Emotional factors in speech based human-machine interaction in the operating room

Salman Can[1], Björn Schuller[2], Michael Kranzfelder[1,3], Hubertus Feussner[1,3]

[1] Research Group MITI, Klinikum rechts der Isar der Technischen Universität München, Germany

[2] Institute for Human-Machine Communication, Technische Universität München, Germany

[3] Klinik und Poliklinik des Klinikums rechts der Isar der Technischen Universität München, Germany

Introduction

Laparoscope positioning systems are introduced to minor access surgery in aim to augment the quality of the operation by eliminating the disadvantages of manual telescope guiding by an assistant surgeon. A user-friendly design of the human-machine interface to control such manipulators plays an important role. Implementation of an intuitive and hands free voice control interface would offer a promising solution. However, speech controlled systems proposed so far did not get acceptance due to too long reaction time, reliability and user dependent interface [1,2]. Two decisive factors for a robust speech interface are the consideration of noisy environment in the operating room and the emotionally colored speech commands [3]. We therefore evaluated the emotional effects on speech commands during laparoscopic interventions to estimate the impact on speech control interfaces.

Material and Methods

For determination of the emotional coloring and its influence to the speech interface we recorded a total of 29 live laparoscopic surgeries at the Klinikum rechts der Isar der TU München. These records were performed under real conditions with noisy environment and unexpected influences. An AKG C 444 L headset was used for the records of seven different surgeons. The recorded operations such as cholecystectomy and fundoplication last between 30 minutes and three hours. These speech records were stored in waveform with a sample rate of 16 kHz and 16 bit per sample. In order to test and train our emotion recognition [4], the “Speech In Minimally Invasive Surgery” (SIMIS) database was created after automatic segmentation, transcription and emotional labeling within five classes for each segment (cf. Figure 1). The recordings are segmented into speech and non-speech patterns, where the latter is classified into background noise, instrument noise, background talk and breath or cough. For the training and testing of the speech recognizer, the labeled emotions were finally classified into happy as positive emotions, angry and impatient as negative emotions and neutral and confused as neutral emotions.

Results

The distribution of speech turns among emotions is shown in Figure 1. Neutral is by far the most common emotion, making up over two-thirds of the total duration. The remaining classes lie between 5 and 10%, with happy being the most common one among them. Angry and impatient are regarded as negative emotions and constitute 14.83% of the speech.

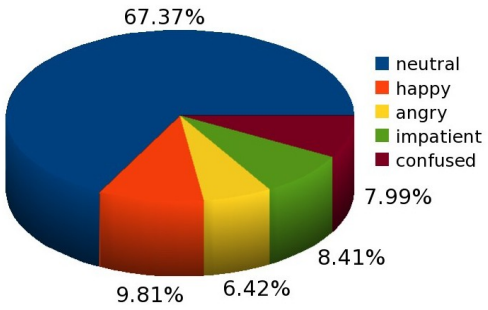

Figure 1: Distribution of emotion classes among speech

It can be assumed that different surgeons express their emotions in different manners. Two series of experiments were accomplished to take this into account. In a first experiment a single surgeon was recorded during 20 operations whereas in the second experiment seven different surgeons were recorded during at least one operation. The classifier’s performance is measured by the values

From: IJCARS, Vol. 5 (Suppl. 1) 2010

- **RR:** Overall recognition rate (Number of correctly classified cases divided by the total number of cases or weighted average).
- **CL:** Mean of the class-wise computed recognition rate (i.e. mean along the diagonal of the confusion matrix in percent or unweighted average).
- **F_1:** Uniformly weighted harmonic mean of RR and CL: 2*CL*RL/(CL+RL).

As classifier for these experiments a Support Vector Machine (SVM) with sequential minimal optimization learning was chosen.

For the first experiment the recognizer was set to separate negative from non-negative speech turns, meaning that neutral and positive turns were clustered together. Since in the real-life scenario the recognizer will not be trained with data from the same operation it is tested on a leave-four-operations-out cross-validation with 5 cycles. This means that speech turns from 16 operations are used for training, while the Support Vector Machine classifier based on 37 (contours as pitch or formants) x 2 (including derivatives) x 19 (functional as mean) acoustic features (cf. [4]) is tested on the data of the remaining four recordings. As result, the RR of the recorded operations varies between 69.43% and 81.69% (μ=75.38%, σ=4.45%), the CL varies between 62.27% and 66.10% (μ=64.27%, σ=1.60%) and thus the F_1 varies between 65.65% and 71.18% (μ=69.33%, σ=2.16%).

For the second experiment the recognizer was again set to separate negative from non-negative speech turns. Leave-one-speaker-out cross-validation was used to evaluate the performance, i.e. the data of every speaker was tested individually with a model trained on the remainder. As result, the RR varies between 70.45% and 94.87% (μ=83.84%, σ=8.31%), the CL varies between 57.88% and 76.42% (μ=66.89%, σ=7.07%) and thus the F_1 varies between 69.31% and 81.03% (μ=73.93%, σ=4.15%).

Conclusion

The experiments showed that it is well solvable to distinguish negative from non-negative speech, if the data for testing and training both stems from the same surgeon. Recognition rates as high as 81.69% were achieved for this case and provide a reliable separation. Further, this information can be used to improve the actual speech recognition for the robot control [5].

A paramount goal for future work in this area should be to further extend the SIMIS database, specifically recording additional operations from surgeons that currently contribute only one or two interventions. This would make it possible to draw more significant conclusions on the subject of speaker independent emotion recognition.

A speech-based camera control system is a possible application of the acoustic emotion recognition. As a further step the recognizer will be implemented into the newly designed speech controlled laparoscope positioning system SoloAssist (AktorMed, Germany) to optimize the human-machine interaction.

References

[1] Allaf, M.E., Jackman, S.V., Schulam, P.G., Cadeddu, J.A., Lee, B.R., Moore, R.G., Kavoussi, L.R.: Laparoscopic Visual Field. Voice vs Foot Pedal Interfaces for Control of the AESOP Robot. In: Surg. Endosc. 12 (12) (Dec 1998) 1415-1418

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

[3] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, and C. Cox, “ASR for emotional speech: clarifying the issues and enhancing performance,” Neural Networks, vol. 18, no. 4, pp. 437–444, 2005.

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

[5] Steidl, S., Batliner, A., Seppi, D., Schuller, B.: On the Impact of Children's Emotional Speech on Acoustic and Language Models, EURASIP Journal on Audio, Speech, and Music Processing (JASMP), Special Issue on "Atypical Speech", Article ID 783954, 14 pages, 2010.