Robot Command Interface Using an Audio-Visual Speech Recognition System

Alexánder Ceballos[1,2], Juan Gómez[2,4], Flavio Prieto[3], and Tanneguy Redarce[4,⋆]

[1] Instituto Tecnológico Metropolitano, Medellín, Colombia
[2] DIEEC, Universidad Nacional de Colombia Sede Manizales, Manizales, Colombia
[3] DIMM, Universidad Nacional de Colombia Sede Bogotá, Bogotá, Colombia
[4] Institut National des Sciences Appliquées de Lyon, Lyon, France
alexanderceballos@itm.edu.co, faprietoo@unal.edu.co,
{juan-bernardo.gomez-mendoza,tanneguy.redarce}@insa-lyon.fr

Abstract. In recent years audio-visual speech recognition has emerged as an active field of research thanks to advances in pattern recognition, signal processing and machine vision. Its ultimate goal is to allow human-computer communication using voice, taking into account the visual information contained in the audio-visual speech signal. This document presents a command's automatic recognition system using audio-visual information. The system is expected to control the laparoscopic robot da Vinci. The audio signal is treated using the Mel Frequency Cepstral Coefficients parametrization method. Besides, features based on the points that define the mouth's outer contour according to the MPEG-4 standard are used in order to extract the visual speech information.

Keywords: Speech recognition, MPEG-4, manipulator.

1 Introduction

The da Vinci system is a laparoscopic surgery system which consists of a control console, a stretcher, four robotical arms and a high performance vision system. The control console can be located at the side of the surgery table or even at an adjacent room, enabling the surgeon to use the system without carrying a face mask. While the surgeon observes 3D images through a stereo vision system, both camera and instruments are controlled by joysticks, and the surgeon switches between them with pedals. When driving the camera, the surgeon loses the instruments control, and sometimes it is necessary to reposition them. In order to avoid this situation, the development of an alternative interface for commanding camera movements is desired.

There are several approaches for commanding an endoscope holder robot proposed in literature. Some of them assist the surgeon in endoscope location by

⋆ This work was partially supported by the ECOS Franco-Colombian program (ECOS-Nord/COLCIENCIAS/ICFES/ICETEX), the Bonpland scholarship program, and the CNRS.

E. Bayro-Corrochano and J.-O. Eklundh (Eds.): CIARP 2009, LNCS 5856, pp. 869–876, 2009.

using joysticks [1], [2], pedals [3], voice commands [3], [4], etc.. Others use visual or force feedback and geometrical constraints in order to track tool's location during the intervention [5].

Automatic speech recognition systems (ASR) is an active research field, mainly because noise in the audio signal propose an unresolved challenge to the recognition systems. Carelessness of the speaker, variation in the frequency and duration of the words, grammar subjects, are other factors that also impose some difficulties when performing the voice command recognition [6], [7], [8].

ASRs proposed in [4] and [9] have a high recognition rate and showed that using voice commands is a admissible approach for controlling the laparoscope holder robot. Nevertheless, those results are unsustainable in noisy environments. In those cases, human beings tends to use also visual information in order to filter speech through lip reading. In fact, it has been considered that to observe the speaker is equivalent to a 15 dB gain in the signal to noise ratio [7], [8].

In our previus work [9], two different approaches for solving the laparoscope command problem were presented. The first one was a Gesture Based Command System, which used a set of mouth movements in order to identify the gesture commands using a state machine. The second one was an only Audio Command System, which used 10 english words in order to fit a state machine.

Audio-visual speech recognition (AVSR) has arisen as an alternative when noise or distortion affect the speech [6]. The selection of acoustic features has been studied widely, and the current efforts are concentrated in the extraction of the visual features and the selection of the audio-visual integration model [10], [11], [12]. With the aim of recognizing a little set of commands to handle the three degrees of freedom of the da Vinci's laparoscope holder, an audio-visual speech recognition system is proposed in this work.

This paper is organized as follows. Section 2 presents the visual features used and the visual feature extraction algorithms. Section 3 describes the model used in the AVSR system. Section 4 shows the experimental tests and results of the system. Conclusions of this work are presented in Section 5.

2 Visual Features

The visual features used in speech recognition can be divided in high level, low level and combined features. Model parameters which define the lip contours are used as high level or shape features [6], [11]. The low level features, or appearance features, are obtained as a result of transformations at pixels level of the mouth region [13], [14] and finally, the combined features mix the shape and the appearance of the mouth concatenating the features or using statistical models [15]. Generally, the visual features vector captures dynamic information including the first and second time derivatives. In addition, because the sampling frequency of the audio is higher that the one of the video, the visual features must be interpolated [15].

The MPEG-4 standard has arisen due to the necessity to standardize the virtual objects of real and synthetic video. It includes video codification, geometric

compression and audio-video synchronization. This standard presents a complex set of Face Definition Parameters (FDPs) which are used for face standardization, and another set which allows the animation of synthetic face models called Face Animation Parameters (FAPs). FAPs serve to describe face movements (model deformations) with respect to the neutral state face model.

The MPEG-4 standard defines 68 FAPs divided in 10 groups, where groups 2 and 8 are used in speech recognition. They describe the movements of inner and outer lip contour, respectively. For visual speech synthesis, Group 1 is used. It defines 14 clearly distinguishables visemes. A viseme is the visual reference pattern of a phoneme, and it can represent to more than one phoneme.

The FAPs are mesured in specific units called FAPUs (Face Animation Parameter Units) [16]. Figure 1 shows the standardized anthropometric measures. The five FAPU represent the distance between the eyes (ES0), the diameter of the iris (IRISD0), the separation between the eyes and the nose (ENS0), the separation between the mouth and the nose (MNS0), and the mouth width (MW0).

In order to extract the high level visual features of the speech, it is necessary to do precise mouth tracking in the video sequences. Lip tracking is still an open subject in artificial vision due to shape, color and texture complexity, and also because of unexpected changes in illumination [17]. This topic has been successfully treated for lateral face views using controlled background and wearing lipstick, but not with frontal views and without lip markers.

For this work an assisted lip tracking algorithm based on appearance and morphologic restrictions defined in the MPEG-4 standard was designed and implemented. The algorithm assumes that all video frames show frontal face views and that speakers do not uses lip markers. Since psychological studies suggest that the most influent visual feature in lip reading is the outer lip contour, only FAPs from Group 8 were tracked (Figure 2). Moreover, in [11] the authors show that using the Group 2, which describes the inner lip contour, does not increase significantly the performance of the automatic recognition speech system, and

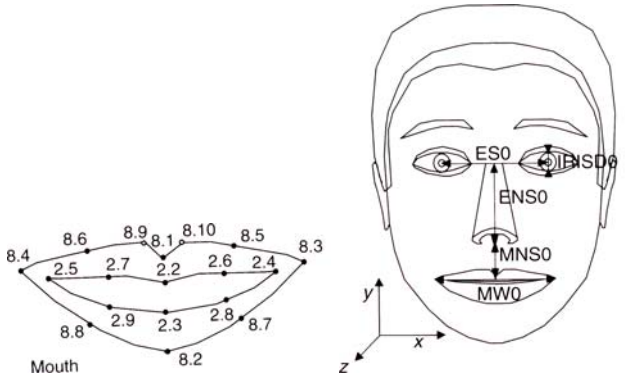

Fig. 1. Groups 2 and 8 of FAPs and the FAPUs measured in a neutral face model

the algorithms are significantly more expensive than those used in outer lip contour tracking.

For calculating the FAPs both magnitude and direction of movement must be preserved, and therefore, they are codified using signed distance functions. Those displacements are standardized using mouth width as normalization factor, which is the FAPU (MW0) for Groups 2 and 8.

Another feature used in this work is mouth roundness. Roundness is found using the Equation 1, in which A corresponds to the area within the outer contour, and d represents the greatest diameter of the mouth region and is equivalent to the mouth width.

$$R = \frac{4A}{\pi d^2} \tag{1}$$

The area is calculated in polar form according to Equation 2, where r_i represents the distance from each one of the 10 points to the mouth center, and $\Delta\Theta_i$ represents the separation angle in between each pair of neighboring points counter clockwise, as shown in Figure 2.

$$A = \sum_{i=1}^{10} r_i^2 \Delta\Theta_i \tag{2}$$

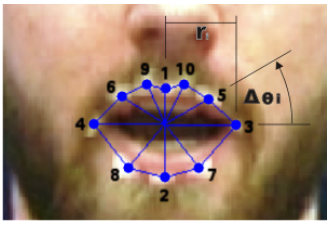

Fig. 2. Outer lip contour defined by the group 8 of the MPEG-4 standard

3 Model Selection

The most popular methods on Automatic Speech Recognition Systems (ASRs) are those based on Hidden Markov Models (HMMs). The HMMs are statistical models whose output is a sequence of symbols. The HMMs deal with the audio sequence as a piecewise statical signal [18], and proved to be more accurate than templates or neural networks at speech recognition [19]. According to recognition task, systems can be classified in the following types: isolated word recognition, where words are separated by pauses; keyword recognition, in which system recognizes certain words in continuous speech; and finally, connected or continuous speech recognition, where the input signal is decoded in a sequence of words, having acknowledged that words are not separated by pauses [20].

HMMs can use either phonemes or words as basic units. There is not direct way to define the number of states for each model, but it has been assumed that using phonemes, three active states is enough [21]. When the models represent words, the model architecture must be assumed in advance. Several configurations must be tested for each word because the system performance strongly depends on the number of states and the probability function of each state.

Figure 3 shows the block diagram of the audio-visual speech recognition system used in our experiments. In this work we used the isolated word recognition approach and words as basic units; we varied the number of states from 3 to 20 active states and used one, two an three probability function of each state. We did not get better results than those obtained using 20 states and one Gaussian per state. We also used the early integration model [10], where the set of the combined visual features from the lip tracking and the audio features is used as the system input.

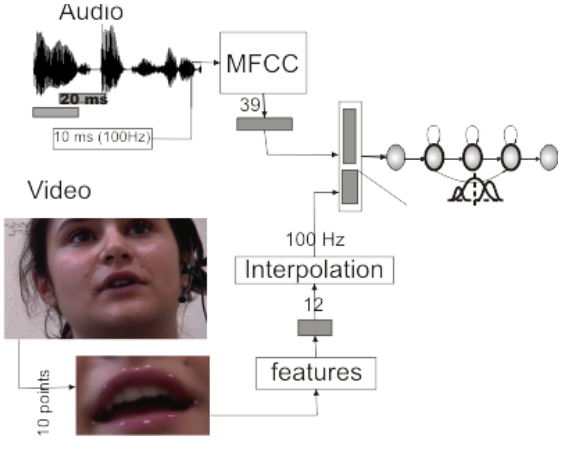

Fig. 3. Block diagram of the AVSR system used in this work

4 Tests and Results

Acquired video data used in this work complies with NTSC standard, whose sampling frequency is of 29.97 frames per second (30 Hertz approximately). Data was recorded in a not controlled enviroment, simulating a realistic situation. There was presence of normal acustic noise sources as computers or other devices. Besides, in the images there was presence of shadows and the light was not controlled. Audio features were extracted using 20 ms windows with overlaps of 50 % between them, which corresponds to 100 Hertz frequency. In order to achieve audio-video synchronization, video features were interpolated from 30 Hertz to 300 Hertz and then subsampled to 100 Hertz.

Principal Components Analysis (PCA) of the FAPs was performed in order to reduce the number of visual features for the audio visual speech recognition

system. At the end, only the first three components of the PCA were used, along with the roundness of the region of the mouth, in the visual feature set. In order to include dynamic information, the first two time derivatives of the visual features were also fed to the system.

The system was trained to recognize six spanish words as commands: “izquierda”, “derecha”, “arriba”, “abajo”, “adelante” and “atrás”. Words’ time fetures are shown in Table 1. Video sequences were acquired from 18 people who all were born in Colombia - 5 women and 13 men.

Table 1. Commands used in the experimets

	Derecha	Izquierda	Adelante	Atrás	Arriba	Abajo
mean (seconds)	0.95	1.05	1.11	0.96	0.90	0.96
standard deviation (seconds)	0.19	0.19	0.24	0.18	0.25	0.27

70% of the data was used to train the system, while the remaining 30% was used for testing. In Table 2 test Word Rate Recognition (WRR) is shown, for the cases in which audio, visual and audio-visual features were taken into account. The best performance was obtained with audio features.

Table 2. Word Rate Recognition using 10 and 20 states

audio	97.70	98.85
video	31.03	5.63
audio + video	90.54	97.30

In order to measure the system roboustness against noise, the audio signal was contaminated with white Gaussian noise. The tests were made to match SNR levels between 20 dB and 0 dB. In Figure 4 it can be seen that the performance of audio only system falls abruptly when the noise is as low as 1:100. Also, performance of the audio-visual system was showed superior for all the SNR levels.

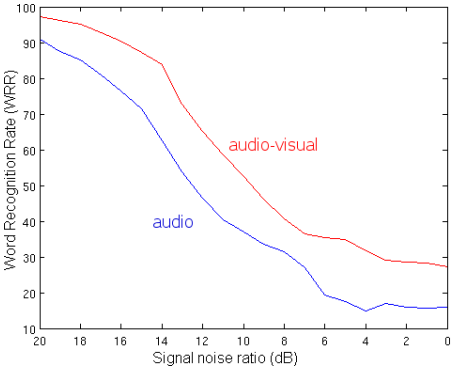

Fig. 4. Audio only and audio-visual WRR vs several SNR levels

5 Conclusion

In this paper we present a speech recognition system for solving the laparoscope command problem. We used audio only, visual only and audio-visual features. Visual features related to mouth shape proved not to be sufficient for solving the recognition task by themselves, but helped when acoustic noise is present in the audio-visual signal. Audio-visual performances exhibited higher errors than the voice based approach when no noise was added, but outperformed in all other cases.

The ASR system based in HMMs using words as basic units in isolated word recognition scheme, which uses both the MFCC as acoustic features and high level visual features based on the standard MPEG-4, presented a WRR near to 100% for recognizing the six spanish words selected as commands. Therefore, the system is reliable for solving the laparoscope holder command task.

Acknowledgments. The authors thank the CHU de Lyon and Doctor Olivier Jegaden for making it possible to access the DaVinci command console.

References

1. Sackier, J., Wang, Y.: Robotically assisted laparoscopic surgery from concept to development. Surgical Endoscopy 8(1), 63–66 (1994)
2. Allen, T.P.K., Goldman, R., Hogle, N.J., Fowler, D.L.: In vivo pan/tilt endoscope with integrated light source, zoom and auto-focusing. Studies in Health Technologies and Informatics, 132–174 (2008)
3. Allaf, M., Jackman, S., Schulam, P., Cadeddu, J., Lee, B., Moore, R., Kavoussi, L.: Voice vs foot pedal interfaces for control of the AESOP robot. Surgical Endoscopy 12, 1415–1418 (1998)
4. Murioz, V., Thorbeck, C.V., DeGabriel, J., Lozano, J., Sanchez-Badajoz, E., Garcia-Cerezoand, A., Toscano, R., Jimenez-Garrido, A.: A medical robotic assistant for minimally invasive surgery. In: IEEE Int. Conf. Robotics and Automation, San Francisco, CA, USA, pp. 2901–2906 (2000)
5. Krupa, A., Gangloff, J., Doignon, C., de Mathelin, M.F., Morel, G., Leroy, J., Soler, L., Marescaux, J.: Autonomous 3-D Positioning of Surgical Instruments in Robotized Laparoscopic Surgery Using Visual Servoing. IEEE transactions on robotics and automation 19(5), 842–853 (2003)
6. Goecke, R.: Current trends in joint audio-video signal processing: A review. In: Eighth International Symposium on Signal Processing and Its Applications (ISSPA 2005), vol. 1, pp. 70–73 (2005)
7. Campbell, R.: Audio-visual speech processing, pp. 562–569. Elsevier, Amsterdam (2006)
8. Campbell, R.: The processing of audio-visual speech: empirical and neural bases. Philosophical Transactions of The Royal Society B 363, 1001–1010 (2008)
9. Gómez, J.B., Ceballos, A., Prieto, F., Redarce, T.: Mouth Gesture and Voice Command Based Robot Command Interface. In: Proceedings of 2009 IEEE International Conference on Robotics and Automation (ICRA 2009), pp. 333–338 (2009)

10. Nefian, A.V., Liang, L., Pi, X., Liu, X., Murphy, K.: Dynamic bayesian networks for audio-visual speech recognition. EURASIP Journal on Applied Signal Processing, 1–15 (2002)
11. Aleksic, P.S., Katsaggelos, A.K.: Comparision of MPEG-4 facial animation parameter groups with respect to audio-visual speech recognition performance. In: IEEE International Conference on Image Processing, ICIP 2005, vol. 3, p. III-501-4 (2005)
12. Kratt, J., Metze, F., Stiefelhagen, R., Waibel, A.: Large vocabulary audio-visual speech recognition using the janus speech recognition toolkit. In: Rasmussen, C.E., Bülthoff, H.H., Schölkopf, B., Giese, M.A. (eds.) DAGM 2004. LNCS, vol. 3175, pp. 488–495. Springer, Heidelberg (2004)
13. Myung, K., Joung, R., Eun, K.: Speech Recognition with Multi-modal Features Based on Neural Networks. In: King, I., Wang, J., Chan, L.-W., Wang, D. (eds.) ICONIP 2006. LNCS, vol. 4233, pp. 489–498. Springer, Heidelberg (2006)
14. Huang, J., Potamianos, G., Connell, J., Neti, C.: Audio-visual speech recognition using an infrared headset. Speech Communication 44, 83–96 (2004)
15. Potamianos, G.: Speech recognition, audio-visual, pp. 800–805. Elsevier, Amsterdam (2006)
16. ISO/IEC: Information technology-generic coding of audio-visual objects, part 2: Visual, ISO/IEC FDIS 14496-2 (final drafts international standard), ISO/IEC JTC1/SC29/WG11 N2502 (1998)
17. Zhilin, W., Aleksic, P., Katsaggelos, A.: Lip tracking for MPEG-4 facial animation. In: Fourth IEEE International Conference on Multimodal Interfaces Processing, vol. 1, pp. 293–298 (2002)
18. Elliot, R.J., Aggoun, L., Moore, J.B.: Applications of mathematics. In: Karatzas, I., Yor, M. (eds.) Hidden Markov Models. Estimation and Control. Springer, New York (1995)
19. Anderson, S., Kewley-Port, D.: Evaluation of speech recognizers for speech training applications. IEEE Transactions on Speech and Audio Processing 3(4), 229–241 (1995)
20. Pasamontes, J.C.: Estrategias de incorporación de conocimiento sintáctico y semántico en sistemas de comprensión de habla continua en espanol. Estudios de Lingüistica Española (2001)
21. Aguilar, R.C.: Diseño y manipulación de modelos ocultos de markov, utilizando herramientas HTK. Ingeniare. Revista chilena de ingeniería 15(1), 18–26 (2007)