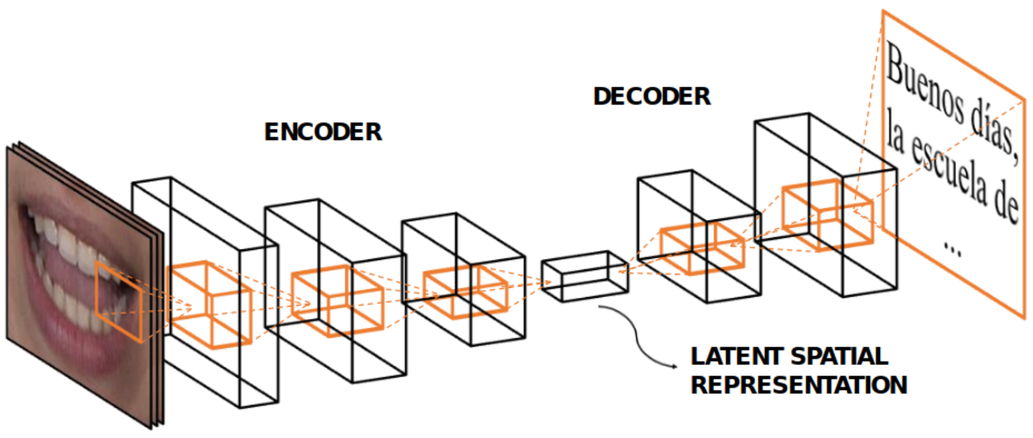
Contributions to Automatic Lipreading for Spanish

PhD Student: **David Gimeno-Gómez** (dagigo1@dsic.upv.es)
Supervisor: **Carlos-D. Martínez Hinarejos** (cmartine@dsic.upv.es)
Doctoral Program in Computer Science

UNIVERSITAT POLITÈCNICA DE VALÈNCIA
Pattern Recognition and Human Language Technology Research Center
PRHLT

Introduction

The importance of **visual information and its relationship with the sounds produced** has been demonstrated [5].



The task is considered as an **open research problem** where different challenges are posed:

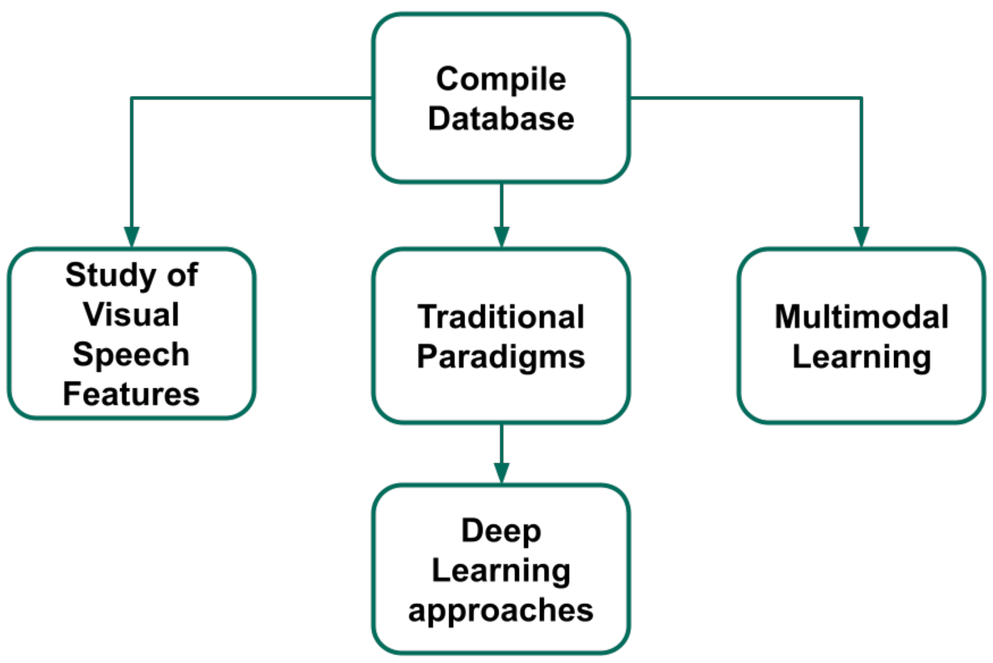
- Difficult Silence Modelling
- Visual Ambiguities
- Co-articulation caused by context influence

Objectives

The main purposes in our thesis are:

- Build an automatic **Visual Speech Recognition System** for the Spanish Language
- Compile an **Audiovisual Database** for Continuous Spanish

Research Development



State of the Art

- **Similar evolution** to that observed in the Acoustic Speech Recognition field [3]:
- **Two main databases** dedicated to English that offer more than 600 hours of data [1]
- Around **70% word recognition accuracy** has been reached [4] using an end-to-end architecture mainly based on Attention Mechanisms

The LIP-RTVE Database

Video Resolution	25 fps	480×270 pixels
Duration	∼13 hours	10,352 overlapped samples
Speakers	**Total:** 323	**Males:** 163 **Females:** 160
Vocabulary	9308 unique words	**Running Words:** 140,123 words

Results

Different preliminary results are reported both for a traditional paradigm and for more recent approaches:

	GMM-HMM	Attention E2E
Speaker-Independent	95.9±0.2	59.3±1.2
Speaker-Dependent	81.4±1.2	32.1±1.2

Applications

- Improve the lives of **people with speech disabilities** who suffer from communication difficulties [2, 6]
- Enhance ASR **performance in adverse scenarios**
- Silent dictation, dubbing and transcribing silent films

However, this field involves certain privacy concerns. Thus, different **ethical aspects must be considered**.

References

[1] Triantafyllos Afouras et al. “Deep audio-visual speech recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018). DOI: 10.1109/TPAMI.2018.2889052.

[2] Bruce Denby et al. “Silent speech interfaces”. In: *Speech Communication* 52.4 (2010), pp. 270–287. DOI: https://doi.org/10.1016/j.specom.2009.08.002.

[3] Adriana Fernandez-Lopez and Federico M Sukno. “Survey on automatic lip-reading in the era of deep learning”. In: *Image and Vision Computing* 78 (2018), pp. 53–72. DOI: https://doi.org/10.1016/j.imavis.2018.07.002.

[4] Pingchuan Ma, Stavros Petridis, and Maja Pantic. “Visual Speech Recognition for Multiple Languages in the Wild”. In: *arXiv preprint arXiv:2202.13084* (2022). URL: https://arxiv.org/abs/2202.13084.

[5] Harry McGurk and John MacDonald. “Hearing lips and seeing voices”. In: *Nature* 264.5588 (1976), pp. 746–748. DOI: 10.1038/264746a0.

[6] Brendan Shillingford et al. “Large-Scale Visual Speech Recognition”. In: *Proc. Interspeech 2019*. 2019, pp. 4135–4139. DOI: 10.21437/Interspeech.2019-1669.