

The LIUM-AVS database : a corpus to test lip segmentation and speechreading systems in *natural* conditions

Philippe Daubias, Paul Deléglise

Laboratoire d'Informatique de l'Université du Maine
Le Mans, France

{Philippe.Daubias,Paul.Deleglise}@lium.univ-lemans.fr

Abstract

We present here a new freely available audio-visual speech database. Contrary to other existing corpora, the LIUM-AVS corpus was recorded in conditions we qualify as *natural*, which are, according to us, much closer to real application conditions than other databases. This database was recorded without artificial lighting using an analog camcorder in camera mode. Images were stored digitally with no compression to keep the highest possible image quality. The LIUM-AVS database comprises two parts:

PBS Phonetically Balanced Sentences in French

LET Spelled letters (also in French)

These two parts contain sequences with both *natural* and *blue* lips. The whole database is released mainly to test and compare lip segmentation approaches on *natural* images, but speech recognition experiments may also be carried using this corpus. For information on obtaining the LIUM-AVS database, please contact us through our webpage (<http://www-lium.univ-lemans.fr/lium/avs-database>).

1. Introduction

Since the first experiments on machine lipreading by Petajan [1], numerous studies about automatic audio-visual speech recognition have been published [2, 3, 4, 5, 6, 7]. Some have a low-level image-based approach and the others a higher level model-based approach. These studies present —sometimes contradictory— results but comparison between them is quite difficult as they were mostly carried on different corpora. One attempt at proposing a common database has recently been made with CUAVE [8], but comparative results of different approaches are still to come.

Some studies nevertheless share the same databases and may be compared: for example, through shape and appearance modelling, Luettin [9] reaches higher lipreading performance with a model-based approach than Movellan [3] with an image-based approach on the —rather small— Tulips1 database. More recently Potamianos [5] has obtained better results with an image-based approach than with a model-based one on the larger AT&T audio-visual database. This result was confirmed on the IBM Viavoice® audio-visual database by Neti et al. [10], as model-based approaches totally failed due to the size of the database. Finally, Heckmann [7] has very recently obtained similar performance with both approaches (DCT image-based features and geometric measurements model-based features) on two different corpora corresponding to the same number spelling tasks, but model-based parameters were obtained using blue-marked lips.

It should then be possible to reach an equivalent performance level with model-based approaches (measurements) than with image-based approaches (DCT coefficients), provided the geometric measurements are very precise and with no errors. It is obvious that the quality of the lip-measurements obtained on *blue* lips cannot be reached yet with unprepared subjects. Error-free measurements without preparation of the subjects are probably even extremely hard to obtain in *natural* conditions, but we think that work towards this goal should be carried out.

It is not clear which approach will suffer the less from the degradation of image quality between “ideal” corpus images and real application images. On the one hand DCT for example is not shift invariant and requires high ROI location precision. Image-based approaches are consequently presumably favored by constant lighting conditions which enables precise ROI location. On the other hand, lip shape and position estimation which is necessary for model-based approaches will also be less precise and more error-prone if shadows or other disturbing phenomenon are present on the images. No conclusions can be drawn *a priori* on the approach to use for lipreading applications in real environments and we have recorded a corpus in conditions we qualify as *natural*, that we believe are more realistic for a real lipreading application. After proving it possible to model lip-appearance automatically with accuracy using a statistical model in these conditions [11], we have decided to make the corpus upon which our experiments were based, available to the speech community through this article. This new audio-visual database may be used for example to test lip-segmentation approaches or imaged-based audio-visual speech recognition in *natural* conditions.

After presenting other existing audio-visual databases in section 2, we will give more details about the contents (section 3) and technical specifications (section 4) of the LIUM-AVS database.

2. Other audio-visual speech databases

As for acoustic only databases, audio-visual databases correspond to a specific speaking task (isolated or continuous speech, small to large vocabulary) in a specific language. Not taking into account that some databases are only recorded in grayscale [3, 12], audio-visual databases also differ regarding the recording conditions :

- subjects may be prepared with a specific make up or not (colored plastic dots placed at particular locations on the face, blue lipstick),
- lighting conditions can be more or less controlled (from powerful frontal lighting to ambient *natural* daylight),

- and the subject’s head movements may be restricted (physically blocked or diminished by the fact that they have to read a fixed prompt).

All prepared-speaker databases are useful for research but real lipreading applications may hardly require the user to wear blue lipstick to use them ! So these databases are unlikely to be used for training or testing systems devoted to real conditions.

Databases with unprepared subjects do also exist. At the exception of the M2VTS corpus [13] which is in French, most large or widely spread audio-visual databases (Tulips1 [3], AVletters [12], BT DAVID [14], AT&T [15], XM2VTSDB [16], CMU [6], IBM Viavoice® [10] and CUAVE [8]) are in English¹. Some databases with unprepared subjects do also exist for other languages such as German [2] or Dutch [18] but they are either not published or not widely spread yet. All these databases contain subjects with bare lips, but they were recorded in “studio conditions” using artificial lighting. Images in these databases are closer to what could be obtained in an office environment, but even if such databases are useful for research and systems comparison, one can regret that their artificial lighting conditions are probably not representative of real lipreading applications as offices are rarely windowless rooms.

To tackle lip location in *natural* daylight conditions, we have developed a method to locate robustly and automatically lips on unprepared subjects with *a posteriori* models [11], but this requires having both *natural* and *blue* sequences of the same content. As no corpus was usable to test this method (not even BT DAVID), we had to record our own. In the following sections, we describe more precisely this database.

3. The LIUM-AVS database contents

The LIUM-AVS database comprises more than 17 000 images of nine subjects (8 males, 1 female), seven of whom were recorded wearing blue lipstick (6 males, 1 female) to allow training and testing of different shape models, and six with no make-up (5 males, 1 female) for training and testing different appearance models. The database is separated in two sub-corpora: the LET corpus which contains spelled letters in French and the PBS corpus in which subjects utter Phonetically Balanced Sentences, also in French. Both parts contain *natural* and cosmetically assisted *blue* lips.

One subject (BJ) was recorded in both subparts of the database. He went through more sessions (*blue* and *natural*) than the others in the LET subpart to permit small speech recognition experiments. He was also recorded in the PBS subpart to enable automatic appearance models training with data specific to this subject. In the following subsections, we describe both subparts of the database in more details.

3.1. LET corpus

This first part of the LIUM-AVS database is dedicated mainly to speech recognition experiments on spelled letters, a small vocabulary but highly confusable task. It was also designed to build automatically and compare lip-shape models of different subjects which were therefore recorded with blue lipstick on their lips. This part comprises five subjects (4 males, 1 female), spelling letters in French. Fig. 1 shows a sample image for each subject recorded in the LET sub-corpus.

Each subject spells from twenty to forty letters in different sessions. For one subject (see top left of Fig. 1), we have also

¹For a more detailed description of these databases, see <http://www-lium.univ-lemans.fr/~daubias/databases> or [17]

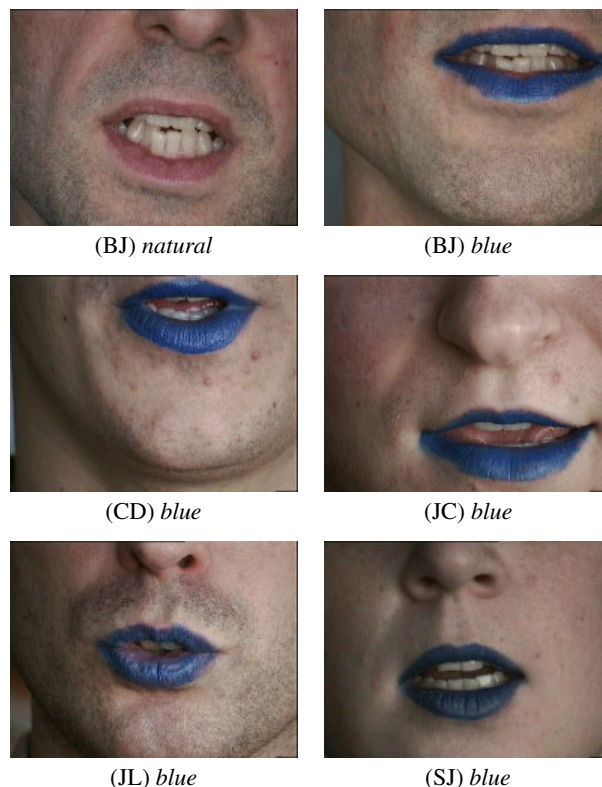


Figure 1: A sample image of each subject from the LET part of database.

recorded 80 series of four letters without cosmetic assistance. These sessions are intended for audio-visual speech recognition experiments (or purely visual speech recognition), and may be used through the so called jack-knife technique as the corpus is quite small compared to the size of the vocabulary (only 12 repetitions for each letter).

3.1.1. Phonetic labelling

We have labelled phonetically the 80 *natural* series of four letters from the LET corpus. This was done automatically, and errors were corrected by hand.

3.2. PBS corpus

This second part of the database was recorded to test a new automatic labelling method presented elsewhere [11], which enables automatic *a posteriori* appearance models training. It comprises both *natural* and *blue* sentences uttered by six subjects (5 males, 1 female), two of whom (1 male, 1 female) were also recorded in the LET part of the corpus. The PBS part thus gives the opportunity to compare the lip shape models of a given subject for different speaking styles.

Four different phonetically balanced sentences in French from BDSONS [19] were chosen randomly by three subjects (NR, SJ, TL) and repeated twice: the first time with blue lipstick and the second time bare lips. Another subject (BJ) recorded six sentences with blue lipstick and the twelve sentences without. Finally two more subjects (GG, SG) uttered once (bare lips) the twelve sentences. Fig. 2 shows a sample image for each subject recorded in the PBS sub-corpus.

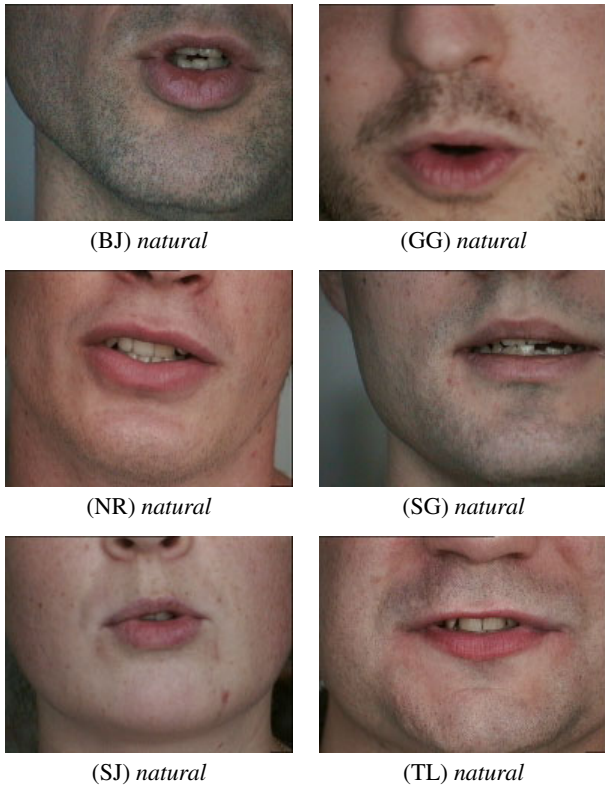


Figure 2: A sample image of each subject from the PBS part of database.

3.2.1. Hand-labelled images

We have manually located the lips on a few *natural* images to compare our automatic labelling method to a “ground truth”. In total about 50 images are labelled per subject for BJ, NR, SJ and TL : the outer-lip boundary is represented with 24 points and the inner-lip contour with 20 points.

4. Technical specifications of the database

As indicated among others by Wojdel [18], to decrease significantly the storage requirement of images, video compression (JPEG, MPEG) degrades their quality, especially their chromaticity. As most face and lip locating systems rely on the use of color (hue in most cases), not to lose any potentially useful information, images should be stored without compression. With the hardware which was available to us at the time of recording, it was only possible to grab 384×288 pixels sized color images at a 25 fps frame rate without compression. As PAL resolution is 768×576 , to comply with the maximum resolution recordable using our hardware, we had to chose between:

- grabbing full face images in PAL and then cropping the mouth region (thus keeping interleaved images),
- zooming to the mouth region and reducing the image size by a factor of two (eliminating all odd columns and lines and making un-interleaved images).

As decreasing the video parameter rate to 25 Hz does not degrade much speechreading capability according to Potamianos [5], the second option was chosen.

The camera was set up manually to have the lip region roughly centered in the image, with a zooming factor allowing lips to cover from one to two third of image width. The distance between the camera and the speaker was variable, and so was the zooming factor. Contrary to other databases, the subjects can move their head freely. Some speakers do move either before or during recording and as a consequence of the important zooming factor, the lips sometimes fall out of the image. In that case (about 10 times on the whole database), the recording was re-done but the subjects were never asked to restrict their head movements during speech. As head tracking systems have been reported [20], we suppose that the face of the subject can be roughly located and that its location can be used to automatically adjust another camera to capture the region we have set manually.

To comply with the *natural* conditions, the corpus was recorded without paying much attention to the lighting conditions. Natural daylight was used, which resulted in shadows on some images under the nose and mouth of the subjects. The windows of the room where the recordings took place were located on the left side of the subjects and the right parts of the images are consequently often better illuminated.

Videos were captured uncompressed at a resolution of 384×288 pixels in 24 bit RGB color, at a frame rate of 25 un-interleaved images per second. For video acquisition, a PAL analog camcorder in camera mode was used. It was connected to a SUN ULTRA2 workstation with an analog video card for digitization. Audio was acquired through a standard SUN microphone and recorded on one single audio channel (mono) at a sample rate of 16 kHz with 16 bit samples. Due to the proximity of other workstations, background noise can be heard on this audio channel. The audio and visual data were recorded in real time on the same workstation (simultaneously) but by two separated processes. Through a set of measurements, we have calculated the delay between video and audio acquisition which enables synchronization of the two channels (video starts 60 ms before audio).

Table 1 gives an overview on the database contents. The LET part is about 3,3 GByte and the PBS part about 2,1 GByte. Information on how to obtain the LIUM-AVS database can be found at the following address : <http://www-lium.univ-lemans.fr/lium/avs-database>

5. Previous experiments on the LIUM-AVS database

The bulk of the database was used for experiments on audio-visual automatic speech recognition (AV ASR). More precisely, AV ASR experiments were carried out with the LET part of the database using the jack-knife (leave one out) technique and were reported in [21]. The PBS part of the database was used

Subpart	LET	PBS
Subject	4 males, 1 female	5 males, 1 female
Utterances	418 letters	66 sentences
Frames	10450	6600
Image size	384×288 pixels	384×288 pixels
Mouth size	$\approx 200 \times 100$ pixels	$\approx 200 \times 100$ pixels
Compression	none	none

Table 1: Characteristics from the LIUM-AVS database.

to build different shape and appearance models of the lips: with ANN-based models and an automatic labelling method, we managed to segment lips from skin and inner mouth with relatively high accuracy [11].

6. Conclusions and planned extensions

In this paper, we have presented a new audio-visual speech corpus. It was recorded without artificial lighting and contain images which are, to our point of view, more representative of what could be obtained in real conditions for a true lipreading application, than other existing audio-visual databases. Although small, it is a versatile database which may be used to test lip segmentation approaches or for small speechreading experiments.

At least 120 more series of four letters should be recorded by the subject BJ to extend the *natural* LET part to 200 series and reach 30 repetitions per letter. More images should also be hand-labelled, in particular for speakers GG and SG, but as this is a time-consuming task and should be redone in case of lip-shape model modification, we have tried to keep hand-labelling as little as possible and are currently still working on automatic labelling methods instead.

7. Acknowledgements

The authors would like to thank all the subjects recorded in the database for their patience.

8. References

- [1] E. D. Petajan, "Automatic lipreading to enhance speech recognition," in *Proc. CVPR*, June 1985, pp. 40–47.
- [2] C. Bregler and Y. Konig, "'Eigenlips" for robust speech recognition," in *Proc. ICASSP*, Adelaide, Australia, Apr. 1994, vol. II, pp. 669–672.
- [3] J. R. Movellan, "Visual speech recognition with stochastic networks," in *ANIPS*, G. Tesauro, D. Touretzky, and T. Leen, Eds., Cambridge, MA, USA, 1995, vol. 7, pp. 851–858, The MIT Press.
- [4] S. Dupont and J. Luetttin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.
- [5] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. ICIP*, Chicago, IL, USA, Oct. 1998, vol. III, pp. 173–177.
- [6] X. Zhang, C. C. Broun, R. M. Mersereau, and M. Clements, "Automatic speechreading with applications to human-computer interfaces," *EURASIP Journal on Applied Signal Processing, special issue on Joint Audio-Visual Speech Processing*, vol. 2002, no. 11, pp. 1228–1247, Nov. 2002.
- [7] M. Heckmann, K. Kroschel, C. Savariaux, and F. Berthommier, "DCT-based video features for audio-visual speech recognition," in *Proc. 7th ICSLP*, Denver, CO, USA, Sept. 16–20 2002, vol. 3, pp. 1925–1928.
- [8] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Proc. ICASSP*, Orlando, FL, USA, May 13–17 2002, vol. II, pp. 2017–2020.
- [9] J. Luetttin and N. A. Thacker, "Speechreading using probabilistic models," *Computer Vision and Image Understanding*, vol. 65, no. 2, pp. 163–178, Feb. 1997.
- [10] C. Neti, G. Potamianos, J. Luetttin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Tech. Rep. Workshop 2000, International Computer Science Institute, Center for Language and Speech Processing (CLSP), The Johns Hopkins University, Baltimore, MD, USA, Oct. 12 2000.
- [11] P. Daubias and P. Deléglise, "Statistical lip-appearance models trained automatically using audio information," *EURASIP Journal on Applied Signal Processing, special issue on Joint Audio-Visual Speech Processing*, vol. 2002, no. 11, pp. 1202–1212, Nov. 2002.
- [12] I. Matthews, J. Bangham, and S. Cox, "Audiovisual speech recognition using multiscale nonlinear image decomposition.," in *Proc. 4th ICSLP*, Philadelphia, PA, USA, Oct. 3–6 1996, vol. 1, pp. 38–41.
- [13] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database," in *Proc. 1st AVBPA*, J. Bign, G. Chollet, and G. Borgefors, Eds., Crans-Montana, Switzerland, Mar. 12–14 1997, pp. 403–410, Springer-Verlag.
- [14] C. C. Chibelushi, S. Gandon, J. S. D. Mason, F. Deravi, and R. D. Johnston, "Design issues for a digital audio-visual integrated database," in *IEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication*, Savoy Place, London, UK, Nov. 1996, pp. 7/1–7/7.
- [15] G. Potamianos, E. Cosatto, H. P. Graf, and D. B. Roe, "Speaker independent audio-visual database for bimodal ASR," in *Proc. 1st ESCA Workshop on Audio-Visual Speech Processing (AVSP'97): Cognitive and Computational Approaches*, C. Benoît and R. Campbell, Eds., Rhodes, Greece, Sep. 26–27 1997, pp. 65–68.
- [16] K. Messer, J. Matas, J. Kittler, J. Luetttin, and G. Maître, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd AVBPA*, Washington, DC, USA, Mar. 22–23 1999, pp. 72–77.
- [17] P. Daubias, *Modèles a posteriori de la forme et de l'apparence des lèvres pour la reconnaissance automatique de la parole audiovisuelle*, Ph.D. thesis, Université du Maine, Le Mans, Dec. 5 2002.
- [18] J. C. Wojdel, P. Wiggers, and L. J. M. Rothkrantz, "An audio-visual corpus for multimodal speech recognition in dutch language," in *Proc. 7th ICSLP*, Denver, CO, USA, Sept. 16–20 2002, vol. 3, pp. 1917–1920.
- [19] R. Descout, J.-F. Ségrinat, O. Cervantes, and R. Carré, "BDSONS : Une base de données des sons du fran cais," in *Proc. 12th ICA*, Toronto, Canada, 1986.
- [20] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel, "Towards unrestricted lip reading," *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, vol. 14, no. 5, pp. 571–586, 2000.
- [21] P. Daubias and P. Deléglise, "Lip-reading based on a fully automatic statistical model," in *Proc. 7th ICSLP*, Denver, CO, USA, Sept. 16–20 2002, vol. 1, pp. 209–212.