LPFAV2: a multi-modal database for developing continuous speech recognisers in assistive technology applications

António Moura[1]*, Vítor Pêra*[2]*, Diamantino Freitas*[2]

[1]**School of Technology and Management, Polytechnic Institute of Bragança**
Quinta de Sta Apolónia, Apartado 134, 5301 – 857 Bragança, Portugal
phone: +351 273 303130, fax: 273 313051, email: moura@ipb.pt, web: www.estig.ipb.pt/

[2]**LSS, DEEC, Faculty of Engineering, University of Porto**
Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
phone: +351 22 508 1808, fax: +351 22 508 1440, email: {vpera, dfreitas}@fe.up.pt, web: www.fe.up.pt/

Abstract. In this poster we report on the motivations, acquisition and contents of a new database intended for developing audio-visual speech recognition systems dedicated to assistive technology applications.

1. INTRODUCTION

The main motivations for the work carried out constructing the speech database here presented can be expressed through the two following research topics: robust speech recognition; and computer technologies for persons with disabilities. In fact, the use of visual features jointly with the acoustic information is becoming increasingly important as a technique to improve the speech recognition robustness [Paterson 02] [Weber 03]. On the other side, the content of many catalogues advertising assistive technology products for persons with disabilities confirm the relevance of speech recognition in this application domain.

One of the most important resources for doing work in speech recognition is a database with the appropriate materials for training and testing the systems under development. In general the size and quality of the database are 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 can be even more difficult due to the supplementary information streams and the huge amount of data. 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 [Pera 03] existed in the particular case of speech recognition applications supporting the Portuguese language. Moreover, the present is the first one that was specially designed for an application where the user has a specific motor impairment, a muscular dystrophy disease. 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. One of the symptoms of that disease consists of general weakness and fatigue. Although the muscles associated with the speech production apparatus are affected too, in general the automatic speech recognition assistive technology can still be very effective.

2. LPFAV2 DATABASE

Application: The LPFAV2 database was designed having in mind a reasonable application domain, either in terms of its support to the intended research in several AVSR technology topics or its practical interest. It must be noticed that the nature of this application, in particular the natural relative immobility of the user, favours the use of a visual information stream complementing the acoustic signal.

The application addressed by this database consists of a speaker dependent voice-controlled interface for a basic scientific calculator. Each sentence, corresponding to a basic mathematic expression that the calculator will process, is spoken in the European Portuguese language in a continuous and natural way.

Corpus: The vocabulary of the application contains 68 different words, divided among four subsets: Numerals, used to compose numbers from zero to the billions range; Mathematic Operators, corresponding to the calculator specified operations; Commands, used to perform special commands; and Connectors, consisting mainly of articulation words in Portuguese. Most of the words occur approximately one hundred times in the entire corpus; just a small group of words are significantly more frequent, occurring a few hundred times each. The LPFAV2 corpus contains more than four hundred sentences with quite variable durations and linguistic-units lengths. However, the designed generative grammar corresponding to the collected spoken sentences confirms that the structure of each sentence is very rigid, complying with a strict set of rules. Stochastic language models were developed too, and the respective perplexities were estimated using the Carnegie Mellon University Statistical Language Modelling toolkit, version 2 [Clarkson 97].

Collection and pre-processing: The LPFAV2 database was recorded in the Laboratory for Speech Processing, Electroacoustics, Signals and Instrumentation[1] (LPF-ESI). A controlled environment was set, in a LPF-ESI room, and proper illumination and recording equipments were used to capture the video signals. Along with the collected multi-modal speech materials, the respective orthographic transcription and time-alignment files are supplied. Concerning the visual stream, some preliminary processing was performed in order to extract the most discriminative information from the users face; according to this, the selected region consists of the lips-frame rectangle. The sentences segmentation and labelling and the image segmentation were carried out for the entire database, including the materials used for the systems development and the test set.

3. CONCLUSION AND FINAL REMARKS

This poster presents a new database created to support the development of multi-modal speech recognition systems. Considering the characteristics of the LPFAV2 database, 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

[Clarkson 97] P.R. Clarkson, R. Rosenfeld, *Statistical Language Modeling Using the CMU-Cambridge Toolkit*, in *Eurospeech '97*, Rhodes, Greece, September 22-25, 1997.

[Paterson 02] E. K. Paterson, *Audio Visual Speech Recognition for Difficult Environments*, PhD Thesis, Clemson University, 2002.

[Pera 03] V. Pera, F. Sá, P. Afonso, R. Ferreira, *Audio-Visual Speech Recognition in a Portuguese Language Based Application*, in *International Conference on Industrial Technology '03*, Maribor, Slovenia, December 10-12, 2003.

[Weber 03] K. Weber, I. Ikbal, S. Bengio, H. Bourlard, *Robust Speech Recognition and Feature Extraction Using HMM2*, Computer Speech & Language, 2003, p17.

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