# SPEECH CONTROL IN SURGERY: A FIELD ANALYSIS AND STRATEGIES

*Björn Schuller*[1]*, Salman Can*[2]*, Hubertus Feussner*[3]
*Martin Wöllmer*[1]*, Dejan Arisc*[1]*, Benedikt Hörnler*[1]

[1]Institute for Human-Machine Communication,
[2]Workgroup MITI,
[3] Department of Surgery, Polyclinic rechts der Isar
Technische Universität München, Germany
schuller@tum.de

**ABSTRACT**

This work introduces a robot driven camera controlled by speech. The SIMIS database of 20 recordings of real life surgical operations serves as basis for analyses and noise modelling. To overcome low recognition performance due to high noise levels during operations, the vocabulary was chosen to be highly limited and multiple noise reduction methods have been investigated. We show that the use of feature enhancement techniques, such as Histogram Equalization or a Switching Linear Dynamic Model capturing the dynamics of speech show a remarkable improvement in recognition accuracy. Considering a severe condition of usage of the recognition system with all appearing noise types, the mean accuracy can be raised from 89.67 % to 91.16 % with SLDM, and to 95.50 % with HEQ enhancement.

***Index Terms*—** Speech recognition, Speech enhancement, Acoustic noise, Robustness, Biomedical equipment safety

## 1. INTRODUCTION

The opposition of laparoscopic surgery to open surgery points out several distinct benefits as reduced pain, shorter hospitalization, and quicker convalescence to the affected patient. However, the surgeon's direct visual control gets lost and the view of the operating field has to be displayed on a screen using a laparoscopic high-resolution camera. It is common practice that, during laparoscopic interventions the assistant surgeon holds the laparoscope for the operating surgeon and positions the scope following the surgeon's instructions. A result of this fact is an unstable and suboptimal camera view, because the telescope is sometimes aimed incorrectly and vibrates due to the assistant's hand trembling. In a long-lasting and complex intervention, considering the worst case, this can result in a patient injury since stress and fatigue start playing a major role. A significant step towards the solution of this problem is the introduction of a telemanipulator system for guiding the telescope, aiming to replace the assistant surgeon. Thereby, the design of a user-friendly and intelligent human-robot interface to control the telemanipulator plays an important role [1].

The majority of laparoscope positioning systems proposed so far are based on input devices such as joysticks, foot pedals, and similar human-robot interfaces. Although it improves the standard of work for the surgeon, he is faced with additional burdens since he already uses his hands or feet to control a variety of other surgical tools. Therefore, the implementation of a voice control interface is an effective approach to overcome these drawbacks since verbal instructions are natural for a human. There have been several laparoscope positioning systems that introduced voice control interfaces [1, 2, 3, 4]. However, these systems could not achieve the required accceptance since long reaction time, limited reliability, and a user dependent interface made its use inappropriate. Emotional factors derived from speech play a huge role in automatic speech recognition to perform an online emotional model adaptation and overcome typical losses arising from emotionally coloured speech. Basing on this fact we introduced the integration of social competence by acoustic emotion recognition [5]. Although robustness could be improved by integrating emotional factors the appearing background noises in the operation room environment result in an insufficient reliability.

Therefore, we developed a novel speech control interface for the newly designed SoloAssistTM (AktorMed, Barbing, Germany). The salient feature of this interface is the presence of an improved noise robustness towards the appearing background noises in medical operation room environments. This work investigates different feature enhancement algorithms, such as Cepstral Mean Subtraction (CMS) , Histogram Equalization (HEQ) [6], and a model based feature enhancement technique using a Switching Linear Dynamic Model (SLDM) as introduced in [7]. Hereby, the investigation took place based on real life recordings of minimal invasive surgeries in a medical operation room in the Clinic r. d. Isar, Munich, Germany.

The organization of this paper is as follows: first, the SIMIS database is introduced in sec. 2, next, stretegies to cope with noise are detailed in sec. 3, in sec. 4 we discuss experiments and results before concluding in sec. 5.

## 2. SIMIS DATABASE

To ensure the SIMIS (*Speech in Minimal Invasive Surgery*) database as used in [5] to supply sufficient data, i.e. to be valuable and reliable, additional 10 live surgical operations were recorded. This was done under the same conditions as in [5] to guarantee identical conditions. The most important part concerning this work was to capture all kinds of noise that appear in a real life operation room. In general, 6 to 10 persons reside in the operation room during a surgery consisting of the operating surgeon, 2 to 3 assistants, and 3 to 7 surgical nurses depending on the complexity of the surgical intervention. These facts play an important role, since they greatly influence the noise level during the operation. The type of surgeries that have been recorded were all minimal invasive operations where the SoloAssistTM positioning robot shall play a decisive role in the future. All recorded operations had a length of approximately one hour on average. The format used was 16 Bit, 16 kHz. The used headset AKG C 444 L has a cardioid type polar pattern and a speech

optimized frequency response. To avoid any additional burden to the operating surgeon the headset was chosen to be wireless with a AKG PT 40 sender and AKG SR 40 receiver.

The results of a semi-automated segmentation are illustrated in Table 1.

| **Time & turns** # type | **Total time** [m:s] | **Speech time** [m:s] | **Speech turns** [#] |
|---|---|---|---|
| 3x fundoplicatio | 211:04 | 56:27 | 1554 |
| 12 x gall | 740:11 | 149:26 | 5064 |
| 4x sigma wedge | 296:22 | 39:26 | 1919 |
| 1x stomach | 71:01 | 15:18 | 306 |
| **Total** | **1 318:38** | **311:45** | **8 843** |

**Table 1**. Semi-automated segmentation results for the SIMIS database giving the total time of each operation type and the corresponding speech time and turns, respectively.

Besides the comprehensive recordings of real life surgeries, the commands to control the SoloAssistTM robot by speech were recorded as well under exact same mic and room conditions, but without an ongoing operation. The robot as described in [5] can be completely controlled by the 15 prompts *soloassist*, *left*, *right*, *up*, *down*, *forward*, *backward*, *moveleft*, *moveright*, *moveup*, *movedown*, *moveforward*, *movebackward*, *stop*, and *quit*. The recordings of the 15 prompts, from now on referred to as keywords, were recorded from 5 speakers (24 to 54 years). Hereby, each of the 15 keywords was spoken 9 times by every speaker, resulting in 675 clean turns. Each speaker was advised to change the speaking style when speaking the prompts. The instruction was to speak every keyword three times in the following ways: *normal* as in everyday life, *faster*,but still well audible and recognizable, and *slower*.

The recordings that were done during the live operations needed to be annotated in a way the ASR engine can benefit the most from, i.e. train suitable models to ensure noise robustness. Table 1 shows that for every recorded operation, only sparse speech time (approximately 14 minutes per operation) exists. However, this speech was used in this work to model extraneous utterances and extraneous words that are not directed to the robot or any camera movements. Transcription was done on word level to receive the desired data. This has the great advantage, that speech used in the operation often recurs and as a result, the annotated turns strongly qualify to be used for a later garbage model.

Not less impact than the extraneous speech has the different appearing background noises during the operation. As the noise level is relatively high during a surgical intervention, it is not possible to model silence as it is done in conventional methods. Most of the recordings represent nonspeech data containing different noise types, thus it is an important task to describe the noise in a proper way. Four different noise types could be chosen that describe the appearing noise most suitable:

1. *Standard background noise (std.bkgrd.)*: This noise type is represented by permanently appearing noise provoked from different machines running during the surgical interventions, such as computers or artificial ventilation. In many cases this noise type can be considered as stationary.
2. *Instrument click noise (instr.click)*: As the name already states, this noise type consists of noise caused by different instruments during the surgery. This noise type is not to be underestimated since it also encloses noises that are produced by depositing surgical tools onto the table which is made of metal in every case, thus producing a relatively high energy noise.
3. *Background talk (bkgrd.talk)*: Noises produced by persons present during the operation, such as surgical assistants, nurses, and students. Usually an amount of 6 to 10 persons are present at one time, thus, the noise level can get high during stress situations.
4. *Stressed breath or cough (str.breath)*: This noise type is characterized by loud breath noises or coughing by the surgeon wearing the microphone or also by assistants standing close. According to the gained experience this was one of the most significant noise types.

Table 2 shows the statistical distribution of the different noise types over the operation recordings. Herein, the total number of occurences and a mean value for one operation can be seen. As the

| **Statistics** type | **Turns** [#] | **∅ Turns/OP** [#] | **Distrib.** [%] | **Time** [m:s] |
|---|---|---|---|---|
| *std.bkgrd.* | 19 855 | 993 | 57.9 | 583:07 |
| *instr.click* | 7 839 | 392 | 22.9 | 230:13 |
| *bkgrd.talk* | 3 015 | 151 | 8.8 | 88:31 |
| *str.breath* | 3 575 | 179 | 10.4 | 105:02 |
| **Total** | **34 284** | **1 715** | **100** | **1 006:53** |

**Table 2**. Statistical distribution of different noise types over operation recordings giving number of turns, the average number of turns per operation, the distribution in percent, and the total time.

recognizer demands for utmost precision, it is inevitable to verify the performance of the ASR engine by means of different testsets. Therefore, different testsets were conceived and created on the basis of superposing the clean keyword turns one to one by different noise types and levels obtained from the recordings. In the following, each testset that was created is explained. For each testset property, every keyword was taken resulting in 675 noisy turns for each noise condition.

The first three different superpositioning conditions were strictly related to the resulting Signal-To-Noise Ratio (SNR). It represents a term for the power ratio between a signal and the background noise and was calculated by

$$SNR = 10 \log_{10} \left( \frac{P_{keyword}}{P_{noise}} \right) dB \quad (1)$$

whereas $P_{keyword}$ denotes the average power of the keyword turn and $P_{noise}$ the average power of the noise turn. Based on Equation 1 three different SNR levels (high, medium, low) testsets were created: with a resulting SNR that lies between -30 and -5 dB (mean -9.4 dB), between -5 and 15 dB (mean 2.9 dB), and between 15 and 35 dB (mean 13.3 dB). The mean power was calculated for every keyword turn and a list with all available nonspeech turns of the recorded operations was checked for the mean energy of each turn (calculated over the length of the recent keyword turn) that results in

the desired SNR range. The list was accessed randomly to avoid using same noises multiple times. Due to the fact that the SNR level is the only information to measure how the ASR engine performs under noisy conditions so far, further four different testsets were conceived. Hereby, the type of the noise appearing during the surgical operation was taken into account. This way, the testing results will give more information about what noise is most detrimental to the recognition performance. The procedure performed to superpose the clean turns was the same as above, except that the list with all nonspeech turns was adjusted to gain four additional testsets with turns superposed with each noise type described in Table 2. These lists were again accessed randomly to avoid having multiple turns superposed with the same noise turn. The resulting testsets are the std.bkgrd. set (mean 11.9 dB), instr. click set (mean 5.5 dB), bkgrd. talk set (mean 7.4 dB), and str. breath set (mean -3.6 dB), again each having 675 superposed turns.

## 3. NOISE ROBUSTNESS

Three different feature enhancement techniques are considered in this work selected basing on our experiences from [8]: simple Cepstral Mean Subtraction (CMS) , well known Histogram Equalization (HEQ) [6] where the histogram of a feature is mapped onto a reference histogram, and a Switching Linear Dynamic Model (SLDM) as introduced in [7] [1]. Unlike CMS and HEQ, the feature enhancement is hereby realized by models for speech and noise.

The modeling of noise is done by a simple Linear Dynamic Model (LDM) obeying the system equation

$$x_t = Ax_{t-1} + b + g_t. \tag{2}$$

The Matrix $A$ and the vector $b$ hereby characterize how the noise process evolves over time while $g_t$ is a zero-mean Gaussian noise source driving the system.

Alternatively the LDM is defined by

$$p(x_t|x_{t-1}) = \mathcal{N}(x_t; Ax_{t-1} + b, C) \tag{3}$$

$$p(x_1^T) = p(x_1) \prod_{t=2}^{T} p(x_t|x_{t-1}) \tag{4}$$

whereas $\mathcal{N}(x_t; Ax_{t-1} + b, C)$ represents a multivariate Gaussian with mean vector $Ax_{t-1} + b$ and covariance matrix $C$, while $T$ is to be understood as the input sequence length.

A SLDM is used to model clean speech. Hereby the matrix $A$ and the vector $b$ depend on a hidden state variable $s_t$ at each time $t$ so that the SLDM can be described as

$$x_t = A(s_t)x_{t-1} + b(s_t) + g_t. \tag{5}$$

This type of model is an appropriate solution to describe the evolution of time-varying systems, e.g. the evolution of speech features over time. Figure 1 shows the graphical representation of the SLDM used to model clean speech.

Analogous to the LDM, the SLDM can be alternatively described with the equations

$$p(x_t, s_t|x_{t-1}) = \mathcal{N}(x_t; A(s_t)x_{t-1} + b(s_t), C(s_t)) \cdot p(s_t) \tag{6}$$

$$p(x_1^T, s_1^T) = p(x_1, s_1) \prod_{t=2}^{T} p(x_t, s_t|x_{t-1}). \tag{7}$$

The parameters $A(s)$, $b(s)$, and $C(s)$ are trained by using standard EM algorithm techniques (see [7]).

[1] We would like to thank Jasha Droppo for providing SLDM binaries.

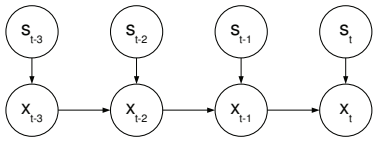


**Fig. 1**. SLDM used to model speech.

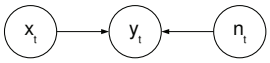


**Fig. 2**. Observation model to relate noisy observations to hidden speech and noise features.

The observation model illustrated in Figure 2 assumes that speech $x_t$ and noise $n_t$ perform a linear mix in the time domain resulting in a non-linear mix considering the cepstral feature space.

The SLDM modeling clean speech, as it shall be used, bears a major problem regarding the computation of a posterior. As the SLDM switches between the hidden states, i.e. can take any of $S$ hidden states at each single time step, this calculation becomes intractable since there are $S^T$ possible state sequences for an amount of $T$ speech frames. In the case that the whole state sequence is known, the problem would be easily solvable since the SLDM would reduce to a time-varying LDM with one single mixture component as a posterior. Since this is not the case in the SLDM because the state sequence is unknown, a total of $S^T$ mixture components, meaning one component for each possible sequence, exist.

To overcome this intractable calculation, the Generalized Pseudo-Bayesian (GPB) algorithm is applied to reduce the search space size. The approximations that are made by GPB base on the assumption that it is not of importance to keep track of distinct state histories, i.e. whose differences occur more than $r$ frames in the past. As a result the posteriors can be reduced from $S^T$ to $S^r$ (choosing $r \ll T$) which is a huge improvement in calculation performance that can be illustrated by setting $r = 1$ meaning that the posterior is reduced to $S$ Gaussian components at each time step.

## 4. EXPERIMENTS AND RESULTS

Prior to experiments with respect to noise a baseline system was optimized: the keyword models performed the best with 16 state models and 3 Gaussian mixtures per state. As a further achievement in performance the keyword models could be reduced to 9 models plus 1 *move* model since multiple keywords were similar with the only difference of a prepended *move* word. The separate *move* model showed the best performance with 12 states and 3 Gaussian mixtures per state.

After extensive investigation the model topology for the additional garbage model was kept at a 10 state model with 3 Gaussian mixture components per state. It was trained on speech recorded during the operation which was of non-keyword type. Additionally a silence model with a *tee-state* short-pause model was created and trained on nonspeech turns from the operation recordings.

The use of three different noise reduction methods has been described above. Table 3 summarizes all the results obtained with the different techniques. The column **MC** hereby represents matched conditions, meaning that the recognizer was trained according to the noise type to be tested on. Even though this mostly results in the best performance, a use in a real life application is not quite suitable as the occurrence of the noise is not easily predictable.

| Acc. [%] | MFC | PLP | NT* | CMS | HEQ | SDM |
|---|---|---|---|---|---|---|
| clean | **98.53** | 98.16 | 97.06 | 87.50 | 97.43 | 92.96 |
| high SNR | 92.59 | 92.96 | 97.06 | 81.99 | **95.96** | 92.52 |
| med. SNR | 90.49 | 90.49 | 95.22 | 79.78 | **95.22** | 90.15 |
| low SNR | 89.34 | 89.63 | 94.12 | 79.04 | **93.75** | 88.56 |
| std.bkgrd | 91.65 | 92.65 | 95.59 | 86.40 | **97.06** | **92.11** |
| instr.klick | 89.34 | 89.63 | 95.96 | 81.62 | **94.12** | **92.22** |
| bkgrd.talk | 89.71 | 89.71 | 94.85 | 80.51 | **94.41** | 88.42 |
| str.breath | 79.41 | 79.62 | 90.07 | 77.57 | **90.81** | **85.84** |
| **mean** | 90.13 | 90.36 | 94.99 | 81.80 | **94.84** | **90.35** |
| **W. mean** | 89.67 | 90.34 | 95.03 | 83.87 | **95.50** | **91.16** |

**Table 3**. Accuracies for different noise reduction methods and noise types (cf. text): optimal model topology, each, and training on clean speech except for training with noisy speech (NT*): general noisy training data (upper half), and matched conditions wrt. the exact noise type (lower half). Test and training are always disjunctive. Note that noisy training data reduces clean speech recognition performance. SLDM (SDM in the table) use 32 states and the first and last 10 frames per phrase for noise estimation. Weighted mean (W. mean) incorporates distribution of the four located noise types. MFC abbreviates MFCC.

Under noisy conditions the recognizer performance strongly degrades. This is especially the case for a high average SNR level and an additive stressed breath or cough noise. CMS showed to be definitely not suitable since it seems to subtract information that is needed for recognition. HEQ and SLDM showed the best performance although HEQ seems to be the enhancement technique to choose since it outperforms all others. Given the percentage values of the noise occurrences in Table 2, a weighted accuracy value ($w.Accuracy$) was additionally calculated for each of the accuracies obtained with the four different testsets. On the basis of the weighted mean values the fact that HEQ and SLDM represent the most suitable feature enhancement technique is pointed out explicitly. Note that HEQ even surpasses matched conditions training in the case of the noise type distribution weighted mean.

## 5. CONCLUSION

In this work a speech-based camera control in minimal invasive surgery with emphasis on noise robustness has been implemented. Guaranteeing the supply of noise robustness, multiple approaches have been applied to the recorded data and its results are presented in this work. To describe the appearing background noises adequately, the recorded operations have been annotated where four main noise types could be distinguished: standard background noise, instrument click noise, background talk noise, and breath/cough noise. With the optimal model topology an accuracy of 98.53 % can be achieved using a clean training and test set with MFCCs serving as features. The accuracy strongly degrades using the noisy testsets where the keywords are superposed with the noise types appearing in the operation room. Changing the feature set from MFCC to PLP showed only little to no improvement, wheras different speech enhancement techniques showed great improvement except for Cepstral Mean Subtraction where the testing performed worse constantly. HEQ thereby showed best improvement from 89.67 % to 95.50 % in the case of noise. SLDM prove only second choice here opposing our experience from [8] in the automotive environment and the *Consonant Challenge 2008*. Based on the gained knowledge about the noise level and influence, a live recognizer working in open-microphone mode was constructed with the optimal model topology obtained. This recognizer is communicating with the surgical robot over the UDP protocol realizing the movement control by speech.

Future work concerning noise reduction shall definitely base on combining various methods. Another modification is the modeling of noise: when using the SLDM for feature enhancement in this work noise is modeled by a simple Gaussian mixture component. A different modeling of noise, especially a different one for the specific noise types occurring, may have interesting impact on the performance as one has to deal with non-stationary noise.

## 6. REFERENCES


[1] K. Seong-Young, J. Kim, K. Dong-Soo, and L. Woo-Jung, "Intelligent interaction between surgeon and laparoscopic assistant robot system," in *IEEE International Workshop on Robot and Human Interactive Communication*, 2005, pp. 60–65.

[2] M.E. Allaf, S.V. Jackman, P.G. Schulam, J.A. Cadeddu, B.R. Lee, R.G. Moore, and L.R. Kavoussi, "Laparoscopic Visual Field. Voice vs Foot Pedal Interfaces for Control of the AESOP Robot," in *Surg. Endosc. 12 (12)*, 1998, pp. 1415–1418.

[3] G. F. Buess, A. Arezzo, M.O. Schurr, F. Ulmer, H. Fisher, L. Gumb, T. Testa, and C. Nobman, "A New Remote-Controlled Endoscope Positioning System for Endoscopic Solo Surgery. The FIPS Endoarm," in *Surg. Endosc. 14 (4)*, 2000, pp. 395–399.

[4] V.F. Munoz, C. Vara-Thorbeck, J.G. DeGabriel, J.F. Lozano, E. Sanchez-Badajoz, A. Garcia-Cerezo, R. Toscano, and A. Jimenez-Garrido, "A medical robotic assistant for minimally invasive surgery," in *IEEE International Conference on Robotics and Automation*, 2000, vol. 3, pp. 2901–2906.

[5] B. Schuller, G. Rigoll, S. Can, and H. Feussner, "Emotion Sensitive Speech Control for Human-Robot Interaction in Minimal Invasive Surgery," in *Proc. 2008 IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2008)*, Munich, Germany, 2008, pp. 453–458, IEEE.

[6] A. de la Torre, A.M. Peinado, J.C. Segura, J.L. Perez-Cordoba, M.C. Benitez, and A.J. Rubio, "Histogram equalization of speech representation for robust speech recognition," in *IEEE Transactions on Speech and Audio Processing*, 2005, vol. 13, pp. 355–366.

[7] J. Droppo and A. Acero, "Noise robust speech recognition with a switching linear dynamic model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, vol. 1.

[8] B. Schuller, M. Wöllmer, T. Moosmayr, and G. Rigoll, "Speech recognition in noisy environments using a switching linear dynamic model for feature enhancement," in *Proc. 9th Interspeech 2008, Brisbane, Australia*. 2008, ISCA.