BL-Database: A French audiovisual database for speech driven lip animation systems

Yannick Benezeth, Grégoire Bachman, Guylaine Le-Jan, Nathan Souviraà-Labastie, Frédéric Bimbot

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

BL-Database: A French audiovisual database for speech driven lip animation systems

Yannick Benezeth — Grégoire Bachman — Guylaine Le Jan — Nathan Souviraà-Labastie — Frédéric Bimbot

N° 7711

August 2011

_ Audio, Speech, and Language Processing _

apport de recherche

ISSN 0249-6399 ISRN INRIA/RR--7711--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
RENNES - BRETAGNE ATLANTIQUE

BL-Database: A French audiovisual database for speech driven lip animation systems

Yannick Benezeth, Grégoire Bachman, Guylaine Le Jan, Nathan Souviraà-Labastie, Frédéric Bimbot

Theme : Audio, Speech, and Language Processing
Perception, Cognition, Interaction
Équipe-Projet Metiss

Rapport de recherche n° 7711 — August 2011 — 9 pages

Abstract: The lack of publicly available annotated databases is a major limitation to research advances in speech processing. We describe in this paper an audiovisual speech database which is being made available to the research community. Our database, called *BL-database (Blue Lips-database)*, consists of 238 utterances spoken by 17 speakers. The recordings have been performed during two sessions. The data of the first session can be used to analyze the 2D movements of the mouth while the data collected by the second session is dedicated to 3D analysis. The audio signal has been phonetically segmented and labeled. Such data is expected to be of great interest to all research groups working on multimodal automatic speech recognition, audio/visual synchronization or speech-driven lip animation.

Key-words: 3D audiovisual speech database, speech-driven lip animation.

Centre de recherche INRIA Rennes – Bretagne Atlantique
IRISA, Campus universitaire de Beaulieu, 35042 Rennes Cedex
Téléphone : +33 2 99 84 71 00 — Télécopie : +33 2 99 84 71 71

BL-Database : Une base de données audiovisuelle en Français pour l'animation labiale à partir de flux de parole

Résumé : Le manque de bases de données annotées et librement accessibles est une limitation importante aux avancées de la recherche sur le traitement de la parole. Nous décrivons dans cet article une base de données audiovisuelle mise à la disposition de la communauté scientifique. Notre base de données audiovisuelle, appelée *BL-database*, est composée de 238 phrases prononcées par 17 locuteurs. Les enregistrements se sont déroulés lors de deux sessions. Les données de la première session peuvent être utilisées pour l'étude des mouvements 2D de la bouche d'un locuteur alors que les données de la deuxième session permettent une analyse des mouvements 3D de la bouche. Nous fournissons également, en plus des données audiovisuelles, la transcription phonétique de la base. Ces données peuvent être utilisées pour des travaux de recherche portant sur la reconnaissance de la parole multimodale, la synchronisation audiovisuelle ou sur l'animation labiale à partir de la parole.

Mots-clés : base de données audiovisuelle 3D, animation labiale.

1 Introduction

Over the past decade, speech databases have been collected, annotated and shared to the research community. Such data are very important to speed-up the development of innovative speech-related algorithms and are also necessary in order to fairly compare new algorithms over the same data. Most of these speech databases are composed of audio files associated with the corresponding phonetic transcription. These databases are primarily designed for the development of automatic speech recognition and speaker authentification systems, and for linguistic research. Most are in English (*e.g.* [1, 2]) but some are also in French [3, 4].

It is important to note that human perception of speech is multimodal. The acoustic signal is the primary choice for recognizing speech but visual information (mainly lips, tongue and jaw movements) also contribute to the perception of speech. The multimodality of speech has been demonstrated with the well-known McGurk effect [5]. Consequently, multimodal speech corpora have been collected. These corpora are usually composed of audio and visual modalities [6, 7] whereas the Mocha corpus [8] is constituted by audio and sensor signals describing the movement of lips and tongue. These multimodal corpora are used to develop multimodal speech recognition or speech-driven lip animation algorithms.

We are interested in this latter category of method. Our objective is to study the correspondence between acoustic signals of speech and mouth movements in order to synthetically reproduce these movements on virtual characters. The work presented in this paper is within the context of the project Rev-TV which aims to create a new category of TV programs in which viewers have the opportunity to interact with the content of a TV program through an immersive and user friendly environment. Viewers will have their virtual representation in a synthetic environment. This avatar will be controlled by several methods (motion and video analysis, speech processing, etc.). In order to obtain natural looking and realistic animation of virtual characters, the lip movements shall be consistent with the speech. It is possible to make such an animation manually, by adjusting frame after frame, the control parameters of the avatar lips. The obtained results are then very realistic, but this method requires a substantial workload and does not allow real time avatar animations. Hence, methods to automatically generate lip movements from speech have been proposed. These methods can be divided into two classes, namely (1) the low-level mapping between acoustic features and mouth positions control parameters and (2) the mouth shape recognition techniques.

Methods in (1) establish the direct correspondence between audio descriptors, *e.g.* Mel Frequency Cepstral Coefficients (MFCC) or Linear Predictive Coding (LPC) and parameters controlling the shape of the mouth (aperture, protrusion, *etc.*). This correspondence can be learned by neural networks, Gaussian mixture models or vector quantization (*e.g.* [9, 10]). Then, methods in (2), are based on the mouth shape recognition from the speech signal. The mouth movements are then monitored with a discrete set of positions: the visemes. A viseme is a shape of the mouth associated with particular sounds. Observation vectors provided by a spectral and/or a temporal analysis are used to estimate the parameters of the mouth shape recognition system. Well-known techniques in automatic speech recognition, such as Hidden Markov models (HMM) [11],

can be used. The recognized acoustic unit can be the phoneme [12] or the viseme [13]. If the recognition system is based on the phoneme recognition, a correspondence between each detected phoneme and its corresponding viseme is performed.

Unlike methods in (2) which only use an unimodal corpus with audio and phonetic transcription, methods in (1) necessitate audiovisual corpora for the supervised learning of the mouth control parameters. In this work, we are interested in speech-driven animation systems based on continuous and 3D mouth control parameters (horizontal and vertical aperture, and protrusion). Consequently, we have collected the *BL-database* in which we have recorded the 3D movements of the mouth of several speakers uttering specific sentences. The experimental protocol is made so that it is possible to recover 3D information of the mouth. Two spatially calibrated cameras and one depth camera are used. The utterances have been selected in order to represent a wide variety of the diphones in the French language. Audio data have been recorded with 2 microphones. In order to easily extract the lips position in the image sequences, people were wearing blue lipstick.

This material differs from most other audio-visual databases mainly in the following aspects:

1. Specific recording materials allow to recover 3D information of the movements of the mouth,
2. The language of the database is French,
3. The speech material is composed of most the diphones of the French Language.

In the following, we present the experimental protocol and the composition of the dataset.

2 BL-Database

The *BL-database* is an audio-visual speech corpus composed of 238 sentences uttered by 17 speakers. The actual speech length by speaker is around 20 minutes. The recordings have been performed during two sessions. For the first recording session, data have been recorded with a front view color camera and two microphones. For the second session, we have used two calibrated cameras, one depth camera and two microphones. The language of the corpus is French. In the following, we present the database content, the recording equipment, the speakers and the speech material.

2.1 Database content

As explained previously, the recordings have been performed during two sessions. The data of the first session can be used to analyze the 2D movements of the mouth while the data collected by the second session have been performed for 3D analysis.

Data and information collected and generated during the 2 recording sessions are:

- Session 1 (dedicated to the 2D analysis): 238 utterances spoken by 4 men and 4 women
  - Audio data from 2 microphones (each audio file corresponds to one sentence uttered by one speaker),
  - Video data recorded by 1 front-view camera (each video file corresponds to one sentence uttered by one speaker),
  - Time-aligned phonetic transcription associated to the audio and video data,
  - Non-nominative speaker information (age, gender).
- Session 2 (dedicated to the 3D analysis): 238 utterances enunciated by 5 men and 4 women
  - Audio data from 2 microphones (each audio file corresponds to one sentence uttered by one speaker),
  - Video data recorded by 2 spatially calibrated cameras and 1 depth camera (each video file corresponds to one sentence uttered by one speaker and recorded by one camera),
  - Time-aligned phonetic transcription associated to the audio and video data,
  - Calibration data (23 images of a planar checkerboard),
  - Non-nominative speaker information (age, gender).

It is possible to use the data of session 2 in addition to the data of session 1 for 2D analysis. Finally, depending on your application, the BL-Database is composed of:

- 17 speakers (about 340 minutes) for 2D analysis,
- 9 speakers (about 180 minutes) for 3D analysis.

2.2 Recording equipment

For each session, audio channels are sampled at 44.1 kHz with 16 bits per sample. We have used one AKG microphone, which is composed of a CK91 capsule with a AKG SE300K preamp, and an omnidirectional Labtec microphone. AKG CK91 is a cardioid condenser capsule. For A/D conversion, a Mackie ONYX 1200F soundcard has been employed. The recordings have been realized in a specific recording room which is reasonably sound-proof and with a low reverberation.

For the first session, video data have been recorded with a front view color camera (a Canon MVX3i) which records 25 interlaced frames per second and has a resolution of 576x720. For the second session, video data has been recorded with the front view color camera spatially calibrated with another color side-view camera and one depth camera. The side view color camera records at 30 frames per second and has a resolution of 640x480. The depth camera, a Microsoft Kinect, records 30 depth images per second at a resolution of 640x480. In order to ensure comparable lightning conditions, the room was shaded and illuminated only by an artificial light.

Cameras and microphones have been arranged on a table near the speaker in order to achieve a high resolution close-up view of the speaker's face and a sufficient signal-to-noise ratio of audio signals. A standard PC display was placed in front of the speaker indicating the sentences to pronounce. The recording setup is illustrated in the Figure 1.

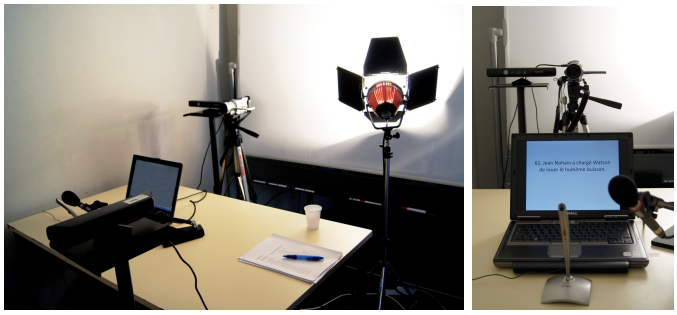

Figure 1: Illustration of the recording setup.

2.3 Speakers

Using only one speaker greatly simplifies the recordings as well as the extraction and interpretation of the data but, in this case, it is not possible to determine general properties that hold for all speakers. Consequently, we design a corpus that contains the spoken utterances of 17 native-French speakers. The speakers was aged from 23 to 48 years-old. Table 1, presents the number of speakers in the database in each session.

Table 1: Content of the Database.

Session / Gender	#1	#2
Male	4	5
Female	4	4

Figure 2 presents two speakers of the database recorded with the front and side-view cameras.

2.4 Speech material

The speech materials is composed of continuous spoken words. The material was developed by G. Gibert at ICP (Grenoble, France). More details about the content of the speech material can be found in [14] (or in French with more details in [15]). The corpus is composed of 238 sentences, containing all diphones of the French langage.

Table 2 presents the 33 phonemes of our dataset. It is possible to observe that the frequency of occurrence of phonemes in the BL-database is closely related to the frequency of occurrence of phonemes in the French language (calculated with the ESTER dataset [3]).

Figure 2: Sample speakers from the database (front and side view).

2.5 Specific instructions

Each recording begins with some explanations on the objectives of the experiment. The speaker is made-up with blue lipstick and is asked to stand in front of the camera. Then, the operator explains to the speaker that he has to pronounce each sentence fluently but without exagerating. In case of mistake, the speaker can repeat the sentence. Before beginning the recording, each speaker is invited to practice his elocution on 10 sample sentences. Even if an operator was present in the room, the speaker control the slides scrolling by himself.

2.6 Post-processing and annotation

The audio signal is first manually segmented into sentences and is then segmented into phonemes with a semi-automatic annotation process. We have used forced alignment, using HTK [1], and have then manually adjusted the phoneme borders.

Each movie is compressed with MPEG4 codec.

3 Conclusion

In this paper, we have presented a new French audiovisual database. This database is expected to be useful for research concerning multimodal automatic speech recognition, audio/visual synchronization or speech-driven lip animation. The dataset has been recorded using specific recording materials (two color calibrated cameras and one depth camera) in order to study 3D information of the movements of the mouth. The corpus is balanced between men and women speakers to determine general audio properties that hold for various speakers. Finally, the corpus has been annotated at the sentence and the phoneme level. The dataset is freely available upon request available at http://bl-database.inria.fr/.

[1] http://www.voxforge.org/home/dev/autoaudioseg

Table 2: A list of the 33 phonemes in the corpus with their occurrence frequency.

IPA symbol	Occurrence	Occurrence frequency	
		BL-Database	ESTER
a	464	7,97	8,12
l	388	6,66	6,35
ʁ	370	6,35	8,20
i	326	5,60	5,74
ø	296	5,08	4,49
t	280	4,81	5,29
e	240	4,12	5,47
s	232	3,98	6,08
ɛ	222	3,81	4,77
d	206	3,54	5,13
y	196	3,37	2.08
ɛ̃	171	2,94	1,58
ã	165	2,83	3,28
o	164	2,82	1,11
p	163	2,80	3,52
n	162	2,78	3,00
k	162	2,78	4,04
u	154	2,64	1,73
b	154	2,64	1,17
m	147	2,52	3,01
j	132	2,27	2,07
ɡ	126	2,16	0,62
ʒ	120	2,06	1,12
v	119	2,04	1,83
ɔ̃	98	1,68	2,10
f	95	1,63	1,38
z	93	1,60	1,62
ʃ	81	1,39	0,50
ɔ	69	1,18	2,56
w	67	1,15	0,81
œ	55	0,94	0,51
ɳ	39	0,67	0,11
ɥ	38	0,65	0,43

References

[1] J.J. Godfrey, E.C. Holliman and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development", IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 517–520, 1992.

[2] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus", in Linguistic Data Consortium, 1993.

[3] S. Galliano, E. Geoffrois, J.F. Bonastre, G. Gravier, D. Mostefa, and K. Choukri, "Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News". In Language Resources and Evaluation Conference, 2006.

[4] M. Avanzi, A.C. Simon, J.P. Goldman and A. Auchlin, "C-PROM. An annotated corpus for french prominence studies", Proceedings of Prosodic Prominence: Perceptual and Automatic Identification, Speech Prosody, 2010.

[5] H. McGurk and J. MacDonald, "Hearing lips and seeing voices", Nature, Vol. 264, no. 5588, pp. 748–756, 1976.

[6] E.K. Patterson, S. Gurbuz, Z. Tufekci and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 2017–2020, 2002.

[7] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu and T. Huang, "AVICAR: Audio-Visual Speech Corpus in a Car Environment", Interspeech, 2004.

[8] A. Wrench and W. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in Proceedings of the 5th Seminar on Speech Production, pp. 305–308, 2000.

[9] T. Frank, M. Hoch and G. Trogemann, "Automated Lip-Sync for 3D-Character Animation", 15th IMACS World Congress on Scientific Computation, Modelling and Applied Mathematics, pp. 24–29, 1997.

[10] S. Nakamura, "Statistical multimodal integration for audio-visual speech processing", IEEE Transactions on Neural Networks, pp. 854–866, Vol. 13(4), 2002.

[11] L.-R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, Vol. 77(2), pp. 257–286, 1989.

[12] J. Park and H. Ko, "Real-Time Continuous Phoneme Recognition System Using Class-Dependent Tied-Mixture HMM With HBT Structure for Speech-Driven Lip-Sync", IEEE Transactions on Multimedia, Vol. 10(7), pp. 1299–1306, 2008.

[13] S.-W. Foo and L. Dong, "Recognition of Visual Speech Elements Using Hidden Markov Models ", Advances in Multimedia Information Processing, Vol. 2532, pp. 153–173, 2002.

[14] G. Gibert, G. Bailly, D. Beautemps, F. Elisei, R. Brun, "Analysis and synthesis of the three-dimensional movements of the head, face and hand of a speaker using Cued Speech", in Journal of the Acoustical Society of America, 118(2), pp 1144-1153, 2005.

[15] G. Gibert, "Conception et évaluation d'un système de synthèse 3D de Langue française Parlée Complétée (LPC) à partir du texte", PhD Thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 2006.

Centre de recherche INRIA Rennes – Bretagne Atlantique
IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399