Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, Technische Universität München, Munich, Germany, August 1-3, 2008

Emotion Sensitive Speech Control for Human-Robot Interaction in Minimal Invasive Surgery

Björn Schuller, Gerhard Rigoll
Institute for Human-Machine Communication
Technische Universität München
D-80333 München, Germany
schuller@tum.de

Salman Can, Hubertus Feussner
Research Group MITI, Surgical and Polyclinic
Klinikum r.d. Isar der TU München
D-81675 München, Germany
salman.can@tum.de

Abstract— **Minimal Invasive Surgery demands for utmost precise and reliable camera control to prevent any harm to the patient during operations. We therefore introduce a robot-driven camera that can be controlled either manually by a joystick, or by speech to ensure free hands and feet, and reduced cognitive workload of the surgeon. Speech control is chosen as simple, yet highly robust command and control application. However, due to high stress, and partially fatigue, emotional factors can play a life decisive role in the operational situation. As any misunderstanding of the surgeon's intent can easily lead to patient injuries by mis-movement of the camera, emotional factors are integrated in the human-robot interaction. In this work we therefore discuss the recording of a 3,035 turns database of spontaneous emotional speech in real life surgical operations. Known to be a challenge, we employ a high dimensional acoustic feature space, and subset optimization for recognition of positive versus negative emotion for interaction adaptation, surgeon self-monitoring, and potential adaptation of acoustic models within speech recognition. Promising 75.5% mean accuracy can be reported in a cross-operation recognition task given the severe condition of usage in real medical operations.**

I. Introduction

Laparoscopic surgery as opposed to open surgery offers distinct benefits as reduced pain, shorter hospitality, and quicker convalescence to the patients. However, the surgeon loses direct visual control so that the view of the operating field has to be displayed on a screen using a laparoscopic camera. During laparoscopic interventions, a camera assistant usually holds the laparoscope for the surgeon and positions the scope according to the surgeon's instructions. Such kind of operation is inefficient for the surgeon since commands are often interpreted and executed erroneously by the assistant. The camera view may be suboptimal and unstable, because the telescope is sometimes aimed incorrectly and vibrates due to the assistant's hand trembling. For an acceptable control a certain amount of experience from the assistant, and a mutual surgeon-assistant understanding are necessary but usually difficult to obtain. 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. A user-friendly design of the human-robot interface to control the telemanipulator thereby plays an important role in this step.

Nishikawa et al. published another user interface solution using the real-time visual tracking of a surgeon's face to guide the laparoscope [7]. Fast reaction time, high positioning accuracy, and easy and intuitive camera guidance are positive outcomes of an experimental study. However, the surgeon felt a little fatigue in the cervix from a lot of rolling face motions [7].

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. Furthermore, he already uses his hands or feet to control a variety of other surgical tools [5], [14]. 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. Up to now, the voice control interface was introduced for several laparoscope positioning systems [1], [4], [6]. However, due to long reaction time, limited reliability, and a user dependent interface these systems could not achieve the required acceptance. This is in particular true, as the emotional factor throughout operations has been widely ignored. We therefore developed a novel speech control interface for the newly designed and produced laparoscope positioning robot SoloAssistTM (AktorMed, Barbing, Germany).

The specialty of this interface is it's integration of social competence by acoustic emotion recognition. In the case of a confused or angry surgeon the interface initializes a security callback dialog to certify that the understood camera direction is correct. Likewise accidental patient injury by wrong movements shall be reduced. Further, this provides an affective (self-)profile of the surgeon for his own analysis or to enable automatic safety-alert functions. Finally, it allows for acoustic model (AM) adaptation within the speech recognizer to improve on robustness: Schuller et al. demonstrated effectiveness of online emotional model adaptation to overcome typical losses arising from emotionally coloured speech [12].

The paper is structured as follows: in section II the robot for the camera control is introduced. Next, in section III requirements and realization for Automatic Speech Recognition (ASR) are discussed. The acoustic features are explained in section IV, and their reduction to relevant ones in section V. In section VI we describe the recording of the SIMIS database of 3,035 speech turns of real operations and spontaneous emotions for tests of the emotion recognition

978-1-4244-2213-5/08/$25.00 ©2008 IEEE

TABLE I

DEFINITION OF THE SPEECH INTERFACE COMMANDS FOR THE SOLOASSISTTM AND INFLUENCE ON CARTESIAN AND INVARIANT POINT CONTROL. *cw.* AND *ccw.* ABBREVIATE CLOCK-, AND COUNTERCLOCKWISE, *rot.* ROTATION, RESPECTIVELY.

Command	left	right	forward	backward	down	up
Cartesian	x-	x+	y-	y+	z-	z+
Inv. point	tilt-	tilt+	cw. rot.	ccw. rot.	zoom+	zoom-

performance in detail. Finally, we provide results in section VII, and draw conclusions in section VIII.

II. LAPAROSCOPE POSITIONING ROBOT

The laparoscope positioning robot 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. Implementation of nonmetallic materials as carbon fiber in the upper part of the system with low level artifacts allows for additional X-Ray applications (c.f. Fig. 1). The SoloAssistTM, resembling to a human arm, has an extended working range with 360 degrees radius in both directions of movement, an inclination of up to 80 degrees, and penetration depth of maximal 250 mm depending on the current telescope length. The system, with a weight of 18 kg, is simple to dock in various rail positions, and can be dismantled quickly and easily. A joystick integrated on a laparoscopic handhold with exchangeable instruments, a small hand panel, and a foot pedal are used input devices so far. For an optimal positioning, the SoloAssistTM robot is calibrated at the trocar point, which serves as a pivot. The defined invariant point allows the calculation of individual axial movements for tilting the telescope and performing circular motion. Furthermore, automatic leveling of the sight facility on the monitor, regardless of the telescope position, is permitted.

Fig. 1. The camera positioning robot SoloAssistTM (AktorMed, Barbing, Germany)

There are two control modes for the SoloAssistTM compared to other laparoscope positioning systems: one is the common Cartesian control (x, y, z), which is usually used to set the trocar point, and the other is the invariant point control, where the x and y axes are replaced by tilting and circular motion referring to the invariant point, and the z axis corresponds to the zoom in and out function of the telescope. Accordingly, the instructions as depicted in Table I were defined for the speech control.

The SoloAssistTM runs on a windows operating system and the visual programming environment ICONNECT (Micro-Epsilon, Ortenburg, Germany). On a separate computer the received speech commands from the headset are processed by passing through the speech recognition software as described in section III. Afterwards the filtered direction commands are transmit to the SoloAssistTM control. We chose the UDP network protocol for fast data transfer between the two computers. For safety reasons during the evaluation, the joystick received higher priority, and the ICONNECT feedbacks the information after receiving the direction commands. The according information flow is visualized in Fig. 2.

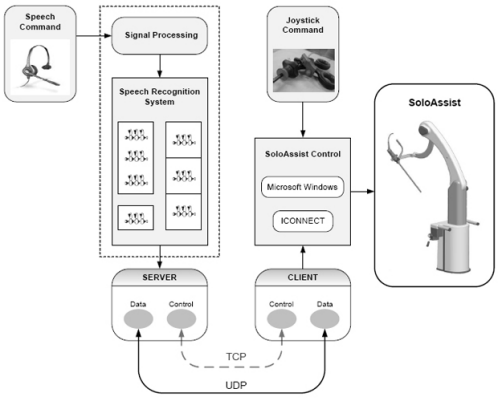

Fig. 2. Speech control interface for the SoloAssist

III. SPEECH CONTROL

A number of requirements must be satisfied as the camera is held by a robot as described during invasive operations:

- The ASR engine must be robust: it will work in a live operation room in which more than one surgeon may speak at a time, and ambient noise arising from further medical machines is given. Emotional coloring of speech clearly needs to be dealt with. As the camera will be inserted into the patient's body, any mistake is not acceptable. Likewise, very high reliability is demanded from the ASR engine.
- The cognitive workload of the surgeon has to be minimized at any time. The user's mental model of the speech control has therefore to be kept utmost simple and intuitive.
- No push-to-talk by manual activation shall be demanded from the surgeon to keep the hands and feet free as named, and reduce cognitive workload arising from the coordination of manual activation and speaking at the right time. The ASR engine will therefore stay in a

listening mode, that is open microphone. However, to avoid confusion with speech directed to assistants during operation, we decided for keyword initializing, here. This means, speech control is activated and deactivated by speech itself, or automatically deactivated, if no suiting command is recognized within a set time.
- The ASR engine has to communicate with the robot in a fast and stable way. Operations are carried out efficiently and seriously. Long and significantly varying delays between speech commands and camera movements are non-tolerable.
- The robot possesses two moving modes: a short precise, and a long distance move. The inserted camera is mostly used to move short distances, to avoid accidental injuries. However, the optional fast move improves the operation's efficiency. This latter move may not be confused, accidentally by ASR.

According to these requirements we decided for an utmost restrained vocabulary command and control design with commands as depicted in Table I. At present, ASR is realized by an HMM recognizer basing on 3 state inner word phoneme tri-phones with Bakis model topology basing on MFCC 1-12 plus energy and derived speed plus acceleration regression coefficients (c.f. section IV). The usage of a phoneme based recognizer allows for flexible exchange of the terms to personalize the vocabulary to a surgeon's preference. The AM is trained on the WSJ corpus [8] and adapted by the SIMIS database as introduced in VI. As we focus on emotion recognition, herein, the reader is referred to [12] for details on AM adaptation. The vocabulary consists of the highly limited 10 terms *camera*, *quit*, *move*, *stop*, *up*, *down*, *left*, *right*, *forward*, and *backward*. Further, we included models for short and long silences, breathing, and beeps deriving from further machines in the operation room. These models were trained on the SIMIS database, too. The grammar is chosen as context free word-loop solution. A short, high-pitched beep sound is played when the speech window opens after key-word recognition. For closure of the open speech time window a lower pitched beep is played to the surgeon. Each recognized command within the open speech window that possesses high confidence prolongs the speech window accordingly to allow for multiple command sequences without the need of repeated keyword initializing. Additionally, the acoustic features and classifier as described in the following sections can be used to verify the speaker, to avoid activation by environmental babble noise.

IV. Acoustic Features

A strictly systematic generation of features was chosen for the construction of a large feature space as basis for subsequent selection of relevant features. Such an approach generally leads to >1k features [17], [3]. Our basis is a set of 37 typical acoustic Low-Level-Descriptors (LLD) well known to carry information about paralinguistic effects [10] shown in Table II. We group the features into the common types duration (DUR), energy (EN), pitch (F0), formants (FX), cepstral (CEP), and voice quality (VQ). Duration features thereby model temporal aspects having milliseconds (ms) as unit. Voice quality is covered by jitter and shimmer, and Harmonics-to-Noise Ratio (HNR); c.f. below.

Duration features model temporal aspects having the basic unit milliseconds, such as position in time or lengths of intervals.

Energy features model intensity, based on the speech signal's amplitude, with implicit or explicit normalisation and perceptual modeling.

Pitch is the acoustic equivalent to the perceptual unit pitch - measured in Hz - and often perceptually modelled e.g. by use of semi-tone intervals.

Formants (i.e. spectral maxima) are known to model spoken content, especially lower ones. Higher ones however also represent speaker characteristics. Each one is fully represented by its position, amplitude, and bandwidth.

Cepstral (Mel-Frequency-Cepstral)-Coefficients (MFCC) base on a homomorphic transform with equidistant band-pass-filters on the Mel-scale. They tend to strongly depend on the spoken content, but have been proven beneficial in practically any speech processing task.

Voice quality is often described by jitter and shimmer - micro-perturbations based on pitch and intensity - and the Harmonics-to-Noise Ratio (HNR).

In order to calculate LLD, first the speech signal is transformed to 16 kHz, 16 bit. In general, a Hamming window function is used, except for the calculation of F0 and HNR, where a Hanning window has been chosen. We use 100 fps with semi-overlapping windows. Energy resembles simple log frame energy. F0 and HNR calculation base on the time-signal ACF with window correction. Formants base on 18-point LPC with root-solving and a pre-emphasis factor $\alpha = 0.7$. F0 and formant trajectories are globally optimized by use of Dynamic Programming. LLD are smoothed by according techniques as semi-tone-interval filters or simple moving average low-pass-filtering to overcome noise. As a next step we add delta coefficients for each LLD.

Following the typical static classification strategy used in the related recognition of emotion [3], we next employ a total of 19 statistical functionals to each of the 37×2 LLD. The obtained multivariate time series of variable length is projected on a single 1406 dimensional feature vector. Here again we decided for a typical selection of common functionals covering the first four statistical moments, quartiles, extremes, ranges, positions, and zero-crossings as depicted in Table II. The three position related functionals lead to a sub-group of features with the physical unit of ms which are treated as duration features, though having a number of diverse LLD as basis. We refrained from inclusion of further duration related features such as those based on e.g. lengths of pauses or syllables, because this information cannot easily be integrated in the strictly systematic generation approach: it is modeled in a general value series rather than in a time series.

Table III shows the obtained distribution of features among the introduced types. A partition of these feature candidates will be selected for classification in the next step.

TABLE II

ACOUSTIC LOW-LEVEL-DESCRIPTORS AND FUNCTIONALS

LLD (37×2)	**Functionals (19)**
Pitch	Mean
Energy	Standard Deviation
Envelope	Zero-Crossing-Rate
Formant 1-5 amplitude	Quartile 1
Formant 1-5 bandwidth	Quartile 2
Formant 1-5 frequency	Quartile 3
MFCC Coefficient 1-16	Quartile 1 - Minimum
HNR	Quartile 2 - Quartile 1
Shimmer	Quartile 3 - Quartile 2
Jitter	Maximum - Quartile 3
ΔPitch	Centroid
ΔEnergy	Skewness
ΔEnvelope	Kurtosis
ΔFormant 1-5 amplitude	Maximum Value
ΔFormant 1-5 bandwidth	Relative Maximum Position
ΔFormant 1-5 frequency	Minimum Value
ΔMFCC coefficient 1-16	Relative Minimum Position
ΔHNR	Maximum Minimum Range
ΔShimmer	Position of 95% Roll-Off-Point
ΔJitter	

TABLE III

DISTRIBUTION OF FREQUENCY OF ACOUSTIC FEATURES AMONG TYPES. DUR: DURATION, EN: ENERGY, F0: PITCH, FX: FORMANTS, CEP: CEPSTRAL, VQ: VOICE QUALITY.

Type	**DUR**	**EN**	**F0**	**FX**	**CEP**	**VQ**
[#]	222	64	32	480	512	96

V. FEATURE SELECTION

In order to improve the performance and speed of processing of emotional speech, the optimal relevant attributes for classification must be selected [11]. This section will describe the according process: an attribute evaluator based on Correlation-based Feature Subset-selection (CFS) is used, herein [18]. Generally, a target function is needed as optimization criterion throughout reduction of the feature space.

CFS evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter correlation are preferred. The evaluator was set to identify locally predictive attributes, which means it could iteratively add attributes with the highest correlation with the class, as long as there is not already an attribute in the subset that possesses a higher correlation with the attribute in question. Further, counts for missing values were distributed across other values in proportion to their frequency.

Exhaustive search is usually not an option in speech emotion recognition [15], especially in our case of a $> 1k$ feature space and 3,035 instance database. We therefore decided for a Sequential Forward Floating Search (SFFS) [9] - the most commonly used search in this field [15], [17], [11]. This means the space of attribute subsets is processed by greedy hill climbing augmented with a backtracking facility. Setting the number of consecutive non-improving nodes allows control of the level of backtracking done. Generally, SFFS may start with the empty set of attributes and search forward, known as best first, start with the full set of attributes and search backward, or start at any point and search in both directions (by considering all possible single attribute additions and deletions at a given point). Herein, the search direction was chosen forward as a small number of final features was expected, to keep complexity low. The final feature set size was determined by a consecutive number of 50 feature additions without further gain in accuracy by polynomial one-vs.-one Support Vector Machine (SVM) classification with Sequential Minimal Optimization learning [18]. It shall be noted that optimization of the feature space needs to be carried out only once prior to the actual online usage. The number of selected features varies form 58 to 114 throughout latter experiments. Likewise, on the average approximately 10 percent of the original 1406 features were selected, meaning 90 percent of them were regarded as redundancy and deleted.

VI. SIMIS DATABASE

In order to test and train the described emotion recognition, a database of emotional speech within the real life situation needed to be recorded. This database of *Speech In Minimal Invasive Surgery* (SIMIS) will be introduced now. The quality and quantity of the database have a strong effect on the final performance. To get the best results, a large database should be collected from the live environment and handled with. This process consists of three main steps: recording of operations, segmentation of speech, and labeling of the emotion classes for each segment.

All kinds of noise such as coughing, machine noises or babble over-talk should be recorded with the speech so that the final results based on this kind of database would be valuable and reliable. 10 live surgeries were recorded with both headset and room microphone in an operation room of the Clinic r.d. Isar of TUM in Munich, Germany, where there were normally one main surgeon and 6 to 10 surgical assistants. These recorded operations were all minimal invasive surgeries such as stomach or gall operations, which took 1 hour on the average. The recording format is 16 bit, and the sample rate is 16 kHz. Active condenser microphones were used, each.

The automatic speech pre-segmentation is based on energy in the time domain. Firstly, the recorded audio file is multiplied with a Hamming window function. The window's width is set to 512 sample points, while the frame length which specifies the intervals between the values is set to 256 sample points. A mean log power value of 5 consecutive frames is calculated. If this value exceeds 50 dB, the current frame is regarded as speech onset. After onset detection, 60 consecutive frames with a log power value less than 21 dB are regarded as speech offset. To prevent loss of speech information, 5 frames are added at both, the start and the end. This very basic segmentation prevents loss of potential speech turns, but demands for a manual check in a subsequent step. This was ensured by one annotator. The thresholds were iteratively optimized on the data. 10 records of different

invasive surgeries such as stomach and gall operations were segmented into speech turns with the strategy described. The results of the semi-automatic segmentation by time and turns is detailed in Table IV. The total time of each surgery took from 36 minutes to 80 minutes. The speech time took from roughly 5 minutes to 17 minutes. The number of segments reached from 159 to 523.

In the next step three experienced male annotators manually labeled these speech segments within 5 emotion classes chosen from an open initial labeling set with respect to frequency of occurrence and the target application: *neutral*, *happy*, *angry*, *impatient*, and *confused*. Turns with majority agreement were attributed to the according majority emotion, usually being a 100% agreement. In the rare case of total divergence, the turns where mapped onto neutral, which likewise functions as a garbage class. This strategy was chosen, as one cannot simply ignore speech turns in the real life application. The final labeling results are also found in Table IV. From these labeling results it can be seen that the "neutral" emotional speech's duration is almost half of the total speech duration, and the negative emotional (angry and impatient) speech's duration is more than a quarter of the total speech showing the high pressure in this work environment. Apart from emotional annotation, one experienced labeler transcribed the spoken content on the word level including the filler models named in section III. For word-level transcription additional 10 operations have been recorded to provide sufficient data for ASR model adaptation. Usually, the commands would be robot-directed. However, sparse human-to-human conversation is also recorded. All turns of the operation leading surgeon captured via the headset are considered, herein.

VII. Experimental Results

Next, the results of the automatic emotion classification as proposed on the SIMIS database are introduced. For benchmark results of the recognition engine on further and public sets as the Berlin Emotional Speech Database (EMO-DB) or Speech Under Simulated and Actual Stress (SUSAS) the reader is referred to [2], [16]. Reported is the F_1 value which is used in the interest of having a unique performance measure. Here, F_1 is defined as the uniformly weighted harmonic mean of RR and CL: $2 \cdot CL \cdot RR/(CL+RR)$. RR is the overall recognition rate (number of correctly classified cases divided by total number of cases or weighted average); CL is the 'class-wise' computed recognition rate, i.e. the mean along the diagonal of the confusion matrix in percent, or unweighted average [3]. The measures for SVM with a parameterization as named in section V are provided.

The original manually labeled emotion classes can be regarded as 3 types according to the dimensional approach in emotion theory: *positive*, *negative*, and *neutral*. Happy thereby is the only positive emotion. Angry, and impatient are considered negative emotions, and neutral and confused are understood as neutral emotions, herein. This is realized by clustering and re-tagging of instances, accordingly. An even more constrained mapping maps the neutral and positive instances into one cluster to further boost performance: thereby we discriminate only whether a feed-back dialog needs to be initialized or not.

TABLE V

Results diverse emotion combinations, SVM, SIMIS database, 10-fold cross validation. A, H, I abbreviate angry, happy, and impatient, respectively. The dimension (dim) of the optimal feature vector, and the number of learning instances per class are also depicted.

Emotion type	**dim** [#]	**H** [#]	**I** [#]	**A** [#]	**RR** [%]	**CL** [%]	F_1 [%]
H, A	87	404	-	265	81.3	80.2	80.9
H, I	82	404	405	-	70.0	70.4	70.2
H, A+I	114	404	670		74.6	72.1	72.9

TABLE VI

Results positive vs. negative emotion, SVM, database SIMIS, leave-one-operation-out.

[%]	μ	σ	**max**
RR	75.5	7.7	92.5
CL	71.4	10.6	92.3
F_1	73.3	9.1	92.4

First, classification results of exemplary single combinations of emotion pairs by deletion and clustering of classes are shown in Table V. The recognition rate varies from 70.0% to 81.3%, and the mean recall rate almost resembles the recognition rate deriving from the more or less balanced distribution among chosen classes. It seems well solvable to distinguish happy and angry or happy and impatient, as opposed to the usual expectation of valance to be hardly separable.

Next, the result for our final use-case is depicted: positive vs. negative, whereby all neutral instances are mapped onto positive and need to be handled. As this highly unbalances the distribution throughout training, random up-sampling throughout classifier learning is chosen as counter-strategy. To keep the conditions throughout evaluation close to the real-life scenario, a leave-one-operation-out cross-validation of 10 cycles is chosen. Likewise, the conditions are maximally varied, each. The performance is depicted in Table VI, and shows the increased difficulty of this task. Mean, standard deviation, and the maximum are provided. From this table one can tell that the maximum recognition rate is considerably high at 92.5%. However, considering the mean of 75.5%, a considerable number of false positives will have to be faced.

VIII. Conclusion and Future Work

In this paper we introduced speech-based camera control in minimal invasive surgery by integration of emotional factors. One of the main achievements of this work is the recording and labeling of 3,035 turns of spontaneous real-life emotional speech. Such data is very sparse in the field of emotion recognition, yet mandatory to obtain a realistic impression of performances. The recordings clearly show the need of handling of emotional speech: only 53%

TABLE IV

DISTRIBUTION OF SPEECH TURNS AMONG EMOTION BY TIME AND TURN NUMBER WITHIN THE SIMIS DATABASE.

Time & turns	**operation**	**speech**		**neutral**		**happy**		**angry**		**impatient**		**confused**	
type	[m:s]	[m:s]	[#]	[m:s]	[#]	[m:s]	[#]	[m:s]	[#]	[m:s]	[#]	[m:s]	[#]
Gall	36:49	6:05	190	2:30	69	1:13	48	0:58	26	0:54	31	0:30	16
Gall	76:14	8:13	308	4:29	151	1:01	56	1:06	34	1:23	57	0:14	19
Gall	34:24	4:45	159	3:18	109	0:24	18	0:30	15	0:09	5	0:24	12
Gall	36:36	8:41	257	6:11	174	1:47	49	0:21	7	0:38	18	0:15	10
Fundoplicatio	54:33	15:05	456	8:26	248	1:01	41	1:57	51	2:30	75	1:11	41
Fundoplicatio	76:25	16:44	523	10:31	331	1:22	57	1:23	37	2:05	54	1:23	44
Sigma wedge	80:08	14:03	201	7:35	97	1:19	21	1:08	19	1:20	19	2:41	45
Sigma wedge	53:59	12:01	340	7:04	189	1:14	43	0:34	22	1:57	53	1:00	33
Sigma wedge	53:51	13:22	295	9:04	204	0:47	22	0:57	15	1:35	31	0:59	23
Stomach	71:01	15:18	306	6:25	121	2:15	48	2:05	39	2:59	62	1:34	35
Total	**574:00**	**114:17**	**3035**	**65:33**	**1509**	**15:45**	**403**	**10:09**	**265**	**15:30**	**405**	**10:11**	**278**

of the surgeon-robot interaction turns were labeled neutral throughout annotation. By a brute-force feature generation and subsequent CFS-SFFS space optimization 75.5% accuracy could be reached for the discrimination of positive and negative emotion by SVM. This figure has to be considered as impressive, given that no utterances were skipped. As opposed to this, usually a pre-selection of "friendly" prototype instances is chosen in practically any other work, even on spontaneous data [3]. This already allows for initialization of safety feed-back dialogs in case of negative emotion, and profiling of the surgeon's emotion during emotion. However, results will have to be interpreted carefully.

Future work will investigate the benefit of the working prototype in a long term usability study in the operation room. Also, we aim at detailed evaluation of ASR in the live usage and the benefit of emotional adaptation. ASR shall also be augmented by speech enhancement through Switching Linear Dynamic Models [13]. Finally, the SIMIS database shall be enlarged and made partly public in a second version. Thereby further valuable information on statistical properties of the duration of emotion periods shall be considered, as it is expected that emotion periods can not be very short in a real situation. Thus, contextual information on the predecent emotion can be of help to further boost accuracies.

IX. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 211486 (SEMAINE). The authors gratefully acknowledge the contribution of the student researchers Jin Yao, Martin Polsky, and Thomas Mikschl.

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] Batliner, A., Schuller, B., Schaeffler, S., Steidl, S.: Mothers, Adults, Children, Pets - Towards the Acoustics of Intimacy. In Proc. IEEE ICASSP 2008, Las Vegas, Nevada, USA, (2008)

[3] Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V.: Combining Efforts for Improving Automatic Classification of Emotional User States. In: Proc. IS-LTC 2006, Ljubliana (2006) 240-245

[4] 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

[5] Hurteau, R., DeSantis, S., Begin, E., Gagner, M.: Laparoscopic Surgery Assisted by a Robotic Cameraman: Concept and Experimental Results. In: Proc. 1994 IEEE Int. Conf. Robotic Automat., San Diego, California, USA (May 1994) 2286-2289

[6] 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. In: Proc. 2000 IEEE Int. Conf. Robotic Automat., San Francisco, California, USA (Apr 2000) 2901-2906

[7] Nishikawa, A., Hosoi, T., Koara, K., Negoro, D., Hikita, A., Asano, S., Miyazaki, F., Sekimoto, M., Miyake, Y., Yasui, M.: Real-time visual tracking of the surgeons face for laparoscopic surgery. In: Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2001) 1-8

[8] Paul, D. B., Baker, J. M.: The Design for the Wall Street Journal-based CSR Corpus, Proc. of the DARPA Speech and Natural Language Workshop, Morgan Kaufmann, Pacific Grove, CA (1992) 357–362

[9] Pudil, P., Novovicova, J., Kittler, J.: Floating search methods in feature selection. In Pattern Recognition Letters, vol. 15, (1994) 1119-1125

[10] Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V.: The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals. In Proc. INTERSPEECH 2007, Antwerp, Belgium (2007) 2253-2256

[11] Schuller, B., Müller, R., Lang, M., Rigoll, G.: Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles. In: Proc. 9th Eurospeech - Interspeech 2005, Lisbon (2005) 805-809

[12] Schuller, B., Stadermann, J., Rigoll, G.: Affect-Robust Speech Recognition by Dynamic Emotional Adaptation. In Proc. ISCA Speech Prosody 2006, Dresden (May 2006)

[13] Schuller, B., Wöllmer, M., Moosmayr, T., Rigoll, G.: Robust Spelling and Digit Recognition in the Car: Switching Models and Their Like. In Proc. DAGA 2008, DEGA, Dresden, Germany (Mar 2008)

[14] Taylor, R.H., Funda, J., Eldridge, B., Gomory, S., Gruben, K., LaRose, D., Talamini, M., Kavoussi. L, Anderson, J.: A Telerobotic Assistant for Laparoscopic Surgery. In: IEEE Eng Med Biol Mag 14 (3) (May-Jun 1995) 279-288

[15] Ververidis, D., Kotropoulos, C.: Emotional Speech Classification Using Gaussian Mixture Models and the Sequential Floating Forward Selection Algorithm. In: Proc. Multimedia and Expo, Amsterdam (2005) 1500-1503

[16] Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G.: Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing. In Proc. of 2nd Int. Conf. on Affective Computing and Intelligent Interaction ACII 2007, Lisbon, ACM, Springer, (2007) 139–147.

[17] Vogt, T., Andre, E.: Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition. In: Proc. Multimedia and Expo, Amsterdam (2005) 474-477

[18] Witten, I. H., Frank. E.: Data mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, San Francisco (2005)