Robust Speech Recognition for Human-Robot Interaction in Minimal Invasive Surgery

Björn Schuller[1], *Christoph Scheuermann*[1], *Salman Can*[2], *Hubertus Feussner*[2,3], *Gerhard Rigoll*[1]

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

1 Introduction

Laparoscopic surgery as opposed to open surgery offers distinct benefits as reduced pain, shorter hospitality, and quicker convalescence to the patients. During laparoscopic interventions, a camera assistant usually holds the laparoscope for the surgeon and positions the scope according to the surgeon's instructions. The camera view may be suboptimal and unstable, because the telescope is sometimes aimed incorrectly and vibrates due to the assistant's hand trembling. The introduction of a telemanipulator system for guiding the telescope, in aim to replace the human assistant, is a significant step toward the solution of this problem. Most laparoscope positioning systems proposed so far use input devices such as joysticks, foot pedals, and similar human-robot interfaces. However, this type of interfaces poses additional burden on surgeons. Implementation of a voice control interface is an effective approach to overcome these drawbacks since the verbal instructions are natural for a human, and the use of neither hands nor feet is required in controlling the laparoscope. Voice control was introduced for several laparoscope positioning systems (c.f. e.g. [4]). However, due to long reaction time, limited reliability, and a user dependent interface these systems could not achieve the required acceptance. To improve this fact we introduced integration of social competence by acoustic emotion recognition in [1]. Although robustness could thereby be improved, appearing background noises in the operation room environment still result in insufficient reliability.

From: Proc. RBC-BME 2008, Moscow, Russia
MIET, Zelenograd, ISBN 978-5-7256-0506-8

Therefore a novel speech control interface providing improved noise robustness in medical room environments achieved by using a Switching Linear Dynamic Model for the newly designed and produced laparoscope positioning system SoloAssistTM (AktorMed, Barbing, Germany) was developed. This paper emphasises on benefits obtained with model and feature enhancement to overcome unwanted movements of the tele-manipulator.

2 Laparoscope Positioning System

The laparoscope positioning system SoloAssistTM is the first mechatronic device with a fluid actuation system allowing enhanced power transmission and positioning compared to other technologies. Integrated pressure sensors for each actuation permit pushing the system manually at any time out of the operating field, which is a significant feature for patient safety. It resembles a human arm with an extended working range of 360° radius in both directions of movement, an inclination of up to 80° and penetration depth of maximal 250 mm depending on the current telescope length. Hereby the direction terms correspond to the speech interface commands for controlling the tele-manipulator by speech. A joystick integrated on a laparoscopic handhold with exchangeable instruments, a small hand panel, and a foot pedal are used input devices so far.

3 Automatic Speech Recognition

First, the *Speech In Minimal Invasive Surgery* (SIMIS) database including background noises and often being very emotional within the real life situation was recorded as introduced in detail in [1]. SIMIS covers 20 live surgeries with both headset and room microphone, both active condenser, in an operation room of the university hospital *rechts der Isar* with normally one main surgeon and 6 to 10 surgical assistants in 16 bit, 16 kHz of different minimal invasive surgeries as stomach and gall operations. These were automatically segmented into speech turns. Each surgery took 36-80 min; speech time from 5-17 min. The number of segments reached from 159 to 523.

Additionally, the system controlling commands were recorded of 5 male speakers saying each of 15 keywords 9 times resulting in 675 clean turns. These were artificially one-to-one interfered with 5 types of noisy non-speech SIMIS

recordings from the same headset providing highly similar conditions. Thus, 3375 noisy turns are available for testing and training. As a result of the superposition 21% of the noisy test utterances have an SNR below 10 dB, 12% are between 10-35 dB and the rest is beyond 35 dB.

The tele-manipulator possesses two moving modes: a short precise, and a long distance move. This fact leads to a highly limited vocabulary of 15 keywords, where the directions with a prepended move command represent the long distance move: *left*, *right*, *up*, *down*, *forward*, *backward*, *moveleft*, *moveright*, *moveup*, *movedown*, *moveforward*, *movebackward*, *stop*, *quit*. For every keyword a Hidden Markov Model (HMM) consisting of 8 states and 3 mixtures per state was chosen or the prepended *move* - command was treated as an extra model (see section 4) to reduce substitution errors, since, e.g. the commands *left* and *moveleft*, are subject to confusion. Furthermore, a word-based garbage model consisting of 10 states and 16 mixtures per state to exclude extraneous speech was trained on the SIMIS-recordings. Additionally the silence and the short pause model which is operating as a tee-state model sharing the middle state of the silence model, was trained on non-speech-recordings of minimal-invasive operations. The grammar is chosen as context free word-loop solution. MFCC 0-12 plus δ and $\delta\delta$ serve as features.

The model used in this work to improve robustness is based on modeling speech and noise as applied in [2]. Together with a model of how speech and noise produce the noisy observations, these models are intended to enhance the noisy speech features. In [3] a Switching Linear Dynamic Model (SLDM) is used to capture the dynamics of clean speech. Similar to HMM based approaches to model clean speech, the SLDM assumes that the signal passes through various states. Conditioned on the state sequence the SLDM furthermore enforces a continuous state transition in the feature space.

The modeling of noise is realised by using a simple Linear Dynamic Model (LDM) obeying the following system equation:

$$x_t = Ax_{t-1} + b + v_t \tag{1}$$

This LDM can be seen as simple multivariate Gaussian and corresponds to exclusively the lower line in Figure 1.

The modeling of speech is realised by a more complex dynamic model which also includes a hidden state variable s_t at each time t. Now A and b

depend on the state variable s_t:

$$x_t = A(s_t)x_{t-1} + b(s_t) + v_t \tag{2}$$

As can be seen in Figure 1, every possible state sequence describes an LDM which is non-stationary due to A and b changing over time. Time-varying systems like the evolution of speech features over time can be described adequately by such models. Hereby it is assumed that there are time dependencies among the continuous variables x_t, but not among the discrete state variables s_t.

Coventional EM techniques are used for training throughout.

A relationship of how speech and noise produce the noisy observations is obtained by a zero observation model with SNR interference which assumes that speech and noise mix linearly in the time domain corresponding to a non-linear mixing in the cepstral domain [3].

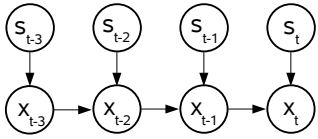

Figure 1: Switching Linear Dynamic Model for speech

4 Experiments and Discussion

To obtain performance on noisy speech we use the 15 keywords described in section 3. The evaluation strategy chosen is a 70/30 - training/test split. Table 1 shows word accuracies for two different model topologies. These are represented by identical models for each keyword and an extra model for the word *move* that can be prepended to each moving direction. For feature enhancement, a global speech SLDM of 32 hidden states was trained whilst the LDM modeling stationary noise was trained for each noisy test sequence using the first and last 10 frames of the utterance. As can be seen in table 1, SLDM slightly outperformed the best performance when using clean training with noisy test utterances. Matched conditions have to be interpreted as upper bench-mark, as they make use of full knowledge about the noise, which cannot easily be provided in a real use-case.

Topology	*clean/clean*	*noisy/noisy*	*clean/noisy*	*clean/noisy(SLDM)*
constant	89.63%	83.83%	77.08%	**78.51%**
sep. move	94.89%	94.02%	90.62%	**90.80%**

Table 1: Accuracies for permutations train/test clean or noisy: constant model parameters (upper row); separated model for the keyword *move* (lower row)

The presented results distinctively show that the use of SLDM increases recognition performance, but has its limits, since in a real life operation room one has to deal with non-stationary noise. Also, comparably lower overall SNR levels lead to lower benefit as e.g. in car noise environment (c.f. [2]).

Future work will investigate the benefit of the working prototype in a long term usability study in the operation room. Furthermore, the ASR shall be augmented by further noise reduction methods and an improved garbage model to reject extraneous speech shall be introduced.

References

[1] Schuller, B., Rigoll, G., Can, S., Feussner, H.: Emotion Sensitive Speech Control for Human-Robot Interaction in Minimal Invasive Surgery, Proc. 2008 IEEE RO-MAN 2008, Munich, Germany, 2008.

[2] Schuller, B., Wöllmer, M., Moosmayr, T., Ruske, G., Rigoll, G.: Switching Linear Dynamic Models for Noise Robust In-Car Speech Recognition, Proc. DAGM 2008, DAGM, Springer LNCS, Munich, Germany, 2008.

[3] Droppo, J., Acero, A.: Noise robust speech recognition with a switching linear dynamic model. Proc. ICASSP 2004, Montreal, Canada, 2004.

[4] Munoz, V.F., Vara-Thorbeck, C., DeGabriel, J.G., Lozano, J.F., Sanchez-Badajoz, E., Garcia-Cerezo, A., Toscano, R., Jimenez-Garrido, A.: A Medical Robotic Assistant for Minimally Invasive Surgery. Proc. 2000 IEEE Int. Conf. Robotic Automat., San Francisco, CA, pp. 2901-2906, Apr 2000.