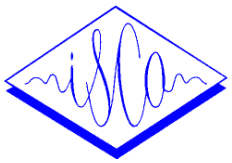

AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4–7, 2003

Czech Audio-Visual Speech Corpus of a Car Driver for In-Vehicle Audio-Visual Speech Recognition

Miloš Železný, Petr Císař

Department of Cybernetics, University of West Bohemia in Pilsen
Univerzitní 8, 306 14 Pilsen, Czech Republic
{zelezny, pcisar1}@kky.zcu.cz

Abstract

This paper presents the design of an audio-visual speech corpus for in-vehicle audio-visual speech recognition. Throughout the world, there exist several audio-visual speech corpora. There are also several (audio-only) speech corpora for in-vehicle recognition. So far, we have not found an audio-visual speech corpus for in-vehicle speech recognition. And, we have not found any audio-visual speech corpora for the Czech language either. Since our aim is to design an audio-visual speech recognizer for in-vehicle recognition, the first thing we had to do was to design, collect, and process the Czech in-vehicle audio-visual speech corpora.

The purpose of in-vehicle speech recognition is usually its utilization for command control of car features, which does not involve driver's hands. Thus, in real deployment, it will be the driver, whose speech will be recognized. Although it is more demanding than to collect the speech of a passenger, we decided to collect the driver's speech for training purposes. This is probably not so important for audio-only speech corpus, but for our purpose we need to collect speech in real conditions, i.e. conditions that include head movements caused by the fact that the driver has to pay attention to the traffic situation.

1. Introduction

Methods of automatic speech recognition are well-developed and used for various purposes in various environments. It is obvious that, depending on the environment, the noise can be more or less present in the input signal. The overall recognition rate of the system is then affected by the presence of the superimposed noise. It is unfortunate that there are many critical applications of speech recognition in just the noisy environment. The performance of such applications is then highly affected by the noise and they may not be able to conform to the requirements, placed on them due to their critical nature.

One of the applications of speech recognition in noisy environment is speech recognition in a car. Although the recognition is usually applied only to auxiliary operations, the most precise recognition is required. However, achievement of such a high recognition rate in such noisy conditions is very difficult [1]. To bypass this obstacle, visual speech recognition (lip-reading) can be utilized in combination with acoustic speech recognition. In this way we can increase the recognition rate of the system and thus make it more robust and more efficient for such a task.

As our aim is to build the in-vehicle audio-visual speech recognition system, we have to follow several steps to succeed in this task. For the successful training of the speech recognition system we need to collect a relatively large speech database. In the case of audio-visual speech recognition the database will contain both the acoustic and visual information. The acoustic part of the database is usually in the form of waveform files, allowing to choose parameterization at a later stage of the training of the system. On the other hand, it is possible to pre-process the visual information. There is certainly much space that cannot be used for the lip shape recognition. It is reasonable to perform image segmentation and store the sequence of segmented lip shapes in the database. This still allows us to choose the way of describing the lip shape later. The last, nevertheless important part of the speech corpus is the phonetic transcription, which is the textual form of the really pronounced utterances.

Figure 1: *Recording of the corpus in a car.*

When collecting only acoustic speech in a car, the car noise is usually the primary reason for collecting speech in real conditions. It is possible to collect the speech of the passenger. The behaviour of the driver and the passenger during a ride is different. To be able to recognize the driver's visual speech, we have to train the system using the driver's speech including movements caused by him or her paying attention to the traffic situation. Recording of the corpus is illustrated in Figure 1.

2. Design

Since there is no known Czech audio-visual speech corpus for audio-visual speech recognition, we decided to design, prepare, collect, and process our own audio-visual speech corpus. We could, of course, use the experience gained in the design of English corpora. There are many acoustic speech corpora, especially for the English language. But only a few of them were collected in such noisy conditions as for example in a car [1, 2]. There are several audio-visual speech corpora for English [3]

or other languages, for example the Dutch corpus [4]. But these were collected in laboratory conditions and thus do not meet our requirements. Moreover, corpora such as Tulips [5] or M2VTS [6] contain relatively few utterances. Our aim was to design a Czech audio-visual speech corpus and collect it in real conditions (during driving), with a car driver speaking, which is quite difficult.

To be able to collect the driver's speech during his driving a car, we had to adopt some unusual procedures different from the commonly used techniques. Firstly, it was impossible for the driver to read a written text. The reading of the text had to be solved in another way. Instead, the text was prompted by the passenger and the driver repeated the text he heard. This was a rather demanding procedure, as more time had to be spent on the recording than usual. But it was the only way of collecting the driver's speech in real conditions. Secondly, we had to solve the problem of the layout of the recording devices in the car. Usually, when recording in laboratory conditions, we try to put microphones close to the speaking person and in the straight direction from his or her mouth. This setup is, of course, impossible in a car, because the recording equipment must not obstruct the driver's view. Lastly, only battery operated devices can be used.

For the design of the corpus we used experiences gained in previous projects. Previously, we designed and collected an audio-visual lip-reading corpus of 500 isolated words from 10 speakers intended for the task of isolated words audio-visual speech recognition [7]. We also designed and collected a corpus for an audio-visual speech synthesis (usually called talking head) [8]. To represent a variety of speakers in utterances, it is necessary to collect data from a certain number of speakers. As a compromise between the desired high number of speakers and the ability to collect such a number of data we decided on 20 speakers. This number is optimal relative to the demands laid on collecting the data from a driver driving a car.

2.1. Selection of text

The selection of text to be pronounced and recorded is an important part of the corpus design process. Text selection is based on principles similar to those in [9]. We divided the text into several parts. The first part of sentences is intended for training of a continuous speech audio-visual recognizer. The second part containing isolated words (special car audio and navigation commands, names of places and numerals) will be used for the training of a recognizer for command control isolated words. The third, and last part contains spelling. The structure of the whole corpus is illustrated in Figure 2.

We collected audio-visual speech data from 12 speakers. The training data set consists of 10 speaker data sets. Moreover, we recorded 4 testing speaker sets: 2 by 2 speakers from training data set and 2 by 2 new speakers whose data did not occur in the training data set. We have a total of 10 training sets and 4 testing sets. All data sets were collected in one car in various traffic conditions (city traffic, motorway, parked car with running engine). We plan to collect the same data set in another car in the near future.

The first part contains 200 sentences for continuous speech recognition for each speaker. Sentences are selected so as to retain the natural distribution of triphones for each speaker. The recording time is approximately 15-20 minutes per speaker. These data are intended for the training and testing of the audio-visual continuous speech recognizer.

The second part consists of isolated words to be used for the

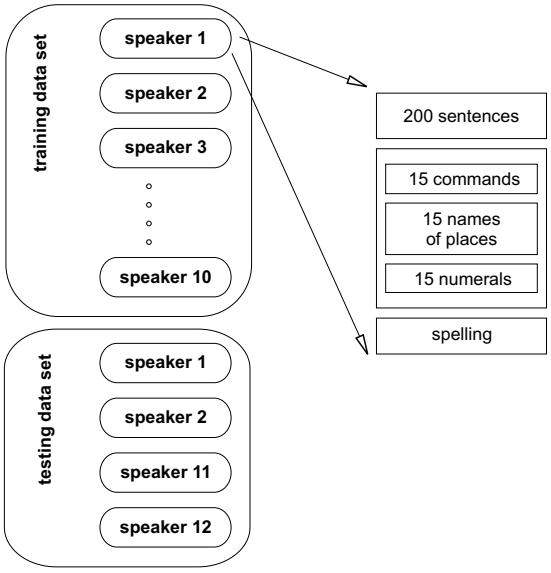

Figure 2: *Structure of the corpus.*

training and testing of a command control isolated word recognizer. This part contains 45 utterances of isolated words per speaker. These words are targeted to the domain of car audio voice control and car navigation. The structure of the this part is 15 commands, 15 names of places and 15 numerals.

The third part contains the spelling. Spelling can be an important way of voice input in case the recognizer fails for some reason. It complements the rest of the corpus with the last choice of speech input when solving the problem of unknown words. This part consist of 7-10 words per speaker, which are prompted as a whole by the passenger and spelled by the driver. Each phoneme should occur at least once in the recording.

Figure 3: *Recording layout in a car.*

3. Recording

When recording audio signals in mobile environments, we can use either a notebook computer with a special sound card for quality audio capturing or some high quality recording device. We have to make a similar decision concerning the visual part. However, notebooks are usually not equipped with quality frame grabbing cards. That is why we decided to capture audio-visual recording by tape recording appliance. The whole

Figure 4: *Rear view of recording.*

Figure 6: *Details of the directions of camera movements.*

recording setup and layout is illustrated in Figures 3 and 4.

For recording the corpus we used the digital tape camcorder SONY TRV-740. This digital camera with a megapixel CCD chip uses the Sony D8 digital tapes. The capture of audio-visual recording can be easily done with the software supplied with the camera and standard IEEE1394 interface. The resolution of the resulting video is 720x576 pixels. The camera is placed on a special holder attached to a control board. The details of the holder are shown in Figure 5.

Figure 5: *Details of the camera holder.*

The holder was specially designed for a particular model of car, which was used for corpus data collection. It makes it possible to move and turn the camera in all needed directions and thus be able to set it up precisely for recording of the driver's face. The movement directions (forward-backward movement, left-right turn and up-down turn) are illustrated in Figure 6.

The digital camera record consists of audio and video parts. We decided that the quality of the audio part of the audio-visual record is sufficient for our purposes. Thus we did not need a separate audio recording device. The parameters of the recording are: codec PCM, sampling frequency 32 kHz, quantization: 12 bits.

3.1. Video part

For the capture of the visual part of the corpus, we had to solve two problems, i.e. choose the appropriate image size and resolution and also the direction of the camera view. When making these decisions, we had to take into consideration several aspects. The highest possible resolution will allow a better description of the lip shape. But there are arguments against it. Trying to catch also the habits of the driver would force us to acquire an image of a larger area than that needed for lip description.

During driving, the driver's head is in a relatively steady position. But, of course, the traffic situation forces the driver to move his head. We assume that the head movement in the plane parallel to the road is very small and that also up and down head movements will be very small. The most important movement then remains turning the head around the neck-head axis. A small turning around the front-back axis of the head may also occur.

The smallest size of the picture has to contain all of the lips. Also using only the image of the lips we would get the highest resolution per lips using the given camera. However, there are other arguments for choosing the size. We have to assume that the driver will move his head according to the traffic situation. To design the mechanical lip tracker is almost impossible. Thus we have to accept the relatively lower resolution per lips using a wider angle of view. For the correct tracking of the lips the image of a whole head may be needed.

The definition of the camera position is not without compromises either. The ideal position of the camera in the centre of the left half of the front car window is obviously not possible. Also positions above or below this point are not usable. Thus, we have to decide between the left car body column (between the front and side window) and the central part of the control board. We decided on the latter position. The reason that supports this decision is, among other things, the size of the camera used. The SONY camcorder placed on a special holder is probably much larger than future cameras serving this purpose. We think that the middle of the control board is the optimal place for a camera under the current constraints.

3.1.1. Double rate

We discussed the layout of the recording appliances in the car and the resolution of the camera. We have mentioned the audio sampling frequency. But so far we have not touched on the video sampling frequency. It seems obvious that the frame rate of the European model of the camcorder is 25 fps.

There are two approaches to audio-visual speech recognition. We can either design two separate recognizers and then combine the results of these recognizers or we can combine the feature vectors and design a joint audio-visual recognizer. We

Figure 7: *Full frame and separated odd and even frames.*

want to retain the optimal settings for both of these recognition schemes. In the case of the joint audio-visual speech recognizer it means to synchronize the feature vector rate for acoustic and visual parts.

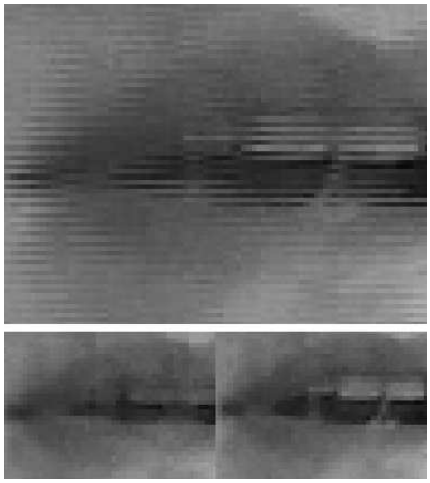

Figure 8: *Detail of an interlaced full frame and separated odd and even frames.*

Let us now assume that the period of acoustic parameterization will be 10 ms. Comparing it with the video frame rate of 25 fps, which corresponds to a period of 40 ms, we can see that there are missing the video parts of the parameter vector. Thus, we need either to repeat the same parameter vector three times or adopt some interpolation algorithm.

To solve this problem at least partially we decided to benefit from the interlaced nature of the video data. The originally interlaced video is deinterlaced into 25 full frames per second. If there is a faster motion in the movie, we can distinguish the odd and even rows in the video file. When we separate the odd and even rows in the video, we can get a video with the double frame rate of 50 fps (corresponding period is then 20 ms). It is, of course, counterbalanced by the decrease in the vertical resolution.

The resulting frame size is 360x288 pixels. The resolution per lips is at least 80x65 pixels. It corresponds to a similar resolution of 80x60 pixels used for example in [10, 11]. Comparing the resulting video with the half resolution with images from other audio-visual corpora, we can see that even after the resolution decrease we can obtain a similar resolution per lips. The comparison of our resulting image and the image from the Tulips1 corpus is presented in Figure 9.

Figure 9: *Resolution comparison of our driver's corpus (left) with the Tulips1 corpus (right).*

3.1.2. Pre-processing

As has already been mentioned, the visual information can be pre-processed to provide enhancement, noise reduction and segmentation of the recorded visual information. Storing the already segmented visual data containing a binary lip shape description is not at the expense of universality, since we can still change the visual parameterization. Experiments with visual lip parameterization will be the most important experiments that we plan to carry out on the collected corpus. Let us note that both original and segmented data are stored in the corpus.

Segmentation of visual information consists of three steps. The first step is the noise reduction by averaging, which reduces the noise contained in the images. The second step is colour processing. We used conversion to the HSV colour model to be able to separate lips from the image. Let us note that the lips were not artificially coloured, as is usual in laboratory audio-visual data collecting. The last step is segmentation, which provides us with the binary output. The segmentation output is illustrated in Figure 10.

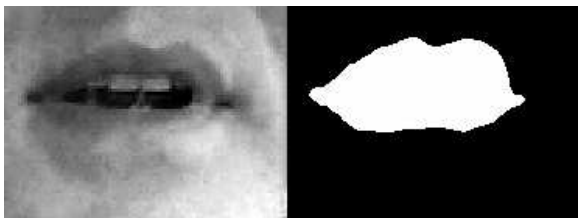

Figure 10: *Illustration of segmentation results. Left - original frame, right - segmented binary image.*

3.2. Audio part

Since the recording made by the camcorder (which, besides the video part, records also the audio part) has a sufficient quality

for our purposes, we decided not to use another recording device for audio recording. Audio recording will be used for the training and testing of both audio-only and audio-visual recognizers. The description of the acoustic part of the recognizer is beyond the scope of this paper.

4. Data evaluation

Parameters of the driver's audio-visual speech corpus are summarized in Table 1. The total length of the corpus is nearly 5 hours. It is divided into training and testing data sets. The testing data set has two parts, which makes it possible to study the dependence of a new speaker (speaker who is not in the training data set) on the recognition rate.

Speaker data set	
Continuous speech	200 sentences
Isolated words	15 commands 15 names of places 15 numerals
Spelling	7-10 words
Data length per speaker	approx. 20 min.

Data sets		
Training data set	10 speakers	3:20 h
Testing data set	4 speakers	1:20 h
Total	12 speakers	4:40 h

Table 1: Parameters of the speech corpus.

5. Conclusion

We have designed an audio-visual speech corpus for automatic audio-visual speech recognition (incorporation of lip-reading into so far audio-only speech recognition) collected in real conditions of a riding car. Since the target application is the control of car features, it is the driver's speech that is collected. We have proposed the size of such a corpus and its structure. The corpus can be divided into training and testing parts. It can also be divided into continuous speech, isolated words and spelling parts.

We have collected audio-visual speech data in the proposed extent. We have pre-processed especially the visual part of the data in order to prepare them for parameterization experiments. The collected corpus will serve as a basis for future research into audio-visual speech recognition. Let us note that the collected corpus is in the Czech language.

6. Future work

Our future work will be directed at least at two domains. Firstly, we will design an audio-visual in-vehicle speech recognizer using the training data from this corpus. We can test the results of such a recognizer using the testing data set of this corpus. Such experiments should prove the ability of visual information to significantly improve the recognition rate in a noisy environment. The next aim of such experiments will be the design of a command control recognizer driven by continuous speech.

Secondly we will use this corpus for experiments on visual speech parameterization. There are several approaches to the parameterization of visual speech which try to describe the shape of lips for recognition. We will study these approaches and try to discover new parameterizations in order to describe the lip shape better and thus improve the audio-visual speech recognition methods.

Finally, we plan to double the amount of the corpus data by recording other speakers' data sets in other cars. We believe that bringing the total length up to 10 hours of speech can improve the quality of the corpus for recognizer training. In this way we also want to avoid the influence of recording in one car only. Furthermore, it would be best to collect data in many different cars, especially different in their noisy conditions.

7. Acknowledgements

This research was supported by the Grant Agency of the Czech Republic, project No. GAČR 102/03/0650 and by the Ministry of Education of the Czech Republic, project No. MSM235200004.

8. References

[1] J. H. L. Hansen, P. Angkititrakul, S. Gallant, U. Yapanel, B. Pellom, W. Ward, and R. Cole, "Cu-move: Analysis & corpus development for interactive in-vehicle speech systems," in *Proceedings of Eurospeech 2001 (CD-ROM)*, Ålborg, Denmark, 2001.

[2] P. Polák, J. Vopička, and P. Sovka, "Czech language database of car speech and enviromental noise," in *Proceedings of Eurospeech 99*, Budapest, Hungary, 1999, vol. 5:2263-6.

[3] G. Potamianos, E. Cosatto, H. P. Graf, and D. B. Roe, "Speaker independent audio-visual database for bimodal asr," in *Proceedings of Eurospeech 2001 (CD-ROM)*, Ålborg, Denmark, 2001.

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

[5] M. S. Gray, T. J. Sejminovski, and J. R. Movellan, "A comparison of image processing techniques for visual speech recognition applications," ftp://ergo.ucsd.edu/pub/tulips1/, 2001.

[6] M. U. Ramos Sánchez, J. Matas, and J. Kittler, "Statictical chromaticity models for lip tracking with b-splines," http://www.tele.ucl.ac.be/PROJECTS/M2VTS/, 1997.

[7] P. Císař and M. Železný, "Feature selection for the czech speaker independent lip-reading," in *Proceedings of ECMS 2003 (CD-ROM)*, Liberec, Czech Republic, 2003.

[8] M. Železný, P. Císař, Z. Krňoul, and J. Novák, "Design of an audio-visual speech corpus for the czech audio-visual speech synthesis," in *Proceedings of ICSLP 2002 (CD-ROM)*, Denver, USA, 2002.

[9] V. Radová and P. Vopálka, "Methods of sentences selection for read-speech corpus design," in *Proceedings of TSD 1999 (CD-ROM)*, Berlin, Germany, 1999.

[10] G. I. Chiou and J.-N. Hwang, "Lipreading from color motion video," in *Proceedings of ICASSP 1996 (CD-ROM)*, Atlanta, USA, 1996.

[11] M. Heckmann, K. Kroschel, Ch. Savariaux, and F. Berthommier, "Dtc-based video features for audio-visual speech recognition," in *Proceedings of ICSLP 2002 (CD-ROM)*, Denver, USA, 2002.