LINDSEY DEAN

PROJECT: VOICE CONTROL OF THE DA VINCI SURGICAL ROBOT

PAPER SELECTION:

- Schuller, Bjorn, Gerhard Rigoll, SalmanCan, and Hubertus Feussner. "Emotion Sensitive Speech Control for Human-Robot Interaction in Minimal Invasive Surgery."*Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication*(2008): 453-58. Print.
- Schuller, Bjorn, SalmanCan, Hubertus Feussner, Martin Wollmer, DejanArisc, and BenediktHornler. "SPEECH CONTROL IN SURGERY: A FIELD ANALYSIS AND STRATEGIES."*Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on* (2009): 1214-217. Print.

INTRODUCTION

Voice control's integration with robotic-assisted minimally-invasive surgery began with ComputerMotion's introduction of the AESOP robot in the nineties. The AESOP robot consisted of a laparoscopic camera fixed to a robotic arm. There were two ways to control to motion of the camera either through a joystick or with voice commands such as "right" and "left". However, many in the medical community complained that the voice control feature caused long reaction times and limited reliability, resulting in the AESOP's discontinuation. For this semester project, Voice Control of the Da Vinci robot, my partner and I aim to implement voice control in a way that will make human-robot interaction more intuitive and less of a distraction. To effectively achieve this we must consider both the sound environment of the operating room and which aspects of the Da Vinci are most appropriate for voice control integration. Since the AESOP robot, speech recognition technology has greatly improved as has research about the conditions being faced in the operating room.

For this presentation I chose two papers from Munich, Germany, "Emotion Sensitive Speech Control for Human-Robot Interaction in Minimal Invasive Surgery (Schuller, Bjorn Salman, Can et al. 2008)" and "Speech Control in Surgery: A Field Analysis and Strategies (Schuller, Bjorn and Salman, Can et al. 2009)." Both papers created a database of minimally-invasive robot-assisted surgeries, called the Speech in Minimally Invasive Surgery (SIMIS) database. Using this database, the authors created and evaluated a model for emotional speech recognition and created and evaluated a model for reducing noise in the operating room for higher rates of word recognition.

BACKGROUND

In these papers, voice control was implemented on the SoloAssistTM surgical robot. This robot closely resembles the AESOP, as it is a laparoscopic camera controlled by a remote joystick. The SoloAssist has two modes for motion, Cartesian and Invariant Point Control. Cartesian adjusts the camera left, right, backward, forward, up, and down and the Invariant Point control mode controls tilting, twisting and zooming of the camera.

EMOTION SENSITIVE SPEECH CONTROL FOR HUMAN-ROBOT INTERACTION IN MINIMALLY INVASIVE SURGERY

Emotion sensitive speech control is the ability of a computer to discern the emotional state of a speaker based on indentifying acoustic features. When integrated into minimally-invasive surgery, this

technology has the potential to allow the surgical-robot to respond to the surgeon according to their emotional state. The authors' example is if a surgeon is irritated or angry the robot could ask for confirmation of their command.

PROCEDURE

The strategy used by the authors was to first examine the statistical distribution of emotional pronunciation during surgeries. To do this they segmented the speech from the ten surgeries in the SIMIS database into turns/words. These words were then evaluated by five experience annotators who labeled each as: happy, confused, impatient, angry or neutral. If there was enough disagreement among annotators the word was designated as neutral thus making neutral act also as the garbage class. Using the final set of labeled words, the authors created a statistical distribution demonstrating the frequency of each emotion during surgery. Neutral was the dominant emotion (53%) and almost as frequently words were colored with emotion. The classes of emotion were then evaluated for discriminating auditory features which were used later by the author's to model emotional recognition.

Each set of emotion was evaluated by thirty seven typical low-level-descriptors (LLD) known to contain paralinguistic information; these descriptors were grouped by type into duration, energy, pitch, formant, cepstral and voice quality. The distribution of frequency of these acoustic features was produced demonstrating most variance found among duration, cepstral, and formant types of acoustic features. In order to select which features best predict emotion the authors used Correlation-based Feature Subset-selection (CFS). CFS evaluates subsets of attributes according to their predictive ability. For each surgery from the SIMIS database, the features with the greatest predictive ability and the highest class-contained correlation and lowest interclass correlation were chosen to comprise the final set used by the ASR to distinguish emotion.

Because the original set of feature subsets chosen was so large, 1046, exhaustive search was not an option therefore the authors used, Sequential Forward Floating Search to process the features for each surgery's classes of emotion. The optimal number of predictive attributes, varied from 58-114 demonstrating that only about 10% of the originally categorized features were necessary to predict emotion. These attributes were then used by the ASR to predict the emotion of the same words it was trained with.

RESULTS

Accuracy was measured by $\frac{2(RR*CL)}{RR+CL}$ where RR is the overall word recognition rate measured by the total number of words correctly recognized over the total number of words, and CL is the class recognition rate defined by the total number of words correctly identified by their emotion class. The highest accuracy was obtained when distinguishing between happy and angry. However, this does not cover the entire span of recorded turns, so the emotions were partitioned into two groups neutral and negative. Neutral contained happy, neutral, and confused whereas negative contained the two negative emotions. When distinguishing positive and negative emotions the trained ASR could distinguish with an average accuracy of 72.9%.

CRITICISMS AND RELEVANCE

My main criticism for this paper is the proposed implementation of emotion sensitive speech recognition. While it illustrates a use for the technology it does not seem make human-robot interaction any easier. In fact, it seems asking an angry surgeon for confirmation of their command may

only make them angrier. Therefore, contrary to their aim of making voice control more effective I believe emotional sensitive speech recognition would end up being more distracting in this sense than useful. For our semester project emotion recognition is not relevant. We aim for a robust speech recognition system which will not discriminate between emotional pronunciations of words. Rather we hope our software will be respond uniformly to a keyword despite changes in volume, emotion.

In the section outlining the criteria for the Automatic Speech Recognition (ASR), it is said that the ASR engine should not have a push to talk button to keep the surgeon's hands free. This is an important demand for our project as well. The authors suggest having the ASR both activate and deactivate through voice. Their model has the microphone inactive while waiting for a specific activating keyword sending the ASR into listening mode where it listens for all keywords. This model seems to be an effective way of fulfilling this need. For the Da Vinci robot since it has many more capabilities more restrictions about which keywords can be recognized at any time will have to be implemented.

The emotions targeted in this paper were happy, confused, angry, impatient and neutral. However I believe that anxiety levels in the voice may better predict a need for confirmation than simply a negative emotion. For further research it may be more useful for the sake of this technology to research which emotions are most correlated with surgeon accidents. Despite all of this the authors did manage to demonstrate through labeling of turns in the SIMIS database that surgeon emotion varies throughout surgery and it is possible with a reasonable amount of accuracy to use acoustic features to distinguish these emotions.

SPEECH CONTROL IN SURGERY: A FIELD ANALYSIS AND STRATEGIES

In experimenting with voice recognition technology, noise has shown to significantly impact the accuracy with which words are recognized (for instance with Windows Vista). In this paper noise in the operating room is qualified and then feature enhancement algorithms were applied to the test sets of noise to see if the accuracy of word recognition could be improved.

PROCEDURE

In "Speech Control in Surgery: A Field Analysis and Strategies," four different categories of noise were qualified into four types from listening to the surgeries in the SIMIS database: background noise, instrument click noise, background talk and cough or breath from speaker. Ten additional surgeries have been added to the database since the previous paper, for a total of twenty recordings of minimally invasive surgeries. The turns of each type of background noise were labeled according to noise type. The statistical distribution created from these labels demonstrates background noise as the most prevalent.

In order to test the effectiveness of the feature enhancement algorithms, the authors created test sets of keywords with noise superposed on top. Two different types of sets were created the first chose the noise to superpose according to the resulting SNR and the second chose noises according to label type. The first three sets were created by selecting noises from any class within a range of decibels that produced either a high, low, or medium signal-to-noise ratio with a specific keyword defined as: $10log_{10}\left(\frac{P_{keyword}}{P_{noise}}\right)dB$ where P denotes power. Low, medium and high SNR sets were produced with mean powers of -9.4, 2.9, and 13.3 dB. A high SNR indicates a great difference between noise and

keyword power and therefore this should have a higher word recognition rate than the set with a low SNR. The second type of test sets superposed noise by type top on top of keywords. The mean power for each set of noises background, click, talk, breath were 11.9dB, 5.5 dB 7.4dB and -3.6dB respectively.

Three feature enhancement algorithms were used to improve word recognition rates by the ASR: simple Cepstral Mean Subtraction (CMS), Histogram Equalization (HEQ), and Switching Linear Dynamic Model (SLDM). The basis of comparison was the set of clean words (no noise) listened to by the unenhanced ASR which had an accuracy of 98.53%. The feature enhancement algorithms were applied to the test sets and the word recognition accuracy was measured for each.

RESULTS

Among the three feature enhancement algorithms, the Histogram Equalization produced the highest accuracy of word recognition (95.50% weighted mean). These are promising results which show that HEQ could be a suitable way to deal with noise in the operating room. It is also important to note that the higher the SNR the worse the accuracy of the ASR which is what we expected. Among the test sets of specific noise type breath proved to be the most detrimental to recognition production only a maximum accuracy of 90.81% when subjected to the HEQ whereas background noise was the easiest to filter our producing a maximum accuracy of 97.06% which is not too far away from the base max of 98.53% and is a considerable improvement compared to the untrained ASR which only produced an accuracy of 91.65%. Therefore the improvements given by the feature enhancement algorithms are significant enough to consider as a viable solution to reduce the effects of noise on voice recognition in the operating room.

RELEVANCE AND CRITICISM

Already in experimenting with voice recognition software, noise has shown to be a problem. For the sake of our semester project we will not have the time to create a SIMIS database and therefore, their results demonstrate an effective approach to minimizing the main noise concerns in the operating room. The paper found that breath was the most detrimental, and given that breath had the highest mean power and they created their sets by superposing words, I understand how they got these result. However, superposing breath on top of a keyword seems to be a contrived situation.

Therefore, it seems that a better strategy to approach the issue of breath would be to see how the speech recognition software handles loud breaths or coughs on their own and whether or not they trigger keyword recognition despite the surgeon not pronouncing a keyword.

I am more concerned from their findings that coughs or breaths may trigger the ASR to recognize a word that has not been said rather than reduce the accuracy of recognition. That is because I think that the way they tested it by superposing the noise on top of a keyword does not actually emulate real life. A surgeon would either say a word or cough, the likelihood of the two happening simultaneously is low.

CONCLUSION

Both papers explore ways to improve speech recognition for robot control in the operating room environment. One of the obstacles necessary to consider is extraneous noise picked up by the ASR. In addition, voice contains a ton of information in the form of auditory features. These features can be evaluated to decide the emotion behind speech. This sort of artificial intelligence may have more

of an application in the future however for the purpose of our semester project we are focusing on emotion un-recognition, meaning regardless of pronunciation a keyword should be uniformly understood.

The main difference between my semester project and these papers is the functionality of voice for controlling the robot. The SoloAssisstTM used in these papers can be manipulated more effectively using a joystick than with voice. Similarly the robotic arms of the Da Vinci are already controlled in an intuitive way with the finger grippers on the master control. Therefore, we aim to implement voice control as a way to control the many user options of the DaVinci rather than the motion of the robotic arms. This is analogous to how voice control has been implemented in luxury vehicles. Voice is used as a way for the driver to control the car's non-steering capabilities so the driver can keep their hands on the wheel. In this analogy the DaVinci would be a luxury vehicle with smooth steering and voice control of windows, radio, odometer, etc and the SoloAssistTM an economy vehicle with voice controlled steering. Obviously the latter does not make much sense and would cause great frustration among operators since a steering wheel is a much easier and less exhausting way to operate a car.

Among our proposed implementations for voice are control of the CISST MultiTask 3D-user interface (dropping markers on the intra-operative image, measuring distance, and access to pre-operative registration plan) and selecting which of the maximum four robotic arms (3 endowrist and 1 camera) the surgeon controls at any time. Currently these functions are controlled by physical gestures, however a command for "adjust camera" more accurately emulates the operating room dynamic where surgeon's rely on commanding nurses to adjust ambient equipment during a procedure. In both of these applications voice can provide an intuitive way of controlling the many functionalities of the Da Vinci system.