Robust Key-Word Spotting in Field Noise for Open-Microphone Surgeon-Robot Interaction

Introduction

Interaction between surgeons and robots can often be considerably eased when providing speech as control modality for assisting robots [1,7]. The advantages of hands-freeness and utmost reduced cognitive workload are thereby best reached when the interaction operates in an "open microphone" mode, i.e. the surgeon may speak at any time without the need of pushing a talk button. However, given the seriousness of a medical operation as environment, failures as still always coming with such ever listening automatic speech recognition (ASR) have to be reduced to the utmost minimum. This is the concern of the present paper in contrast to our earlier work [4], where the spotting process in a noisy scenario was not considered. The setting however remains the same: the investigated ASR offers the possibility for an operating surgeon in minimal invasive surgery to move a laparoscopic camera inserted into the patient's body by speech via a robot arm ("SoloAssist™") as seen in fig. 1.

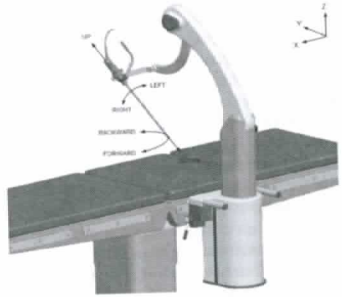

Fig. 1: SoloAssist™ robot arm for camera control.

Material and Methods

We first use our SIMIS (*"Speech in Minimal Invasive Surgery"*) database as presented in [4] to provide for typical noises as present in the field situation: 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 SoloAssist™ positioning robot (cf. fig. 1) 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 an AKG PT 40 sender and an AKG SR 40 receiver. The total recording time amounts to 21:58:38 h, the speech time after automatic energy-based segmentation and manual transcription to 5:11:45 h resembling 8,843 turns. Four dominant noise types were observed and annotated [4]: *standard background noise* permanently provoked from different machines running during the surgical interventions, such as computers or artificial ventilation - it can often be considered as stationary and was judged as pre-dominant in 57.9% of the cases; *instrument click noise* caused by different instruments during the surgery enclosing noises that are produced by depositing surgical tools onto the table made of metal in every case, thus producing a relatively high energy noise - its total observations amount to 22.9%; *background talk* produced by persons present during the operation, such as surgical assistants, nurses, and students - this noise level can get high during stress situations and pre-occupies 8.8% in our transcriptions; *stressed breath or cough* characterized by loud breath noises or coughing by the surgeon wearing the microphone or also by assistants standing close being annotated in 10.4% as major noise.

Besides the comprehensive recordings of real life surgeries, the commands to control the SoloAssist™ robot by speech were recorded as well under exact same microphone and room conditions, but without an ongoing operation. The considered robot as described in [5] can be completely controlled by 15 prompts (*"soloassist"* to open a command window by keyword initialization, six directions as shown in fig. 1 for discrete incremental movement, six times these directions headed by *"move"* for continuous movement, *"stop"* to abort it, and *"quit"* to close the command window). The recordings of these keywords stem 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*.

For automatic recognition of each of the keywords continuous left-right Hidden Markov Models (HMM) are trained based on 3 x (**δ/δδ**) MFCC 0-12. As the vocabulary for the given task is highly limited, the models were chosen to be whole word models for the keyword models as well as for the global garbage model. Additionally a silence model with a tee-state short-pause model was created and trained on non-speech turns from the operation recordings. 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 as multiple keywords were similar with the only difference of a leading *"move"* command (cf. above). This separate model was found

optimal with 12 states and 3 Gaussian mixtures per state. The garbage model is trained on transcribed data on a word basis given the time borders by forced alignment.

To cope with the named noises, three different feature enhancement techniques are further considered in this work selected basing on our experiences from [4,6]: simple Cepstral Mean Subtraction (CMS), well known Histogram Equalization (HEQ) [2] where the histogram of a feature is mapped onto a reference histogram, and a Switching Linear Dynamic Model (SLDM) as introduced in [3]: unlike CMS and HEQ, feature enhancement is realized by models for speech and noise.

Results

As it is utterly important that no false acceptances of keywords occur during a usage of the recognizer in a real life application, the optimal garbage model topology was first determined. This was done by using speech recorded from the operation as a test set and adjusting the garbage model on the basis of the results. The test set included 494 utterances with a total of 1,291 words recorded during operations. The total speech time of these utterances amounts to 9:50 min. As a measure of performance of the garbage model the number of false acceptances of keywords was used, i.e. how many non-keyword turns are recognized falsely as keywords. The optimal model topology was found to be 10 states and 8 mixtures per state. All tests refer to this model constellation.

To gain further performance results, the testing on non-keyword test data was spread out. Using the same test set, the models were tested on results of false acceptances of keywords. It has to be stated that no keyword existed within the test set. Tab. 1 shows results using all feature enhancement techniques giving the percentage of garbage words falsely accepted as keywords (line 1, *"speech"*, column *"FA"*). As can be seen, the least false acceptances appear when using SLDM or HEQ as feature enhancement. The worst results are obtained with CMS, but this was to be expected given the simplicity of its enhancement strategy. Thus, subtracting the mean for each garbage turn also results in an information loss leading to a disturbed recognition relation.

In Tab. 1 so far results were discussed for falsely recognized keywords with actual garbage words sent to the recognizer. Another important fact is how the recognizer performs vice versa, i.e. how many keywords are falsely recognized as garbage. However, the consequences of accepting a garbage turn falsely as a keyword are probably worse since this may result in an undesired movement of the SoloAssist™ robot. Despite that fact the performance of keywords vs. garbage has to operate in a reasonable manner to assure good usability in a real life application, where stress or frustration caused by overheard instructions of the speech interface may play a decisive role. Again, Tab. 1 shows the spotting results for the special noise superposed keyword turns and for the clean turns. As a test set the clean keywords and the keywords superposed with specific noise types were used. Results are presented for no feature enhancement plus the three enhancement techniques. Thereby, the columns labeled *"SUB"* correspond to the percentage of substitution errors that were made between the keywords when recognized as a keyword to the number that were correctly recognized as keywords, and the ones labeled *"FR"* to the percentage of false rejections, i.e. of those that were falsely recognized as garbage.

When using non-enhanced features, the results show only few substitution errors among the keywords if the incidence that the turn was recognized correctly as a keyword is the case. The main problem of the non-enhanced features, when using noisy test sets, is the high amount of keywords that are falsely recognized as garbage and vice versa, especially in the case of the turns superposed with stressed breath or coughing. The spotting results for CMS enhanced features (Tab. 1) show an interesting fact: both, the turns correctly recognized as keywords, and also the substitution errors among the keywords are in an acceptable range. As a result, the conclusion can be made that most of the errors occurring are caused by false insertions of keywords. The most impressive fact of the spotting results with HEQ enhanced features (Tab. 1) are the number of false acceptances of the garbage class that always result to 0 for every test set. Although this seems to be the perfect method to choose, it has to be stated that the substitution errors between the keywords are considerably high. In the last noise reduction method the spotting results for SLDM enhanced features show explicitly that the substitution errors stay in a quite identical region. Although the results in Tab. 1 are worse than the ones where HEQ features are used, it can be concluded that the insertion errors are quite low for SLDM features considering the accuracies given. This is an important fact as it is - as discussed - of higher priority to reject a keyword than to falsely accept it.

The last part is considering the appearing noises in a live operation room and their confusion with keywords: it is of utmost importance that the recognizer is robust against a sudden occurring noise type when used in an open microphone manner, where the case that a simple noise is recognized as speech is an unwanted issue. Thus, for all noise reduction methods tests were performed where non-speech turns of the four different annotated types were sent to the recognizer to observe how many turns are falsely accepted as speech turns. The test set consists of 796 turns for the noises respecting the distribution as observed during operations and described earlier. Tab. 1 shows the classification results for non-enhanced features and all noise reduction methods, respectively (column *"FA"* represents the percentage of falsely recognized non-speech turns as speech). Using no feature enhancement results in a high number of non-speech turns recognized as speech. Tab. 1 further shows that this is especially the case for standard background noise where 18% of the turns were recognized as speech, being an impracticable property. CMS does not show a significant improvement concerning the false acceptance

Error [%]	none			CMS			HEQ			SLDM		
	FA	SUB	FR	FA	SUB	FR	FA	SUB	FR	FA	SUB	FR
clean (speech)	1.4	1.5	0.0	5.3	0.7	2.2	0.8	3.7	0.0	0.6	3.7	3.3
standard background	18.0	2.2	1.1	1.5	3.3	4.8	0.0	3.0	0.0	1.1	5.9	1.9
instrument click	4.9	1.9	7.0	1.1	1.5	2.2	0.0	2.2	0.0	0.0	3.0	4.1
background talk	14.3	0.7	5.9	12.6	0.4	3.0	4.3	5.6	0.0	5.7	7.0	4.4
stressed breath	21.7	3.7	14.8	16.9	5.2	8.5	0.0	9.3	0.0	1.2	6.7	7.8
mean	**12.1**	**2.0**	**5.8**	**7.5**	**2.2**	**4.1**	**1.0**	**4.8**	**0.0**	**1.7**	**5.3**	**4.3**
weighted mean	**15.1**	**2.2**	**4.3**	**4.0**	**2.8**	**4.4**	**0.4**	**3.7**	**0.0**	**1.3**	**5.4**	**3.2**

Tab. 1: Error rates in % by false accept of non-keyword speech or noise per type (FA), and – depending on the noise situation – substitution of keywords (SUB), and false reject (FR) of keywords. Considered are no enhancement (none) and CMS, HEQ, and SLDM enhanced features for clean and noise-overlaid keywords (columns SUB and FR) and speech and noises as "impostors" (column FA). The weighted mean takes the percentage of noises into account.

of noise except for the standard background noise with only 1.5% of the turns falsely recognized as speech. This improvement is explained by the stationary character of this noise type, which can be handled well by CMS. However, for the other non-stationary noise types the false accepts as speech are still too high. Again, HEQ shows a great improvement. As Tab. 1 demonstrates, there are no falsely classified turns except for 4.3% when considering background talk as the noise type. As the name already states, this is actually a limited sort of false recognition since actually a person *is* talking. If this person is standing close to the microphone it is to be expected that the turn is taken as speech. As there is no feature enhancement technique that can overcome this fact, one has to rely on the garbage model to classify the turn correctly as garbage. The Switching Linear Dynamic Model enhanced features also show great improvement towards correct rejection of noise turns: it outperforms all other enhancement techniques except HEQ. As a conclusion, HEQ proves superior to all considered techniques in most tests except for the garbage speech vs. keyword test illustrated in Tab. 1, first line, where the SLDM represented the best method. Thus, an interesting approach would be a combination of the two techniques, which has not been investigated in this work. However, HEQ's (and SLDM's) good performances are owed to higher substitution rates, which are also unwanted due to wrong robot movements.

Discussion

In this work a speech-based camera control in minimal invasive surgery with emphasis on noise robustness towards open microphone usage has been discussed. Feature enhancement was considered as front-end to a standard recognizer. In future efforts the garbage model used to model extraneous speech is subject to improvement. Although it performs well with only few false positives of control commands, a phoneme-based approach modeling out-of-vocabulary words is a reasonable alternative. At present, the garbage model consists of one global model trained on all speech out of the operation recordings annotated on the word level. However, a tri-phone approach could result in a more stable performance using training data recorded from multiple surgeons to overcome speaker dependency. Further, threshold dependencies need to be investigated.

Literature

[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 *Surgical Endoscopy* 12 (12), 1998, pp. 1415–1418.

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

[3] Droppo, J.; Acero, A.: "Noise robust speech recognition with a switching linear dynamic model," in *Proc. ICASSP*, IEEE, 2004, vol. 1.

[4] Schuller, B.; Can, S.; Feussner, H.; Woellmer, M.; Arsić, D.; Hörnler, B.: "Speech Control in Surgery: a Field Analysis and Strategies," in *Proc. Int. Conf. on Multimedia and Exp*, IEEE, 2009, New York, NY.

[5] Schuller, B.; Rigoll, G.; Can, S.; Feussner, H.: "Emotion Sensitive Speech Control for Human-Robot Interaction in Minimal Invasive Surgery," in *Proc. RO-MAN*, IEEE, 2008, pp. 453-458, Munich, Germany.

[6] Schuller, B.; Wöllmer, M.; Moosmayr, T.; Rigoll, G.: "Recognition of Noisy Speech: A Comparative Survey of Robust Model Architecture and Feature Enhancement," in *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, ID 942617, 17 pages.

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

Affiliation of the first Author

Björn Schuller
Technische Universität München
München, D-80333
Germany
schuller@tum.de