AV@CAR: A Spanish Multichannel Multimodal Corpus for In-Vehicle Automatic Audio-Visual Speech Recognition

Alfonso Ortega[1], Federico Sukno[2], Eduardo Lleida[1],

Alejandro Frangi[2], Antonio Miguel[1], Luis Buera[1], Ernesto Zacur[2]

[1]Communication Technologies Group and [2]Computer Vision Group
Aragon Institute of Engineering Research (I3A)
University of Zaragoza, Spain
{ortega,fsukno,lleida,afrangi,amiguel,lbuera,zacur}@unizar.es

Abstract

This paper describes the acquisition of the multichannel multimodal database AV@CAR for automatic audio-visual speech recognition in cars. Automatic speech recognition (ASR) plays an important role inside vehicles to keep the driver away from distraction. It is also known that visual information (lip-reading) can improve accuracy in ASR under adverse conditions as those within a car. The corpus described here is intended to provide training and testing material for several classes of audiovisual speech recognizers including isolated word system, word-spotting systems, vocabulary independent systems, and speaker dependent or speaker independent systems for a wide range of applications. The audio database is composed of seven audio channels including, clean speech (captured using a close talk microphone), noisy speech from several microphones placed on the overhead of the cabin, noise only signal coming from the engine compartment and information about the speed of the car. For the video database, a small video camera sensible to the visible and the near infrared bands is placed on the windscreen and used to capture the face of the driver. This is done under different light conditions both during the day and at night. Additionally, the same individuals are recorded in laboratory, under controlled environment conditions to obtain noise free speech signals, 2D images and 3D + texture face models.

1. Introduction

Automotive industry is one of the main areas of application of automatic speech recognition (ASR). Interaction with several electronic devices (mobile phones, personal digital assistants, car navigation systems, etc.) must be made by means of hands-free interfaces as the driver must keep his or her hands on the steering wheel, and their attention on the road for safe driving. Nevertheless, ASR in car is a challenging task due to the high level of noise present inside the car, coming from the engine, the wind or the road and also because the distance between the microphone and the mouth of the speaker is relatively large. To cope with these problems, it is well known that visual information can improve the accuracy of the recognition systems using lip-reading techniques under adverse conditions. For an updated state of the art review, please refer to (Potamianos, 2003).

Databases which combine both audio and video data are very useful in this research area. There are several of them, mostly in English (Potamianos, 2001, Messer, 1999), but there are also examples in other languages, like (Wojdel, 2002), for Dutch, or (Bailly, 2002), recorded in Spanish, French, English and Italian. Patterson et al, includes in CUAVE (Patterson, 2002), simultaneous speakers, and references other English audio-visual databases.

However, most of these corpora were collected under laboratory conditions. There are no known audiovisual Spanish corpora recorded in a car environment, so there is a need to obtain such a database to develop and test multimodal ASR system in this language. There is one Czech Audio-Visual Speech Corpus recorded while driving (Zelezný, 2003) but it has only one audio channel and was recorded using an expensive digital tape camcorder.

In order to design robust multimodal ASR systems, high quality training and testing databases are fundamental. Large databases collected under realistic conditions inside a car allow the development of several techniques as noise removal, channel adaptation and others that can reduce error rates in the car environment. This corpus is intended to provide training and testing material for several classes of audiovisual speech recognizers including isolated word systems, word-spotting systems, vocabulary independent systems, and speaker dependent or independent systems for a wide range of applications.

This paper is organized as follows. Section 2 describes the AV@CAR database according to the different parts it is composed of, laboratory and car recordings, video and audio parts. Section 3 presents the procedures used in the acquisitions and the tasks that compose the database and finally the conclusions are presented in section 4.

2. Database Description

The audiovisual corpus AV@CAR can be divided into two main parts. The first one is recorded inside a car while driving, and the second one is collected in a noise free, lighting controlled environment.

The car part of the database is composed of seven audio channels, one video channel and information about the speed of the car, the conditions of the road, the weather, the traffic as well as information about the speaker and the lighting conditions. This allows defining different acoustic and visual configurations and perform speaker and environment adaptation for automatic audio-visual speech recognition (Buera, 2004).

On the other hand, laboratory recordings were made, using five audio channels and one video channel, plus a three dimensional pictures session for each driver.

Car Audio Corpus

In order to acquire seven synchronized audio channels and the information about the speed of the car, the use of a notebook with a standard sound car was declined.

Instead of that, the Audio PCM+ board (Bittware, USA) was used, with eight 24-bit input channels and eight 24-bit output channels. This board allows for high speed data transfer from or to the host PC where the corpus is stored. The recordings were made using a DC-Power supplied PC connected to a 12 V battery isolated form the electrical system of the car to avoid electrical noise.

Figure 1: Microphone positions over the rear seats.

We used Q501T microphones (AKG, Austria) for the in-vehicle acquisition because of its high pass filter characteristic that make them appropriate for the car environment. Six microphones were placed in the overhead of the car, three of them over the front seats and the rest over the rear seats.

Clean speech is captured by means of a close-talk (head-worn) microphone C444L (AKG, Austria).

One input audio channel is used to acquire engine noise thanks to an electret microphone placed into the engine compartment.

Synchronized car speed information is also added to the AV@CAR corpus while being acquired using one audio input of the acquisition board.

Figure 2: Microphone positions over the front seats.

In order to obtain data that allow developing and testing noise and music cancellation systems with a noise or music reference, the output of the audio equipment is also recorded in some tasks instead of the rear microphones.

Lab Audio Corpus.

For the laboratory sessions inexpensive electret microphones were used along with a voice array microphone VA 2000 (GN NetCom, Denmark), a close-talk (head-worn) microphone C 477 W R (AKG, Austria) and one Q501T (AKG, Austria).

In order to capture far-field speech, the two electret microphones are placed in the upper corners of a 1.86x2.86x2.11 m cabin. The NetCom array and the Q501T microphone are placed in front of the speaker, 1 meter far from him.

The audio part of the database is sampled at 16 KHz and 16 bits for each channel.

Figure 3: Camera position besides the rear mirror.

Car Video Corpus

For the video database, a small V-1204A camera (Marshall Electronics, USA) sensitive to the visible and the near infrared bands is placed on the windscreen besides the rear mirror and used to capture the face of the driver. The camera board includes six infra red LEDs which guarantee enough illumination without any other lighting source. This makes it possible to deal with different recording conditions, both during the day and at night.

Figure 4: Frame example recorded in driving conditions.

The images provided are black and white, digitized to a spatial resolution of 768x576 pixels, 8-bit pixel depth,

and frame rate of 25 fps, using a DT3120 Frame Grabber (Data Translation Inc., USA). This somewhat high resolution is justified considering that the camera does not have to capture only the head, but it also has to allow for free movements of the driver within the field of view of the camera while still getting the whole face.

Lab Video Corpus

The laboratory recordings are divided into two parts. In the first one, the speaker is recorded while repeating some of the tasks done inside the car, with the same camera model (V-1204A), but with controlled lighting conditions and a frontal view of the face, in contrast to the one obtained from the windscreen camera, which is displaced to the upper right corner of the individual. (see figures 4 and 5).

Figure 5: Frame example recorded in laboratory conditions.

In the second part, several pictures are taken to the speaker by means of a three dimensional capture system from Vision RT Ltd (London, UK), composed of two sets of three cameras each. This system takes 6 simultaneous pictures of the individual, 4 to reconstruct 3D geometry and the 2 remaining to capture black and white texture information, generating a 3D textured mesh of the face.

Figure 6: ·3D Acquisition Equipment.

Each person is asked to stand with different poses and facial expressions, based on the gesture classification of (Ekman, 1975) and (Martínez, 1998). This information is intended to be useful when comparing the car images (top-right view variable according to driver movements and gestures) with the frontal ones taken in the laboratory and present in most of the facial databases.

The 3D session is also recorded using a 1352-5000 (Cohu Inc., USA) color video camera provided with a Navitar TV zoom lens (12.5-75 mm, F 1.8) and digitalized with 768x576 pixels, 24-bit pixel depth and 25 fps.

Audio-Video Synchronism

One of the challenging tasks of this database was to ensure synchronization between the audio part and the video part. To make this independent from hardware delays or errors, an array of eight red LEDs is captured by the camera in a corner of each frame. These LEDs are lighted up and turned off sequentially, every 5 ms, according to a synch signal generated by the audio board. This also produces an acoustic signal through the loudspeakers of the car or the lab registered by the microphones at the beginning of each recording.

The LEDs array is still blinking during the whole video every 1 second intervals controlled by the audio board, to allow synchronism verification.

3. Procedures and Recording Tasks.

The AV@CAR corpus is divided into three main groups, data available for training and adaptation purposes acquired inside the car, application dependent data intended for testing recorded inside the car and studio-like acquisition in a noise free, lighting controlled environment. Each one of these groups is composed of different tasks.

The training and adaptation parts recorded inside the car are composed of four tasks.

a) Read a long text with the car in parking conditions.
b) Repeat 25 phonetically rich sentences in parking conditions.
c) Repeat several phonetically rich sentences while driving the car.
d) Noise only recordings under different conditions of the weather, the road or the traffic.

The second part of the corpus is also composed of four tasks all of them recorded while driving the car.

a) Repeat specific application words and sentences (for the use of a mobile phone, a car navigation application or a remote e-mail application.)
b) Word-spelling
c) Digits and numbers.
d) Names of cities, regions and streets.

The third part of the database is acquired in a studio-like controlled environment where the same people who previously drove the car are recorded. Several tasks done inside the car are repeated in this controlled environment.

The main purpose of this part is to collect data without the noise and light variations of the car, and better camera positioning. in order to acquire frontal face images. This corpus recorded in laboratory conditions will facilitate the comparison to the results with other standard databases.

Additionally, 3D textured meshes and color video are taken of every speaker with different poses and expressions:

a) Frontal view
b) Left profile
c) Right profile
h) Happiness
i) Surprise
j) Yawn

d) Upper view
e) Button view
f) With transparent glasses
g) With sun glasses
k) Anger
l) Disgust
m) Fear
n) Sadness

For the first two parts of this corpus, the procedures followed during the acquisition are very important. As in driving conditions it is not possible for the driver to read a written text[1], every word or sentence to be said by the driver is played through the loudspeakers of the car followed by an acoustic signal that indicates the beginning of the recording. After this acoustic signal, a person located in the rear seats (who will drive the car in the next session) repeats the sentence or word to the driver and after that, the driver says the word or sentence.

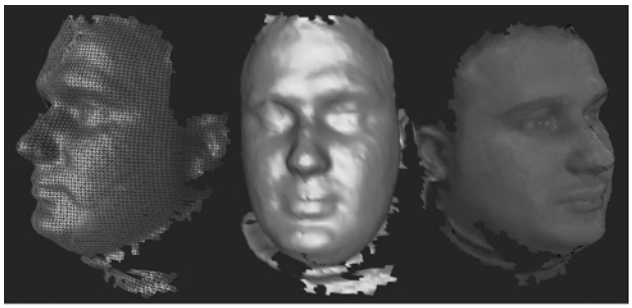

Figure 7: 3D textured (profiles) and non-textured (frontal) meshes of the face.

The audio-visual acquisition application starts recording just before the acoustic signal sounds in order to collect also the voice of the passenger who is seated in the rear part of the car.

Data is collected from 20 speakers, 10 men and 10 women whose ages range from 25 to 50.

Figure 8: Recording of the corpus in the car.

4. Conclusions

In this paper we have presented the audio-visual corpus AV@CAR. The main purpose of this database is to provide useful material to design and test audio-visual automatic speech recognition system in the car environment. Due to the high level of noise inside the cabin, visual information can improve the accuracy of the ASR system. In addition to the car recordings, studio-like acquisition is also performed in order to obtain data in a noise free, light controlled environment. In the laboratory part of this database, several 3D textured meshes of each driver's face are recorded.

[1] In Spain, law forbids the use of screens and displays that could be a source of distraction for the driver.

References

Potamianos, G.; Cosatto, E.; Graf, H.P. and Roe, D. B. "Speaker Independent audio-visual database for bimodal ASR", in Proceedings of Eurospeech 2001 (CD-ROM), Aalborg, Denmark, 2001.

Wojdel, J.C.; Wiggers, P. and Rothkrantz, L.J.M., "An audio-visual corpus for multimodal speech recognition in dutch language", in Proceedings of ICSLP 2002 (CD-ROM), Denver USA, 2002.

Zelezný, M. and Císar, P., "Czech Audio-Visual Speech Corpus of a Car Driver for In-Vehicle Audio-Visual Speech Recognition" In Proceedings of AVSP 2003, St. Jorioz, France 2003.

Buera, L.; Lleida, E.; Miguel, A. and Ortega, A. "Multi-Environment Model based Linear Normalization for Speech Recognition in Car Conditions" In Proceedings of ICASSP 2004, Montreal, Canada, 2004.

Bailly-Baillire, E.; Bengio, S.; Bimbot, F.; Hamouz, M.; Kittler, J.; Mariéthoz, J.; Matas, J.; Messer, K.; Popovici, V.; Porée, F.; Ruiz, B.; Thiran, J. P; "The BANCA Database and Evaluation Protocol", in 4th International Conference AVBPA,. Springer-Verlag, 2003.

Patterson, E.; Gurbuz, S; Tufekci, Z.; Gowdy, J.; "CUAVE: A new audio-visual database for multimodal human-computer interface research", in Proceedings ICASSP, Orlando, FL, USA, 2002.

Messer, K.; Matas, J.; Kittler, J.; Luettin, J.; Maitre, G.; "XM2VTSDB: The extended M2VTS database", in 2nd International Conference AVBPA, Washington D.C, 1999.

Ekman, P.; Friesen, W.; "Understanding the face, A guide to recognising emotions from facial expressions", Prentice Hall, 1975.

Martínez, A; Benavente, R.; "The AR face database. Technical report", Computer Vision Center, Barcelona, Spain, 1998.

Potamianos, G.; Neti, C.; Gravier, G.; Garg, A.; Senior, A. W.; "Recent advances in the automatic recognition of audio-visual speech", in Proceedings of the IEEE vol 91, no. 9, September 2003.

Acknowledgments

This work was partially funded by grants TIC2002-04495-C02 and TIC2002-04103-C03-01 from the Spanish Ministry of Science and Technology, and by Vision RT Ltd (UK). FS is supported by a BSCH grant; LB is supported by an FPU grant from the Spanish Ministry of Education. AFF holds a Ramón y Cajal Research Fellowship.