Speech in Minimal Invasive Surgery – Towards an Affective Language Resource of Real-life Medical Operations

Björn Schuller[1], **Florian Eyben**[1], **Salman Can**[2], **Hubertus Feussner**[2,3]

[1] Institute for Human-Machine Communication, Technische Universität München, Germany
[2] Research Group MITI, Klinikum rechts der Isar der Technische Universität München, Germany
[3] Klinik und Poliklinik des Klinikums rechts der Isar der Technische Universität München, Germany
schuller@tum.de

Abstract

There is a clear desire to collect language resources of utmost realism with respect to spontaneity of speech and naturalness of emotion. This goal is difficult to obtain as a setting is needed that provides sufficient 'emotional moments' while these shall not be disturbed by the speakers' awareness of recording. An obvious setting seems to be collecting surgeon's speech during real-life operations: considering the responsibility of patients' health and life, emotion can be assumed to be present and at the same time natural – there simply is not the time of wasting thoughts on the fact that one is being recorded.

1. Introduction

There is an ever present demand for realistic emotion recorded in a natural context and environment due to a number of prohibitive factors as fore mostly privacy of recorded subjects, their awareness of the recording situation, presence of noises and an ever ambiguous ground truth (Devillers et al., 2005; Douglas-Cowie et al., 2003; Schuller et al., 2009b; Ververidis and Kotropoulos, 2003; Zeng et al., 2009).

A number of efforts have been undertaken leading to some of the most popular databases of emotional speech with 'realistic' emotions. Yet, often these were recorded in experimental settings rather than in true every-day life situations, as in Wizard-of-Oz experiments (e. g. the SmartKom database (Steininger et al., 2002) of user interaction with a smart information kiosk, the FAU Aibo database (Steidl, 2009) of Child-Robot Interaction, or the "Sensitive Artificial Listener" (SAL) database (Douglas-Cowie et al., 2007) of Human-Chatbot conversation. The "Audio-Visual Interest Corpus" (AVIC) (Schuller et al., 2009c) is an example of human-to-human conversational speech (a presenter and a subject experiencing different levels of interest throughout a product presentation) not recorded in a simulation, but still in an experimental framework: the subjects were invited without their own original interest to watch the presentation. The "Vera-Am-Mittag" (VAM) corpus is a popular representative of data from broadcasts – here a reality TV show – yet, appearing in a TV show is hardly to be considered as every-day life situation for the target subjects, and in addition it is not fully known to which degree such a show might be scripted in advance. Finally, the "Speech Under Simulated and Actual Stress" (SUSAS) set (Hansen and Bou-Ghazale, 1997) is a well-known example of subjects recorded in their actual work situation – steering a military chopper – yet, still a pre-defined protocol was followed, as these had to speak pre-defined one-word commands (a 20 items vocabulary) in distinct situations as starting, landing, etc..

Considering the above respects, we decided to establish a language resource of surgeon speech during real-life operations in their usual operation room: this is their work routine, emotion is present in a 'less than 90 % neutral' distribution and these can be assumed to be sufficiently natural, as the surgeons arguably simply do not have time to waste a thought on being recorded when lives are at stake, and in addition repeated sessions rather than a single dialog (as e. g. often the case for call centre data, e. g. (Burkhardt et al., 2005)) can be collected: our "Speech in Minimal Invasive Surgery" (SIMIS) database consists of speech recorded during medical operations and has so far mostly been used for improvement of Automatic Speech Recognition (ASR) for control of assisting surgery robots (Schuller et al., 2008; Schuller et al., 2009a). However, in this paper we introduce details on its labelling with respect to affect to provide a further language resource of affective speech for research purposes.

It consists of 29 live recordings of surgery which we will call Set A in the ongoing. This whole Set A has been textually transcribed and annotated with five affective states by a single male labeller '$L1$'. During the work reported in this paper, 6 new life recordings - referred to as Set B - were added to the original SIMIS Database, which thus now consists of 35 live recordings which we will call Set $A + B$. They were all textually transcribed and annotated within the same set of emotions by one female labeller '$L2$' stemming from the same age group (20–30 years) as the labeller $L1$.

The desire behind this effort is to establish a more reliable ground truth of emotion annotation and the fact that prior to this study the SIMIS database comprised recordings of male German surgeons only. To draw more significant conclusions on the subject of speaker-independent emotion-recognition, it was desirable to include female speakers as well as non-native speakers. In the present version, both, a female person and a Turkish surgeon have been recorded.

In this paper we will provide details on the recordings (Section 2.), segmentation (Section 3.), annotation (Section 4.), and linguistic analysis with respect to emotion classes.

Figure 1: The operating room of the Clinic r. d. Isar where all surgeries were recorded

Figure 2: The operating room of the Clinic r. d. Isar during a surgery

2. Recordings

To work with real life emotions is crucial in our case, because we want to create a reliable database, which can be used in real-life-situations. Moreover, it is important to record a variety of speakers to achieve a reliable speaker independent emotion-recognition.

The Clinic *Rechts der Isar* (abbreviated r. d. Isar in the ongoing) of TUM in Munich, Germany (shown in Figure 1) was selected for the recordings.

Usually during an operation there are 6 to 10 people present in the operating room. The surgeon, 2 to 3 assistants and 3 to 6 nurses conditioned by the complexity of the operation and the experience of the surgeon. The operating room during surgery is shown in Figure 2.

During an operation there is a great amount of background noise: several assisting machines run during the operation, telephones are ringing, the nurses talk to each other and sometimes a radio is playing (Schuller et al., 2009a). Additionally, the entire room is tiled to fulfil the hygiene-rules, and so there are diffuse acoustical reflections that may result in increased background-noise level.

For speech capturing we decided for the AKG C 444 L wireless headset. This device possesses a cardioid pattern. The sagger of the microphone is optimised for speech in the near field. The low-frequency transmission is reduced, so the typical near field-effect of the velocity microphone is shaken out. Because of that the speak-distance of $2\,cm$ between 80 and $5\,kHz$ is nearly linear. Greater speak distances require a greater compensation of low-frequencies.

An AKG PT 40 sender transmitted the data along a quartz stabilised carrier frequency in the UHF domain, and an AKG SR 40 received it. The data was stored with $16\,bit$ per sample and a sample rate of $16\,kHz$ to hard disk drive.

Table 1: Operation types, number of recordings, and average duration. Abbreviations: Avg.: Average.

Operation	#	**Avg. Duration** *[min.]*
Set A		
Gall	17	57
Fundoplicatio	6	103
Sigma Wedge	6	108
Total	**29**	**77**
Set B		
Gall	3	43
Umbilical hernia	1	84
Vakusil	1	23
Thyroid	1	119
Total	**6**	**59**
Set $A + B$		
Gall	20	55
Fundoplicatio	6	103
Sigma Wedge	6	108
Umbilical hernia	1	84
Vakusil	1	23
Thyroid	1	119
Total	**35**	**74:04**

2.1. Set A

Prior to this study the SIMIS database consisted of 29 live recordings with a total duration of $2\,240 : 59\,min$ (i. e. over $37\,h$), and a total speech time of $350 : 01\,min$ (i. e. nearly $6\,h$ – for details on segmentation refer to 3.). The operations were recorded from seven surgeons while three different surgery-types have been recorded. The specific length of an operation is based on its complexity - the average duration is shown in Table 1. The surgeons were solely male and native Germans (cf. Table 2).

2.2. Set B

During the study presented in this paper 6 new live recordings with a duration of $356\,min$ (i. e. nearly $6\,h$) were added. Different surgeries with an average duration of $59\,min$ have been recorded (cf. Table 1). The new recordings include five male and one female speakers, one of them with Turkish as his mother language, the others as before

Table 2: Details of the recorded surgeons. Abbreviations: TRT: Total Recording Time (in minutes); # rec.: number of recorded operation sessions.

ID	Native	Gender	Age [years]	# rec.	TRT *[min.]*
			Set A		
S 00	German	male	54	20	1 289
S 01	German	male	46	3	421
S 02	German	male	38	2	211
S 03	German	male	36	1	126
S 04	German	male	35	1	72
S 05	German	male	33	1	67
S 06	German	male	29	1	56
			Set B		
S 00	German	male	54	1	30
S 01	German	male	46	2	137
S 06	German	male	29	1	55
S 07	German	female	34	1	84
S 08	Turkish	male	39	1	23
			Set $A + B$		
S 00	German	male	54	21	1 319
S 01	German	male	46	3	558
S 02	German	male	38	2	211
S 03	German	male	36	1	126
S 04	German	male	35	1	72
S 05	German	male	33	1	67
S 06	German	male	29	1	111
S 07	German	female	34	1	84
S 08	Turkish	male	39	1	23

native German speakers (cf. Table 2).

2.3. Set $A + B$

The extended SIMIS database (Set $A + B$) consists of 35 recordings of 9 surgeons during 6 surgery-types with an average duration of $74\,min$ shown in table 1 and has a duration of $2\,597\,min$ (i.e. over $43\,h$).
Table 2 gives an overview of the surgeons and the amount of data recorded from each of them.

3. Segmentation

As mentioned above, these recordings contain not only speech, but considerable amounts of background noise, too. The most common noise types during surgery are: standard background noise, instrument click noise, background talk, stressed breath or cough from the surgeon (a detailed analysis of the distribution of noise types concerning Set A is found in (Schuller et al., 2009a)). To generate a database for speech-based emotion-recognition, we need to extract turns that contain speech of the recorded surgeon. To test and train emotion recognition systems in future studies, automatic segmentation and silence removal were thus performed.
For this paper the Set $A+B$ recordings were segmented by applying a root-mean-square energy threshold of 0.01 (the samples were normalised to the range $[-1; +1]$. Thereby a minimum turn length of $0.16\,sec$ and a minimum silence length of $0.3\,sec$ was enforced. Several hundred speech turns with an average duration of about $2\,sec$ were obtained from each recording session. The number of segments per recording reached from 86 to 913, depending on the amount of speech and its scattering, while the total speech time took from about $5\,min$ to $39\,min$ (cf. Table 3). For the 35 operations in the SIMIS database a total of 11 077 speech turns were attained, as also shown in Table 3.

Table 3: Segmentation Results. Abbreviations: TRT: Total Recorded Time, TST: Total Speech Time, ST: Speech Turns.

Operation	#	TRT *[min.]*	TST *[min.]*	ST #
		Set A		
Gall	17	975	162	3 952
Fundoplicatio	6	616	72	2 245
Sigma Wedge	6	650	90	2 676
Total	**29**	**2 241**	**324**	**8 873**
		Set B		
Gall	3	130	24	867
Umbilical hernia	1	84	10	292
Vakusil	1	23	5	132
Thyroid	1	119	32	913
Total	**6**	**356**	**71**	**2 204**
		Set $A + B$		
Gall	20	1 105	186	4 819
Fundoplicatio	6	616	72	2 245
Sigma Wedge	6	650	90	2 676
Umbilical hernia	1	84	10	292
Vakusil	1	23	5	132
Thyroid	1	119	32	913
Total	**35**	**2 597**	**395**	**11 077**

4. Annotation

During the annotation the turns were assigned manually one of the following five classes of emotions: *angry* (ANG), *confused* (CON), *happy* (HAP), *impatient* (IMP) and *neutral* (NEU). The content of Set A has been labelled by two labellers, $L1$ and $L2$. At any stage it was required that no turns were skipped or omitted, since the envisioned applications demand that every speech turn has to be dealt with. If we a look at the emotion distribution by percentage of the emotion-classes annotated by the first annotator and compare it to the percentage of the second annotator as depicted in Table 4, we can see clear deviations, though *neutral* is by far the most common emotion. Besides, labeller $L1$ labelled the non-neutral emotion classes (*angry*, *confused*, *happy*, *impatient*) relatively in balance ($\pm$ max. 3.3 %). Moreover, labeller $L2$ chose the emotion class *impatient* to be the second most frequent while *happy*, *angry* and *confused* have been annotated at relatively equal frequency ($\pm$ max. 0.8 %). The agreement of both annotators will be in-

Table 4: Distribution of the emotions (ANG, CON, HAP, IMP, NEU) in percent for the diverse sets (A and B) and per annotator ($L1$ and $L2$).

Set	**Labeller**	ANG	CON	HAP	IMP	NEU
A	$L1$	6.4	9.8	8.0	8.4	67.4
A	$L2$	2.7	2.8	3.7	13.4	77.1
B	$L2$	1.5	5.2	4.3	17.3	64.1

Table 5: Confusions among annotators in Set A ($L1$ to the right). κ=0.56.

#	ANG	CON	HAP	IMP	NEU
ANGER	**151**	8	8	196	167
CONFUSED	6	**142**	11	44	598
HAPPY	8	14	**132**	46	626
IMPATIENT	25	7	17	**335**	388
NEUTRAL	56	105	142	682	**4959**

vestigated by Cohen's Kappa (no straight forward ordinal relation exists among classes. Naturally, one could be established, though, e. g. by arousal or valence dimensions).

Inter-Labeller-Agreement

To measure the agreement between the labellers we will use Cohen's Kappa (Cohen, 1968). The measurement of the agreement follows a confusion matrix among labellers which is shown in figure 5.

5 579 tracks out of Set A have been annotated in agreement. In 3 291 cases confusion occurs. The allotment of agreement of the labellers p_o is compared by the *accidental*-agreement p_e.

$$Kappa = \frac{p_0 - p_e}{1 - p_e} \quad (1)$$

The difference between p_0 and p_e represents the contingent of cases in which *accidental*-agreement occurs. It is normalised by $1 - p_e$ that is expected by chance. The measured Kappa value for this setting resembles 0.56, which can be considered as good, given the highly subjective nature of spontaneous emotions.

The Kappa value can be improved by combining some of the 5 classes to reduce the problem to a three class valence problem. We thereby map angry and impatient to the class *negative* (NEG), confused and neutral to the class *neutral* (NEU), and happy to the class *positive* (POS). A kappa value of 0.62 is obtained for this reduced 3 class set with the according confusions shown in 6.

Table 6: Confusions among annotators in Set A ($L1$ to the right) with the emotions reduced to 3 valence-motivated classes (explanation in the text). κ=0.62.

#	NEG	NEU	POS
NEGATIVE	**707**	570	25
NEUTRAL	788	**5804**	153
POSITIVE	54	640	**132**

Table 7: Linguistic Statistics (Set A), labeller $L1$. Number of turns per emotion class (N_t), average turn length in words (l_μ), standard deviation of turn length in words (l_σ), size of active vocabulary in words (n_{voc}) and the number of words specific to a single emotion (n_{emo}).

Emotion	N_t	l_μ	l_σ	n_{voc}	n_{emo}
ANGER	530	5.6	4.0	836	203
CONFUSED	801	4.7	2.9	897	203
HAPPY	826	5.1	4.1	1 119	364
IMPATIENT	772	4.2	3.4	677	106
NEUTRAL	5 944	5.5	3.8	4 140	2 861

Table 8: Linguistic Statistics (Set A), labeller $L2$. Number of turns per emotion class (N_t), average turn length in words (l_μ), standard deviation of turn length in words (l_σ), size of active vocabulary in words (n_{voc}) and the number of words specific to a single emotion (n_{emo}).

Emotion	N_t	l_μ	l_σ	n_{voc}	n_{emo}
ANGER	246	4.8	3.6	482	100
CONFUSED	274	5.0	3.4	479	88
HAPPY	310	5.5	4.3	567	119
IMPATIENT	1 303	4.8	3.6	1 245	335
NEUTRAL	6 740	5.4	4.0	4 410	3 246

5. Linguistic statistics

For all speech turns the spoken text is transcribed by $L1$. For set A, we analysed the mean number of words per turn (l_μ), the standard deviation of l_μ (σ_μ), and the size of the active vocabulary (n_{voc}) for each of the five emotion classes. The active vocabulary size is thereby the number of unique words which occur in all turns belonging to one emotion class. We also report the number of vocabulary items which are unique to one emotion class, i. e. occur in only in turns with the respective emotion label (this number is referred to as n_{emo}). The results are shown in table 7 for emotion classes as assigned by labeller $L1$ and in table 8 for emotion classes as assigned by labeller $L2$.

No clear tendency for each class can be deduced from tables 7 and 8, with the exception of the class *Impatient*. Impatient turns marked by labeller $L1$ are clearly shorter than neutral turns (5.3/5.4 words) and the average length of all turns (2.2 seconds with 5.3 words, and a standard deviation of 4.0 words). Second shortest for labeller $L1$ are confused turns. For confused and impatient turns assigned by $L1$ a smaller turn length standard deviation is notable. This may be an indication that $L1$ did take turn length into account when assigning an emotion. The vocabulary used within impatient turns ($L1$) is also notably smaller than for the other three emotions (besides neutral). This indicates that impatient turns may be short turns composed of fewer, simpler command words, or words are repeated by surgeons when they are impatient. For $L2$ the situation seems more balanced, and the conclusions drawn from the $L1$ statistics must be carefully analysed. Impatient (now along with angry) turns remain shorter than turns of the other four classes.

The total vocabulary size of the set A is 5 078 words. Only 4 140 ($L1$) or 4 410 ($L2$) words of these 5 078 words are found in neutral turns. Thus, we can conclude that high percentage of words used in emotionally coloured turns is not used in neutral turns, and therefore characterises emotionally coloured turns. This is further supported by the numbers in the last column (n_{emo}) of tables 7 and 8, which state that roughly one fifth of the vocabulary used in emotionally coloured turns is used only in turns of the respective emotion and does not appear in turns assigned to any other emotion class. For future classification experiments it may thus be very beneficial to include linguistic information as features.

6. Conclusion

Our study again proves the challenge of dealing with emotions in a real-life scenario. Moderate inter labeller agreement is observed which is typical for such spontaneous and naturalistic emotion classification tasks, e. g. (Schuller et al., 2009b). A slight correlation between turn-length (measured in words) and impatient or angry turns has been found. Due to the fact that approximately one fifth of the vocabulary used in emotionally coloured turns of one class is specific to that class, linguistic analysis seems promising for future classification experiments.

Besides investigating automatic classification performance on the full, non-prototypical SIMIS data, future work will fore mostly need to add further labeller tracks to the resource and deal with suited ways to find a mapping to less complex tasks facing the 'full realism' of noisy real-life speech.

7. References

F. Burkhardt, M. van Ballegooy, R. Englert, and R. Huber. 2005. An emotion-aware voice portal. In *Proc. Electronic Speech Signal Processing ESSP*.

J. Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

L. Devillers, L. Vidrascu, and L. Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks – Special Issue on "Emotion and Brain"*, 18(4):407–422.

E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. 2003. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1-2):33–60.

E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilan, A. Batliner, N. Amir, and K. Karpousis. 2007. The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In Ana Paiva, Rui Prada, and Rosalind W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 488–500, Berlin-Heidelberg. Springer.

J.H.L. Hansen and S. Bou-Ghazale. 1997. Getting started with susas: A speech under simulated and actual stress database. In *Proc. EUROSPEECH-97*, volume 4, pages 1743–1746, Rhodes, Greece.

B. Schuller, G. Rigoll, S. Can, and H. Feussner. 2008. Emotion sensitive speech control for human-robot interaction in minimal invasive surgery. In *Proc. 17th Intern. Symposium on Robot and Human Interactive Communication (RO-MAN 2008)*, pages 453–458, Munich, Germany. IEEE.

B. Schuller, S. Can, H. Feussner, M. Wöllmer, D. Arsic, and B. Hörnler. 2009a. Speech control in surgery: a field analysis and strategies. In *Proc. ICME*, pages 1214–1217, New York, NY, USA. IEEE.

B. Schuller, S. Steidl, and A. Batliner. 2009b. The INTERSPEECH 2009 Emotion Challenge. In *Proc. Interspeech*, pages 312–315, Brighton, UK. ISCA.

Björn Schuller, Ronald Müller, Florian Eyben, Jürgen Gast, Benedikt Hörnler, Martin Wöllmer, Gerhard Rigoll, Anja Höthker, and Hitoshi Konosu. 2009c. Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application. *Image and Vision Computing Journal (IMAVIS), Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*. 17 pages, in print.

S. Steidl. 2009. *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Logos Verlag, Berlin. (PhD thesis, FAU Erlangen-Nuremberg).

S. Steininger, F. Schiel, O. Dioubina, and S. Raubold. 2002. Development of user-state conventions for the multimodal corpus in smartkom. In *Proc. Workshop on Multimodal Resources and Multimodal Systems Evaluation*, pages 33–37, Las Palmas.

D. Ververidis and C. Kotropoulos. 2003. A review of emotional speech databases. In *Proc. Panhellenic Conference on Informatics (PCI)*, pages 560–574, Thessaloniki, Greece.

Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58.