LREC 2022
Marseille

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Pattern Recognition and Human Language Technology Research Center

PRHLT

LIP-RTVE: An Audiovisual Database for Continuous Spanish in the Wild

David Gimeno-Gómez
Carlos-D. Martínez Hinarejos

Overview

1. **Introduction**
2. **Related Work**
3. **The LIP-RTVE Database**
4. **Baseline Performance**
5. **Conclusions & Future Work**

Introduction

Introduction

- Multi-sensory nature of speech
- Increasing interest in Visual Speech Recognition (VSR)
- Lack of in-the-wild Spanish Audiovisual resources

Introduction

- Multi-sensory nature of speech
- Increasing interest in Visual Speech Recognition (VSR)
- Lack of in-the-wild Spanish Audiovisual resources

Introduction

- Multi-sensory nature of speech
- Increasing interest in Visual Speech Recognition (VSR)
- Lack of in-the-wild Spanish Audiovisual resources

Related Work

Related Work

Table 1: Details regarding other audiovisual databases in the literature.

Database	Language	Duration	Nature
LRS2-BBC	English	224 hours	In the Wild
LRS3-TED	English	475 hours	In the Wild
VLRF	Español	3 hours	Recording Studio
CMU-MOSEAS	Español	18 hours	In the Wild

The LIP-RTVE Database

The LIP-RTVE Database

Source data

- Collected from a subset of the RTVE database
  - TV broadcast programmes
  - Wide range of speakers
  - Spontaneous speech phenomena

The LIP-RTVE Database

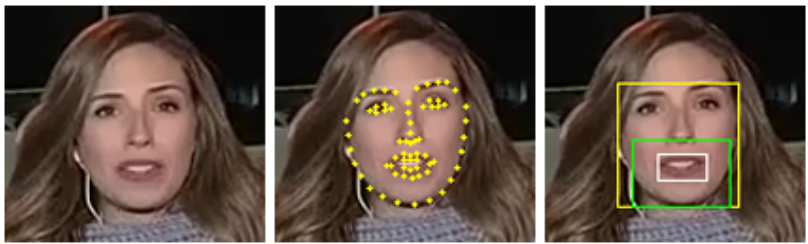
Methodology

1. **Automatically trim scenes** where at least one face appears on scene
2. **Supervise scenes** according to different criteria
3. **Split long scenes** in smaller ones
4. **Manually transcribe** each scene
5. Automatic extraction of **Regions of Interest** (ROIs)

The LIP-RTVE Database

Region of Interest Extraction

Figure 1: The Region of Interest extraction process.
White box: *fitMouth*. Green box: *wideMouth*. Yellow box: *faceROI*.



The LIP-RTVE Database

Statistics

Table 2: Overall details regarding the compiled LIP-RTVE Audiovisual 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

The LIP-RTVE Database

Challenges

Spontaneous phenomena

- Background noise
- Mistakes, hesitations
- Head movements
- Different lighting conditions

Inherent to VSR

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

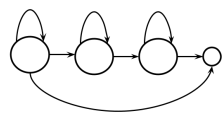
Baseline Performance

Baseline Performance

Automatic Speech Recognition System

- Traditional **GMM-HMM** system using the Kaldi toolkit
- Non-standard topology for VSR

Figure 2: The HMM's topology employed in VSR.



- External **4-gram Language Model** using the SRILM toolkit

Baseline Performance

Feature Extraction

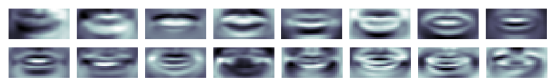
Acoustic Speech

- 39-dimensional MFCC$+\Delta+\Delta\Delta$
- Extracted at 100 fps

Visual Speech

- 16-component Eigenlips (PCA)
- Extracted at 25 fps

Figure 3: The computed eigenlips for VSR.



Baseline Performance

Data sets partition

Table 3: Details on the different scenarios and partitions and their statistics with respect to the external language model.

Dataset		No. Speakers	Utterances	Running Words	Vocabulary	Language Model	
						Perplexity	OOV words
SI	TRAIN	29	7142	99449	7524	98.9	755
	DEV	151	1638	20541	2932	107.1	191
	TEST	143	1572	20133	2983	104.2	193
SD	TRAIN	323	7355	96174	8244	100.5	782
	DEV	219	1597	22670	4316	98.5	192
	TEST	123	1400	21259	4133	105.4	165

Baseline Performance

Data sets partition

Table 3: Details on the different scenarios and partitions and their statistics with respect to the external language model.

Dataset		No. Speakers	Utterances	Running Words	Vocabulary	Language Model	
						Perplexity	OOV words
SI	TRAIN	29	7142	99449	7524	98.9	755
	DEV	151	1638	20541	2932	107.1	191
	TEST	143	1572	20133	2983	104.2	193
SD	TRAIN	323	7355	96174	8244	100.5	782
	DEV	219	1597	22670	4316	98.5	192
	TEST	123	1400	21259	4133	105.4	165

Baseline Performance

Data sets partition

Table 3: Details on the different scenarios and partitions and their statistics with respect to the external language model.

Dataset		No. Speakers	Utterances	Running Words	Vocabulary	Language Model	
						Perplexity	OOV words
SI	TRAIN	29	7142	99449	7524	98.9	755
	DEV	151	1638	20541	2932	107.1	191
	TEST	143	1572	20133	2983	104.2	193
SD	TRAIN	323	7355	96174	8244	100.5	782
	DEV	219	1597	22670	4316	98.5	192
	TEST	123	1400	21259	4133	105.4	165

Baseline Performance

Results & Discussion

Table 4: Baseline results (%WER) for each modality and scenario.

Dataset		Modality	
		Audio-only	**Video-only**
SI	**DEV**	16.9±0.8	95.9±0.3
	TEST	15.3±0.8	95.9±0.2
SD	**DEV**	9.5±0.6	82.9±1.1
	TEST	8.0±0.5	81.4±1.2

State of the art in the LRS3-TED ∼30%WER

Conclusions & Future Work

Conclusions

- A **new audiovisual database** for continuous Spanish has been compiled
- Our contribution:
  - covers the **lack of in-the-wild Spanish resources**
  - encourages advances in **Spanish VSR**
- A **suitable benchmark** for different scenarios has been defined
- **Baseline performances** have been obtained with traditional GMM-HMMs

Conclusions

- A **new audiovisual database** for continuous Spanish has been compiled
- Our contribution:
  - covers the **lack of in-the-wild Spanish resources**
  - encourages advances in **Spanish VSR**
- A **suitable benchmark** for different scenarios has been defined
- **Baseline performances** have been obtained with traditional GMM-HMMs

Conclusions

- A **new audiovisual database** for continuous Spanish has been compiled
- Our contribution:
  - covers the **lack of in-the-wild Spanish resources**
  - encourages advances in **Spanish VSR**
- A **suitable benchmark** for different scenarios has been defined
- **Baseline performances** have been obtained with traditional GMM-HMMs

Conclusions

- A **new audiovisual database** for continuous Spanish has been compiled
- Our contribution:
  - covers the **lack of in-the-wild Spanish resources**
  - encourages advances in **Spanish VSR**
- A **suitable benchmark** for different scenarios has been defined
- **Baseline performances** have been obtained with traditional GMM-HMMs

Conclusions

- A **new audiovisual database** for continuous Spanish has been compiled
- Our contribution:
  - covers the **lack of in-the-wild Spanish resources**
  - encourages advances in **Spanish VSR**
- A **suitable benchmark** for different scenarios has been defined
- **Baseline performances** have been obtained with traditional GMM-HMMs

Conclusions

- A **new audiovisual database** for continuous Spanish has been compiled
- Our contribution:
  - covers the **lack of in-the-wild Spanish resources**
  - encourages advances in **Spanish VSR**
- A **suitable benchmark** for different scenarios has been defined
- **Baseline performances** have been obtained with traditional GMM-HMMs

Future Work

- Increase the size of the LIP-RTVE database
- Experiment with end-to-end approaches

LREC 2022
Marseille

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Pattern Recognition and Human Language Technology Research Center

PRHLT

LIP-RTVE: An Audiovisual Database for Continuous Spanish in the Wild

David Gimeno-Gómez
Carlos-D. Martínez Hinarejos