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

Impact of noise and other factors on speech recognition in anaesthesia

Alexandre Alapetite[*]

Risø National Laboratory, Systems Analysis Department, SYS-110, P.O. Box 49, Frederiksborgvej 399, DK-4000 Roskilde, Denmark

ARTICLE INFO

Article history:
Received 19 July 2006
Received in revised form
8 September 2006
Accepted 20 November 2006

Keywords:
Anaesthesia
Electronic medical records
Voice recognition
Speech recognition
Noise
Error rates

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 3600 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 the interaction of several parameters that are traditionally studied separately.

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 [1] 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

[*] Tel.: +45 46 77 51 82; fax: +45 46 77 51 99.
E-mail address: alexandre.alapetite@risoe.dk.

1386-5056/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2006.11.007

background noise) 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

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. Some background noises were recorded in an OR (Herlev University Hospital of Copenhagen) during real anaesthesias with surgery and X-rays, using a multi-directional microphone placed in the proximity of the anaesthesiologist. 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. The 60 s L_{eq}[1] 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.

The collected sound files were edited and samples selected. Samples of the same type of noise were concatenated to create longer sequences with the same type of noise. The nine “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, 48–63 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, 58–82 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 61–73 dB(A);
(9) “Ventilation3”: Same as “Ventilation1” but 20 dB(A) louder, giving 71–83 dB(A), slow measure 75 dB(A).

2.1.1. *Reproducing sounds*

Samples were reproduced with a computer plugged to an audio amplifier (Sony STR-GX290) with two loudspeakers (Jamo Compact 1000, 65 Hz to 20 kHz, 90–120 W), positioned 1.5 m apart and pointing toward participants about 2 m 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.

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).

2.2. *Experiment*

2.2.1. *Speech recognition software*

The lab experiment was made with the speech recognition system Philips[2] 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.[3] The speech recognition workflow is the same as detailed in [2].

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 [1]. 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 [3] describing the set of possible commands. The grammar must contain the phonetic transcription of the terms used, for which the “Phonetic Transcriber component” can help.

Philips Speech Magic is now available in various languages, 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.

2.2.2. *Hardware*

Two similar laptop computers were used, running identical software. 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[4] (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),

[1] L_{eq}: equivalent continuous sound pressure.

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

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

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

model PC145-USB[5] from Sennheiser Communications (unidirectional, 80–15,000 Hz, −38 dB, ∼2 kΩ). Sennheiser indicates that this model is suited for voice 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” [5]. This effect is the tendency to alter the voice in noisy environments, and is known to affect speech recognition performance [6].

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. 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 were 2 m to the left of the participants. The two microphones were approximately at the same distance from the loudspeakers.

2.2.4. *Participants*

Eight subjects participated in this experiment (four males, four females, 27–62 years of age). The participants were Risø staff with no medical background. One of the participants had limited prior experience with speech recognition, 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.

2.2.5. *Test material*

The 100 most frequently recorded comments in Køge Hospital’s anaesthesia journal system from 2004 were identified and used as the basis for command mode training and testing in this study. The distribution of frequencies is interesting: the most frequent comment was used 9495 times, the 43rd 105 times, the 982nd 2 times and the rest only once. During dictations, each comment was followed by the Danish word for “full stop”.

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 twice: once with a silent background ∼32 dB(A) and once with the background noise “Ventilation1”. Half of the participants trained first with the silent background and then the noisy one, the other half in opposite order. The system was set up not to improve its general model across users.

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. 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:

$$8\,\text{participants} \times 9\,\text{background noises} \times \sim 50\,\text{sentences}$$
$$\simeq 3600\,\text{dictations}$$

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):

$$\sim 3600\,\text{dictations} \times 2\,\text{microphones} \times (1\,\text{command mode}$$
$$+2\,\text{free text modes}) \simeq 21,600\,\text{recognition samples}.$$

2.3. Statistics

The results and analysis presented below are based on descriptive statistics and regression analysis (Table 1, binary logistics regression where the dependent variable is binary: recognition is successful or not) using SPSS[6] version 14.

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.

2.3.1. *Recognition rate*

For calculating the recognition rate of any speech recognition engine, one of the most common metrics is the word error

[5] [http://www.oticon.com/eprise/main/Sennheiser Communications/com/Products/CNT05_VBLG?ProductId=PC145].

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

Table 1 – Regression model (binary logistic)

Variables	β	d.f.	Sig.
Mode(1): *free text mode*	−.144	1	.208[a]
Microphone(1): *microphone 2*	.162	1	.174[a]
Training_with_noise(1): *with noise*	−1.039	1	.000
Person_id: (*woman ∼average*)		7	.000
Person_id(1): *woman*	.178	1	.026
Person_id(2): *man*	−.322	1	.000
Person_id(3): *man*	−.982	1	.000
Person_id(4): *man*	.264	1	.000
Person_id(5): *woman*	.006	1	.932
Person_id(6): *man*	−.307	1	.000
Person_id(7): *woman*	−.625	1	.000
Session_id: (*silence*)		8	.000
Session_id(1): ventilation1	−.213	1	.069
Session_id(2): *alarms*	−.503	1	.000
Session_id(3): *scratch*	−1.520	1	.000
Session_id(4): *aspiration*	−1.116	1	.000
Session_id(5): *discussion*	−.751	1	.000
Session_id(6): *metal*	−.475	1	.000
Session_id(7): *ventilation2*	−1.051	1	.000
Session_id(8): *ventilation3*	−1.414	1	.000
Session_order: (*first session*)		9	.000
Session_order(1)	.067	1	.594
Session_order(2)	.675	1	.000
Session_order(3)	.762	1	.000
Session_order(4)	1.310	1	.000
Session_order(5)	.414	1	.004
Session_order(6)	.727	1	.000
Session_order(7)	.932	1	.000
Session_order(8)	.585	1	.000
Session_order(9): *last sessions*	1.201	1	.000
Mode(1) by training_with_noise(1)	1.214	1	.000
Mode[a] session_order		9	.000
Mode(1) by session_order(1)	−.070	1	.647
Mode(1) by session_order(2)	−.869	1	.000
Mode(1) by session_order(3)	−.863	1	.000
Mode(1) by session_order(4)	−1.737	1	.000
Mode(1) by session_order(5)	−.524	1	.001
Mode(1) by session_order(6)	−.745	1	.000
Mode(1) by session_order(7)	−1.168	1	.000
Mode(1) by session_order(8)	−.781	1	.000
Mode(1) by session_order(9)	−1.541	1	.000
Microphone[a] session_id		8	.000
Microphone(1) by session_id(1)	−.024	1	.884
Microphone(1) by session_id(2)	.038	1	.811
Microphone(1) by session_id(3)	.778	1	.000
Microphone(1) by session_id(4)	.552	1	.000
Microphone(1) by session_id(5)	.533	1	.001
Microphone(1) by session_id(6)	.154	1	.339
Microphone(1) by session_id(7)	.519	1	.001
Microphone(1) by session_id(8)	.908	1	.000
Constant	2.256	1	.000

[a] Mode and microphone are also used as combined variables. Their effect is significant.

rate (WER) or its complement, the word recognition rate (WRR), but both have limitations. To facilitate comparisons with other articles, WRR will be reported for some of the results.

$$\mathrm{WRR} = 1 - \mathrm{WER} = \frac{N - L}{N}$$

where N is the number of words in the reference and L is the Levenshtein distance at the word level (i.e., substitutions + deletions + insertions).

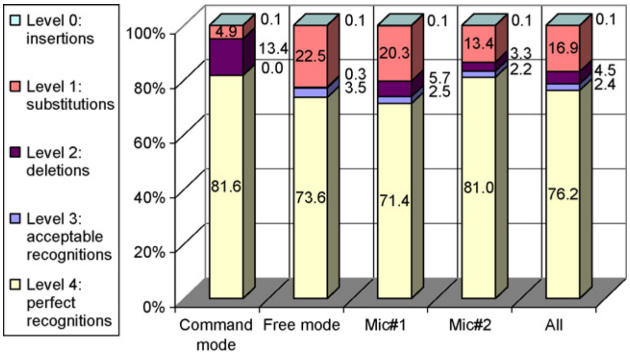


Fig. 1 – Classification of correct and failed recognitions (per sentence).

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*" [7] 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 Fig. 1). 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.

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 Fig. 1).

2.3.2. *Danish language*

The natural language of this study was Danish, a language that, like German, joins compound nouns. For instance, "the general department" is written "stamafdelingen" so if "the child department" ("børneafdelingen") was recognised instead, that would give 0 good recognitions and 1 false recognition in Danish, but two good recognitions and one false recognition in English. This illustrates that WER 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) [8].

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" [9]. This makes Danish vowels, which in addition have long and short versions (total of 28) [10], potentially more difficult to distinguish than English ones, with a direct impact on current speech recognition engines that typically prioritise vowels. 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 (∼5.5 M speakers), little research has been published about tuning speech recognition to its specificities (such as the glottal catch "stød").

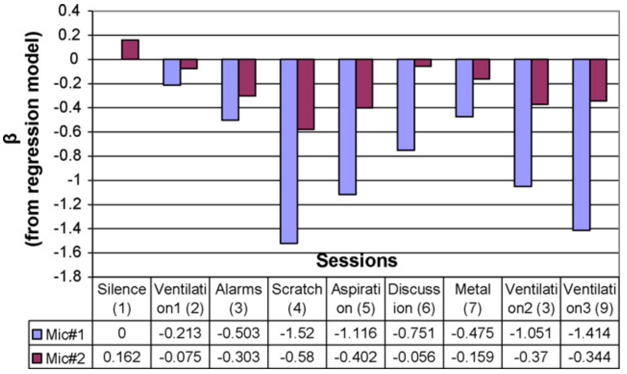
3. Results

Fig. 1 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.

3.1. *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 Fig. 1 (levels 3 + 4). This advantage of microphone 2 is present for all sessions (cf. Fig. 2). 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). The regression model (Table 1) shows a significant difference for microphone type when combined with the type of background noise, as reported below (Fig. 2); the combined effect of the type of microphone and the type of background noise is significant for most cases ($p < .001$). 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").

This contrasts with a recent study [11] that finds no significant difference between two types of microphone (unidirectional headset, versus built-in omni-directional microphone of a laptop). A possible explanation may be that during the experiment reported in [11], some noises were mixed afterwards (i.e., not recorded simultaneously with the speech), and possibly not replayed at a sufficiently high volume.



	Silence (1)	Ventilation1 (2)	Alarms (3)	Scratch (4)	Aspiration (5)	Discussion (6)	Metal (7)	Ventilation2 (3)	Ventilation3 (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

Fig. 2 – Effect of microphone type combined with session types (noises). The reference (0) is microphone 1 with a quiet background.

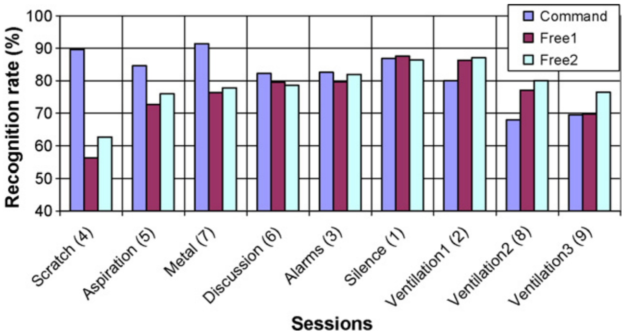


Fig. 3 – Recognition rates detailed per recognition modes and session types.

3.2. *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. Fig. 3 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").

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 Fig. 4). In contrast, one participant shows the opposite effect (−12.55 points for Woman4).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, $p < .001$ for most cases (the order of the sessions – see "time effect" in Table 5 – has only a very small impact).

3.2.1. *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%).

This is confirmed by the regression analysis (Tables 1 and 2); however, since in command mode there are only four participants for each type of training this result should be treated with caution.

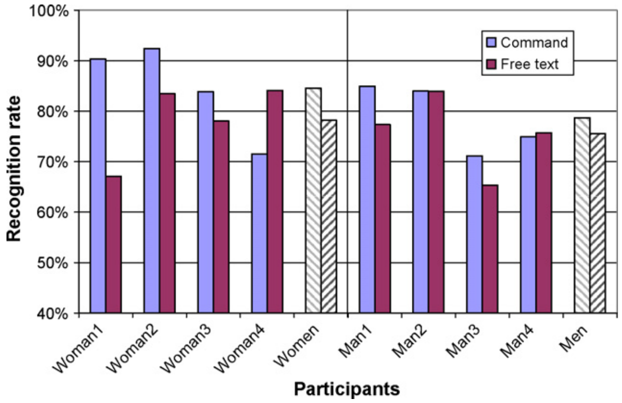


Fig. 4 – Recognition rates detailed per recognition mode and person (gender).

Table 2 – Effect of training type combined with recognition mode

	Recognition mode	
	Command	Free
β from regression model		
Training		
Without noise	0	−.144
With noise	−1.039	.031

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 2). 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 3). 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.

These results are similar to previous studies [12]. A system using free text mode should therefore be trained with the type of background noise that will typically be present during use.

3.2.2. *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 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 [8]. In the present experiment the confidence score was 1.0 in 5813 (80.77%) of the 7197 command-mode recognitions. For these 5813 recognitions, the recognition rate was 98.16%. The recognition rate decreases rapidly for lower confidence scores, showing that it can reliably be used as a threshold.

3.3. *Background noises*

Relative to the recognition rates obtained with a silent background (86.82%), the recognition rates obtained with “Ventilation1” are not significantly inferior (84.4%, $\beta = -.213$ for mic#1, $\beta = -.075$ for mic#2), see Fig. 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 intervenes. This is in agreement with another study [4], which reports that ambient noise (hospital ward, emergency room) had no effect on recognition accuracy.

Table 3 – Recognition rates in free text mode detailed per training type and sessions (noises)

Session	Free text mode	
	Training: silence (Free1) (%)	Training: ventilation1 (Free2) (%)
Silence	87.53	86.43
Ventilation1	86.34	87.09
Other seven sessions	74.09	77.15

The seven other types of background noises gave significantly lower recognition rates than the session with a silent background ($\beta \leq -.475$ down to -1.52 for mic#1, $\beta \leq -.056$ down to $-.58$ for mic#2, with mic#1 in silence as reference). In Table 1, differences between most noises are significant ($p < .001$), also when combined with the type of microphone. The limited impact, for the best microphone, of people talking in the background is encouraging.

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).

3.3.1. *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 and each of the nine types of background noises was produced for 1 min, with the same experimental setup. 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. ≤ .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 .025 ($N = 2$) for microphone 2 but reached .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.

3.4. *Participant differences*

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 ($p < .001$, 18.1 points, see Fig. 4). A previous study [13] reports inter-subject variabilities as large as 40 points (55–95% word accuracy) for 39 endocrinology authors.

3.5. *Test material*

The experiment has shown a very high inter-command variability of the recognition rate ($p < .001$ between many of them, even when taking a reference close to the mean). 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%.

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.

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, possibly because they are phonetically distinct.

3.6. *Additional training*

The experiment presented here has been done with minimal training. Max Manus reports that it requires 10 h 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 [4] reports that “*Accuracy improves with error correction by at least 5 percent over two weeks*”. Another more detailed study [14] (using IBM Via-Voice 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 6th 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 min of training.

3.7. *Redundant cross-matching validation*

Speech recognition in noisy environments is a long-standing problem, and many solutions have been tried [15]. 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 [16] that is using an alignment and voting module. Previous experiments [17] 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 7–12 points if the voting was perfect. Other experiments have combined multiple microphones [18] 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.

Table 4 summarises the results. 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

Table 4 – Recognition rates with cross-matching validations

	Mic. 1	Mic. 2	Best mic. (potential)	Best mic. (effective)
Command mode	78.33%	84.96%	86.79% (+1.83)	86.41%[b] (+1.45)
Free text mode	71.76%	82.41% 83.43%[a]	84.88% (+2.47)	N/A
Best mode[a] (potential)		96.30% (+11.34)	96.67%[b] (+11.71)	
Best mode[a] (effective)		N/A		

[a] Using free text mode trained with background noise.

[b] Using effective combination for command mode.

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.

3.7.1. *Discussion on cross-matching validation*

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 considerable potential in combining these two types of recognition.

3.8. *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): 1158 of 1272 words recognised (91.04%), Levenshtein word distance of 155, WRR = 87.41%.
- Microphone 2 (88.30% accuracy on 401 sentences): WRR = 88.60%.

Keeping in mind that the experiment was made in Danish and that enrolments were very short (about 15 min), it is possible to compare the above reported recognition rate obtained with free text mode with a previous study [19] evaluating continuous speech recognition in the medical domain (in English, enrolment in less than 60 min). In this study IBM ViaVoice 98 with General Medicine Vocabulary performed best (90.9–93% word accuracy) followed by the L&H Voice Xpress for Medicine, General Medicine Edition, version 1.2 (84.9–86.6%) and then Dragon Systems NaturallySpeaking Medical Suite, version 3.0 (84.8% to 14.1% to 85.9%). Another study [13] obtained an average of 84.5% word accuracy and another one [20] even reached 98% with one highly trained speaker in French and in a narrow medical field.

4. Descriptive statistics summary

To provide an overview, Table 5 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 Table 1 provides the statistical analysis results, Table 5 gives a less precise but perhaps more illustrative overview.

5. Discussion

5.1. *Participants*

The experiment would have been more realistic if participants had been medical staff. Undeniably, there were some medical words that 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.

5.2. *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

Table 5 – 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

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.

5.3. *Laboratory*

The reverberation observed in the small room where the experiment was conducted is known to affect speech intelligibility [21] 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.

5.4. *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” [2]. 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.

6. 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.

Summary points

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.

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

[1] A. Alapetite, V. 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, September 29–October 1, 2005, Chania, Crete, Greece; Section II on research and applications in the medical domain, pp. 189–196. ACM International Conference Proceeding Series, vol. 132. University of Athens, pp. 197–204.

[2] A. Zafar, B. Mamlin, S. Perkins, A.M. Belsito, J. Marc Overhage, C.J. McDonald, A simple error classification system for understanding sources of error in automatic speech recognition and human transcription, Int. J. Med. Informat. 73 (2004) 719–730, doi:10.1016/j.ijmedinf.2004.05.008.

[3] T. Giorgino, I. Azzini, C. Rognoni, S. Quaglini, M. Stefanelli, R. Gretter, D. Falavigna, Automated spoken dialog system for hypertensive patient home management, Int. J. Med. Informat. 74 (2005) 159–167, doi:10.1016/j.ijmedinf.2004.04.026.

[4] Atif Zafar, J. Marc Overhage, Clement J. McDonald, Continuous speech recognition for clinicians, J. Am. Med. Informat. Assoc. 6 (3) (1999) 195–204.

[5] E. Lombard, Le signe de l'élévation de la voix, Ann. Maladies Oreille, Larynx, Nez, Pharynx 31 (1911) 101–119.

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

[7] 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. Informat. Assoc. 2 (1) (Jan–Feb 1995) 46–57.

[8] 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.

[9] K.S. Anja, O.-S. 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 (1999). Abstract in the J. Acoust. Soc. Am. (1999) 105 (2) 1097. doi:10.1121/1.425143.

[10] C.P. Sobel, A generative phonology of Danish, Ph.D. Thesis, City University of New York, 1981.

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

[12] H.-G. Hirsch, D. Pearce, The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions, in: Proceedings of the ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium", September 18–20, Paris, France, 2000.

[13] D.N. Mohr, D.W. Turner, G.R. Pond, J.S. Kamath, K.B. De Vos, P.C. Carpenter, Speech recognition as a transcription aid: a randomized comparison with standard transcription, J. Am. Med. Informat. Assoc. 10 (1) (2003) 85–93, doi:10.1197/jamia.M1130.

[14] M.M. Al-Aynati, K.A. Chorneyko, Comparison of voice-automated transcription and human transcription in generating pathology reports, Arch. Pathol. Lab. Med. 127 (6) (2003) 721–725.

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

[16] G.F. Jonathan, A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER), in: Proceedings of the IEEE'1997 Workshop Automatic Speech Recognition and Understanding, 1997.

[17] M. Matsushita, H. Nishizaki, T. Utsuro, Y. Kodama, S. Nakagawa, Evaluating multiple LVCSR model combination in NTCIR-3 speech-driven web retrieval task, in: Proceedings of the Eurospeech'2003, the 8th European Conference on Speech Communication and Technology, 2003, pp. 1205–1208.

[18] C.Y.-K. Lai, P. Aarabi, Multiple-microphone time-varying filters for robust speech recognition, in: Proceedings of ICASSP'2004, International Conference on Acoustics, Speech, and Signal Processing, 2004.

[19] Eric G. Devine, Stephan A. Gaehde, Arthur C. Curtis, Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports, J. Am. Med. Informat. Assoc. 7 (5) (2000) 462–468.

[20] André Happe, Bruno Pouliquen, Anita Burgun, Marc Cuggia, Pierre Le Beux, Automatic concept extraction from spoken medical reports, Int. J. Med. Informat. 70 (2003) 255–263, doi:10.1016/S1386-5056(03)00055-8.

[21] Stanley A. Gelfand, Shlomo Silman, Effects of small room reverberation upon the recognition of some consonant features, J. Acoust. Soc. Am. 66 (1) (July 1979) 22–29, doi:10.1121/1.383075.