 (2000) 295–307.

[4] C.B. DeVos, D.A. Martin, P.A. John, An evaluation of an automated anesthesia record keeping system, Biomed. Sci. Instrum. 27 (1991) 219–225.

[5] W.M. Detmer, S. Shiffman, J.C. Wyatt, C.P. Friedman, C.D. Lane, L.M. Fagan, A continuous-speech interface to a decision support system. II. An evaluation using a Wizard-of-Oz experimental paradigm, J. Am. Med. Inf. Assoc. 2 (1) (1995) 46–57.

[6] Y. Sanjo, T. Yokoyama, S. Sato, K. Ikeda, R. Nakajima, Ergonomic automated anesthesia recordkeeper using a mobile touch screen with voice navigation, J. Clin. Monit. Comput. 15 (1999) 347–356.

[7] C. Lovis, R.H. Baud, P. Planche, Power of expression in the electronic patient record: structured data or narrative text? Int. J. Med. Inf. 58–59 (2000) 101–110, doi:10.1016/S1386-5056(00)00079-4.

[8] N. Cowan, The magical number 4 in short-term memory: a reconsideration of mental storage capacity, Behav. Brain Sci. 24 (2001) 87–114, doi:10.1017/S0140525X01003922 (Cambridge University Press).

[9] I. Kozine, Simulation of human performance in time-pressured scenarios, in: Proceedings of the Institution of Mechanical Engineers, IMechE, vol. 221, part O, J. Risk Reliability (2007) 141–152, doi:10.1243/1748006XJRR48.

[10] A. Sears, J. Feng, K. Oseitutu, C.-M. Karat, Hands-free, speech-based navigation during dictation: difficulties, consequences, and solutions, Hum. Comput. Interact. 18 (2003) 229–257, doi:10.1207/S15327051HCI1803_2.

[11] S. Shiffman, W.M. Detmer, C.D. Lane, L.M. Fagan, A continuous-speech interface to a decision support system. I. Techniques to accommodate for misrecognized input, J. Am. Med. Inf. Assoc. 2 (1) (1995) 36–45.

[12] D. Østergaard, National medical simulation training program in Denmark, Crit. Care Med. 32 (February (Suppl. 2)) (2004) S58–S60, doi:10.1097/01.CCM.0000110743.55038.94.

[13] D.M. Gaba, A. DeAnda, A Comprehensive anesthesia simulation environment: re-creating the operating room for research and training, Anesthesiology 69 (1988) 387–394.

[14] Y. Liu, Queuing network modeling of human performance of concurrent spatial and verbal tasks, IEEE Trans. Syst. Man Cybernetics Part A 27 (2) (1997) 195–207, doi:10.1109/3468.554682.

[15] A.W. Kushniruk, V.L. Patel, Cognitive and usability engineering methods for the evaluation of clinical information systems, J. Biomed. Inf. 37 (1) (2004) 56–76, doi:10.1016/j.jbi.2004.01.003.

[16] A. Happe, B. Pouliquen, A. Burgun, M. Cuggia, P. Le Beux, Automatic concept extraction from spoken medical reports, Int. J. Med. Inf. 70 (2003) 255–263., doi:10.1016/S1386-5056(03)00055-8.

[17] J. Sauer, Prospective memory: a secondary task with promise, Appl. Ergonomics 31 (2) (2000) 131–137, doi:10.1016/S0003-6870(99)00042-3.