Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2750–2758
Marseille, 20-25 June 2022

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

David Gimeno-Gómez, Carlos-D. Martínez-Hinarejos
Pattern Recognition and Human Language Technologies Research Center,
Universitat Politècnica de València, Camino de Vera, s/n, 46022, València, Spain
{dagigo1, cmartine}@dsic.upv.es

Abstract

Speech is considered as a multi-modal process where hearing and vision are two fundamentals pillars. In fact, several studies have demonstrated that the robustness of Automatic Speech Recognition systems can be improved when audio and visual cues are combined to represent the nature of speech. In addition, Visual Speech Recognition, an open research problem whose purpose is to interpret speech by reading the lips of the speaker, has been a focus of interest in the last decades. Nevertheless, in order to estimate these systems in the currently Deep Learning era, large-scale databases are required. On the other hand, while most of these databases are dedicated to English, other languages lack sufficient resources. Thus, this paper presents a semi-automatically annotated audiovisual database to deal with unconstrained natural Spanish, providing 13 hours of data extracted from Spanish television. Furthermore, baseline results for both speaker-dependent and speaker-independent scenarios are reported using Hidden Markov Models, a traditional paradigm that has been widely used in the field of Speech Technologies.

Keywords: Audiovisual Database, Speech Recognition, Lipreading, Computer Vision

1. Introduction

Despite the fact that speech perception is commonly considered as a purely auditory process, the truth is that it is a process involving multiple senses, as well as high-level knowledge related with grammar and semantics (Dupont and Luettin, 2000). In fact, the study carried out by McGurk and MacDonald (1976) demonstrated the importance of visual information and its relationship with the sounds produced. Nevertheless, for deaf or hearing-impaired people, who are totally or partially dependent on their sense of sight, speech understanding poses a great challenge since, as Duchnowski et al. (2000) support, only 30% of speech information is visible. For this reason, different areas of research have focused their efforts on speech recognition when the auditory sense is not functional, such as lipreading (Kaplan et al., 1987; Rodríguez Ortiz, 2008), cued speech (Cornett, 1967), or silent speech interfaces (Denby et al., 2010).

Regarding the field of Speech Technologies in its origins, Automatic Speech Recognition (ASR) systems were focused only on processing the acoustic signal. Nowadays, this type of systems reaches high-quality performances (Chan et al., 2016). However, these approaches suffer a deterioration in quality when the audio signal is damaged or corrupted (Juang, 1991). As a consequence, in order to deal with this issue, the research was impulsed towards Audio-Visual Speech Recognition (AVSR) approaches (Potamianos et al., 2003; Dupont and Luettin, 2000). In this way, as the above mentioned studies demonstrate, it was shown that the combination of acoustic and visual cues could represent the nature of speech more robustly. On the other hand, as Fernandez-Lopez and Sukno (2018) present, in the last decades there has been an increasing interest in the Visual Speech Recognition (VSR) task, an open research problem where the automatic system aims to interpret speech by reading the lips of the speaker. In fact, the current state-of-the-art in this challenging task has been achieved thanks to the incorporation of acoustic cues during the training phase (Ma et al., 2022; Afouras et al., 2018).

Notwithstanding, it is well-known that data is a fundamental pillar for this area of research. Thus, two relevant aspects must be taken in consideration. First, in the currently Deep Learning (DL) era, large-scale audiovisual databases are required in order to estimate the immense amount of parameters that form this type of systems. On the other hand, there is a language unbalanced proportion, since while most of these databases are dedicated to English, other languages lack of sufficient resources (Fernandez-Lopez and Sukno, 2018; Zadeh et al., 2020).

Contributions: due to the above presented reasons, this paper presents an audiovisual database to deal with unconstrained natural Spanish, following the so-called *in the wild* philosophy. More precisely, around 13 hours of data extracted from Spanish broadcast television have been semi-automatically collected. In this way, we intended to ensure the quality of the compiled data. On the other hand, baseline results for both speaker-dependent and speaker-independent scenarios are reported using Hidden Markov Models (HMMs), a traditional paradigm that has been widely used in the field of Speech Technologies (Gales and Young, 2008). Furthermore, different input modalities, i.e., acoustic, visual or audiovisual features, have been studied. In fact, an audio-only approach was considered as the lower bound for our proposed task.

2. Related Work

As Fernandez-Lopez and Sukno (2018) suggest, advances achieved in the field of audiovisual Speech

LIP-RTVE: an Audiovisual Database for Continuous Spanish in the Wild			
Audio Resolution	16 kHz mono-channel	16 bit-depth	WAV format
Video Resolution	25 frames/second	RGB images	PNG format
Duration	∼13 hours	10,352 overlapped samples	1,168,087 frames
Speakers	**Total:** 323	**Males:** 163	**Females:** 160
ROIs Average Size	**FitMouth:** 27×16 pixels	**WideMouth:** 45×30 pixels	**FaceROI:** 55×58 pixels
Vocabulary	9308 unique words	**Running Words:** 140,123 words	
Phonemes	24 unique phonemes	**Running Phonemes:** 654,368 phonemes	
Characters	28 unique characters	**Running Characters:** 801,830 characters	
Speech Rate (words/second)	**Min:** 0.58	**Median:** 2.94	**Max:** 9.73
Words per Utterance	**Min:** 1	**Median:** 12	**Max:** 62
Phonemes per Utterance	**Min:** 4	**Median:** 55	**Max:** 270
Characters per Utterance	**Min:** 4	**Median:** 68	**Max:** 343
Seconds per Utterance	**Min:** 0.97	**Median:** 4.00	**Max:** 15.97

Table 1: Overall details regarding the compiled LIP-RTVE Audiovisual Database.

Technologies have been conditioned, among other reasons, by the available audiovisual databases at the time. In its origins, these databases began by collecting data in order to deal with simple tasks like alphabet or digit recognition, such as AVLetters (Matthews et al., 2002) and CUAVE (Patterson et al., 2002) corpora, respectively. Nonetheless, in the last decade numerous large-scale publicly available audiovisual databases which address natural speech recognition have been compiled (Fernandez-Lopez and Sukno, 2018). Thus, due to the nature of our contribution, this section is focused on this type of databases, with special emphasis on those that follow the so-called *in the wild* philosophy.

With respect to databases that have been recorded in a controlled setting, the 4-hours RM-3000 (Howell et al., 2016) database is focused on natural speech, but the main inconvenient is that this corpus only has one speaker. Another database we have to mention, despite having also been recorded in controlled conditions, is the TCD-TIMIT (Harte and Gillen, 2015) corpus, which offers around 7 hours of data collected from 62 different speakers. On the other hand, in relation with realistic scenarios, we must mention the LRS2-BBC (Chung et al., 2017; Afouras et al., 2018), MV-LRS (Chung and Zisserman, 2017), and LRS3-TED (Afouras et al., 2018) corpora, all of them automatically collected from mass media. In this way, each one of these databases offers hundreds of recorded hours, providing an adequate support for training architectures based on DL techniques. However, a noteworthy fact is that all these resources are dedicated to English.

Regarding Spanish, our language of interest, there has been an increase in available resources in the last years. Nevertheless, it is not comparable with the vast amount of data mentioned above. First, we must mention the VLRF (Fernandez-Lopez et al., 2017) corpus, where 25 speakers provide around 3 hours of natural sentences but recorded in controlled conditions. Furthermore, speakers were asked to strive to vocalize in an appropriate and expressive way. On the other hand, the recent multilingual CMU-MOSEAS (Zadeh et al., 2020) database covers 4 low-resources languages, including Spanish. Concretely, although a large amount of data was compiled, only about 18 hours of samples were annotated for each one of these languages. Aditionally, this is an interesting corpus, as it provides a multi-modal point of view, supplying information related with the emotions and subjectivity expressed by the speaker. Finally, Córdova-Esparza et al. (2019) defined, inspired by the large-scale English-based corpora previously mentioned, a process to automatically collect audiovisual data from Youtube videos. In this way, a database with around 100,000 annotated samples and the employed automatic collector software were made publicly available.

3. LIP-RTVE Database

The compiled audiovisual database is composed of around 13 hours of semi-automatically collected and annotated data, whose main overall details and statistics are depicted in Table 1.

Nonetheless, as it is reflected along this section, it is necessary to mention that the LIP-RTVE database was conceived at the first instance as a corpus focused on the Automatic Lipreading or VSR task. Thus, our purpose was to increase the available language resources to support the research regarding VSR for unconstrained and natural Spanish in the wild. However, in our experiments (see Section 5), different input modalities were studied, among which the audio-only approach was considered as the lower bound for our proposed task.

3.1. Source Data

In order to provide an appropriate support to estimate robust automatic systems against realistic scenarios, we decided to extract our corpus from TV broadcast programmes. Thus, we compiled it from a subset of the RTVE database (Lleida et al., 2018) which has been employed in the Albayzín evaluations (Lleida et al., 2019). Concretely, despite the fact that this database is made up of different programmes broadcast by *Radio Televisión Española*, we compiled our corpus only from the news programme 20H.

Thereby, the corpus belongs to the so-called *in the wild* philosophy, offering a large number of speakers in a wide range of scenarios, either inside a record studio or in outdoor locations, where the speaker does not always maintain a frontal plane but can sometimes adopt tilted postures. Furthermore, it includes variations on intra-personal aspects, light conditions, or in distance from the speaker to the camera. It is remarkable that not all the compiled speakers are well-trained television professionals, but a considerable number of them are interviewees who speak naturally, making mistakes or hesitating. In fact, these and other types of spontaneous speech phenomena, as it is described in Section 3.6, were identified throughout the entire database.

On the other hand, regarding the details of the recording setting, this subset contains MP4-format files with a 48 kHz two-channel stereo audio resolution and videos recorded with a resolution of 480×270 pixels at 25 fps.

One aspect that must be noted is that the RTVE database is protected by a Non-Disclouse Agreement (NDA)[1], which implies the signing of this license to be able to access our source data. In any case, this license allows to freely use the audiovisual material for research purposes.

3.2. Methodology

The MP4-format files provided by the source data had to be pre-processed since, on numerous occasions, voice-over was used or more than one speaker appeared on the scene, aspects that were not suitable to deal with the VSR task. Therefore, we defined a methodology to obtain samples that were appropriate for both ASR and VSR at the same time.

For this reason, in order to ease the collecting process, our first step was to implement an automatic software to obtain extracts from the MP4 files where at least one face appeared on the scene. This stage was made possible thanks to the use of the face detection tools described in Section 3.3. Then, once these extracts were obtained, we selected those where a unique speaker was talking for a maximum of 15 seconds. Nevertheless, those scenes where other people appeared in the background did not pose a problem and were accepted as new samples of the database, since the Region of Interest (ROI) extraction process was implemented to capture the face that occupies the largest area in the scene. This process was manually supervised, since in certain situations the largest face did not always correspond to the person speaking.

Subsequently, one aspect which must be mentioned is that we split each long sample into smaller ones, as long as the speaker made pauses in his or her speech that allowed us to make an adequate division of the message. In this way, at expense of building a corpus with overlap, we were able to increase the amount of available data.

Finally, each sample of the database was manually annotated, obtaining its corresponding transcription. The pre-processing details regarding these transcriptions are described in Section 3.5.

3.3. Region of Interest Extraction

When facing VSR it is necessary to apply Computer Vision techniques in order to extract our ROIs. In this case, we are talking about the face of the speaker, a region where it is contained the information related with face expressions that would allow us to address the lipreading task. Thus, by using open-source resources[2,3] (Deng et al., 2020; Bulat and Tzimiropoulos, 2017), an automatic software to identify 68 facial landmarks (Sagonas et al., 2016) was implemented and released together with the database.

Once these landmarks were found, by selecting some of them, we were able to define the three types of ROIs depicted in Figure 1. Henceforth, these ROIs, from the smallest to the largest size, are referenced as *fitMouth*, *wideMouth*, and *faceROI*, whose average sizes are reflected in Table 1. As we have previously suggested in Section 3.2, in each frame the face occupying the largest area on the scene was selected in order to avoid confusion with people in the background.

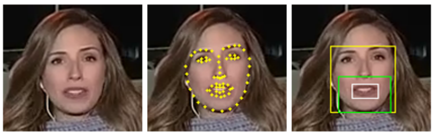

Figure 1: The Region of Interest extraction process. White box: *fitMouth*. Green box: *wideMouth*. Yellow box: *faceROI*.

The reason why we decided to define ROIs with different sizes was due to the differences that exist between the approaches based on end-to-end DL structures (Chan et al., 2016; Chung et al., 2017; Afouras et al., 2018; Ma et al., 2022) and those based on the traditional paradigm of HMMs (Gales and Young, 2008; Thangthai et al., 2015). The latter is made up of several modules where each of them is independent from the other. This fact implies that visual speech features will be static during the training phase of the speech module and, therefore, a smaller and specific ROI should be more convenient. Conversely, by employing end-to-end approaches, all their parameters, including those in charge of extracting the visual speech features, are estimated according to the mistakes found during the

[1] http://catedrartve.unizar.es/rtvedatabase.html

[2] https://github.com/hhj1897/face_alignment

[3] https://github.com/hhj1897/face_detection

Dataset		Duration	Speakers			Utterances	Running Words	Vocabulary	Language Model	
			Males	Female	Total				Perplexity	OOV words
SI	**TRAIN**	∼9 hours	10	19	29	7142	99449	7524	98.9	755
	DEV	∼2 hours	86	65	151	1638	20541	2932	107.1	191
	TEST	∼2 hours	67	76	143	1572	20133	2983	104.2	193
SD	**TRAIN**	∼9 hours	163	160	323	7355	96174	8244	100.5	782
	DEV	∼2 hours	100	119	219	1597	22670	4316	98.5	192
	TEST	∼2 hours	55	68	123	1400	21259	4133	105.4	165

Table 2: Details regarding the training (TRAIN), development (DEV), and testing (TEST) data sets defined in LIP-RTVE for both speaker-independent (SI) and speaker-dependent (SD) scenarios. For each data set, the perplexity and the number of Out Of Vocabulary (OOV) words were computed based on the language model described in Section 4.5.

message decoding phase. In this way, end-to-end approaches are able to identify or select relevant features in a wider ROI which might additionally provide more useful information (Zhang et al., 2020).

3.4. Audio Files

The acoustic signal from each sample was transformed into a 16 kHz mono-channel with 16 bit-precision WAV file, as detailed in Table 1. This process was made possible by using the open-source library FFmpeg (Tomar, 2006). In the same way as with the ROIs, a software was released to process the acoustic signals.

3.5. Transcriptions

As a pre-processing, after lowercasing all the text, all punctuation marks as well as accents were removed. Finally, transcriptions were coded using the UTF-8 standard.

3.6. Identified Challenges

The database presents all those challenges that could be expected in a realistic scenario, as it is reflected in Table 1. Some details we must comment are that, in certain samples, the speaker may speak too quickly, as it is often the case on news programmes. Another feature is that there are samples with considerable differences in length, from samples where we can find numerous words to samples where the speaker only pronounces a word. Additionally, there is a remarkable unbalance between the participation of the each one of speakers in terms of seconds.

Regarding acoustic signals, there are many occasions where we found background noises that could complicate the understanding of the message. Nevertheless, as we have mentioned at the beginning of this section, the LIP-RTVE database was primarily designed to address the VSR task. For this reason, we must highlight the following identified lipreading-related challenges:

- Complex silence modelling (Thangthai, 2018). We could consider that the speaker is silent when his or her mouth is closed or when there is no lip movements. The former is not always true and, additionally, there are certain phonemes, such as the sound /p/, that are produced by bringing the lips together. Regarding the lack of lip movements, there are sounds that are mainly produced from the throat with an imperceptible participation of the tongue or lips.

- Visual ambiguities, since several phonemes can be associated with one viseme, i.e. the basic speech unit in the video domain (Fisher, 1968). In other words, there is no one-to-one correspondence between both entities. The clearest example would be the ambiguity that exists when visually discerning between the phonemes /p/, /b/, and /m/.

- Co-articulation caused by context influence. As Fernandez-Lopez and Sukno (2017) suggest in their study, there are phonemes whose visual correspondence can suffer noticeable changes depending on their surrounding context.

- Finally, certain aspects, such as wetting the lips, poor vocalization, errors and rectifications, or even lowering the head to read notes, could hinder the correct learning of the system in some way.

Thus, we must be aware of the challenges that the lack of the auditory sense implies in the automatic speech recognition field.

3.7. Public Release

Unfortunately, several details related with the NDA license of the RTVE database must be considered before sharing our contribution with the rest of the research community. For this reason, the entire LIP-RTVE database has not yet been publicly released. Nevertheless, as the data is processed, all the details and resources needed to obtain our database in a license-respecting way will be available on the authors' Github Repository[4] as soon as possible.

[4] `https://github.com/david-gimeno/LIP-RTVE`

4. Experimental Setup

4.1. Data Sets

The LIP-RTVE database offers two partitions both for a speaker-dependent (SD) and speaker-independent (SI) scenario. Each partition, in order to define an experimental benchmark, was split in specific training (TRAIN), development (DEV) and testing (TEST) sets. Nonetheless, due to the nature of the source data, there is a significant unbalance regarding the participation of each speaker in the compiled corpus. For this reason, with the intention of providing the best possible learning to the automatic system, the speakers with the longest appearances were allocated first to the training set until reaching the 70% of the total data. Then, the remaining samples were randomly assigned to the DEV or TEST set, gathering, for each of them, around 15% of data. The main details regarding these data sets for each scenario are depicted in Table 2. We must mention that in the SD scenario different aspects were taken into account in order to provide a non-overlapping partition.

4.2. Acoustic Features

The standard representation in the field of ASR was applied on acoustic signals. More specifically, the 39-dimensional Mel Frequency Cepstral Coefficients (MFCC) and their corresponding first- and second-order dynamic differential parameters ($\Delta+\Delta\Delta$) (Gales and Young, 2008) were extracted at 100 fps.

4.3. Visual Features

Unlike acoustic ASR, there is no consensus on which is the best option to represent the nature of visual speech (Fernandez-Lopez and Sukno, 2018). Thus, as this concept has been widely studied in the field of VSR (Thangthai et al., 2017; Lan et al., 2009), we decided to extract the appearance-based features known as eigenlips. By selecting 25 random frames from each training sample, we computed the Principal Component Analysis (PCA) technique (Wold et al., 1987), reducing each frame into 16 components. These eigenlips are shown in Figure 2, where we can observe how each component focuses on different aspects, such as lip contours or zones where we can find teeth and tongue.

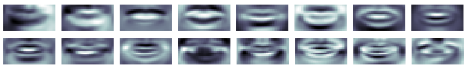

Figure 2: The eigelips obtained in the speaker-independent scenario.

On the other hand, as it was explained in Section 3.3, due to the nature of traditional ASR paradigms, as it is the case of our experiments (see Section 4.4), we considered to use the *fitMouth* ROIs as the better option to extract the visual speech features. More specifically, all these ROIs were normalized to a resolution of 32×16 pixels, converted to gray-scale images and, in addition, a histogram equalization was applied to them.

4.4. Automatic Speech Recognition System

The ASR system employed in our research was designed in the Kaldi toolkit (Povey et al., 2011), where several workflows or recipes to build different paradigms in the field of Speech Technologies are provided. Concretely, we defined a traditional HMM-based system in combination with Gaussian Mixture Models (GMMs) (Gales and Young, 2008), taking as a reference the Wall Street Journal (WSJ) recipe[5]. In order to facilitate the understanding of the results reported in Section 5, we must briefly describe the different stages that compound the estimation process of a GMM-HMM system. In this way, we distinguish the following phases:

- **MONO**: a context-independent GMM-HMM is estimated from scratch applying, over the raw features, the Cepstral Mean and Variance Normalization (CMVN) technique and $\Delta+\Delta\Delta$ coefficients (Gales and Young, 2008).
- **DELTAS**: in this phase, a context-dependent GMM-HMM is trained, employing a decision tree-based triphone state clustering (Young et al., 1994). The input features remain identical to the previous step.
- **LDA+MLLT**: in this stage, the Linear Discriminant Analysis (LDA) (Rao, 1965) and Maximum Likelihood Linear Transform (MLLT) (Gopinath, 1998) techniques are applied to compute the known as HiLDA features (Potamianos et al., 2001), whose purpose is to reduce the feature dimensionality and capture contextual information. Thus, the GMM-HMM is re-estimated.
- **SAT**: the last GMM-HMM is obtained by applying a Speaker Adaptive Training (SAT) (Anastasakos et al., 1997) based on the feature space Maximum Likelihood Linear Regression (fM-LLR) method (Gales, 1998).

Finally, the decoding phase is based on a Weighted Finite-State Transducer (Mohri et al., 2008) which integrates the morphological model, phonetic context-dependencies, the lexicon, and the language model.

4.5. Lexicon and Language Models

In order to estimate both models, around 80k sentences were collected from other news programmes broadcast by RTVE during the same period of time.

In this way, the lexicon model was built, integrating a vocabulary of 45247 unique words. The foundations of this model are based on a vocabulary of 24 phonemes defined according to Spanish phonetic rules (Quilis, 1997) in addition to the default *silence* phones of Kaldi. On the other hand, a 4-gram word-based language model was estimated using the SRLIM toolkit (Stolcke,

[5] `https://github.com/kaldi-asr/kaldi/tree/master/egs/wsj/s5`

2002). Nonetheless, as it is detailed in Table 2, the high perplexity and the considerable number of Out Of Vocabulary (OOV) words offered by this language model over both DEV and TEST data sets must be taken into consideration along our experiments.

4.6. Tool Setup

As we have commented in Section 4.4, the configuration of our recognition system is mainly based on the WSJ recipe. For the training phase, default parameters were kept. For decoding, we set a value of 13.0 to pruning beam and 6.0 to lattice beam. The language model covers scale factors between 1 and 20, while the speech model scale factor has a value of 0.08333. On the other hand, based on the BABEL recipe[6], word insertion penalty values between -5.0 and 5.0 were studied. All these decoding parameters were evaluated in each experimental trial but only the lowest word error rate was considered.

4.7. Evaluation

All the results presented along our experiments are evaluated by the well-known Word Error Rate (WER) with 95% confidence intervals obtained by the bootstrap method as described in (Bisani and Ney, 2004).

5. LIP-RTVE Baseline Performance

A context-dependent GMM-HMM system, whose details are described in Section 4.4, was employed in our baseline experiments. Different modalities, as Tables 3 and 4 reflect, were studied over the data sets defined in Section 2. Concretely, we report the recognition performance obtained over the DEV and TEST sets along the training phases for both a SD and SI scenario.

Regarding our audio-only experiments, henceforth considered as the lower error bound for our task, we employed the 39-dimensional MFCCs (described in Section 4.2) and the standard three-state HMM's topology. From the results reported in Table 3, the first aspect we must highlight is how, in both scenarios, the system recognition performance improves as we progress through the training stages; especially, since we build the first context-dependent system (DELTAS). Another aspect we must consider is that, as it might be expected, results for the SD scenario provide better recognition rates than those for the SI scenario.

Dataset		**Training phases**			
		MONO	**DELTAS**	**LDA+MLLT**	**SAT**
SI	**DEV**	40.7±1.1	20.9±0.9	18.8±0.9	16.9±0.8
	TEST	40.4±1.2	20.0±0.9	16.7±0.9	15.3±0.8
SD	**DEV**	38.9±1.0	14.2±0.7	11.5±0.6	9.5±0.6
	TEST	37.5±1.1	12.2±0.6	10.1±0.5	8.0±0.5

Table 3: Audio-only baseline results (WER) for each training phase in both a speaker-independent (SI) and speaker-dependent (SD) scenario.

With respect to video-only experiments, the eigenlips described in Section 4.3 were employed. Nevertheless, we first must consider that visual data presents a lower sample rate than audio data. Therefore, our first experiments were focused on defining the optimal HMM's topology, either adding transitions and/or reducing the number of states. Thus, the topology depicted in Figure 3 was found as the better approach to fit the temporary nature of our visual data.

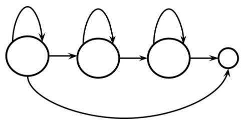

Figure 3: The HMM's topology employed in video-only experiments.

When dealing with the VSR task, as it was mentioned in Section 3.6, their inherent challenges must be taken into consideration. Thus, as we can see in Table 4, the SD scenario reflects, although the recognition rates are not comparable, an evolution similar to that observed in the audio-only experiments, where as we progress through the training phases, the error rate decreases. However, SI experiments did not reach acceptable results. This behaviour is in accordance with the work carried out by Cox et al. (2008), where the authors studied the different challenges posed by the SI setting. In any case, our results demonstrate that further research is necessary in the VSR task, either improving the quality of the extracted visual features or employing more powerful automatic systems.

Dataset		**Training phases**			
		MONO	**DELTAS**	**LDA+MLLT**	**SAT**
SI	**DEV**	96.5±0.3	95.9±0.2	96.0±0.3	95.9±0.3
	TEST	96.5±0.4	96.2±0.2	96.3±0.3	95.9±0.2
SD	**DEV**	96.0±0.3	90.4±0.7	88.0±0.8	82.9±1.1
	TEST	95.6±0.2	90.1±0.7	87.5±0.8	81.4±1.2

Table 4: Video-only baseline results (WER) for each training phase in both a speaker-independent (SI) and speaker-dependent (SD) scenario.

Finally, regarding the audiovisual approach, different feature fusion methods (Potamianos et al., 2003) were

[6] `https://github.com/kaldi-asr/kaldi/blob/master/egs/babel/s5/local/score_combine.sh`

explored. Nonetheless, these experiments did not improve the quality of the obtained audio-only results.

6. Conclusions

This paper has described a new audiovisual database to deal with unconstrained natural Spanish, following the so-called *in the wild* philosophy. More precisely, the compiled LIP-RTVE database offers around 13 hours of data semi-automatically collected from Spanish broadcast television. Thus, our contribution attempts to cover the relative lack of *in the wild* Spanish resources (Zadeh et al., 2020) and the increased interest in VSR in recent decades (Fernandez-Lopez and Sukno, 2018), since the LIP-RTVE database was primarily conceived to address this challenging task. On the other hand, in order to establish an experimental benchmark, both a SI and SD partition were defined. Baseline performances were obtained by employing the traditional GMM-HMM paradigm (Gales and Young, 2008).

Regarding future work, we must first address the public distribution of our corpus, which will be released as soon as possible in a format that respects the NDA license of its source data. Once this issue is solved, we consider organizing competitions in order to encourage research on our database, especially in the VSR task which remains an open problem (Fernandez-Lopez and Sukno, 2018). On the other hand, our future research is focused on exploring more robust and accurate systems where Deep Learning techniques would be incorporated. Concretely, in addition to study the combination of HMMs with Deep Neural Networks (Hinton et al., 2012), we consider experimenting with end-to-end architectures (Chan et al., 2016; Chung et al., 2017; Afouras et al., 2018; Ma et al., 2022), whose possible benefits were indicated in Section 3.3. Finally, we plan to increase the size of the LIP-RTVE database.

Acknowledgements

This work was partially supported by Generalitat Valenciana under project DeepPattern (PROMETEO/2019/121) and by Ministerio de Ciencia under project MIRANDA-DocTIUM (RTI2018-095645-B-C22).

Bibliographical References

Afouras, T., Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2018). Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi: 10.1109/TPAMI.2018.2889052.

Anastasakos, T., McDonough, J., and Makhoul, J. (1997). Speaker adaptive training: A maximum likelihood approach to speaker normalization. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1043–1046. IEEE.

Bisani, M. and Ney, H. (2004). Bootstrap estimates for confidence intervals in asr performance evaluation. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 409–412. IEEE.

Bulat, A. and Tzimiropoulos, G. (2017). How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030.

Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964.

Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2017). Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453. IEEE.

Córdova-Esparza, D.-M., Terven, J., Romero, A., and Herrera-Navarro, A. M. (2019). Audio-visual database for spanish-based speech recognition systems. In Lourdes Martínez-Villaseñor, et al., editors, *Advances in Soft Computing*, pages 452–460, Cham. Springer International Publishing.

Cornett, O. (1967). Cued speech. *American annals of the deaf*, 112:3–13.

Cox, S. J., Harvey, R. W., Lan, Y., Newman, J. L., and Theobald, B.-J. (2008). The challenge of multi-speaker lip-reading. In *AVSP*, pages 179–184. Citeseer.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., and Brumberg, J. S. (2010). Silent speech interfaces. *Speech Communication*, 52(4):270–287.

Deng, J., Guo, J., Ververas, E., Kotsia, I., and Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5202–5211.

Duchnowski, P., Lum, D. S., Krause, J. C., Sexton, M. G., Bratakos, M. S., and Braida, L. D. (2000). Development of speechreading supplements based on automatic speech recognition. *IEEE transactions on biomedical engineering*, 47(4):487–496.

Dupont, S. and Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141–151.

Fernandez-Lopez, A. and Sukno, F. M. (2017). Optimizing phoneme-to-viseme mapping for continuous lip-reading in spanish. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics*, pages 305–328. Springer.

Fernandez-Lopez, A. and Sukno, F. M. (2018). Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing*, 78:53–72.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of speech and hearing research*, 11(4):796–804.

Gales, M. and Young, S. (2008). *The application of hidden Markov models in speech recognition*. Now Publishers Inc.

Gales, M. (1998). Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech & Language*, 12(2):75–98.

Gopinath, R. A. (1998). Maximum likelihood modeling with gaussian distributions for classification. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 2, pages 661–664. IEEE.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.

Juang, B. (1991). Speech recognition in adverse environments. *Computer Speech & Language*, 5(3):275–294.

Kaplan, H., Bally, S. J., and Garretson, C. (1987). *Speechreading: A way to improve understanding*. Gallaudet University Press.

Lan, Y., Harvey, R., Theobald, B., Ong, E.-J., and Bowden, R. (2009). Comparing visual features for lipreading. In *International Conference on Auditory-Visual Speech Processing 2009*, pages 102–106.

Lleida, E., Ortega, A., Miguel, A., Bazán-Gil, V., Pérez, C., Gómez, M., and de Prada, A. (2019). Albayzin 2018 evaluation: the iberspeech-rtve challenge on speech technologies for spanish broadcast media. *Applied Sciences*, 9(24):5412.

Ma, P., Petridis, S., and Pantic, M. (2022). Visual speech recognition for multiple languages in the wild. *arXiv preprint arXiv:2202.13084*.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

Mohri, M., Pereira, F., and Riley, M. (2008). Speech recognition with weighted finite-state transducers. In *Springer Handbook of Speech Processing*, pages 559–584. Springer.

Potamianos, G., Luettin, J., and Neti, C. (2001). Hierarchical discriminant features for audio-visual lvcsr. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 1, pages 165–168. IEEE.

Potamianos, G., Neti, C., Gravier, G., Garg, A., and Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. Paper no. EPFL-CONF-192584.

Quilis, A. (1997). *Principios de fonología y fonética españolas*, volume 43 of *Cuadernos de lengua española*. Arco libros.

Rao, C. R. (1965). *Linear Statistical Inference and is Applications*. John Wiley & Sons, New York.

Rodríguez Ortiz, I. d. l. R. (2008). Lipreading in the prelingually deaf: What makes a skilled speechreader? *The Spanish Journal of Psychology*, 11(2):488–502.

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2016). 300 faces in-the-wild challenge: database and results. *Image and Vision Computing*, 47:3–18.

Stolcke, A. (2002). Srilm – an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, pages 901–904.

Thangthai, K., Harvey, R. W., Cox, S. J., and Theobald, B.-J. (2015). Improving lip-reading performance for robust audiovisual speech recognition using dnns. In *AVSP*, pages 127–131.

Thangthai, K., Bear, H. L., and Harvey, R. (2017). Comparing phonemes and visemes with dnn-based lipreading. In *Proc. British Machine Vision Conference*, pages 4–7.

Thangthai, K. (2018). *Computer lipreading via hybrid deep neural network hidden Markov models*. Ph.D. thesis, University of East Anglia.

Tomar, S. (2006). Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Young, S. J., Odell, J. J., , and Woodland, P. C. (1994). Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of the Workshop on Human Language Technology, HLT '94*, pages 307–312, USA. Association for Computational Linguistics.

Zadeh, A. B., Cao, Y., Hessner, S., Liang, P. P., Poria, S., and Morency, L.-P. (2020). Cmu-moseas: A multimodal language dataset for spanish, portuguese, german and french. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1801–1812.

Zhang, Y., Yang, S., Xiao, J., Shan, S., and Chen, X. (2020). Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 356–363.

Language Resource References

Afouras, T., Chung, J. S., and Zisserman, A. (2018). LRS3-TED: a large-scale dataset for visual speech recognition. In *arXiv preprint arXiv:1809.00496*.

Chung, J. S. and Zisserman, A. (2017). Lip reading in profile. In *British Machine Vision Conference*.

Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2017). Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453. IEEE.

Fernandez-Lopez, A., Martinez, O., and Sukno, F. M. (2017). Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database. In *12th International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 208–215. IEEE.

Harte, N. and Gillen, E. (2015). TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615.

Howell, D., Cox, S., and Theobald, B. (2016). Visual units and confusion modelling for automatic lipreading. *Image and Vision Computing*, 51:1–12.

Lleida, Eduardo and Ortega, Alfonso and Miguel, Antonio and Bazán, Virginia and Pérez, Carmen and Zotano, M and de Prada, Alberto. (2018). *RTVE2018 database description. Online Available: `http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf`.*

Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., and Harvey, R. (2002). Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):198–213.

Patterson, E. K., Gurbuz, S., Tufekci, Z., and Gowdy, J. N. (2002). CUAVE: A new audio-visual database for multimodal human-computer interface research. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II–2017–II–2020. IEEE.

Zadeh, A. B., Cao, Y., Hessner, S., Liang, P. P., Poria, S., and Morency, L.-P. (2020). CMU-MOSEAS: A multimodal language dataset for spanish, portuguese, german and french. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1801–1812.