# A New Multi-modal Database for Developing Speech Recognition Systems for an Assistive Technology Application

António Moura[1], Diamantino Freitas[2], and Vitor Pera[2]

[1] School of Technology and Management, Polytechnic Institute of Bragança
Quinta de Sta Apolónia, Apartado 134, 5301-857 Bragança, Portugal
Email: moura@ipb.pt
[2] Faculty of Engineering, University of Porto
Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
Email: dfreitas@fe.up.pt, vpera@fe.up.pt

**Abstract.** In this paper we report on the acquisition and content of a new database intended for developing audio-visual speech recognition systems. This database supports a speaker dependent continuous speech recognition task, based on a small vocabulary, and was captured in the European Portuguese language. Along with the collected multi-modal speech materials, the respective orthographic transcription and time-alignment files are supplied. The package also includes data on stochastic language models and the generative grammar associated to the collected spoken sentences. The application addressed by this database, which consists of voice control of a basic scientific calculator, has the particularity of being designed for a person with a specific motor impairment, namely muscular dystrophy. This specificity is a remarkable characteristic, given the lack of such kind of data resources for developing assistive systems based on audio-visual speech recognition technology.

## 1 Introduction

The main motivations for the work carried out constructing the LPFAV2[1] speech database, here presented, can be expressed through the two following research topics: robust speech recognition; and computer technologies for persons with physical disabilities. In general, the use of visual features jointly with the acoustic information is becoming increasingly important as a technique to improve the speech recognition robustness [1,2]. On the other side, examining the contents of many catalogues advertising assistive technology products for persons with disabilities one can confirm the relevance of speech recognition in this application domain. Based on social justice values, it is obvious that the development of technologies effectively useful for this community in particular must be encouraged. From the technological point of view, the audio-visual approach to the development of speech recognisers to be used by persons with muscular dystrophy, in this particular case, presents two essential advantages (both quite obvious): in general the produced acoustic signal is

[1] LPFAV2 stands for: Laboratório de Processamento da Fala – Audio Visual 2 (this is the second database for AVSR created at this speech processing laboratory).

Petr Sojka, Ivan Kopeček, and Karel Pala (Eds.): TSD 2004, LNAI 3206, pp. 385–392, 2004.

weak, so the acoustic robustness is a critical problem; besides, in many applications the user maintains a stable pose, therefore robust visual features can be extracted.

One of the most important requirements for doing work in speech recognition is a database with the appropriate materials for training and testing the systems under development. The size of the database is crucial to achieve the intended results, so collecting and processing the required data to build a useful database is not trivial. In the case of multi-modal databases, this problem usually becomes harder due to the multiple information streams and the huge amount of data [3,4]. Therefore, the limited number of available audio-visual databases is not surprising. To the best of the authors' knowledge, before the LPFAV2 was created only one audio-visual database existed in the particular case of speech recognition applications supported by the Portuguese language [5]. Furthermore, the LPFAV2 is the first one that was specially designed for an application where the user has a specific motor impairment.

Besides supporting research on general issues internal to audio-visual speech recognisers, this database can also be a valuable resource to study relevant topics specific to applications involving the previously referred physical disabilities. Obviously, being a single-speaker database, the scope of those studies presents some limitations and a careful approach must be followed since one cannot expect that some conclusions hold for other speakers too. There are plans to upgrade the LPFAV2 extending its potential, for instance, one important issue that is intended to address in the future is related to the progressive muscle weakness that often affects the user. Since the symptoms can worsen as time goes on, it is important to study the main implications of this for the recognition task, so that useful adapting techniques can be implemented.

The rest of this paper is organised as follows. Section 2 presents a brief analysis of the recognition task that gives justification to this database, which capture is shortly described in Section 3. Section 4 summarizes several topics on the corpus, such as the vocabulary and the associated generative grammar and stochastic language models. The LPFAV2 package is shortly described in Section 5, including short descriptions of the most important aspects involved in the development of the respective materials. The final conclusions are drawn in Section 6.

## 2 The Recognition Task

One of the areas of application of audio-visual speech recognition (AVSR) with high potential is in assistive technology, namely to facilitate the life of speech and motion impaired computer users. The LPFAV2 database was designed having in mind an effectively useful application. From a range of hypothesis, the speech interface module to operate a scientific calculator was selected for target application. One of the main reasons for this selection is the relatively small complexity of the recognition task, allowing to shorten the development cycle. The usefulness of this application could be confirmed by a student at this school, suffering from muscular dystrophy, who made himself ready for the recording sessions. This user presents a severe motor impairment at the level of the upper limbs and a moderately perceptible low intensity speech. One of the symptoms of this disease consists of general weakness and fatigue. Although the muscles associated with the speech production system are affected too, in general the automatic speech recognition assistive technology can still be

very effective, such as it was already stressed in Section 1. The decision concerning to the user population of this application was conditioned by practical restrictions, so the existing LPFAV2 package only provides the materials for developing speaker dependent systems. A continuous speech recognition task was established. The database speech materials were collected with the subject uttering each sentence, corresponding to a mathematic expression that can be executed by the calculator, in a typical read-like way. Even so, a brief analysis of the recorded material reveals the existence of different types of miss-fluency, such as prosodic discontinuities, hesitations and repetitions. The vocabulary of the application is small, with approximately 70 words including the mathematic operators supported by the calculator and all the numbers, from zero to the billion ranges. The recognition is based on the captured speech and face image of the user, so the joint decoding of both information streams can be performed in order to compute the demanded numeric result.

In conclusion, this application fulfills the designed main specifications and presents a quite suitable complexity for the intended AVSR technological approach.

## 3 The Audio-Visual Signal Acquisition

The LPFAV2 database was recorded in the Laboratory for Speech Processing, Electroacoustic, Signals and Instrumentation (LPF-ESI)[1]. A controlled environment was settled, and proper illumination and recording equipments were used.

Three recording sessions, spaced at intervals of a few days, were carried out by December 2003. Some small differences found in the recorded data along all the sessions are not significant. In order to capture the speech signal as clean as possible and also to avoid the natural light variability, all sessions happened at night, during the weekend. The effective duration of the whole captured audio-visual materials is approximately 125 minutes, corresponding to, roughly, 700 sentences. The recordings were accomplished almost continuously along sequences with 25 sentences, then a few minutes break was made before the following subset.

High quality colour video recordings were made using a Canon mini-DV XM-1 3CCD digital video camera recorder. The video files were captured in Digital Video (DV) format, at 25 frames per second, with 720x576 pixels resolution. The sound was synchronously captured with an external microphone, a Shure Beta 58 unit, and was encoded into the PCM format with 16bit 22.05KHz, resulting in a signal-to-noise ratio (SNR) of approximately 25dB. Both data streams were transferred into a computer in real time, trough a Firewire connection. All video files, originally with 3700 Kbps data rate, were MPEG-4 encoded, allowing an average 1/10 compressing ratio without significant quality loss.

In order to get high quality image frames, avoiding shadows and reflexes, a proper illumination was implemented. Three holophotes (Lowel, Totta and Omilight) were used, one of them equipped with reflector. Attention was also paid to the background, which was chosen monochromatic to simplify the image analysis.

The recordings were made in the open central part of a typical laboratory room, with area approximately $67m^2$. A schematic representation of the system used to collect the database is shown on Fig. 1.

[1] `http://lpf-esi.fe.up.pt/`

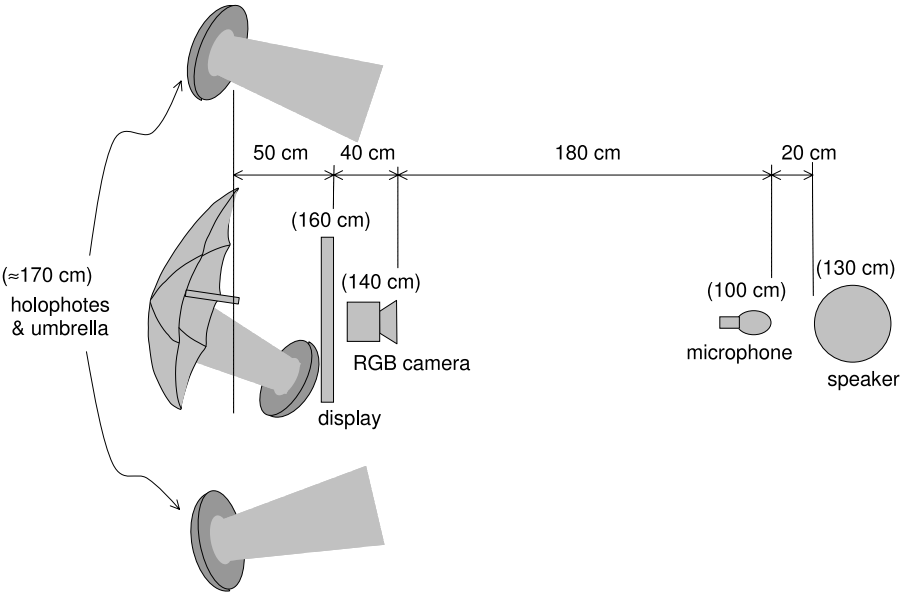


**Fig. 1.** Diagram of the LPFAV2 database collection set-up (inside round brackets is the elevation of each element's centre relatively to the floor)

## 4 The Corpus

The LPFAV2 corpus was developed to support the recognition task outlined in Section 2. Considering the entire training and testing data, the number of recorded sentences is 652, amounting to 12239 word realizations. The structure of each sentence corresponds to the natural reading of a mathematic expression to be processed by a typical scientific calculator.

### 4.1 Vocabulary

The vocabulary of the application contains 68 different words, divided into four subsets: Numerals (N), used to compose numbers from zero to billion ranges; Mathematic Operators (MO), corresponding to the specified calculator operations; Commands (CMM), used to perform special commands; and Connectors (CNN), consisting mainly of articulation words in Portuguese. Table 1 presents the list, in Portuguese, of the whole vocabulary.

Most of the words occur approximately one hundred times in the entire corpus. Given the nature of this application, it was not trivial to achieve this result [6,7]. Just a small group of words are significantly more frequent, occurring a few hundred times each.

### 4.2 Generative Grammar

The nature of the designed application leads to sentences presenting a very rigid syntactic structure, complying with a restrict set of rules. The generative grammar extracted from the

**Table 1.** The LPFAV2 vocabulary

| Group | Words |
|---|---|
| Numerals | /zero/ /um/ /dois/ /três/ /quarto/ /cinco/ /seis/ /sete/ /oito/ /nove/ /dez/ /onze/ /doze/ /treze/ /catorze/ /quinze/ /dezasseis/ /dezassete/ /dezoito/ /dezanove/ /vinte/ /trinta/ /quarenta/ /cinquenta/ /sessenta/ /setenta/ /oitenta/ /noventa/ /cem/ /cento/ /duzentos/ /trezentos/ /quatrocentos/ /quinhentos/ /seiscentos/ /setecentos/ /oitocentos/ /novecentos/ /mil/ /milhão/ /milhões/ |
| Mathematic Operators | /mais/ /menos/ /vezes/ /dividir/ /raíz/ /quadrada/ /cúbica/ /índice/ /quadrado/ /cubo/ /elevado/ /logaritmo/ /base/ /inverso/ |
| Commands | /igual/ /apaga/ /última/ /tudo/ /abrir/ /fechar/ /parênteses/ |
| Connectors | /vírgula/ /e/ /a/ /ao/ /por/ /de/ |

collected spoken sentences is represented in Fig. 2, using a tree-based structure, so that in the terminals any word in the sentence is assigned to one of the four classes (N, MO, CMM and CNN) shown in Table 1.

$$\textit{sentence} \rightarrow \left\{ \begin{array}{l} (\mathrm{OCOM})\ \mathrm{FOMO}\ \mathrm{MO}\ \mathrm{FOMO}\ (\mathrm{OCOM}\ \mathrm{MO}\ \mathrm{FOMO}) \\ \mathrm{OMO}\ (\mathrm{MO}\ (\mathrm{OCOM})\ \mathrm{FOMO}\ (\mathrm{MO}\ \mathrm{FOMO}\ \mathrm{OCOM})) \\ \mathrm{FOMO} \end{array} \right\} \mathrm{OCOM}$$

$$\mathrm{OCOM} \rightarrow \mathrm{CMM} \left\{ \begin{array}{l} \mathrm{CMM} \\ \mathrm{CNN} \end{array} \right\}$$

$$\mathrm{FOMO} \rightarrow \left\{ \begin{array}{l} \mathrm{OMO} \\ \mathrm{F} \end{array} \right\}$$

$$\mathrm{OMO} \rightarrow \left\{ \begin{array}{l} \mathrm{MO} \left\{ \begin{array}{l} \mathrm{MO}\ (\mathrm{N})\ (\mathrm{CNN}) \\ \mathrm{CNN} \end{array} \right\} \\ (\mathrm{CNN})\ \mathrm{MO}\ (\mathrm{CNN}) \end{array} \right\}$$

$$\mathrm{F} \rightarrow \mathrm{ONUM} \left( \left\{ \begin{array}{l} \mathrm{CNN} \left\{ \begin{array}{l} \mathrm{N} \\ \mathrm{MO} \end{array} \right\} \\ \mathrm{MO}\ \mathrm{CNN}\ \mathrm{N} \end{array} \right\} \right)$$

$$\mathrm{ONUM} \rightarrow \mathrm{ON}\ (\mathrm{N})\ (\mathrm{ON})\ (\mathrm{N})\ (\mathrm{ON})\ (\mathrm{CNN})\ (\mathrm{ON})\ (\mathrm{N})\ (\mathrm{ON})$$

$$\mathrm{ON} \rightarrow \mathrm{N}\ (\mathrm{CNN})\ (\mathrm{N})\ (\mathrm{CNN})\ (\mathrm{N})$$

**Fig. 2.** Rule-based grammar of the LPFAV2 corpus (classes between round brackets are optional and classes delimited by the same vertical brace are mutually exclusive)

Deliberately, a small subset of the recorded sentences were designed incorrectly, in the sense that they do not apply with the referred rules. Some of these sentences are not intended to perform any mathematical operations but can occur in a realistic interface. They were designed mainly with the purpose of allowing the study of specific reactions to the surprise they naturally cause in the speaker.

### 4.3 Stochastic Language Modelling

Using the Carnegie Mellon University Statistical Language Modelling toolkit, version 2 [8], several statistic results were computed in order to evaluate the task complexity.

The entire corpus was used to achieve these results. The unigram perplexity was estimated around 15,51 (entropy 3,96). Two bigrams were also generated, considering different discounting techniques: Linear and Witten Bell [9]. The number of different word-pairs found in the corpus was 769. The perplexity estimations for the Linear and Witten Bell methods were 6,44 (entropy 2,69) and 6,90 (entropy 2,79) respectively. Considering that the sentences have a very rigid structure, even simple language models such as these lead to quite low perplexities. The effectiveness of these smoothing techniques could be confirmed comparing these perplexities with other estimations obtained using jacknife techniques [10].

## 5 The Database Package

The database package is so structured to contain audio-visual speech material for acoustic-visual modelling and textual material for language modelling, besides documentation and other auxiliary information.

The entire database is recorded in six CDs, divided into two subsets just differing on the visual contents. The recorded audio signals and the several text files remain exactly the same in both parts. In one of the subsets, with three CDs, the video files (.AVI) contain the original image frames, capturing the whole speaker's face, such as they were collected. In the other three CDs, just the so-called region of interest (ROI), the rectangular area enclosing the lips, is recorded. This operation was performed because the discriminative information outside that rectangle is comparatively much smaller.

Obviously, this was done in order to pave the way for an expeditious usage of the database. This image segmentation was carried out for the entire database, including the materials used for the systems development and the test set. The extraction of the ROI was a relatively easy task, given the nature of this application, with the user maintaining a quite steady position and a frontal image of the face being captured. The implemented approach uses an algorithm based on image symmetry properties [5].

For each part, two CDs hold the training materials and the other contains the data to run the tests. Each CD has four directories, by name: AVI_FILES, WAV_FILES, TXT_FILES, and DOC_FILES.

The AVI_FILES directory contains the .AVI files, such as referred above, each one corresponding to the respective sentence. The name of each file depends on the recording date, the script number and the sentence number; for instance, 13122003_2_10.AVI was recorded December 13th 2003 reading the 10th sentence from the 2nd script. Using the AVI2WAV software, the sound track was extracted from each .AVI file and was saved in a respective file at the WAV_FILES directory.

The TXT_FILES directory contains .TXT files, each one holding the orthographic transcription and segmentation of each word of a recorded sentence. A qualified researcher processed all the sentences, manually segmenting and labelling each word with the help of the Adobe Premiere 6.5 software. The standard procedure to define the segmentation boundaries was based in two steps: first, the ROI from the image signal was inspected to

define the initial boundaries; then, the acoustic signal was inspected to refine the definitive boundaries. Although the segmentation was performed at a relatively high level, a set of reliable criteria was established in order to assure consistency even in the more difficult cases. Such as expected, quite often the boundaries are not well defined due to the cross-word co-articulation phenomenon. It is already projected to carry out the segmentation and labelling of the LPFAV2 at the phonetic level, expanding the usability of this database.

The DOC_FILES directory contains diverse informative files about the LPFAV database: the file READ.ME condenses most of this information; LISTA_CD.TXT lists the name of all video-files in the package; other text files contain the data needed to develop different linguistic models to this recognition task.

## 6 Conclusion

In this paper, a new database created to support the development of multi-modal speech recognition systems was presented. This database establishes a small-vocabulary speaker dependent application, supporting a continuous speech recognition task based in the European Portuguese language. This application has the singularity of being used by a person with muscular dystrophy. Both the speech and the user's face image are captured in order to allow their joint decoding. The available package also includes the respective orthographic transcription and time-alignment files. Raw data and other information needed to develop language models are supplied too.

Considering the characteristics of the LPFAV2 package, which was designed having in mind an application domain that naturally combines two important research topics – the speech recognition robustness and assistive technologies for persons with disabilities – it can be a valuable contribution to the research effort in these areas.

## References

1. Paterson, E. K.: Audio Visual Speech Recognition for Difficult Environments. Ph.D. thesis, Clemson University (2002).
2. Weber, K., Ikbal, S., Bengio, S., and Bourlard, H.: Robust Speech Recognition and Feature Extraction Using HMM2. Computer Speech & Language, **17** (2003) 2–3.
3. Bailly-Baillire, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., Matas, J., Messer, K., Popovici, V., Porée, F., Ruiz, B., and Thiran, J.-P.: The BANCA Database and evaluation protocol. 4th International Conference on Audio- and Video-Based Biometric Person Authentication (2003).
4. Warren, P.: NZSED: building and using a speech database for New Zealand English. New Zealand Journal, 1(6) (2002).
5. Pera, V., Sá, F., Afonso, P., and Ferreira, R.: Audio-Visual Speech Recognition in a Portuguese Language Based Application. Proceedings of the International Conference on Industrial Technology, Maribor, Slovenia (2003).
6. Aiello, D., Cerrato, L., Delogu, C., and Carlo, A. D.: The acquisition of a speech corpus for limited domain translation. Proceedings of the EuroSpeech, Budapest, Hungary (1999).
7. Trancoso, I., Viana, M. C., Duarte, I., and Matos, G.: Corpus de diálogo CORAL. In Actas do PROPOR 1998 - III Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada, Porto Alegre, Brazil (1998).

8. Clarkson, P.R. and Rosenfeld, R.: Statistical Language Modeling Using the CMU-Cambridge Toolkit. Proceedings of the EuroSpeech, Rhodes, Greece (1997).
9. Peng, F., Schuurmans, D.: Combining Naive Bayes and $n$-Gram Language Models for Text Classification. 25$^{th}$ European Conference on IR Research, Pisa, Italy, (2003).
10. Efron, A.: The jacknife, the bootstrap and other resampling plans. Regional Conference Series in Applied Mathematics, Philadelphia, U.S.A. (1982).